Ann Inst Stat Math (2008) 60:843–864
DOI 10.1007/s10463-007-0128-2

Estimating a mean matrix: boosting efficiency by multiple affine shrinkage

Rudolf Beran

Received: 20 July 2006 / Revised: 18 December 2006 / Published online: 19 April 2007
© The Institute of Statistical Mathematics, Tokyo 2007

Abstract The unknown matrix M is the mean of the observed response matrix in a multivariate linear model with independent random errors. This paper constructs regularized estimators of M that dominate, in asymptotic risk, least squares fits to the model and to specified nested submodels. In the first construction, the response matrix is expressed as the sum of orthogonal components determined by the submodels; each component is replaced by an adaptive total least squares fit of possibly lower rank; and these fits are then summed. The second, lower risk, construction differs only in the second step: each orthogonal component is replaced by a modified Efron-Morris fit before summation. Singular value decompositions yield computable formulae for the estimators and their asymptotic and estimated risks. In the asymptotics, the row dimension of M tends to infinity while the column dimension remains fixed. Convergences are uniform when signal-to-noise ratio is bounded.

Keywords Total least squares · Efron-Morris fit · Penalized least squares · Regularization · Rank reduction

1 Introduction

Consider the multivariate linear model

$$Y = M + E, \quad M = XB. \tag{1}$$

This research was supported in part by National Science Foundation Grant DMS 0404547.

R. Beran (✉)
Department of Statistics, University of California, Davis, CA 95616, USA
e-mail: beran@wald.ucdavis.edu

Here the observation matrix Y and the constant matrix M are both $n \times q$ with $n \geq q$. The $n \times d$ design matrix X is known and has rank $p \leq \max\{d, n\}$ while the $d \times q$ matrix B is unknown. The elements of the $n \times q$ error matrix E are independent, identically distributed random variables, each having a $N(0, \sigma^2)$ distribution with unknown positive variance σ^2. A basic problem is to estimate M and σ^2 under this model. Section 5 will extend the results to a multivariate linear model with correlated errors.

The Frobenius norm of a matrix D is defined by $|D|^2 = \mathrm{tr}(DD') = \mathrm{tr}(D'D)$. Let $\hat{M}$ denote any estimator of M. The quality of $\hat{M}$ is assessed through the normalized quadratic loss

$$L(\hat{M}, M) = p^{-1}|\hat{M} - M|^2. \tag{2}$$

The risk of $\hat{M}$ is then

$$R(\hat{M}, M, \sigma^2) = \mathrm{E}L(\hat{M}, M), \tag{3}$$

the expectation being evaluated under model (1). Let X^+ denote the pseudoinverse of X. The *least squares estimator* $\hat{M}_{\mathrm{ls}} = XX^+Y$ of M has risk $q\sigma^2$. It is known from Stein (1956) that $\hat{M}_{\mathrm{ls}}$ is inadmissible under quadratic loss when pq exceeds 2. This paper develops regularized estimators of M that trade off bias against variance so as to achieve lower risk, asymptotically in p, than $\hat{M}_{\mathrm{ls}}$ under model (1). The proposed estimators also dominate specified least squares submodel fits used in their construction.

We begin by considering a set of nested least squares submodel fits to the linear model. It is *not* assumed that any of the proper submodels is true. All risk calculations in this paper are done under model (1), except those for those done in Sect. 5 under a more general error model. Suppose that $X_1, X_2, \ldots X_{s-1}$ are submodel design matrices whose range spaces are nested as follows:

$$\mathcal{R}(X_1) \subset \mathcal{R}(X_2) \subset \cdots \subset \mathcal{R}(X_{s-1}) \subset \mathcal{R}(X). \tag{4}$$

Define

$$P_1 = X_1X_1^+, \quad P_2 = X_2X_2^+ - X_1X_1^+, \ldots, \quad P_s = XX^+ - X_{s-1}X_{s-1}^+. \tag{5}$$

The $\{P_k\}$ are orthogonal projections that are mutually orthogonal: each P_k is symmetric, idempotent and $P_iP_j = 0$ whenever $i \neq j$. Because $\sum_{k=1}^s P_k = XX^+$ and $M = XB$ under model (1), it follows that

$$M = \sum_{k=1}^s P_kM, \quad \hat{M}_{\mathrm{ls}} = \sum_{k=1}^s P_kY = \sum_{k=1}^s P_k\hat{M}_{\mathrm{ls}}. \tag{6}$$

Rank reduction. The first fundamental idea in this paper is to estimate each P_kM more efficiently by reducing, in a data-based manner, the rank of $P_kY = P_k\hat{M}_{\mathrm{ls}}$.

The rank-reduced estimators of the $\{P_k M\}$ are then summed to obtain the *adaptive projection estimator* of M. More fully, suppose that the singular value decomposition of $P_k Y$ is

$$P_k Y = \hat{U}_k \hat{L}_k \hat{V}_k' = \sum_{j=1}^{q} \hat{l}_{kj} \hat{u}_{kj} \hat{v}_{kj}', \tag{7}$$

where $\hat{U}_k = [\hat{u}_{k1}, \ldots \hat{u}_{kq}]$ is $n \times q$, $\hat{V}_k = [\hat{v}_{k1}, \ldots, \hat{v}_{kq}]$ is $q \times q$, $\hat{U}_k' \hat{U}_k = \hat{V}_k' \hat{V}_k = \hat{V}_k \hat{V}_k' = I_q$ and $\hat{L}_k = \mathrm{diag}(\hat{l}_{k1}, \ldots, \hat{l}_{kq})$ with $\hat{l}_{k1} \geq \cdots \geq \hat{l}_{kq} \geq 0$. Let $\hat{\sigma}^2$ be a consistent estimator of σ^2. For each k, define

$$\hat{\tau}_k = p^{-1} \hat{\sigma}^2 \,\mathrm{tr}(P_k), \quad \breve{W}_k = p^{-1} (P_k Y)' P_k Y - \hat{\tau}_k I_q. \tag{8}$$

Note for later use that $\sum_{k=1}^{s} \hat{\tau}_k = \hat{\sigma}^2$. Equation (7) implies the spectral representation

$$\breve{W}_k = \sum_{j=1}^{q} \breve{w}_{kj} \hat{v}_{kj} \hat{v}_{kj}', \quad \breve{w}_{kj} = p^{-1} \hat{l}_{kj}^2 - \hat{\tau}_k. \tag{9}$$

Let $\breve{w}_{kj+} = \max\{\breve{w}_{kj}, 0\}$ and $\breve{w}_{kj-} = \min\{\breve{w}_{kj}, 0\}$. Define $\hat{w}_{kj} = \breve{w}_{kj+}$ and

$$\hat{W}_k = \breve{W}_{k+} = \sum_{j=1}^{q} \hat{w}_{kj} \hat{v}_{kj} \hat{v}_{kj}', \quad \breve{W}_{k-} = \sum_{j=1}^{q} \breve{w}_{kj-} \hat{v}_{kj} \hat{v}_{kj}'. \tag{10}$$

Note that $\hat{W}_k$ is positive semidefinite. The *adaptive projection estimator* of M is

$$\hat{M}_{\mathrm{pro}} = \sum_{k=1}^{s} P_k Y \sum_{j:\hat{w}_{kj} > \hat{\tau}_k} \hat{v}_{kj} \hat{v}_{kj}' = \sum_{k=1}^{s} \sum_{j:\hat{w}_{kj} > \hat{\tau}_k} \hat{l}_{kj} \hat{u}_{kj} \hat{v}_{kj}', \tag{11}$$

the last expression using (7). Because $\hat{M}_{\mathrm{pro}} = \sum_{k=1}^{s} P_k \hat{M}_{\mathrm{pro}} = X X^+ \hat{M}_{\mathrm{pro}}$, it satisfies the linear model constraint on M in model (1). The derivation of estimator (11) from the risk considerations in Sects. 2 and 3 justifies its name.

