

ADAPTIVE ESTIMATORS OF A MEAN MATRIX: TOTAL LEAST SQUARES VERSUS TOTAL SHRINKAGE

RUDOLF BERAN*

University of California, Davis

Abstract

An unknown constant matrix M is observed with additive random error. The basic problem considered is to devise an estimator of M that trades off bias against variance so as to achieve relatively low quadratic risk. This paper develops an adaptive total least squares estimator and an adaptive total shrinkage estimator of M that minimize estimated risk over certain large classes of linear estimators. It is shown that the asymptotic risk of the adaptive total least squares estimator is the smallest attainable among reduced rank total least squares fits to the data matrix. The asymptotic risk of the adaptive total shrinkage estimator is shown to be smaller still. A close link is established between total shrinkage and the Efron-Morris estimator of M . In the asymptotics, the row dimension of M tends to infinity while the column dimension stays fixed. The risks converge uniformly when the signal-to-noise ratio and the measurement error variance are both bounded. A second problem treated is estimation of M under the assumption that a linear relation holds among its columns. In this formulation of the errors-in-variables linear regression model, rank constrained adaptive total least squares asymptotically dominates the usual total least squares estimator of M and rank constrained adaptive total shrinkage is better still.

*This research was supported in part by National Science Foundation Grant DMS 0404547. Address correspondence to Rudolf Beran, Department of Statistics, University of California at Davis, Davis, CA 95616, USA; e-mail: beran@wald.ucdavis.edu

1 INTRODUCTION

Let $M = [m_1, \dots, m_p]'$ be an unknown $p \times q$ constant matrix, where $p \geq q$, that is observed with additive random error. Recorded is the $p \times q$ data matrix $X = [x_1, \dots, x_p]'$ modeled by

$$X = M + E. \quad (1)$$

Here $E = [e_1, \dots, e_p]'$ is a random $p \times q$ random error matrix that has the following properties: the column vectors $\{e_i: 1 \leq i \leq p\}$ are independent, and for every i , $E(e_i) = 0$ and $\text{Cov}(e_i) = \sigma^2 I_q$. Typically both M and $\sigma^2 > 0$ are unknown. Equation (1) is a multivariate trend model that relates successive vector observations, the rows of the observed X , to successive vector means, the rows of the unknown M . A basic problem is to estimate M and σ^2 .

The Frobenius (or Euclidean) norm of a matrix B is defined by $|B|^2 = \text{tr}(BB') = \text{tr}(B'B)$. Let $\hat{M} = [\hat{m}_1, \dots, \hat{m}_p]'$ denote any estimator of $M = [m_1, \dots, m_p]'$. The quality of \hat{M} is assessed through the quadratic loss

$$L(\hat{M}, M) = (pq)^{-1} |\hat{M} - M|^2 = (pq)^{-1} \sum_{i=1}^p |\hat{m}_i - m_i|^2. \quad (2)$$

The risk of \hat{M} is then

$$R(\hat{M}, M, \sigma^2) = \text{EL}(\hat{M}, M), \quad (3)$$

the expectation being computed under the model (1).

The risk of the classical unbiased estimator X of M is just σ^2 . This paper develops better estimators of M that trade off bias against variance so to reduce quadratic risk. The fundamental idea is to estimate M by shrinking the rank or, more generally, the singular values of the data matrix X . It is important to note that the risk analysis in the basic problem will make no assumption on the rank of M , even though some of the estimators devised for M have less than full rank.

The *total least squares (TLS) estimator of order k* for M is

$$\hat{M}_{TLS}(k) = \underset{M: \text{rank}(M) \leq k}{\text{argmin}} |X - M|^2. \quad (4)$$

Suppose that the singular value decomposition of X is

$$X = \hat{U} \hat{L} \hat{V}' = \sum_{j=1}^q \hat{l}_j \hat{u}_j \hat{v}_j', \quad (5)$$

where $\hat{U} = [\hat{u}_1, \dots, \hat{u}_q]$ is $p \times q$, $\hat{V} = [\hat{v}_1, \dots, \hat{v}_q]$ is $q \times q$, $\hat{U}'\hat{U} = \hat{V}'\hat{V} = \hat{V}\hat{V}' = I_q$ and $\hat{L} = \text{diag}(\hat{l}_1, \dots, \hat{l}_q)$ with $\hat{l}_1 \geq \dots \geq \hat{l}_q \geq 0$. By the Eckart-Young matrix approximation theorem,

$$\hat{M}_{TLS}(k) = \sum_{j=1}^k \hat{l}_j \hat{u}_j \hat{v}_j'. \quad (6)$$

Derivations of (6) as solution to the minimization problem (4) are given in Golub and Van Loan (1980, 1996) and in Van Huffel and Vandewalle (1991). Existence of numerically stable algorithms for computing the singular value decomposition is an important advantage of representation (6) for $\hat{M}_{TLS}(k)$.

How should we choose k in (6) to minimize the quadratic risk of $\hat{M}_{TLS}(k)$ as an estimator of M ? The answer necessarily takes into account the noise level σ^2 . Let $\hat{\sigma}^2$ be a consistent estimator of σ^2 . Sections 3 and 4 find an asymptotically best data-based choice \hat{k} of k :

$$\hat{k} = \#\{j: \hat{\pi}_j > 1/2\}, \quad \hat{\pi}_j = 1 - p\hat{\sigma}^2/\hat{l}_j^2. \quad (7)$$

It is understood that $\hat{\pi}_j = -\infty$ if $\hat{l}_j = 0$. The corresponding *adaptive TLS estimator* is then

$$\hat{M}_{TLS} = \hat{M}_{TLS}(\hat{k}) = \sum_{j: \hat{\pi}_j > 1/2} \hat{l}_j \hat{u}_j \hat{v}_j'. \quad (8)$$

In notation akin to (5), suppose that the singular value decomposition of M is

$$M = ULV' = \sum_{j=1}^q l_j u_j v_j', \quad (9)$$

where $U = [u_1, \dots, u_q]$ is $p \times q$, $V = [v_1, \dots, v_q]$ is $q \times q$, $U'U = V'V = VV' = I_q$ and $L = \text{diag}(l_1, \dots, l_q)$ with $l_1 \geq \dots \geq l_q \geq 0$. Let

$$\pi_j = \frac{l_j^2}{p\sigma^2 + l_j^2} = 1 - \frac{p\sigma^2}{p\sigma^2 + l_j^2}. \quad (10)$$

It is shown in Section 4 that the risk of \hat{M}_{TLS} converges asymptotically, as p tends to infinity, to

$$q^{-1}\sigma^2 \left[\#\{j: \pi_j > 1/2\} + \sum_{j: \pi_j \leq 1/2} \frac{\pi_j}{1 - \pi_j} \right]. \quad (11)$$

Moreover, $\min_k R(\hat{M}_{TLS}(k), M, \sigma^2)$ also converges to the value (11). In this manner, the adaptive TLS estimator \hat{M}_{TLS} behaves asymptotically like the best of the competing TLS estimators $\{\hat{M}_{TLS}(k): 0 \leq k \leq q\}$. The result makes no assumption on the rank of the unknown M .

It is evident from (8) that \hat{M}_{TLS} applies a shrinkage factor that is either 1 or 0 to each summand in the singular value decomposition (5) of the unbiased estimator X . Can using other shrinkage factors reduce asymptotic risk below that of \hat{M}_{TLS} ? The answer is yes. It is shown in Sections 3 and 4 that the *adaptive total shrinkage* (TS) estimator

$$\hat{M}_{TS} = \sum_{j: \hat{\pi}_j > 0} \hat{\pi}_j \hat{l}_j \hat{u}_j \hat{v}_j' \quad (12)$$

has smaller risk asymptotically than the adaptive TLS estimator of M . Indeed, the risk of \hat{M}_{TS} converges asymptotically, as p tends to infinity, to

$$q^{-1}\sigma^2 \sum_{j=1}^q \pi_j, \quad (13)$$

which cannot exceed (11) and can be smaller. Section 5.1 extends low risk estimation of M to the more general setting where $\text{Cov}(e_i) = \sigma^2 K$ with K a known positive definite matrix.

A second problem is to estimate M and σ^2 under the restriction that $\text{rank}(M)$ is $q - 1$. Model (1) plus this rank restriction defines an errors-in-variables linear regression model in which the columns of M satisfy the linear relation $Mv_q = 0$. The extensive literature on fitting this model uses a variety of other labels, including “total least squares”, “orthogonal regression”, and “measurement error modeling”. For results and further references, see Anderson (1976), Fuller (1987), Gleser (1981), Golub and Van Loan (1980), Van Huffel and Vandewalle (1991), Van Huffel (2004). This literature develops the estimator \hat{v}_p for v_p and the total least squares estimator $\hat{M}_{TLS}(q-1)$ for M . Section 5.2 constructs rank constrained adaptive TLS and TS estimators of M that achieve smaller asymptotic risk than $\hat{M}_{TLS}(q-1)$ under the errors-in-variables model. Again, the key is to trade off bias against variance so as to reduce risk.

2 ORACLE ESTIMATORS

This section studies classes of candidate linear estimators for M , constructing within each class an estimator that minimizes the quadratic risk (3). The best such estimators are *oracle* estimators in that they depend on functions of the unknown parameters M and σ^2 . The labeling of the oracle estimators foreshadows their linkage, in Sections 3 and 4, with the adaptive TLS and TS estimators described in the Introduction.

2.1 Oracle linear estimators

Let A be an arbitrary $q \times q$ matrix. This subsection considers the candidate linear estimator XA for M . The quadratic risk (2) of XA is

$$R(XA, M, \sigma^2) = (pq)^{-1} \text{E}|XA - M|^2 = (pq)^{-1} \text{E} \sum_{i=1}^p |Bx_i - m_i|^2, \quad (14)$$

where $B = A'$. This risk is a strictly convex function of B . Expanded algebraically,

$$R(XA, M, \sigma^2) = (pq)^{-1} \sum_{i=1}^p [\sigma^2 \text{tr}(B'B) + m_i' m_i - 2m_i' B m_i + m_i' B' B m_i]. \quad (15)$$

We find the matrix \tilde{A} that minimizes (15). Let

$$W = p^{-1}M'M = p^{-1} \sum_{i=1}^p m_i m_i'. \quad (16)$$

The matrix derivatives

$$\frac{\partial \text{tr}(B'B)}{\partial B} = 2B, \quad \frac{\partial m_i' B m_i}{\partial B} = m_i m_i', \quad \frac{\partial m_i' B' B m_i}{\partial B} = 2B m_i m_i' \quad (17)$$

(see Section A.15 in Rao and Toutenberg (1995)) entail that

$$\frac{\partial R(XA, M, \sigma^2)}{\partial B} = 2q^{-1}[\sigma^2 B - W + BW]. \quad (18)$$

Setting this risk derivative equal to zero and simplifying yields the minimizing matrix

$$\tilde{A}' = W(\sigma^2 I_q + W)^{-1} = I_q - \sigma^2(\sigma^2 I_q + W)^{-1} = \tilde{A}. \quad (19)$$

Applying (9) to (16) and (19) yields the expressions

$$W = p^{-1} \sum_{j=1}^q l_j^2 v_j v_j', \quad \tilde{A} = \sum_{j=1}^q \pi_j v_j v_j', \quad (20)$$

with π_j as in (10). \tilde{A} is evidently a symmetric $q \times q$ matrix whose eigenvalues all lie in $[0, 1]$.

Thus, the linear estimator XA with smallest quadratic risk is $X\tilde{A}$. It is an *oracle* estimator because its definition involves the unknown parametric function W and σ^2 . By substitution of \tilde{A} into (14), the risk of the oracle linear estimator is seen to be (13). Section 3 constructs an adaptive estimator whose risk converges asymptotically to this desirable value.

2.2 Oracle symmetric affine shrinkage estimators

Let \mathcal{A}_S denote the set of all $q \times q$ symmetric matrices whose eigenvalues lie in $[0, 1]$. By the analysis of the preceding subsection, it is reasonable to limit the search for low risk candidate linear estimators to the symmetric affine shrinkage class $\{XA: A \in \mathcal{A}_S\}$. Let \mathcal{A}_P denote those symmetric matrices whose eigenvalues are either 0 or 1. This is the class of orthogonal projections into R^q . For $0 \leq k \leq q$, let $\mathcal{A}_P(k) \subset \mathcal{A}_P$ denote the orthogonal projections that have eigenvalue 1 with multiplicity k and eigenvalue 0 with multiplicity $q - k$. Evidently

$$\bigcup_{k=0}^q \mathcal{A}_P(k) = \mathcal{A}_P \subset \mathcal{A}_S. \quad (21)$$

For every $A \in \mathcal{A}_S$, the risk (14) of the linear estimator XA simplifies to

$$R(XA, M, \sigma^2) = r(A, W, \sigma^2), \quad (22)$$

where

$$\begin{aligned} r(A, W, \sigma^2) &= q^{-1} \text{tr}[\sigma^2 A^2 + (I_q - A)^2 W] \\ &= q^{-1} \text{tr}[(A - \tilde{A})^2 (\sigma^2 I_q + W)] + q^{-1} \sigma^2 \text{tr}(\tilde{A}). \end{aligned} \quad (23)$$

Of interest for subsequent developments are the following oracle estimators, obtained by minimizing risk over various closed subsets of \mathcal{A}_S :

- The *oracle total shrinkage estimator* of M is $\tilde{M}_{TS} = X \tilde{A}_{TS}$, where

$$\tilde{A}_{TS} = \underset{A \in \mathcal{A}_S}{\text{argmin}} r(A, W, \sigma^2). \quad (24)$$

- The *oracle order k total least squares estimator* of M is $\tilde{M}_{TLS}(k) = X \tilde{A}_{TLS}(k)$, where

$$\tilde{A}_{TLS}(k) = \underset{A \in \mathcal{A}_P(k)}{\text{argmin}} r(A, W, \sigma^2). \quad (25)$$

- The *oracle total least squares estimator* of M is $\tilde{M}_{TLS} = X \tilde{A}_{TLS}$, where

$$\tilde{A}_{TLS} = \underset{A \in \mathcal{A}_P}{\text{argmin}} r(A, W, \sigma^2) = \tilde{A}_{TLS}(\tilde{k}), \quad \tilde{k} = \underset{0 \leq k \leq q}{\text{argmin}} r(\tilde{A}_{TLS}(k), W, \sigma^2). \quad (26)$$

The labels given to the oracle estimators are justified by their linkage to adaptive TLS and TS estimators in Sections 3 and 4. The next theorem provides explicit formulae for these oracle estimators and their risks. The following properties of the $\{\pi_j : 1 \leq j \leq q\}$, defined by (10), assist understanding of this theorem:

- The $\{\pi_j\}$ are nonincreasing in j and all lie in $[0, 1]$.
- $p\sigma^2(1 - \pi_j)^{-1} = p\sigma^2 + l_j^2$.
- $p^{-1}|l_j u_j v_j'|^2 / \sigma^2 = (p^{-1} l_j^2) / \sigma^2 = \pi_j / (1 - \pi_j)$ is the mean square signal-to-noise ratio of the j -th summand in the singular value decomposition of M .
- $\pi_j > 1/2$ if and only if the foregoing signal-to-noise ratio exceeds 1.
- $\pi_j = 0$ if and only if the foregoing signal-to-noise ratio is 0.

THEOREM 2.1. *The following expressions hold:*

$$\begin{aligned} \tilde{A}_{TS} &= \sum_{j=1}^q \pi_j v_j v_j' = \tilde{A} \\ R(\tilde{M}_{TS}, M, \sigma^2) &= q^{-1} \sigma^2 \sum_{j=1}^q \pi_j = q^{-1} \sigma^2 \text{tr}(\tilde{A}), \end{aligned} \quad (27)$$

with \tilde{A} defined in (19), and

$$\begin{aligned}\tilde{A}_{TLS} &= \sum_{j: \pi_j > 1/2} v_j v_j' \\ R(\tilde{M}_{TLS}, M, \sigma^2) &= q^{-1} \sigma^2 \left[\#\{j: \pi_j > 1/2\} + \sum_{j: \pi_j \leq 1/2} \frac{\pi_j}{1 - \pi_j} \right],\end{aligned}\tag{28}$$

and

$$\begin{aligned}\tilde{A}_{TLS}(k) &= \sum_{i=1}^k v_i v_i' \\ R(\tilde{M}_{TLS}(k), M, \sigma^2) &= q^{-1} \sigma^2 \left[k + \sum_{j=k+1}^q \frac{\pi_j}{1 - \pi_j} \right].\end{aligned}\tag{29}$$

Proof. The first line in (27) follows from (23) and (20) or from Section 2.1.

Consider symmetric matrices A that take the form

$$A = \sum_{j=1}^q f_j v_j v_j',\tag{30}$$

where each $f_j \in [0, 1]$. For such A , (20) and the second line of (23) imply

$$r(A, W, \sigma^2) = q^{-1} \sigma^2 \sum_{j=1}^q [(f_j - \pi_j)^2 (1 - \pi_j)^{-1} + \pi_j].\tag{31}$$

The second line in (27) follows by applying (31) to the first line of (27).