Let $\hat{h}_k = \#\{j : \hat{w}_{kj} > \hat{\tau}_k\})$. By the Eckart–Young matrix approximation theorem, the kth summand on the right side of (11), $\sum_{j:\hat{w}_{kj} > \hat{\tau}_k} \hat{l}_{kj} \hat{u}_{kj} \hat{v}_{kj}'$, gives, among matrices of rank not exceeding $\hat{h}_k$, the best approximation in Frobenius norm to $P_k Y$. In other words, the kth summand is the total least squares approximation to $P_k Y$ of rank not exceeding $\hat{h}_k$. For accounts of the total least squares problem and of its solution through the singular value decomposition, see Golub and Van Loan (1980, 1996), Van Huffel and Vandewalle (1991), and Van Huffel (2004). The latter references relate total least squares to the extensive statistical literature on errors-in-variables regression, including Gleser (1981) and Fuller (1987). Section 4 shows that $\hat{M}_{\mathrm{pro}}$ implements the asymptotically minimum risk strategy for estimating M through rank reduction of each $P_k Y$ as p tends to infinity.

Suppose, for $1 \le k \le s$, that the singular value decomposition of $P_k M$ is

$$P_k M = U_k L_k V_k' = \sum_{j=1}^{q} l_{kj} u_{kj} v_{kj}', \tag{12}$$

where $U_k = [u_{k1}, \ldots u_{kq}]$ is $n \times q$, $V_k = [v_{k1}, \ldots, v_{kq}]$ is $q \times q$, $U_k' U_k = V_k' V_k = V_k V_k' = I_q$ and $L_k = \mathrm{diag}(l_{k1}, \ldots, l_{kq})$ with $l_{k1} \ge \cdots \ge l_{kq} \ge 0$. For each k, define

$$\tau_k = p^{-1}\sigma^2 \,\mathrm{tr}(P_k), \quad W_k = p^{-1}(P_k M)' P_k M. \tag{13}$$

Note for later use that $\sum_{k=1}^{s} \tau_k = \sigma^2$. Equation (12) implies the spectral representation

$$W_k = \sum_{j=1}^{q} w_{kj} v_{kj} v_{kj}', \quad w_{kj} = p^{-1} l_{kj}^2. \tag{14}$$

It is shown in Sect. 4 that the risk $R(\hat{M}_{\mathrm{pro}}, M, \sigma^2)$ converges asymptotically, as p tends to infinity, to

$$\sum_{k=1}^{s} \sum_{j=1}^{q} \min\{\tau_k, w_{kj}\} \le q\sigma^2 = R(\hat{M}_{\mathrm{ls}}, M, \sigma^2). \tag{15}$$

The convergence is uniform when $p^{-1}|M|^2$ is bounded. It will be seen that $\hat{M}_{\mathrm{pro}}$ asymptotically dominates each of the submodel least squares estimators $\{X_t X_t^+ Y : 1 \le t \le s\}$ for M.

Symmetric affine shrinkage. It is evident from the right side of (11) that $\hat{M}_{\mathrm{pro}}$ applies a shrinkage factor that is either 0 or 1 to each summand in the singular value decomposition (7) of $P_k Y$. Can a more sophisticated shrinkage strategy reduce asymptotic risk below that of $\hat{M}_{\mathrm{pro}}$? The answer is yes. The second fundamental idea in this paper is to estimate each $P_k M$ more efficiently by applying data-based, symmetric affine shrinkage to the right side of each $P_k Y = P_k \hat{M}_{\mathrm{ls}}$. These affinely shrunk estimators of the $\{P_k M\}$ are then summed to estimate M.

The *adaptive symmetric affine estimator* of M is

$$\hat{M}_{\mathrm{sym}} = \sum_{k=1}^{s} P_k Y \sum_{j=1}^{q} \hat{w}_{kj} (\hat{\tau}_k + \hat{w}_{kj})^{-1} \hat{v}_{kj} \hat{v}_{kj}' = \sum_{k=1}^{s} \sum_{j=1}^{q} \hat{w}_{kj} (\hat{\tau}_k + \hat{w}_{kj})^{-1} \hat{l}_{kj} \hat{u}_{kj} \hat{v}_{kj}', \tag{16}$$

the last expression using (7). Because $\hat{M}_{\mathrm{sym}} = \sum_{k=1}^{s} P_k \hat{M}_{\mathrm{sym}} = XX^+ \hat{M}_{\mathrm{sym}}$, it satisfies the linear model constraint on M in model (1). In view of (16), $\hat{M}_{\mathrm{sym}}$ applies a shrinkage factor that lies between 0 and 1 to each summand in the singular value decomposition (7) of $P_k Y$. The derivation of estimator $\hat{M}_{\mathrm{sym}}$ from the risk considerations in Sects. 2 and 3 justifies its name. Section 3 further shows that (16) is equivalent

to estimating each $P_k M$ by a modified Efron and Morris (1972) estimator based on $P_k Y$ and then summing over all k.

Section 4 shows that the risk $R(\hat{M}_{\text{sym}}, M, \sigma^2)$ converges asymptotically, as p tends to infinity, to

$$\sum_{k=1}^{s} \sum_{j=1}^{q} \tau_k w_{kj} (\tau_k + w_{kj})^{-1}. \tag{17}$$

The convergence is uniform when $p^{-1}|M|^2$ is bounded. It follows algebraically from (17) and (15) that the asymptotic risk of $\hat{M}_{\text{sym}}$ lies between the asymptotic risk of $\hat{M}_{\text{pro}}$ and one half that asymptotic 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). These best candidate 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 Sect. 3, with the adaptive projection and symmetric affine shrinkage estimators described in Sect. 1.

2.1 Oracle linear estimators

For $1 \le k \le s$, let A_k be an arbitrary $q \times q$ matrix and let A denote the concatenated matrix $A = [A_1, A_2, \ldots, A_s]$. As candidate estimators for M, we first consider the linear estimators

$$\hat{M}(A) = \sum_{k=1}^{s} P_k Y A_k = \sum_{k=1}^{s} P_k \hat{M}_{\text{ls}} A_k. \tag{18}$$

Because $\hat{M}(A) = \sum_{k=1}^{s} P_k \hat{M}(A) = XX^+ \hat{M}(A)$, it satisfies the linear model constraint on M in (1). The loss (2) of candidate estimator $\hat{M}(A)$ is

$$L(\hat{M}(A), M) = p^{-1}|\hat{M}(A) - M|^2 = p^{-1} \sum_{k=1}^{s} |P_k Y A_k - P_k M|^2. \tag{19}$$

The corresponding risk is

$$\begin{aligned} R(\hat{M}(A), M, \sigma^2) &= \mathrm{E} L(\hat{M}(A), M) \\ &= \sum_{k=1}^{s} \mathrm{tr}[\tau_k A_k A_k' + W_k (A_k - I_q)(A_k - I_q)']. \end{aligned} \tag{20}$$

Using derivative formulae compiled by Lütkepohl (1996),

$$\partial R(\hat{M}(A), M, \sigma^2)/\partial A_k = 2(\tau_k A_k + W_k A_k - W_k). \tag{21}$$

The derivative in (21) vanishes if and only if A_k equals

$$\tilde{A}_k = (\tau_k I_q + W_k)^{-1} W_k = I_q - \tau_k(\tau_k I_q + W_k)^{-1} = W_k(\tau_k I_q + W_k)^{-1}, \tag{22}$$

a symmetric matrix whose eigenvalues all lie in [0, 1]. Let $\tilde{A} = [\tilde{A}_1, \ldots, \tilde{A}_s]$. Because the risk $R(\hat{M}(A), M, \sigma^2)$ is convex in A, it is minimized by setting $A = \tilde{A}$.