For every $A \in \mathcal{A}_P(k)$, (23) gives

$$r(A, W, \sigma^2) = q^{-1} [\sigma^2 k + \text{tr}(I_q - A)^2 W].\tag{32}$$

This and (25) entail

$$\tilde{A}_{TLS}(k) = \underset{A \in \mathcal{A}_P(k)}{\text{argmin}} |W^{1/2} - AW^{1/2}|^2.\tag{33}$$

In this equation, $\text{rank}(AW^{1/2}) \leq k$ under the constraint $\text{rank}(A) = k$. By the Eckart-Young matrix approximation theorem and (20), the best approximation in norm of rank $\leq k$ to $W^{1/2}$ is

$$p^{-1/2} \sum_{j=1}^k l_j v_j v_j'.\tag{34}$$

This quantity equals $\tilde{A}_{TLS}(k)W^{1/2}$ when $\tilde{A}_{TLS}(k)$ is given by the first line of (29), thereby justifying that formula. The second line follows from (31) because $\tilde{A}_{TLS}(k)$ has the form (30).

Because \tilde{A}_{TLS} minimizes risk over the matrices $\{\tilde{A}_{TLS}(k): 0 \leq k \leq q\}$, finding it amounts to minimizing the risk (31) over $\{f_j\}$ values that are either 0 or 1. The minimum risk is achieved by setting $f_j = 1$ when $\pi_j > 1/2$ and $f_j = 0$ otherwise. The expressions in (28) follow. In other words, the optimal choice of k in constructing the oracle TLS estimator is $\tilde{k} = \#\{j: \pi_j > 1/2\}$. Note that the first \tilde{k} summands in the singular value decomposition of M are precisely those summands whose signal-to-noise ratio exceeds 1. This makes \tilde{M}_{TLS} an intuitively plausible thresholding estimator. □

3 ADAPTIVE ESTIMATORS

This section devises adaptive estimators that are realizable data-based approximations to the oracle estimators discussed in Section 2. The oracle construction is modified by replacing the risk function $r(A, W, \sigma^2)$, which contains unknown parameters, with an estimated risk function. The resulting adaptive estimators coincide with the TS and TLS estimators described in the Introduction. It will be seen in Section 4 that the risk of each adaptive estimator converges to that of its oracle counterpart as p tends to infinity.

3.1 Estimated risk and adaptation

Let $\hat{\sigma}^2$ be an L_1 -consistent estimator of σ^2 . Two constructions of $\hat{\sigma}^2$ are described in Section 4.1. It will be seen in proving Theorem 4.3 that $\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbf{E}|p^{-1}X'X - W - \sigma^2 I_q|^2 = 0$ for every finite $c > 0$. This suggests estimating W by

$$\hat{W} = p^{-1}X'X - \hat{\sigma}^2 I_q = p^{-1} \sum_{i=1}^p x_i x_i' - \hat{\sigma}^2 I_q. \quad (35)$$

Let

$$\check{A} = \hat{W}(\hat{\sigma}^2 I_q + \hat{W})^{-1} = I_q - \hat{\sigma}^2(\hat{\sigma}^2 I_q + \hat{W})^{-1}. \quad (36)$$

Applying singular value decomposition (5) yields

$$\hat{W} = p^{-1} \sum_{j=1}^q (\hat{l}_j^2 - p\hat{\sigma}^2) \hat{v}_j \hat{v}_j', \quad \check{A} = \sum_{j=1}^q \hat{\pi}_j \hat{v}_j \hat{v}_j', \quad (37)$$

for $\hat{\pi}_j = 1 - p\hat{\sigma}^2/\hat{l}_j^2$ as in (7). The *estimated risk* of the candidate symmetric affine shrinkage estimator XA is defined to be

$$\begin{aligned} \hat{r}(A) &= r(A, \hat{W}, \hat{\sigma}^2) = q^{-1} \text{tr}[\hat{\sigma}^2 A^2 + (I_q - A)^2 \hat{W}] \\ &= q^{-1} \text{tr}[(A - \check{A})^2 (\hat{\sigma}^2 I_q + \hat{W})] + q^{-1} \hat{\sigma}^2 \text{tr}(\check{A}). \end{aligned} \quad (38)$$

Corresponding to the oracle estimators discussed in Section 2.2 are the following adaptive estimators, obtained by minimizing estimated risk over various closed subsets of \mathcal{A}_S :

- The *adaptive total shrinkage estimator* of M is $\hat{M}_{TS} = X\hat{A}_{TS}$, where

$$\hat{A}_{TS} = \operatorname{argmin}_{A \in \mathcal{A}_S} \hat{r}(A). \quad (39)$$

- The *adaptive order k total least squares estimator* of M is $\hat{M}_{TLS}(k) = X\hat{A}_{TLS}(k)$, where

$$\hat{A}_{TLS}(k) = \operatorname{argmin}_{A \in \mathcal{A}_P(k)} \hat{r}(A). \quad (40)$$

- The *adaptive total least squares estimator* of M is $\hat{M}_{TLS} = X\hat{A}_{TLS}$, where

$$\hat{A}_{TLS} = \operatorname{argmin}_{A \in \mathcal{A}_P} r(A, W, \sigma^2) = \hat{A}_{TLS}(\hat{k}), \quad \hat{k} = \operatorname{argmin}_{0 \leq k \leq q} \hat{r}(\hat{A}_{TLS}(k)). \quad (41)$$

The next theorem shows that these adaptive estimators coincide with TLS and TS estimators discussed in the Introduction and provides formulae for their estimated risks. The following properties of the $\{\hat{\pi}_j : 1 \leq j \leq q\}$, defined by (7), assist understanding of this theorem:

- The $\{\hat{\pi}_j\}$ are nonincreasing in j with values that do not exceed 1 but can be negative, unlike the values of the $\{\pi_j\}$.
- $p\hat{\sigma}^2(1 - \hat{\pi}_j)^{-1} = \hat{l}_j^2$.
- $p^{-1}\hat{l}_j^2 - \hat{\sigma}^2$ is an asymptotically unbiased estimator for $p^{-1}l_j^2$ as p tends to infinity (cf. the reasoning for Theorem 4.3 and, more generally, Van Huffel and Vandewalle (1981)).
- Hence, $(p^{-1}\hat{l}_j^2)/\hat{\sigma}^2 - 1 = \hat{\pi}_j/(1 - \hat{\pi}_j)$ estimates the mean square signal-to-noise ratio $(p^{-1}l_j^2)/\sigma^2 = p^{-1}|l_j u_j v_j'|^2 = \pi_j/(1 - \pi_j)$ of the j -th summand in the singular value decomposition of M .
- $\hat{\pi}_j > 1/2$ if and only if the foregoing estimated signal-to-noise exceeds 1.
- $\hat{\pi}_j \geq 0$ if and only if the foregoing estimated signal-to-noise ratio is non-negative.

Unlike its oracle counterpart \tilde{A} , the matrix \check{A} defined in (37) has eigenvalues less than or equal to 1 rather than in $[0, 1]$. Define

$$\check{A}_+ = \sum_{j: \hat{\pi}_j \geq 0} \hat{\pi}_j \hat{v}_j \hat{v}_j', \quad \check{A}_- = \sum_{j: \hat{\pi}_j < 0} \hat{\pi}_j \hat{v}_j \hat{v}_j'. \quad (42)$$

Note that $\check{A}_+ \in \mathcal{A}_S$, $\check{A} = \check{A}_+ + \check{A}_-$, and $\check{A}_+ \check{A}_- = 0$. The adaptive TS estimator involves \check{A}_+ .