2.2 Oracle affine shrinkage and projection estimators

Let $\mathcal{A}_{\text{sym}}$ denote the set of all symmetric $q \times q$ matrices whose eigenvalues lie in [0, 1]. The concatenated matrix $A = [A_1, \ldots, A_s]$ then lies in $\mathcal{A}^s_{\text{sym}}$. By the preceding paragraphs, it is reasonable to limit the search for low risk linear estimators to the *symmetric affine shrinkage* candidate estimators $\{\hat{M}(A) \colon A \in \mathcal{A}^s_{\text{sym}}\}$. Let $\mathcal{A}_{\text{pro}}$ be the subset of matrices in $\mathcal{A}_{\text{sym}}$ that are orthogonal projections. These projections are the elements of $\mathcal{A}_{\text{sym}}$ whose eigenvalues are either 0 or 1.

Let $W = [W_1, W_2, \ldots, W_s]$ and let $\tau = (\tau_1, \ldots, \tau_s)$. For every $A \in \mathcal{A}^s_{\text{sym}}$, the risk (20) of $\hat{M}(A)$ simplifies to

$$R(M(A), M, \sigma^2) = \sum_{k=1}^{s} \rho(A_k, \tau_k, W_k) = r(A, \tau, W) \quad \text{(say)}, \tag{23}$$

where

$$\begin{aligned} \rho(A_k, \tau_k, W_k) &= \operatorname{tr}[\tau_k A_k^2 + (I_q - A_k)^2 W_k] \\ &= \operatorname{tr}[(A_k - \tilde{A}_k)^2(\tau_k I_q + W_k)] + \tau_k \operatorname{tr}(\tilde{A}_k). \end{aligned} \tag{24}$$

Of interest for subsequent developments are the following oracle estimators, obtained by minimizing risk over $\mathcal{A}_{\text{sym}}$ and over $\mathcal{A}_{\text{pro}}$:

- The *oracle symmetric affine shrinkage estimator* of M is $\tilde{M}_{\text{sym}} = \hat{M}(\tilde{A}_{\text{sym}})$, where

$$\tilde{A}_{\text{sym}} = \operatorname*{argmin}_{A \in \mathcal{A}^s_{\text{sym}}} r(A, \tau, W) = [\tilde{A}_{\text{sym},1}, \ldots, \tilde{A}_{\text{sym},s}]. \tag{25}$$

- The *oracle projection estimator* of M is $\tilde{M}_{\text{pro}} = \hat{M}(\tilde{A}_{\text{pro}})$, where

$$\tilde{A}_{\text{pro}} = \operatorname*{argmin}_{A \in \mathcal{A}^s_{\text{pro}}} r(A, \tau, W) = [\tilde{A}_{\text{pro},1}, \ldots, \tilde{A}_{\text{pro},s}]. \tag{26}$$

The next theorem provides explicit formulae for the oracle estimators and their risks.

Theorem 1 *The following expressions hold:*

$$\begin{aligned}\tilde{A}_{\mathrm{sym,k}} &= \tilde{A}_k = \sum_{j=1}^{q} w_{kj}(\tau_k + w_{kj})^{-1} v_{kj} v_{kj}' \\ \tilde{M}_{\mathrm{sym}} &= \sum_{k=1}^{s} P_k Y \tilde{A}_{\mathrm{sym,k}} \\ R(\tilde{M}_{\mathrm{sym}}, M, \sigma^2) &= \sum_{k=1}^{s} \tau_k \operatorname{tr}(\tilde{A}_{\mathrm{sym,k}}) = \sum_{k=1}^{s}\sum_{j=1}^{q} \tau_k w_{kj}(\tau_k + w_{kj})^{-1}.\end{aligned} \tag{27}$$

Moreover,

$$\begin{aligned}\tilde{A}_{\mathrm{pro,k}} &= \sum_{j: w_{kj} > \tau_k} v_{kj} v_{kj}' \\ \tilde{M}_{\mathrm{pro}} &= \sum_{k=1}^{s} P_k Y \tilde{A}_{\mathrm{pro,k}} \\ R(\tilde{M}_{\mathrm{pro}}, M, \sigma^2) &= \sum_{k=1}^{s}\sum_{j=1}^{q} \min\{\tau_k, w_{kj}\}.\end{aligned} \tag{28}$$

Proof The three equations in (28) follow easily from (22), (23), (24), and the spectral representation (14).

Let $\mathcal{A}_{\mathrm{pro}}(t)$ denote the set of matrices in $\mathcal{A}_{\mathrm{pro}}$ whose rank is t. To find $\tilde{A}_{\mathrm{pro}}$, we first minimize each risk component $\rho(A_k, \tau_k, W_k)$ over all $A_k \in \mathcal{A}_{\mathrm{pro}}(t)$, then minimize further over t. For each k, using (24), let

$$\tilde{A}_{\mathrm{pro},k}(t) = \operatorname*{argmin}_{B \in \mathcal{A}_{\mathrm{pro}}(t)} \rho(B, \tau_k, W_k) = \operatorname*{argmin}_{B \in \mathcal{A}_{\mathrm{pro}}(t)} |W_k^{1/2} - B W_k^{1/2}|^2. \tag{29}$$

Evidently, $\operatorname{rank}(B W_k^{1/2}) \le t$ under the constraint $\operatorname{rank}(B) = t$. By the Eckart–Young matrix approximation theorem and the spectral representation (14), the minimum norm approximation to $W_k^{1/2}$ of rank not exceeding t is $\sum_{j=1}^{t} w_{kj}^{1/2} v_{kj} v_{kj}'$. It follows that $\tilde{A}_{\mathrm{pro},k}(t) = \sum_{j=1}^{t} v_{kj} v_{kj}'$ achieves the minimum in (29). From this, the second line of (24), and (14),

$$\begin{aligned}\rho(\tilde{A}_{\mathrm{pro},k}(t), \tau_k, W_k) = &\sum_{j=1}^{q} \{[I(j \le t) - w_{kj}(\tau_k + w_{kj})^{-1}]^2 (\tau_k + w_{kj})\} \\ &+ \tau_k \operatorname{tr}(\tilde{A}_k).\end{aligned} \tag{30}$$

This, in turn, is minimized by selecting t so that the indicator $I(j \le t)$ equals 1 when $w_{kj}(\tau_k + w_{kj})^{-1}$ exceeds 1/2 and equals 0 otherwise. The expression for $\tilde{A}_{\mathrm{pro},k}$ in (28) follows.

Using that expression, (18), (23), and the spectral representation (14) yields the other two equations in (28). □

2.3 Remarks on candidate and oracle estimators

Remark 1 It follows from the risk formulae in Theorem 1 that the risks of the oracle affine projection and symmetric affine shrinkage estimators satisfy the inequalities

$$\begin{aligned} 2^{-1} R(\tilde{M}_{\text{pro}}, M, \sigma^2) &\le R(\tilde{M}_{\text{sym}}, M, \sigma^2) \le R(\tilde{M}_{\text{pro}}, M, \sigma^2) \\ &\le R(\hat{M}_{\text{ls}}, M, \sigma^2) = q\sigma^2. \end{aligned} \tag{31}$$

The oracle estimators $\tilde{M}_{\text{pro}}$ and $\tilde{M}_{\text{sym}}$ are not realizable because they depend on functions of the unknown parameters M and σ^2. However, Sect. 3 will construct data-based approximations to both oracle estimators that achieve the oracle risks asymptotically as p tends to infinity.