THEOREM 3.1. *The following expressions hold:*

$$\begin{aligned}\hat{A}_{TS} &= \sum_{j: \hat{\pi}_j > 0} \hat{\pi}_j \hat{v}_j \hat{v}'_j = \check{A}_+ \\ \hat{r}(\hat{A}_{TS}) &= q^{-1} \hat{\sigma}^2 \left[\sum_{j=1}^q \hat{\pi}_j + \sum_{j: \hat{\pi}_j < 0} \frac{\hat{\pi}_j^2}{1 - \hat{\pi}_j} \right] = q^{-1} \hat{\sigma}^2 \left[\sum_{j: \hat{\pi}_j > 0} \hat{\pi}_j + \sum_{j: \hat{\pi}_j < 0} \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} \right]\end{aligned}\quad (43)$$

and

$$\begin{aligned}\hat{A}_{TLS} &= \sum_{j: \hat{\pi}_j > 1/2} \hat{v}_j \hat{v}'_j \\ \hat{r}(\hat{A}_{TLS}) &= q^{-1} \hat{\sigma}^2 \left[\#\{j: \hat{\pi}_j > 1/2\} + \sum_{j: \hat{\pi}_j \leq 1/2} \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} \right],\end{aligned}\quad (44)$$

and

$$\begin{aligned}\hat{A}_{TLS}(k) &= \sum_{i=1}^k \hat{v}_i \hat{v}'_i \\ \hat{r}(\hat{A}_{TLS}(k)) &= q^{-1} \hat{\sigma}^2 \left[k + \sum_{j=k+1}^q \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} \right].\end{aligned}\quad (45)$$

Consequently,

$$\hat{M}_{TS} = \sum_{j: \hat{\pi}_j > 0} \hat{\pi}_j \hat{l}_j \hat{u}_j \hat{v}'_j, \quad \hat{M}_{TLS} = \sum_{j: \hat{\pi}_j > 1/2} \hat{l}_j \hat{u}_j \hat{v}'_j \quad (46)$$

Proof. Let

$$\hat{T} = \hat{\sigma}^2 I_q + \hat{W} = p^{-1} \sum_{j=1}^q \hat{l}_j^2 \hat{v}_j \hat{v}'_j = \hat{\sigma}^2 \sum_{j=1}^q (1 - \hat{\pi}_j)^{-1} \hat{v}_j \hat{v}'_j, \quad (47)$$

From the second line of (38),

$$\hat{A}_{TS} = \operatorname{argmin}_{A \in \mathcal{A}_S} \operatorname{tr}[(A - \check{A})^2 \hat{T}] \quad (48)$$

Observe that

$$\begin{aligned}\operatorname{tr}[(A - \check{A})^2 \hat{T}] &= \operatorname{tr}[\{(A - \check{A}_+) - \check{A}_-\}^2 \hat{T}] \\ &= \operatorname{tr}[(A - \check{A}_+)^2 \hat{T}] - \operatorname{tr}[A \check{A}_- \hat{T}] - \operatorname{tr}[\check{A}_- A \hat{T}] + \operatorname{tr}[\check{A}_-^2 \hat{T}].\end{aligned}\quad (49)$$

For every $A \in \mathcal{A}_S$,

$$-\operatorname{tr}[A \check{A}_- \hat{T}] = -p^{-1} \operatorname{tr} \left[A \sum_{j: \hat{\pi}_j < 0} \hat{\pi}_j \hat{l}_j^2 \hat{v}_j \hat{v}'_j \right] = -p^{-1} \left[\sum_{j: \hat{\pi}_j < 0} \hat{\pi}_j \hat{l}_j^2 \hat{v}'_j A \hat{v}_j \right] \geq 0 \quad (50)$$

because A is positive semidefinite. Similarly, $-\text{tr}[\check{A}_- A \hat{T}] \geq 0$. Moreover, $\text{tr}[(A - \check{A}_+)^2 \hat{T}] = |(A - \check{A}_+)|^2 \hat{T}^{1/2}|^2 \geq 0$. Hence, (49) implies

$$\min_{A \in \mathcal{A}_S} \text{tr}[(A - \check{A})^2 \hat{T}] \geq \text{tr}[\check{A}_-^2 \hat{T}]. \quad (51)$$

On the other hand,

$$\min_{A \in \mathcal{A}_S} \text{tr}[(A - \check{A})^2 \hat{T}] \leq \text{tr}[(\check{A}_+ - \check{A})^2 \hat{T}] = \text{tr}[\check{A}_-^2 \hat{T}]. \quad (52)$$

From (51) and (52), $\min_{A \in \mathcal{A}_S} \text{tr}[(A - \check{A})^2 \hat{T}] = \text{tr}[\check{A}_-^2 \hat{T}]$ and the minimum is achieved when $A = \check{A}_+$. This proves the first line in (43).

Consider symmetric matrices A that take the form

$$A = \sum_{j=1}^q f_j \hat{v}_j \hat{v}_j', \quad (53)$$

where each $f_j \in [0, 1]$. For such A , (37) and the second line of (38) imply

$$\hat{r}(A) = q^{-1} \hat{\sigma}^2 \sum_{j=1}^q [(f_j - \hat{\pi}_j)^2 (1 - \hat{\pi}_j)^{-1} + \hat{\pi}_j]. \quad (54)$$

The second line in (43) follows by applying (54).

For every $A \in \mathcal{A}_P(k)$, the first line in (38) and (47) imply

$$\hat{r}(A) = q^{-1} \{ \hat{\sigma}^2 (2k - q) + \text{tr}[(I_q - A)^2 \hat{T}] \}. \quad (55)$$

This and (40) entail

$$\hat{A}_{TLS}(k) = \underset{A \in \mathcal{A}_P(k)}{\text{argmin}} |\hat{T}^{1/2} - A \hat{T}^{1/2}|^2. \quad (56)$$

In this equation, $\text{rank}(\hat{T}^{1/2} A) \leq k$ under the constraint $\text{rank}(A) = k$. By the Eckart-Young matrix approximation theorem and (47), the best approximation in norm of $\text{rank} \leq k$ to $\hat{T}^{1/2}$ is

$$p^{-1/2} \sum_{j=1}^k \hat{l}_j \hat{v}_j \hat{v}_j'. \quad (57)$$

This quantity equals $\hat{A}_{TLS}(k) \hat{T}^{1/2}$ when $\hat{A}_{TLS}(k)$ is given by the first line of (45), thereby justifying that formula. The second line follows from (54) because $\hat{A}_{TLS}(k)$ has the form (53).

Because \hat{A}_{TLS} minimizes estimated risk over the matrices $\{\hat{A}_{TLS}(k) : 0 \leq k \leq q\}$, finding it amounts to minimizing the estimated risk (54) over $\{f_j\}$ values that are either 0 or 1. The minimum risk is achieved by setting $f_j = 1$ when $\hat{\pi}_j > 1/2$ and $f_j = 0$ otherwise. The expressions in (44) follow. In other words, the optimal choice of k in constructing the

adaptive TLS estimator is $\hat{k} = \#\{j: \hat{\pi}_j > 1/2\}$. Note that the first \hat{k} summands in the singular value decomposition of M are precisely those summands whose estimated signal-to-noise ratio exceeds 1. This makes \hat{M}_{TLS} an intuitively plausible thresholding estimator.

The expressions (46) for \hat{M}_{TLS} and \hat{M}_{TS} follow from their definitions and the already proved parts of Theorem 3.1. \square

3.2 Total shrinkage and the Efron-Morris estimator

In the case of Gaussian errors, empirical Bayes estimators of M may be constructed as follows. Suppose that the conditional distribution of x_i given m_i is $N(m_i, \sigma^2 I_q)$ and that the $\{x_i\}$ are conditionally independent. Suppose that the $\{m_i\}$ are independent and that the distribution of m_i is $N(0, W)$. Under quadratic loss, the Bayes estimator of M is then the oracle estimator $\tilde{M}_{TS} = X\tilde{A}_{TS}$.

Since $\hat{A}_{TS} \in \mathcal{A}_S$ is a consistent estimator of \tilde{A}_{TS} in the Bayes model, the adaptive TS estimator

$$\hat{M}_{TS} = X\hat{A}_{TS} = \sum_{j: \hat{\pi}_j > 0} \hat{\pi}_j \hat{l}_j \hat{u}_j \hat{v}_j', \quad (58)$$

derived in Theorem 3.1, is an empirical Bayes estimator of M . Another consistent estimator of \tilde{A}_{TS} in the Bayes model is \check{A} , which need not belong to \mathcal{A}_S for finite p . This generates the empirical Bayes estimator

$$\hat{M}_{EB} = X\check{A} = X[I_p - p\hat{\sigma}^2(X'X)^{-1}] = \sum_{j=1}^q \hat{\pi}_j \hat{l}_j \hat{u}_j \hat{v}_j'. \quad (59)$$

Related estimators, albeit more complex, have been used by Green, Berman, Switzer and Craig (1985) to analyze multiband satellite image data. The Efron-Morris (1976) estimator of M , assuming σ^2 known, is

$$\hat{M}_{EM} = X[I_q - (p - q - 1)\sigma^2(X'X)^{-1} - \frac{\sigma^2(q^2 + q - 2)}{\text{tr}(X'X)}I_q]. \quad (60)$$

The estimator \hat{M}_{TS} generalizes the positive-part James-Stein (1961) estimator to q -dimensional means, while the estimator \hat{M}_{EB} in (59) similarly generalizes the James-Stein estimator. The Efron-Morris estimator \hat{M}_{EM} is a refinement of \hat{M}_{EB} whose efficacy in reducing risk depends on the Gaussian error assumption. Under Assumptions A1 and A2 of Section 4, which do *not* assume Gaussian errors, the risks of \hat{M}_{TS} , \hat{M}_{EB} , and \hat{M}_{EM} converge together as p tends to infinity, as do the estimators themselves in terms of the normalized loss metric. Each of these symmetric affine shrinkage estimators has a computationally stable expression in terms of the singular value decomposition of X . Moreover, the singular value representations reveal the relationship between these shrinkage estimators and the adaptive total least squares estimator \hat{M}_{TLS} .

4 ASYMPTOTIC THEORY

This section develops asymptotic theory in which row dimension q is fixed as column dimension p tends to infinity. It is found that the loss of \hat{M}_{TS} or \hat{M}_{TLS} converges to its respective risk. Secondly, the risk (or loss) of \hat{M}_{TS} or \hat{M}_{TLS} converges to that of its oracle counterpart. Thirdly, the estimated risk of \hat{M}_{TS} or \hat{M}_{TLS} converges to its risk (or loss). The convergences are uniform when the signal-to-noise ratio $(p\sigma^2)^{-1}|M|^2$ and the variance σ^2 are both bounded. For simplicity, the notation often omits the dependence of quantities on p .

4.1 Convergence of loss, risk, and estimated risk

Let \mathcal{A} denote the set of all symmetric matrices $A = \{a_{ij}\}$ such that $\max_{i,j} |a_{ij}| \leq 1$. Evidently $\mathcal{A}_P \subset \mathcal{A}_S \subset \mathcal{A}$. The following Assumptions support the asymptotic results:

A1 The error vectors $\{e_i: 1 \leq i \leq p\}$ are independent, identically distributed, such that $E(e_i) = 0$, $\text{Cov}(e_i) = \sigma^2 I_q$, where $\sigma^2 > 0$ is unknown.

A2 Under the model of Assumption A1, the variance estimator $\hat{\sigma}^2$ satisfies

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} E|\hat{\sigma}^2 - \sigma^2| = 0 \quad (61)$$

for every finite $c > 0$.

THEOREM 4.1. *Suppose Assumptions A1 and A2 hold. Let $T(A)$ denote either the loss $L(XA, M)$ or the estimated risk $\hat{r}(A)$ of candidate estimator XA . Then, for every finite $c > 0$ and every finite $\sigma^2 > 0$,*

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} E[\sup_{A \in \mathcal{A}} |T(A) - r(A, W, \sigma^2)|] = 0. \quad (62)$$

This theorem shows that the loss, risk, and estimated risk of candidate estimator XA converge together asymptotically. The uniformity of this convergence over all $A \in \mathcal{A}$ makes estimated risk a trustworthy surrogate for the true loss or risk.

THEOREM 4.2. *Suppose Assumptions A1 and A2 hold. Then, for every finite $c > 0$ and for every finite $\sigma^2 > 0$,*

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} |R(\hat{M}_{TS}, M, \sigma^2) - r(\hat{A}_{TS})| = 0. \quad (63)$$

Moreover, for S equal to either the loss $L(\hat{M}_{TS}, M)$ or the risk $R(\hat{M}_{TS}, M, \sigma^2)$,

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} E|\hat{r}(\hat{A}_{TS}) - S| = 0. \quad (64)$$

Replacing the subscript TS by TLS in these statements yields valid assertions about \hat{M}_{TLS} .

By (63), the risk of the adaptive TS estimator \hat{M}_{TS} converges to the risk of the oracle TS estimator \tilde{M}_{TS} , which achieves minimum risk over the class of symmetric affine shrinkage estimators $\{XA: A \in \mathcal{A}_S\}$. By (64), the plug-in estimator $\hat{r}(\hat{A}_{TS})$ of the risk of \hat{M}_{TS} converges to the actual risk or loss of \hat{M}_{TS} . In this manner, we can gauge directly from the data how well adaptation has controlled risk. Theorem 3.1 gives a convenient, computationally stable expression for $\hat{r}(\hat{A}_{TS})$. Analogous statements hold for the parts of Theorems 4.2 and 3.1 that concern the adaptive TLS estimator \hat{M}_{TLS} .

Variance estimation. It remains to construct suitably consistent estimators of σ^2 . When enough repeated observations are available on rows of M , the least squares estimator of σ^2 satisfies Assumption A2. In the absence of replication, we may consider the *smallest singular value* estimator

$$\hat{\sigma}_{SSV}^2 = \hat{l}_q^2/p. \quad (65)$$

The next theorem shows that $\hat{\sigma}_{SSV}^2$ is asymptotically biased upwards in general and satisfies Assumption A2 under a restriction on the magnitude of the smallest singular value of M .

THEOREM 4.3. *Suppose Assumption A1 holds and that the q components of each e_i have finite fourth moments. Then, for every finite $c > 0$ and for every finite $\sigma^2 > 0$,*

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|\hat{\sigma}_{SSV}^2 - p^{-1}l_q^2 - \sigma^2|^2 = 0. \quad (66)$$

Let $\{\epsilon_p \geq 0: p \geq 1\}$ be any sequence such that $\lim_{p \rightarrow \infty} \epsilon_p = 0$. Then

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c, p^{-1}l_q^2 \leq \epsilon_p} \mathbb{E}|\hat{\sigma}_{SSV}^2 - \sigma^2| = 0. \quad (67)$$

Thus, $\hat{\sigma}_{SSV}^2$ is a consistent estimator of σ^2 provided l_q^2/p tends to zero as p tends to infinity. If $\hat{\sigma}_{SSV}^2$ is used to construct the adaptive TS and TLS estimators, then Theorems 4.1 and 4.2 continue to hold with one modification: the supremum over all M such $p^{-1}|M|^2 \leq c$ must be replaced by the supremum over M such $p^{-1}|M|^2 \leq c$, $p^{-1}l_q^2 \leq \epsilon_p$. This recognizes that (67) holds for $\hat{\sigma}_{SSV}^2$ rather than Assumption A2.

When σ^2 is estimated by $\hat{\sigma}_{SSV}^2$, then the \hat{k} in (7) that defines the adaptive TLS estimator simplifies elegantly to $\hat{k} = \#\{j: \hat{l}_j^2 > 2\hat{l}_q^2\}$. The interpretation of \hat{k} remains as described in the proof of Theorem 3.1: the first \hat{k} summands in the singular value decomposition of M are precisely those summands whose estimated signal-to-noise ratio exceeds 1.

4.2 Numerical experiment

The pertinence of the asymptotic theory to finite sample sizes can be checked on artificial data. We report one experiment with \hat{M}_{TS} and \hat{M}_{TLS} . Let $\{u_i: 1 \leq i \leq 4\}$ be independent

100×1 random vectors, the elements of each being independent Uniform $[0, 1]$ random variables. Set

$$\begin{aligned} c_1 &= 100u_1, & c_2 &= 70u_2, & c_3 &= 20u_3 \\ c_4 &= 10u_4, & c_5 &= .1c_1 + .9c_2 + 1. \end{aligned} \tag{68}$$

Let M be the result of rounding each entry in the matrix $[c_1, c_2, c_3, c_4, c_5]$ to the nearest integer. Let E be a 100×5 matrix of independent Normal $(0, \sigma^2)$ random variables, with $\sigma^2 = 25$, each rounded to the nearest integer. Set $X = M + E$. This construction satisfies Assumption A1 with $p = 100$ and $q = 5$.

For an instance of such artificial data constructed in S-Plus (code available from the author), the variance estimate $\hat{\sigma}_{SSV}^2 = 23.58$ approximates $\sigma^2 = 25$, the vectors

$$\begin{aligned} \pi &= (.996, .967, .679, .287, .005)' \\ \hat{\pi} &= (.996, .969, .673, .158, .000)' \end{aligned} \tag{69}$$

are close, while

$$\begin{aligned} l &= (803.29, 271.84, 72.72, 31.75, 3.38)' \\ \hat{l} &= (814.80, 273.54, 84.88, 52.90, 48.56)' \end{aligned} \tag{70}$$

The estimated singular values exhibit an upward bias that is caused by the errors E . Theorem 4.3 explains this for the smallest singular value and similar analysis is available for the others.

Because the data has been constructed, it is possible to compute oracle estimators and to compare their loss and risk with that of the unbiased estimator X . In particular, using risk expressions in Theorem 2.1,

$$\begin{aligned} L(\tilde{M}_{TS}, M) &= 15.09, & R(\tilde{M}_{TS}, M, \sigma^2) &= 14.67 \\ L(\tilde{M}_{TLS}, M) &= 17.28, & R(\tilde{M}_{TLS}, M, \sigma^2) &= 17.04 \\ L(X, M) &= 24.00, & R(X, M, \sigma^2) &= \sigma^2 = 25.00. \end{aligned} \tag{71}$$

The rank of the oracle TLS estimator is 3. As expected from limits in Theorem 4.2, the loss and risk are close to one another for each oracle estimator and are smaller for the oracle TS estimator than for the oracle TLS estimator. Both the oracle TLS estimator and the oracle TS estimator dominate the unbiased estimator X substantially.