Remark 2 The class of candidate estimators $\{\hat{M}(A)\colon A \in \mathcal{A}^s_{\text{sym}}\}$ considered in Sect. 2.2 substantially enlarges the underlying set of least squares submodel fits specified in Sect. 1. Indeed, if $A_k = I_q$ for $1 \le k \le t < s$ and A_k vanishes otherwise, then $\hat{M}(A) = \sum_{k=1}^t P_k Y = X_t X_t^+ Y$, the least squares estimator for M under the constraint $M \in \mathcal{R}(X_t)$. More generally, when some eigenvalues of A_k are strictly less than 1, then $P_k Y A_k = P_k \hat{M}_{\text{ls}} A_k$ shrinks toward 0 the corresponding components of the projection $P_k \hat{M}_{\text{ls}}$, after the latter quantity is expanded in terms of the eigenvectors of A_k. Consequently, if $A_k = I_q$ for $1 \le k \le t < s$ and $A_k \ne I_q$ otherwise but still lies in $\mathcal{A}_{\text{sym}}$, then $\hat{M}(A)$ shrinks $\hat{M}_{\text{ls}}$ toward $\sum_{k=1}^t P_k Y = X_t X_t^+ Y$, the least squares estimator of M under the constraint $M \in \mathcal{R}(X_t)$.

Remark 3 The candidate estimators $\{\hat{M}(A)\colon A \in \mathcal{A}_s\}$ are regularized estimators of M because they can be derived as generalized penalized least squares estimators or limits thereof. Indeed, let the $\{C_k\colon 1 \le k \le s\}$ be specified penalty matrices of row dimension q and let $C = [C_1, C_2, \ldots, C_s]$. The *affinely penalized least squares candidate estimator* $\hat{M}_{\text{pls}}$ minimizes

$$\begin{aligned} T(M) &= |Y - M|^2 + \sum_{k=1}^s |P_k M C_k|^2 \\ &= \operatorname{tr}(Y'Y) - 2\operatorname{tr}(M'Y) + \operatorname{tr}(M'M) + \sum_{k=1}^s \operatorname{tr}(M C_k C_k' M' P_k) \end{aligned} \tag{32}$$

over all M that satisfy the linear model constraint $M = XB$ in (1). This generalization of ordinary penalized least squares has s penalty terms in which the matrices $\{C_k\colon 1 \le k \le s\}$ replace scalar penalty weights.

We will show that

$$\hat{M}_{\text{pls}} = \sum_{k=1}^{s} P_k Y (I_q + C_k C_k')^{-1}. \tag{33}$$

To verify this, observe that $M = XB = XX^+M$ for some $d \times q$ matrix B if and only if $M = XX^+G = \sum_{k=1}^{s} P_k G$ for some $n \times q$ matrix G. By substituting the last expression for M into the right side of (32) and then using derivative formulae in Lütkepohl (1996), we obtain

$$\partial T(M)/\partial G = 2\left(-\sum_{k=1}^{s} P_k Y + \sum_{k=1}^{s} P_k G + \sum_{k=1}^{s} P_k G C_k C_k'\right). \tag{34}$$

The derivative in (34) vanishes if and only if $\sum_{k=1}^{s} P_k[G(I_q + C_k C_k') - Y] = 0$ or, equivalently, if and only if $P_k G(I_q + C_k C_k') = P_k Y$ for every k. Hence, the minimizing value $\hat{G}$ of G satisfies $P_k \hat{G} = P_k Y (I_q + C_k C_k')^{-1}$ for every k. It follows that $\hat{M}_{\text{pls}} = \sum_{k=1}^{s} P_k \hat{G}$ is given by (33).

As C_k varies, $(I_q + C_k C_k')^{-1}$ generates all matrices in $\mathcal{A}_{\text{sym}}^s$ whose eigenvalues lie in (0, 1]. Equations (33) and (18) indicate that each of the candidate estimators $\{\hat{M}(A)\colon A \in \mathcal{A}_{\text{sym}}^s\}$ can be expressed either as a penalized least squares estimator or as a limit of such.

3 Adaptive estimators

This section devises adaptive estimators that are realizable data-based approximations to the oracle estimators derived in Sect. 2. The oracle construction is modified by replacing the unknown parameters τ and W in the risk function $r(A, \tau, W)$ with estimators. The resulting adaptive estimators coincide with the estimators $\hat{M}_{\text{pro}}$ and $\hat{M}_{\text{sym}}$ discussed in the Introduction. It will be seen in Sect. 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 in a sense to be made precise in Theorem 3. The strategy is to estimate the unknown risk function $r(A, \tau, W)$ by plugging in the estimators $\hat{\tau}$ and $\hat{W}$ defined by (8) and (10).

Let

$$\hat{A}_k = (\hat{\tau}_k I_q + \hat{W}_k)^{-1} \hat{W}_k = I_q - \hat{\tau}_k (\hat{\tau}_k I_q + \hat{W}_k)^{-1} = \hat{W}_k (\hat{\tau}_k I_q + \hat{W}_k)^{-1}, \tag{35}$$

a matrix that lies in $\mathcal{A}_{\text{sym}}$. Let $\hat{A} = [\hat{A}_1, \ldots, \hat{A}_s]$. For every $A \in \mathcal{A}_{\text{sym}}^s$, define the *estimated risk* of candidate estimator $\hat{M}(A)$ by analogy with (23) and (24):

$$r(A, \hat{\tau}, \hat{W}) = \sum_{k=1}^{s} \rho(A_k, \hat{\tau}_k, \hat{W}_k), \tag{36}$$

where

$$\begin{aligned}\rho(A_k, \hat{\tau}_k, \hat{W}_k) &= \operatorname{tr}[\hat{\tau}_k A_k^2 + (I_q - A_k)^2 \hat{W}_k] \\ &= \operatorname{tr}[(A_k - \hat{A}_k)^2 (\hat{\tau}_k I_q + \hat{W}_k)] + \hat{\tau}_k \operatorname{tr}(\hat{A}_k).\end{aligned} \tag{37}$$

Corresponding to the oracle estimators discussed in Sect. 2.2 are the following adaptive estimators, obtained by minimizing estimated risk over $\mathcal{A}_{\text{sym}}$ and over $\mathcal{A}_{\text{pro}}$:

- The *adaptive symmetric affine estimator* of M is $\hat{M}_{\text{sym}} = \hat{M}(\hat{A}_{\text{sym}})$, where

$$\hat{A}_{\text{sym}} = \operatorname*{argmin}_{A \in \mathcal{A}_{\text{sym}}^s} r(A, \hat{\tau}, \hat{W}) = [\hat{A}_{\text{sym},1}, \ldots, \hat{A}_{\text{sym},s}]. \tag{38}$$

- The *adaptive projection estimator* of M is $\hat{M}_{\text{pro}} = \hat{M}(\hat{A}_{\text{pro}})$, where

$$\hat{A}_{\text{pro}} = \operatorname*{argmin}_{A \in \mathcal{A}_{\text{pro}}^s} r(A, \hat{\tau}, \hat{W}) = [\hat{A}_{\text{pro},1}, \ldots, \hat{A}_{\text{pro},s}]. \tag{39}$$

The estimated risks of these two estimators are, respectively,

$$\hat{R}(\hat{M}_{\text{sym}}) = r(\hat{A}_{\text{sym}}, \hat{\tau}, \hat{W}), \quad \hat{R}(\hat{M}_{\text{pro}}) = r(\hat{A}_{\text{pro}}, \hat{\tau}, \hat{W}). \tag{40}$$

The next theorem shows that these adaptive estimators coincide with the adaptive projection and adaptive symmetric affine estimators discussed in Sect. 1 and provides formulae for their estimated risks.