Using estimated risk expressions in Theorem 3.1, the loss and estimated risk of the adaptive estimators are

$$\begin{aligned} L(\hat{M}_{TS}, M) &= 15.38, & \hat{r}(\hat{A}_{TS}) &= 13.18 \\ L(\hat{M}_{TLS}, M) &= 17.42, & \hat{r}(\hat{A}_{TLS}) &= 15.03. \end{aligned} \tag{72}$$

The rank of the adaptive TLS estimator is 3, like that of its oracle counterpart. The loss and estimated risk approximate one another for each adaptive estimator and are smaller for the adaptive TS estimator than for the adaptive TLS estimator. Moreover, the loss of the adaptive TS or TLS estimator is only slightly larger than the loss of its oracle counterpart. These findings are again consistent with limits in Theorem 4.2. Both the adaptive TLS estimator and the adaptive TS estimator dominate the unbiased estimator X substantially.

4.3 Proofs

We first state three propositions that will be used in proving Theorem 4.1. Let the $\{\lambda_i(B)\}$ denote the eigenvalues of any symmetric matrix B and let $\lambda_{max}(B)$ denote the largest eigenvalue. The spectral norm of any matrix B , symmetric or not, is defined by

$$|B|_s = \sup_{x \neq 0} \frac{|Bx|}{|x|}. \quad (73)$$

Let $l_{max}(B)$ denote the largest singular value of B .

PROPOSITION 4.4.

- a) $|\cdot|_s$ is a matrix norm.
- b) If matrices A and B have compatible dimensions, $|AB|_s \leq |A|_s |B|_s$.
- c) If u is a row or column vector, $|u|_s = |u|$.
- d) If vectors u , v and matrix A have compatible dimensions, $|u'Av| \leq |u||v||A|_s$.
- e) $|B|_s = \lambda_{max}^{1/2}(B'B) = l_{max}(B) = l_{max}(B') = \lambda_{max}^{1/2}(BB') = |B'|_s$.
- f) $|B|_s \leq |B|$.
- g) If B is symmetric, then $|B|_s = \lambda_{max}^{1/2}(B^2) = \max_i |\lambda_i(B)|$ and $|B^2|_s = |B|_s^2$.
- h) If B is $q \times q$ symmetric, then $q^{-1} |\text{tr}(B)| \leq |B|_s$.

These properties follows readily from the definitions given.

PROPOSITION 4.5. Let $\{Y_p: p \geq 1\}$ be random elements of $C([-1, 1]^d)$, the set of all continuous real-valued functions on $[-1, 1]^d$. Let $\text{plim}_{p \rightarrow \infty}$ denote the limit in probability as $p \rightarrow \infty$. Suppose that

$$\text{plim}_{p \rightarrow \infty} Y_p(a) = 0 \quad \forall a \in [-1, 1]^d \quad (74)$$

and that

$$\lim_{\delta \rightarrow 0} \limsup_{p \rightarrow \infty} \text{P} \left[\sup_{|a-b| \leq \delta} |Y_p(a) - Y_p(b)| \geq \epsilon \right] = 0. \quad (75)$$

Then

$$\text{plim}_{p \rightarrow \infty} \sup_{a \in [-1, 1]^d} |Y_p(a)| = 0. \quad (76)$$

The assumptions of Proposition 4.5 imply the weak convergence in $C([-1, 1]^d)$ of $\{Y_p: p \geq 1\}$ to the zero element. From this, (76) follows. See Wichura (1971) for a short proof of the weak convergence.

PROPOSITION 4.6. *Suppose that $\{y_p: p \geq 1\}$, y , $\{z_p: p \geq 1\}$, and z are non-negative random variables such that $\text{plim}_{p \rightarrow \infty} y_p = y$, $y_p \leq z_p$ a.s. for every p , $\text{E}z < \infty$, and $\lim_{p \rightarrow \infty} \text{E}|z_p - z| = 0$. Then $\lim_{p \rightarrow \infty} \text{E}|y_p - y| = 0$.*

Indeed, the conditions on the $\{z_p\}$ and z imply that the $\{z_p\}$ are uniformly integrable. Hence, so are the $\{y_p\}$. The result follows (cf. Neveu (1965), p. 52).

Proof of Theorem 4.1. For any symmetric $q \times q$ matrix $A = \{a_{ij}\}$, let $a = \{\{a_{ij}: 1 \leq i \leq j\}, 1 \leq j \leq q\}$ denote the diagonal and subdiagonal elements of A , organized as a column vector of dimension $d = q(q+1)/2$. The function $A(a)$ that maps $a \in [-1, 1]^d$ to $A \in \mathcal{A}$ is one-to-one and onto. In the proof, we identify A with $A(a)$ and treat the loss, risk, and estimated risk of a candidate estimator XA as functions of a . For convenient notation, let $r(a) = r(A, W, \sigma^2)$ as defined in (23), let $\hat{r}(a) = \hat{r}(A)$ as defined in (38), and let $T(a) = T(A)$ as in the Theorem statement.

The strategy is to show that $T(a) - r(a)$ converges in probability to zero for every $a \in [-1, 1]^d$, then use Proposition 4.5 to show that $\sup_{a \in [-1, 1]^d} |T(a) - r(a)|$ converges in probability to zero, and finally invoke Proposition 4.6 to establish (62). Repeatedly used are the constraint $p^{-1}|M|^2 \leq c$ and the following properties of $B = B(a) = A^2$ and $\bar{B} = \bar{B}(a) = (I_q - A)^2$. For $i \leq j$, the operator ∇_{ij} denotes partial differentiation with respect to the element a_{ij} of a .

$$|B|_s \leq q^2 \quad |\bar{B}|_s \leq (q+1)^2 \quad (77)$$

$$|\nabla_{ij} B|_s \leq 2^{3/2}q \quad |\nabla_{ij} \bar{B}|_s \leq 2^{3/2}(q+1). \quad (78)$$

These bounds use Proposition 4.4 and the identity $\nabla_{ij} B = \nabla_{ij} A \cdot A + A \cdot \nabla_{ij} A$.

We first prove the case $T(a) = \hat{r}(a)$ of (62). Define $\tilde{r}(a)$ by replacing $\hat{\sigma}^2$ with σ^2 in the definition (38) of $\hat{r}(a)$. Because of the inequality

$$|\tilde{r}(a) - \hat{r}(a)| \leq |\hat{\sigma}^2 - \sigma^2| q^{-1} [|\text{tr}(B)| + |\text{tr}(\bar{B})|] \leq |\hat{\sigma}^2 - \sigma^2| [|B|_s + |\bar{B}|_s] \quad (79)$$

and Assumption A2, we may replace $\hat{r}(a)$ by $r(a)$ in the subsequent argument.

Pointwise consistency. Let

$$\begin{aligned} Y_p(a) &= \tilde{r}(a) - r(a) = q^{-1} \text{tr}[\bar{B}(\hat{W} - W)] \\ &= q^{-1} [2p^{-1} \sum_{i=1}^p m'_i \bar{B} e_i + \{p^{-1} \sum_{i=1}^p e'_i \bar{B} e_i - \sigma^2 \text{tr}(\bar{B})\}]. \end{aligned} \quad (80)$$

By the law of large numbers, $\text{plim}_{p \rightarrow \infty} p^{-1} \sum_{i=1}^p e_i' \bar{B} e_i = \text{E}(e_i' \bar{B} e_i) = \sigma^2 \text{tr}(\bar{B})$. Moreover, $\text{E}(p^{-1} \sum_{i=1}^p m_i' \bar{B} e_i) = 0$ and

$$\begin{aligned} \text{Var}(p^{-1} \sum_{i=1}^p m_i' \bar{B} e_i) &= \sigma^2 p^{-2} \sum_{i=1}^p m_i' \bar{B}^2 m_i \leq \sigma^2 p^{-2} \sum_{i=1}^p |m_i|^2 |\bar{B}|_s^2 \\ &= \sigma^2 p^{-2} |M|^2 |\bar{B}|_s^2 \leq \sigma^2 p^{-1} c |\bar{B}|_s^2. \end{aligned} \quad (81)$$