Theorem 2 *The following expressions hold:*

$$\begin{aligned}\hat{A}_{\text{sym,k}} &= \hat{A}_k = \sum_{j=1}^{q} \hat{w}_{kj} (\hat{\tau}_k + \hat{w}_{kj})^{-1} \hat{v}_{kj} \hat{v}_{kj}' \\ \hat{M}_{\text{sym}} &= \sum_{k=1}^{s} P_k Y \sum_{j=1}^{q} \hat{A}_{\text{sym},k} = \sum_{k=1}^{s} \sum_{j=1}^{q} \hat{w}_{kj} (\hat{\tau}_k + \hat{w}_{kj})^{-1} \hat{l}_{kj} \hat{u}_{kj} \hat{v}_{kj}' \\ \hat{R}(\hat{M}_{\text{sym}}) &= \sum_{k=1}^{s} \hat{\tau}_k \operatorname{tr}(\hat{A}_{\text{sym},k}) = \sum_{k=1}^{s} \sum_{j=1}^{q} \hat{\tau}_k \hat{w}_{kj} (\hat{\tau}_k + \hat{w}_{kj})^{-1},\end{aligned} \tag{41}$$

with $\hat{A}_k$ defined in (35). *Moreover,*

$$\begin{aligned}
\hat{A}_{\text{pro},k} &= \sum_{j:\ \hat{w}_{kj} > \hat{\tau}_k} \hat{v}_{kj}\hat{v}'_{kj} \\
\hat{M}_{\text{pro}} &= \sum_{k=1}^{s} P_k Y \hat{A}_{\text{pro},k} = \sum_{k=1}^{s} \sum_{j:\hat{w}_{kj} > \hat{\tau}_k} \hat{l}_j \hat{u}_{kj} \hat{v}'_{kj} \\
\hat{R}(\hat{M}_{\text{pro}}) &= \sum_{k=1}^{s} \sum_{j=1}^{q} \min\{\hat{\tau}_k, \hat{w}_{kj}\}.
\end{aligned} \tag{42}$$

Proof Because $\hat{W}_k$ was constructed to lie in $\mathcal{A}_{\text{sym}}$, just like W_k, the argument that proved Theorem 1 also works for Theorem 2. The rightmost expressions for $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{pro}}$ use the singular values decomposition of $P_k Y$, given in (7). □

It follows from the estimated risk formulae in Theorem 2 that

$$2^{-1}\hat{R}(\hat{M}_{\text{pro}}) \le \hat{R}(\hat{M}_{\text{sym}}) \le \hat{R}(\hat{M}_{\text{pro}}) \le \hat{R}(\hat{M}_{\text{ls}}) = q\hat{\sigma}^2. \tag{43}$$

Section 4 shows that each estimated risk in (43) converges asymptotically, as p tends to infinity, to the corresponding actual risk, which asymptotically equals the oracle risk. In this sense, the estimated risk inequalities (43) approximate the oracle risk inequalities (31).

3.2 Examples

The examples of model (1) in this section have two purposes:

- To relate, by specialization, the adaptive estimator $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{pro}}$ to earlier work on shrinkage estimators;
- To illustrate ways in which $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{pro}}$ go beyond the earlier procedures. The case study in Sect. 4.3 develops this point on data.

Example 1 (Multiple scalar shrinkage). Consider the scalar response case, $q = 1$. We write $Y = y$, $M = m$, both column vectors, and $A_k = a_k$, a scalar for each k. Then, $\hat{W}_k$ is the non-negative scalar $\hat{w}_k = [p^{-1}(P_k y)'(P_k y) - \tau_k]_+$. The adaptive affine shrinkage estimator of m is

$$\hat{m}_{\text{sym}} = \sum_{k=1}^{s} \hat{w}_k(\hat{\tau}_k + \hat{w}_k)^{-1} P_k y \tag{44}$$

and it has estimated risk

$$\hat{R}(\hat{m}_{\text{sym}}) = \sum_{k=1}^{s} \hat{\tau}_k \hat{w}_k(\hat{\tau}_k + \hat{w}_k)^{-1}. \tag{45}$$

It may be verified that, as p tends to infinity, $\hat{m}_{\text{sym}}$ converges to a multiple scalar shrinkage estimator constructed by Stein (1966).

In this scalar case, the adaptive projection estimator of m is the vector

$$\hat{m}_{\text{pro}} = \sum_{k:\, \hat{w}_k > \hat{\tau}_k} P_k y, \tag{46}$$

with estimated risk

$$\hat{R}(\hat{m}_{\text{pro}}) = \sum_{k=1}^{s} \min\{\hat{\tau}_k, \hat{w}_k\}. \tag{47}$$

The projection estimator $\hat{m}_{\text{pro}}$ is a multiple submodel selection estimator that simplifies $\hat{m}_{\text{sym}}$ as follows: the shrinkage factor $\hat{w}_k(\hat{\tau}_k + \hat{w}_k)^{-1}$ in (44) is replaced by 1 whenever it exceeds $1/2$ and is replaced by 0 otherwise. The asymptotics in Sect. 4 establish that the risk estimators (45) and (47) both converge to the respective true risks under model (1).

Example 2 (Efron–Morris estimator and total least squares). Let $n = p$, $X = I_p$, and $s = 1$. For simplicity, suppose that σ^2 is known to be 1, so that $\hat{\sigma}^2 = 1$. Then $P_1 = XX^+ = I_p$, $\tau_1 = 1$, and $\hat{W}_1$ adjusts $\breve{W}_1 = p^{-1}Y'Y - I_q$ to be positive semidefinite. In this case,

$$\hat{M}_{\text{sym}} = Y\hat{W}_1(I_q + \hat{W}_1)^{-1}. \tag{48}$$

is a modification of the estimator $Y\breve{W}_1(I_q + \breve{W}_1)^{-1} = Y[I_q - p(Y'Y)^{-1}]$. For p much larger than q, this second expression nearly coincides with the Efron–Morris estimator

$$\hat{M}_{\text{EM}} = Y[I_q - (p - q - 1)(Y'Y)^{-1}], \tag{49}$$

developed from a different perspective by Efron and Morris (1972). Under asymptotics where p tends to infinity while response dimension q stays fixed, $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{EM}}$ have the same asymptotic risk.

Over certain neighborhoods of $M = 0$ that are defined in terms of the eigenvalues of W_1, the estimator $\hat{M}_{\text{sym}}$ is asymptotically minimax as p tends to infinity. The least squares estimator $\hat{M}_{\text{ls}}$ is not. This result in Beran (1999) applies the Pinsker (1980) theorem to a canonical transformation of the present example.

In the present example, $\hat{M}_{\text{pro}}$, given by (43) with $s = 1$, is the adaptive total least squares estimator for M of rank $\hat{h}_1 = \#\{j : \hat{w}_{1j} > \hat{\tau}_1\}$. The total least squares problem and its solution through the singular value decomposition were described by Golub and Van Loan (1980, 1996), by Van Huffel and Vandewalle (1991), and by Van Huffel (2004). The asymptotics in Sect. 4 provide a risk rationale for choosing the rank of the total least squares fit to be $\hat{h}_1$. As in Example 1, $\hat{M}_{\text{pro}}$ is a simplification of the modified Efron–Morris estimator. The risk penalty incurred by $\hat{M}_{\text{pro}}$ relative to $\hat{M}_{\text{sym}}$ can be estimated asymptotically as p increases. For another treatment of this example, see Beran (2007).

Example 2 reveals, more generally, that $\hat{M}_{\text{sym}}$ is the sum of s modified Efron–Morris estimators, one fitted to each $P_k Y$. Similarly, $\hat{M}_{\text{pro}}$ is the sum of s adaptive total least squares estimators, one fitted to each $P_k Y$.

Example 3 (Two-way MANOVA model). Consider the two-way layout with q-variate responses where factors 1 and 2 have, respectively, p_1 and p_2 levels. For each factor level combination (i, j), the observed responses are $1 \times q$ row vectors $\{y_{ijh} : 1 \le h \le n_{ij}\}$. The model asserts that

$$y_{ijh} = \beta_{ij} + e_{ijh}, \quad 1 \le h \le n_{ij}, \tag{50}$$

where β_{ij} is a constant $1 \times q$ vector and e_{ijh} is a random $1 \times q$ error vector.