Thus, for every $a \in [-1, 1]^d$,

$$\text{plim}_{p \rightarrow \infty} Y_p(a) = 0. \quad (82)$$

Uniform consistency. For any a, b in $[-1, 1]^d$,

$$\begin{aligned} Y_p(a) - Y_p(b) &= q^{-1} \sum_{j \leq k} (a_{jk} - b_{jk}) [2p^{-1} \sum_{i=1}^p m_i' \nabla_{jk} \tilde{B} e_i + p^{-1} \sum_{i=1}^p e_i' \nabla_{jk} \tilde{B} e_i \\ &\quad - \sigma^2 \text{tr}(\nabla_{jk} \tilde{B})], \end{aligned} \quad (83)$$

where $\nabla_{jk} \tilde{B} = \nabla_{jk} \bar{B}(\tilde{a})$ for some \tilde{a} on the line segment that joins a and b . Thus

$$\begin{aligned} \sup_{|a-b| \leq \delta} |Y_p(a) - Y_p(b)| &\leq \delta q^{-1} \sum_{j \leq k} [2p^{-1} \sum_{i=1}^p m_i' \nabla_{jk} \tilde{B} e_i + |p^{-1} \sum_{i=1}^p e_i' \nabla_{jk} \tilde{B} e_i| \\ &\quad + \sigma^2 |\text{tr}(\nabla_{jk} \tilde{B})|]. \end{aligned} \quad (84)$$

Moreover, using Proposition 4.4,

$$\begin{aligned} q^{-1} |\text{tr}(\nabla_{jk} \tilde{B})| &\leq |\nabla_{jk} \tilde{B}|_s \\ q^{-1} \text{E} |p^{-1} \sum_{i=1}^p m_i' \nabla_{jk} \tilde{B} e_i| &\leq q^{-1} p^{-1} \sum_{i=1}^p |m_i| \text{E} |e_i| |\nabla_{jk} \tilde{B}|_s \leq q^{-1/2} \sigma c^{1/2} |\nabla_{jk} \tilde{B}|_s \\ q^{-1} \text{E} |p^{-1} \sum_{i=1}^p e_i' \nabla_{jk} \tilde{B} e_i| &\leq q^{-1} p^{-1} \sum_{i=1}^p \text{E} |e_i|^2 |\nabla_{jk} \tilde{B}|_s = \sigma^2 |\nabla_{jk} \tilde{B}|_s. \end{aligned} \quad (85)$$

Applying Markov's inequality and (78) to the right side of (85) establishes existence of a finite constant C , not depending on p , such that

$$\text{P} \left[\sup_{|a-b| \leq \delta} |Y_p(a) - Y_p(b)| \geq \epsilon \right] \leq C\delta. \quad (86)$$

Hence,

$$\lim_{\delta \rightarrow 0} \limsup_{p \rightarrow \infty} \text{P} \left[\sup_{|a-b| \leq \delta} |Y_p(a) - Y_p(b)| \geq \epsilon \right] = 0. \quad (87)$$

Limits (82) and (87) plus Proposition 4.5 yield

$$\text{plim}_{p \rightarrow \infty} \sup_{a \in [-1, 1]^d} |Y_p(a)| = 0. \quad (88)$$

L_1 uniform consistency. Let $y_p = \sup_{a \in [-1,1]^d} |Y_p(a)|$. Using (80),

$$\begin{aligned} y_p &\leq q^{-1} \sup_{a \in [-1,1]^d} \left[\left| 2p^{-1} \sum_{i=1}^p m'_i \bar{B} e_i \right| + \left| p^{-1} \sum_{i=1}^p e'_i \bar{B} e_i \right| + \sigma^2 |\text{tr}(\bar{B})| \right] \\ &\leq \sup_{a \in [-1,1]^d} \left[\left| 2q^{-1} p^{-1} \sum_{i=1}^p |m_i| |e_i| \right| + q^{-1} p^{-1} \sum_{i=1}^p |e_i|^2 + \sigma^2 \right] |\bar{B}|_s. \end{aligned} \quad (89)$$

Let $w_p = p^{-1} \sum_{i=1}^p |e_i|^2$. It follows by the Cauchy-Schwarz inequality and (77) that

$$y_p \leq (q+1)^2 (2q^{-1} c^{1/2} w_p^{1/2} + q^{-1} w_p + \sigma^2) \equiv z_p. \quad (90)$$

By the law of large numbers, w_p converges in probability to its expectation $q\sigma^2$. By Vitali's theorem,

$$\begin{aligned} \lim_{p \rightarrow \infty} \mathbb{E} |w_p - q\sigma^2| &= 0 \\ \lim_{p \rightarrow \infty} \mathbb{E} |w_p^{1/2} - q^{1/2} \sigma| &\leq \lim_{p \rightarrow \infty} \mathbb{E}^{1/2} [w_p^{1/2} - q^{1/2} \sigma]^2 = 0. \end{aligned} \quad (91)$$

Let z be the constant obtained from z_p , the right hand side of (90), by replacing each instance of w_p with $q\sigma^2$. It follows from (91) that $\lim_{p \rightarrow \infty} \mathbb{E} |z_p - z| = 0$. This convergence, inequality (90), and Proposition 4.6 imply that (88) can be strengthened to

$$\lim_{p \rightarrow \infty} \mathbb{E} \left[\sup_{a \in [-1,1]^d} |Y_p(a)| \right] = 0. \quad (92)$$

This completes the proof of (62) when $T(A) = \hat{r}(A)$.

The argument for the case $T(A) = L(XA, M)$ of (62) is similar. The loss of XA is

$$\begin{aligned} L(XA, M) &= (pq)^{-1} \sum_{i=1}^p |Ax_i - m_i|^2 \\ &= q^{-1} p^{-1} \sum_{i=1}^p [e'_i B e_i + m'_i \bar{B} m_i + 2m_i D e_i], \end{aligned} \quad (93)$$

where $D = B - A$. Let $Y_p(a) = L(XA, M) - r(a)$. Then

$$Y_p(a) = q^{-1} \left[2p^{-1} \sum_{i=1}^p m'_i D e_i + \left\{ \sum_{i=1}^p e'_i B e_i - \sigma^2 \text{tr}(B) \right\} \right]. \quad (94)$$

Because the right side of (94) has the same structure as the last expression in (80), an argument parallel to the one that follows (80) completes the proof. Indeed, using (77) and (78),

$$|D|_s \leq q^2 + q \quad |\nabla_{ij} D|_s \leq 2^{3/2} q + 2^{1/2}. \quad (95)$$

□

Proof of Theorem 4.2. We show that (62) implies

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|Z - r(\tilde{A}_{TS}, W, \sigma^2)| = 0, \quad (96)$$

where Z can be $L(X\hat{A}_{TS}, M)$ or $L(X\tilde{A}_{TS}, M)$ or $\hat{r}(\hat{A}_{TS})$. The three limits to be proved in (63) and (64) are immediate consequences of (96).

First, (62) with $T(A) = \hat{r}(A)$ entails

$$\begin{aligned} \lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|\hat{r}(\hat{A}_{TS}) - r(\tilde{A}_{TS}, W, \sigma^2)| &= 0 \\ \lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|\hat{r}(\hat{A}_{TS}) - r(\hat{A}_{TS}, W, \sigma^2)| &= 0. \end{aligned} \quad (97)$$

Hence, (96) holds for $Z = \hat{r}(\hat{A}_{TS})$ and

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|r(\hat{A}_{TS}, W, \sigma^2) - r(\tilde{A}_{TS}, W, \sigma^2)| = 0. \quad (98)$$

Second, (62) with $T(A) = L(XA, M)$ gives

$$\begin{aligned} \lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|L(X\hat{A}_{TS}, M) - r(\hat{A}_{TS}, W, \sigma^2)| &= 0 \\ \lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|L(X\tilde{A}_{TS}, M) - r(\tilde{A}_{TS}, W, \sigma^2)| &= 0. \end{aligned} \quad (99)$$

These limits together with (98) establish the remaining two cases of (96). The argument continues to hold when the subscript TS is replaced by TLS . \square

Proof of Theorem 4.3. First we show that

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|p^{-1}X'X - p^{-1}M'M - \sigma^2 I_q|^2 = 0. \quad (100)$$

Write $M = \{m_{ij}\}$ and $E = \{e_{ij}\}$. Let δ_{jk} denote the Kronecker delta. Then,

$$\begin{aligned} p^{-1}X'X - p^{-1}M'M - \sigma^2 I_q &= \left\{ p^{-1} \sum_{i=1}^p e_{ij}e_{ik} - \sigma^2 \delta_{jk} + p^{-1} \sum_{i=1}^p m_{ij}e_{ik} \right. \\ &\quad \left. + p^{-1} \sum_{i=1}^p e_{ij}m_{ik} : 1 \leq j, k \leq q \right\}. \end{aligned} \quad (101)$$

Under Assumption A1 and the finite fourth moment condition on the components of each e_i ,

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E} \left[p^{-1} \sum_{i=1}^p e_{ij}e_{ik} - \sigma^2 \delta_{jk} \right]^2 = 0. \quad (102)$$

As well,

$$\sup_{p^{-1}|M|^2 \leq c} \mathbb{E}[p^{-1} \sum_{i=1}^p m_{ij} e_{ik}]^2 \leq p^{-1} c \sigma^2, \quad \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}[p^{-1} \sum_{i=1}^p e_{ij} m_{ik}]^2 \leq p^{-1} c \sigma^2. \quad (103)$$

Limit (100) now follows.