Let $n = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} n_{ij}$ and $p = p_1 p_2$. Form the $n \times q$ response matrix Y by stacking the row vectors $\{y_{ijh}\}$ in the following order: replication label h varies most quickly, the level i of factor 1 varies next most quickly, and the level j of factor 2 varies most slowly. Form the error matrix E analogously. Form the $p \times q$ constant matrix B by stacking the row vectors $\{\beta_{ij}\}$ in the following order: the level i of factor 1 varies most quickly, and the level j of factor 2 varies most slowly. Let X be the $n \times p$ data incidence matrix, a matrix with orthogonal columns whose elements are either 0 or 1, that links each component of B to the corresponding responses in Y. Then Y in this example satisfies model (1)

For $i = 1, 2$, consider the $p_i \times 1$ vector $u_i = p_i^{-1/2}(1, \ldots, 1)'$ and the orthogonal projections $J_i = u_i u_i'$ and $H_i = I_{p_i} - J_i$. Let

$$Q_0 = J_2 \otimes J_1, \quad Q_1 = J_2 \otimes H_1, \quad Q_2 = H_2 \otimes J_1, \quad Q_{12} = H_2 \otimes H_1. \tag{51}$$

The two-way MANOVA decomposition of the cell means B into overall means, main effects, and interactions is the identity $B = Q_0 B + Q_1 B + Q_2 B + Q_{12} B$. Customary nested submodel design matrices for two-way MANOVA are

$$X_1 = X Q_0, \quad X_2 = X(Q_0 + Q_1) \text{ or } X(Q_0 + Q_2), \quad X_3 = X(Q_0 + Q_1 + Q_2). \tag{52}$$

In this example, $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{pro}}$ provide superior estimation of $M = XB$.

Example 4 (Multivariate regression). Let $x = (x_1, x_2, \ldots, x_n)'$ be a non-random covariate vector associated with the rows of the response matrix Y. In model (1), let X be the matrix $[x^0, x^1, \ldots, x^{p-1}]$, the powers being computed componentwise. For $1 \le k \le s - 1$, let X_k be the submatrix consisting of the first c_k columns of X,where $1 \le c_1 \le \cdots \le c_{s-1} \le p - 1$. By construction, $\hat{M}_{\text{sym}}$ asymptotically dominates in risk the least squares polynomial fits of degrees $c_1 - 1, c_2 - 1, \ldots, p - 1$. The polynomial basis in this example can be replaced by other bases, such as discrete spline bases.

4 Asymptotic theory

This section develops conditions under which the adaptive estimators behave asymptotically in loss and risk like their oracle counterparts and under which the estimated risks of the adaptive estimators converge to their actual risks. The usefulness of $\hat{M}_{\text{sym}}$ and $\hat{M}_{\text{pro}}$ for data analysis is explored in the case study of Sect. 4.3. The assumption that the rows of the error matrix E are independent $N(0, \sigma^2 I_q)$ random vectors is important to the proofs below because it ensures the distributional property that follows (56). For the special case when $s = 1$ and $P_1 = I_n$, analogs of Theorems 4.1 and 4.2 have been proved by Beran (2007), using an empirical process approach that does not require a multivariate normality assumption. However, the approach of that paper does not suffice to handle the case $s > 1$.

4.1 Convergence of loss, risk, and estimated risk

The first theorem establishes that the loss and the estimated risk of candidate estimator $\hat{M}(A)$ both converge, uniformly over $A \in \mathcal{A}^s_{\text{sym}}$, to the estimator's risk as p tends to infinity.

Theorem 3 *Suppose that model* (1) *holds and that* $\hat{\sigma}^2$ *satisfies, for every finite* $c > 0$,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0. \tag{53}$$

Then, for every finite $c > 0$ *and fixed integers* q *and* s,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E} \sup_{A\in\mathcal{A}^s_{\text{sym}}} |L(\hat{M}(A), M) - r(A, \tau, W))| = 0 \tag{54}$$

and

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E} \sup_{A\in\mathcal{A}^s_{\text{sym}}} |r(A, \hat{\tau}, \hat{W}) - r(A, \tau, W))| = 0. \tag{55}$$

Proof Each $n \times n$ projection matrix P_k has spectral decomposition $P_k = U_k U_k'$, where U_k is $n \times t_k$ with $t_k = \mathrm{tr}(P_k)$ and $U_k' U_k = I_{t_k}$. Let

$$Y_k = U_k' Y, \quad M_k = U_k' M, \quad E_k = U_k' E. \tag{56}$$

Under model (1), the elements of the $t_k \times q$ matrix E_k are independent identically distributed $N(0, \sigma^2)$ random variables. Moreover, $W_k = p^{-1} M_k' M_k$ and $\tau_k = p^{-1}\sigma^2 t_k$.

Let $\hat{F}_k = p^{-1} E_k' M_k$ and $\hat{G}_k = p^{-1} E_k' E_k - \tau_k I_q$. Using the respective definitions (8) and (13) of $\check{W}_k$ and W_k gives

$$\begin{aligned} \check{W}_k - W_k &= p^{-1} Y_k' Y_k - \hat{\tau}_k I_q - p^{-1} M_k' M_k \\ &= \hat{F}_k + \hat{F}_k' + \hat{G}_k + (\tau_k - \hat{\tau}_k) I_q. \end{aligned} \tag{57}$$

Because the elements of E_k are independent $N(0, \sigma^2)$ variables and $|M_k|^2 \le |M|^2$,

$$\sup_{p^{-1}|M|^2<c} \mathrm{E}|\hat{F}_k|^2 = O(p^{-1}), \qquad \sup_{p^{-1}|M|^2<c} \mathrm{E}|\hat{G}_k|^2 = O(p^{-1}). \tag{58}$$

From (53) and the definitions (8) and (13) of $\hat{\tau}_k$ and τ_k,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|\hat{\tau}_k - \tau_k| = 0. \tag{59}$$

Recall the definitions in (10) of $\hat{W}_k = \breve{W}_{k+}$ and $\breve{W}_{k-}$. Because the matrix W_k is positive semidefinite, $\mathrm{tr}(W_k \breve{W}_{k-}) = \sum_{j=1}^q \breve{w}_{kj-} \hat{v}_{kj}' W_k \hat{v}_{kj} \le 0$. Consequently,

$$\begin{aligned} |\breve{W}_k - W_k|^2 &= |\breve{W}_{k+} - W_k|^2 + |\breve{W}_{k-}|^2 - 2\,\mathrm{tr}(W_k \breve{W}_{k-}) \\ &\ge |\breve{W}_{k+} - W_k|^2 + |\breve{W}_{k-}|^2 \\ &\ge |\breve{W}_{k+} - W_k|^2 = |\hat{W}_k - W_k|^2. \end{aligned} \tag{60}$$

From (57) to (60),

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|\hat{W}_k - W_k| \le \lim_{p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|\breve{W}_k - W_k| = 0. \tag{61}$$

Let $\{B_i : 1 \le i \le h\}$ be any $q \times q$ matrices. Then, $|\,\mathrm{tr}[\prod_{i=1}^h B_i]| \le \prod_{i=1}^h |B_i|$ (cf. Lütkepohl 1996, p. 111). If $A_k \in \mathcal{A}_{\mathrm{sym}}$, its spectral decomposition implies that $|A_k^2| \le q^{1/2}$, $|(I_q - A_k)^2| \le q^{1/2}$, and $|A_k(I_q - A_k)| \le 4^{-1} q^{1/2}$. It follows from (23) that

$$\begin{aligned} |r(A, \tau, \hat{W}) - r(A, \tau, W)| &= \left| \sum_{k=1}^s \mathrm{tr}[(I_q - A_k)^2 (\hat{W}_k - W_k)] \right| \\ &\le \sum_{k=1}^s |(I_q - A_k)^2| |\hat{W}_k - W_k| \le q^{1/2} \sum_{k=1}^s |\hat{W}_k - W_k|. \end{aligned} \tag{62}$$

On the other hand,

$$|r(A, \hat{\tau}, \hat{W}) - r(A, \tau, \hat{W})| = \left| \sum_{k=1}^s (\hat{\tau}_k - \tau_k)\,\mathrm{tr}(A_k^2) \right| \le q|\hat{\sigma}^2 - \sigma^2|. \tag{63}$$

Inequalities (61), (62), (63) and theorem condition (53) together imply limit (55).

Let $J = |L(\hat{M}(A), M) - r(A, \tau, W)|$. From (19) and (23),

$$J = \left| \sum_{k=1}^{s} \left\{ p^{-1}|Y_k A_k - M_k|^2 - \operatorname{tr}\left[\tau_k A_k^2 + (I - A_k)^2 p^{-1} M_k' M_k\right] \right\} \right|$$
$$= \left| \sum_{k=1}^{s} \operatorname{tr}\left[A_k^2 \hat{G}_k + 2(A_k^2 - A_k)\hat{F}_k'\right] \right| \le q^{1/2} \sum_{k=1}^{s} \left[|\hat{G}_k| + 2^{-1}|\hat{F}_k|\right]. \quad (64)$$

This and (58) imply limit (54). □

From Theorem 3 follows

Theorem 4 *Suppose that the assumptions of Theorem 3 hold. Let T denote any one of $L(\hat{M}_{\text{sym}}, M)$ or $L(\tilde{M}_{\text{sym}}, M)$ or $\hat{R}(\hat{M}_{\text{sym}})$. Then, for every finite $c > 0$ and for every fixed integers q and s,*

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E}|T - R(\tilde{M}_{\text{sym}}, M, \sigma^2)| = 0. \quad (65)$$

Consequently,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E}|L(\hat{M}_{\text{sym}}, M) - L(\tilde{M}_{\text{sym}}, M)| = 0 \quad (66)$$

and

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} |R(\hat{M}_{\text{sym}}, M, \sigma^2) - R(\tilde{M}_{\text{sym}}, M, \sigma^2)| = 0. \quad (67)$$

Replacing the subscript "sym" by the subscript "pro" in these assertions is valid.

Proof Recall that the adaptive affine symmetric shrinkage estimator $\hat{M}_{\text{sym}} = \hat{M}(\hat{A})$ with $\hat{A} = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_k]$ while the oracle affine shrinkage estimator $\tilde{M}_{\text{sym}} = \tilde{M}(\tilde{A})$ with $\tilde{A} = [\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_k]$. That (65) holds for $T = L(\tilde{M}_{\text{sym}}, M)$ is immediate from (54).

Limit (55) implies that, for $A = \hat{A}$ and for $A = \tilde{A}$,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E}|r(A, \hat{\tau}, \hat{W}) - r(A, \tau, W)| = 0 \quad (68)$$

and, through the respective minimizing properties of $\hat{A}$ and $\tilde{A}$,

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E}|r(\hat{A}, \hat{\tau}, \hat{W}) - r(\tilde{A}, \tau, W)| = 0. \quad (69)$$

Because $r(\tilde{A}, \tau, W) = R(\tilde{M}_{\text{sym}}, M, \sigma^2)$ and $r(\hat{A}, \hat{\tau}, \hat{W}) = \hat{R}(\hat{M}_{\text{sym}})$, (69) establishes that limit (65) holds for $T = \hat{R}(\hat{M}_{\text{sym}})$.

Moreover, (68) and (69) entail

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|r(\hat{A},\tau,W) - r(\tilde{A},\tau,W)| = 0. \tag{70}$$

On the other hand, limit (54) gives

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|L(\hat{M}(\hat{A}),M) - r(\hat{A},\tau,W)| = 0. \tag{71}$$

Combining (71) with (70) yields (65) for $T = L(\hat{M}_{\mathrm{sym}}, M)$. Limit (66) is immediate from (65) and implies (67).

Finally, because $\mathcal{A}_{\mathrm{pro}} \subset \mathcal{A}_{\mathrm{sym}}$, the argument just given applies as well to the adaptive and oracle projection estimators. □

By (66) and (67), the loss or risk of the adaptive estimator $\hat{M}_{\mathrm{sym}}$ or $\hat{M}_{\mathrm{pro}}$ converges asymptotically to the loss or risk of the corresponding oracle estimator. Moreover, by (65), the plug-in risk estimator $\hat{R}(\hat{M}_{\mathrm{sym}})$ or $\hat{R}(\hat{M}_{\mathrm{pro}})$ converges asymptotically to the actual loss or risk of the corresponding adaptive estimator. Thus, it is meaningful to compare the estimated risks of $\hat{M}_{\mathrm{sym}}$ and $\hat{M}_{\mathrm{pro}}$, given in Theorem 2, with one another and with the estimated risk $q\hat{\sigma}^2$ of the least squares estimator $\hat{M}_{\mathrm{ls}} = XX^+Y$.

4.2 Variance estimation

This section considers three estimators for the variance σ^2, giving conditions for each under which it is L_1-consistent. The first estimator is useful when n substantially exceeds p. The other two estimators express alternative approaches available when $n = p$ and $X = I_p$.

Least squares variance estimator. Let $\hat{E} = Y - \hat{M}_{\mathrm{ls}} = (I_n - XX^+)Y$ be the matrix of residuals after the least squares fit. Write $\hat{E} = [\hat{e}_1, \ldots, \hat{e}_n]'$, where $\hat{e}_i'$ denotes the ith row of $\hat{E}$. The classical *least squares estimator* of σ^2 is

$$\hat{\sigma}^2_{\mathrm{ls}} = [q(n-p)]^{-1}|\hat{E}|^2 = [q(n-p)]^{-1}\sum_{i=1}^{n}|\hat{e}_i|^2. \tag{72}$$

Because the residual matrix $\hat{E} = (I_n - XX^+)E$, it does not depend on M. The estimator $\hat{\sigma}^2_{\mathrm{ls}}$ is unbiased for σ^2 and has the L_1-consistency property

$$\lim_{n-p\to\infty} \sup_{p^{-1}|M|^2\le c} \mathrm{E}|\hat{\sigma}^2_{\mathrm{ls}} - \sigma^2| = 0. \tag{73}$$

First difference variance estimator. Express $Y = [y_1, \ldots . y_n]'$ and $M = [m_1, \ldots, m_n]'$ in terms of their rows. When $n = p$, $X = I_p$, and row-to-row variation

in M is expected to be slow, it is reasonable to consider the *first difference variance estimator*

$$\hat{\sigma}^2_{\text{dif}} = [2q(p-1)]^{-1} \sum_{i=2}^{p} |y_i - y_{i-1}|^2. \tag{74}$$

The bias of this estimator is $b(M) = \sum_{i=2}^{p} |m_i - m_{i-1}|^2$. Let $\{\epsilon_p\}$ be any sequence such that $\lim_{p\to\infty} \epsilon_p = 0$. Then $\hat{\sigma}^2_{dif}$ has the L_1-consistency property

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c,\, b(M) \le \epsilon_p} \mathrm{E}|\hat{\sigma}^2_{\text{dif}} - \sigma^2| = 0. \tag{75}$$

Variance estimation based on higher-order differences of the $\{y_i\}$ is analogous to that just outlined.