Next,

$$\begin{aligned} |p^{-1} \hat{l}_q^2 - p^{-1} l_q^2 - \sigma^2| &\leq |\min_{|v|=1} p^{-1} v' X' X v - \min_{|v|=1} (p^{-1} v' M' M v - \sigma^2 v' v)| \\ &\leq \sup_{|v|=1} |v' (p^{-1} X' X - p^{-1} M' M - \sigma^2 I_q) v| \\ &\leq |p^{-1} X' X - p^{-1} M' M - \sigma^2 I_q|, \end{aligned} \quad (104)$$

the last inequality using parts d and f of Proposition 4.4. This inequality plus (100) imply limit (66) and hence also (67). \square

5 FURTHER RESULTS

Subsection 5.1 addresses low risk adaptive estimation of M for a more general covariance structure on the rows of the error matrix E . Subsection 5.2 gives low risk adaptive estimators of M when the rank of M is constrained, as in errors-in-variables linear regression models.

5.1 Generalized TLS and TS estimators

Suppose that Assumptions A1 and A2 are replaced with the more general

B1 The error vectors $\{e_i: 1 \leq i \leq p\}$ are independent, identically distributed, such that $\mathbb{E}(e_i) = 0$, $\text{Cov}(e_i) = \sigma^2 K$, where K is a known, positive definite, $q \times q$ matrix and $\sigma^2 > 0$ is unknown.

B2 Under the model of Assumption B1, the variance estimator $\hat{\sigma}^2$ satisfies

$$\lim_{p \rightarrow \infty} \sup_{p^{-1}|M|^2 \leq c} \mathbb{E}|\hat{\sigma}^2 - \sigma^2| = 0 \quad (105)$$

for every finite $c > 0$.

Consider the problem of estimating M under the loss function

$$\begin{aligned} L(\hat{M}, M) &= (pq)^{-1} \text{tr}[(\hat{M} - M)K^{-1}(\hat{M} - M)'] \\ &= (pq)^{-1} \sum_{i=1}^p (\hat{m}_i - m_i)' K^{-1} (\hat{m}_i - m_i). \end{aligned} \quad (106)$$

To solve this problem, let $Y = XK^{-1/2}$, $N = MK^{-1/2}$, and $F = EK^{-1/2}$. Then $F = [f_1, \dots, f_p]' = [K^{-1/2}e_1, \dots, K^{-1/2}e_p]'$ and

$$Y = N + F. \quad (107)$$

The random vectors $\{f_i\}$ are independent, each having mean vector 0 and covariance matrix $\sigma^2 I_q$.

Estimating M under loss (105) and Assumptions B1 and B2 is thus isomorphic to estimating N under the loss $L(\hat{N}, N) = (pq)^{-1}|\hat{N} - N|^2$ and Assumptions A1 and A2. To obtain the appropriate adaptive TS estimator:

- Use the data transformed matrix Y and variance estimator $\hat{\sigma}^2$ to construct adaptive estimator \hat{N}_{TS} of N as in Section 3;
- Map \hat{N}_{TS} into the estimator $\hat{M}_{TS} = \hat{N}_{TS}K^{1/2}$.

The appropriate TLS estimator is constructed analogously. The theorems in preceding sections map immediately into counterparts that hold for loss (106) under Assumptions B1 and B2. Note that the construction (65) of $\hat{\sigma}_{SSV}^2$ is here applied to the transformed data matrix Y .

5.2 Estimating rank-constrained M

When $\text{rank}(M)$ is $q - 1$, then $l_q = 0$, $M = \sum_{j=1}^{q-1} l_j v_j v_j'$ and the columns of M satisfy the linear constraint $Mv_q = 0$. Assumptions A1, A2 and the rank condition thus define an errors-in-variables linear regression model. The large literature on this topic, cited in Section 1, provides the estimator \hat{v}_p for v_p and the estimator $\hat{M}_{TLS}(q - 1)$ for M . The point of this subsection is to construct better estimators of M , estimators that have lower risk under the rank constraint.

Let $\mathcal{A}_{S,q-1} \subset \mathcal{A}_S$ denote the set of all $q \times q$ symmetric matrices whose smallest eigenvalue vanishes and whose other eigenvalues lie in $[0, 1]$. Under the rank constraint on M , $\pi_q = 0$. Consequently, the oracle TS estimator in (27) now has the form XA with $A \in \mathcal{A}_{S,q-1}$. Let $\mathcal{A}_{P,q-1} \subset \mathcal{A}_S$ denote the set of all $q \times q$ symmetric matrices whose smallest eigenvalue vanishes and whose other eigenvalues are either 0 or 1. Under the rank constraint on M , the oracle TLS estimator in (27) has the form XA with $A \in \mathcal{A}_{P,q-1}$.

Much as in Theorem 3.1, minimizing $\hat{r}(A)$ over $A \in \mathcal{A}_{S,q-1}$ yields the rank constrained adaptive TS estimator

$$\hat{M}_{TS,q-1} = \sum_{j \leq q-1: \hat{\pi}_j > 0} \hat{\pi}_j \hat{l}_j \hat{u}_j \hat{v}_j'. \quad (108)$$

The estimated risk of $\hat{M}_{TS,q-1}$ is obtained by deleting the $\hat{\pi}_q$ term in the second line of (43). On the other hand, minimizing $\hat{r}(A)$ over $A \in \mathcal{A}_{P,q-1}$ yields the rank constrained adaptive TLS estimator

$$\hat{M}_{TLS,q-1} = \sum_{j \leq q-1: \hat{\pi}_j > 1/2} \hat{l}_j \hat{u}_j \hat{v}_j'. \quad (109)$$

The estimated risk of $\hat{M}_{TLS,q-1}$ is obtained by deleting the $\hat{\pi}_q$ term in the second line of (44).

The asymptotic theory of Section 4 carries over to the foregoing rank constrained adaptive and oracle estimators. Under Assumptions A1, A2 and the condition that $\text{rank}(M)$ is $q-1$, it is found that the loss of $\hat{M}_{TS,q-1}$ or $\hat{M}_{TLS,q-1}$ converges to the respective risk. Secondly, the risk (or loss) of $\hat{M}_{TS,q-1}$ or $\hat{M}_{TLS,q-1}$ converges to that of its oracle counterpart. Thirdly, the estimated risk of $\hat{M}_{TS,q-1}$ or $\hat{M}_{TLS,q-1}$ converges to the respective risk (or loss). Moreover, Theorem 4.3 applies with $l_q^2 = 0$. Thus, $\hat{\sigma}_{SV}^2$ has the desired L_1 -consistency property (67).

REFERENCES

- Anderson, T.W. (1976) Estimation of linear functional relationships: approximate distributions and connection with simultaneous equations in econometrics (with discussion). *Journal of the Royal Statistical Society, Series B* 38, 1–36.
- Efron, B. & C. Morris (1976) Multivariate empirical Bayes and estimation of covariance matrices. *Annals of Statistics* 4, 22–32.
- Fuller, W.A. (1987) *Error Measurement Models*. John Wiley.
- Gleser, L.J. (1981) Estimation in a multivariate “errors in variables” regression model: large sample results. *Annals of Statistics* 9, 24–44.
- Golub, G.H. & C.F. Van Loan (1980) An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis* 17, 883–893.
- Golub, G.H. & C.F. Van Loan (1996) *Matrix Computations* (third edition). Johns Hopkins University Press.
- Green, A.A., M. Berman, P. Switzer & M.D. Craig (1985) A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing* 26, 65–74.
- James, W. & C. Stein (1961) Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman ed.) 1, pp. 361–380. University of California Press.

- Neveu, J. (1965) *Mathematical Foundations of the Calculus of Probability*. Holden-Day.
- Rao, C.R. & H. Toutenberg (1995) *Linear Models. Least Squares and Alternatives*. Springer.
- Van Huffel, S. (2004) Total least squares and errors-in-variables modeling: bridging the gap between statistics, computational mathematics and engineering. In *Compstat 2004: Proceedings in Computational Statistics* (J. Antoch, ed.) pp. 539–555. Physica-Verlag.
- Van Huffel, S. & J. Vandewalle (1991) *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM Publications.
- Wichura, M.J. (1971) A note on the weak convergence of stochastic processes. *Annals of Mathematical Statistics* 42, 1769-1772.