Smallest singular values variance estimator. When $n = p$, $X = I_p$, and M is expected to be nearly singular, it is reasonable to consider the *smallest singular values variance estimator*

$$\hat{\sigma}^2_{\text{ssv}} = p^{-1} \sum_{k=1}^{s} \hat{l}^2_{kq}. \tag{76}$$

Recalling (7), (12), and the definition (13) of τ_k, observe that

$$\begin{aligned}
&|p^{-1}\hat{l}^2_{kq} - p^{-1} l^2_{kq} - \tau_k| \\
&= \left| \min_{|v|=1} p^{-1} v'(P_k Y)' P_k Y v - \min_{|v|=1} [p^{-1} v'(P_k M') P_k M v - \tau_k v' v] \right| \\
&\le \sup_{|v|=1} |v'[p^{-1}(P_k Y)' P_k Y - p^{-1}(P_k M)' P_k M - \tau_k I_q] v| \\
&\le |p^{-1}(P_k Y)' P_k Y - p^{-1}(P_k M)' P_k M - \tau_k I_q|.
\end{aligned} \tag{77}$$

By a calculation akin to (57) and (58), it follows from (76) and (77) that

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c} \mathrm{E} \left| \hat{\sigma}^2_{\text{ssv}} - p^{-1} \sum_{k=1}^{s} l^2_{kq} - \sigma^2 \right| = 0. \tag{78}$$

Let $\{\epsilon_p\}$ be any sequence such that $\lim_{p\to\infty} \epsilon_p = 0$. Then $\hat{\sigma}^2_{\text{ssv}}$ is L_1-consistent in the sense that

$$\lim_{p\to\infty} \sup_{p^{-1}|M|^2 \le c,\, p^{-1}\sum_{k=1}^{s} l^2_{kq} \le \epsilon_p} \mathrm{E}|\hat{\sigma}^2_{\text{ssv}} - \sigma^2| = 0. \tag{79}$$

Variance estimators based on the smallest singular values of only some of the $\{P_k Y\}$, can be defined and analyzed similarly.

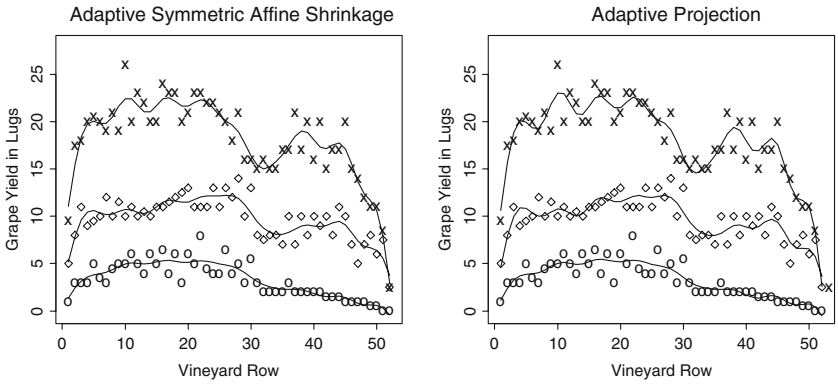

Fig. 1 Linearly interpolated adaptive symmetric affine shrinkage and adaptive projection fits to the trivariate grape yields. The $\{P_k\}$ are defined in terms of a third-difference penalty basis

Theorems 3 and 4 are readily adjusted to use the consistency properties in (73) or (75) or (79) in place of (53).

4.3 Viticultural case study

The data matrix Y in this case study is 52×3. Row i of Y reports the grape yields harvested in three different years from row i of a vineyard with 52 rows. The data is taken from Chatterjee et al. (1995). The grape yields, measured in lugs of grapes harvested from each row, are plotted in Fig. 1, using a different plotting character for each of the three years. Both year-to-year and row-to-row changes in viticulture affect the observed yields. The analysis seeks to bring out patterns in the row harvest yields that persist across years.

Let D denote a matrix with n columns such that, for any column vector $x \in R^n$, the penalty function $\pi(x) = |Dx|^2 = x'D'Dx$ measures the roughness of x. For instance, D may be the t-th difference operator, defined below, when the components of x are equally spaced. If adjacent coordinates of x vary slowly in terms of the difference operator, then the penalty $\pi(x)$ is small. Construct an orthogonal *penalty basis* for R^n as follows:

- Find the unit vector $\gamma \in R^n$ that minimizes the penalty $\pi(\gamma)$. This smoothest vector is γ_n, where the $\{\gamma_j : 1 \le j \le n\}$ are the eigenvectors of $D'D$, ordered so that the associated eigenvalues go from largest to smallest. In case of tied eigenvalues, the corresponding eigenvectors are selected and ordered by imposing an additional rule.
- Find the unit vector $\gamma \in R^n$ that, subject to the constraint $\gamma \perp \gamma_n$, minimizes the penalty $\pi(\gamma)$. This second smoothest vector is γ_{n-1}.
- Continue sequential constrained minimization in order to obtain the orthogonal penalty basis matrix $U = [u_1, u_2, \ldots, u_n]$, where $u_i = \gamma_{n-i+1}$.

The columns of U are the eigenvectors of $D'D$, ordered from the smallest to the largest eigenvalue of $D'D$.

Consider the $(n-1) \times n$ matrix $\Delta(n) = \{\delta_{i,j}\}$ in which $\delta_{i,i} = 1$, $\delta_{i,i+1} = -1$ for every i and all other entries are zero. Define the tth *difference operator* $D_t(n)$ through

$$D_1(n) = \Delta(n), \quad D_t(n) = \Delta(n-t+1)D_{t-1}(n) \text{ for } 2 \le t \le n-1. \tag{80}$$

In the present case study, where $n = 52$ and $q = 3$, we set D to be the third difference operator $D_3(52)$. The penalty function π then penalizes departures from locally quadratic behavior and the columns of U form a discrete spline basis.

To specify the multivariate linear model and the submodels, set $p = 20$ and $s = 12$. Define X to be the first 20 columns of U. Because these columns are orthonormal, $X^+ = X'$. Under this model, the least squares estimate (60) of σ^2 is $\hat{\sigma}^2_{\text{ls}} = 1.839$, with 96 degrees of freedom. For $1 \le k \le 11$, define X_k to be the first k columns of X. Then $P_k = u_k u_k'$ for $1 \le k \le 11$ and $P_{12} = XX' - X_{11}X_{11}' = \sum_{k=12}^{20} u_k u_k'$.

Figure 1 plots the adaptive estimates of mean grape yields for each year, adding linear interpolation between adjacent estimated row means to display their trend. The adaptive symmetric affine shrinkage and adaptive projection estimates are similar visually. Both estimates reveal shared patterns in the fitted harvests for each of the three years. Large dips in estimated mean grape yield occur in the outermost rows and near row 33; and smaller fluctuations occur elsewhere.

The estimated risk $\hat{R}(\hat{M}_{\text{sym}})$ of the adaptive symmetric affine shrinkage estimator is 1.504 while the larger estimated risk $\hat{R}(\hat{M}_{\text{pro}})$ of the adaptive projection estimator is 1.833. Both estimated risks are less than one third of the estimated risk $\hat{R}(\hat{M}_{\text{ls}}) = 3\hat{\sigma}^2_{\text{ls}} = 5.517$ of the least squares estimator. The ordering of the estimated risks agrees with the inequalities (43). Moreover, Theorem 4 provides grounds for regarding the estimated risks as useful guides to the relative performance of the three estimators.

In this example, the shrinkage factors $\{\hat{w}_{kj}(\hat{\tau}_k + \hat{w}_{kj})^{-1} \colon 1 \le k \le 12, 1 \le j \le 3\}$ that define $\hat{M}_{\text{sym}}$ through (16) are as follows. For $j = 1$, their values, as k ranges from 1 to 12 are: 1.00, 1.00, 1.00, 0.94, 0.99, 0.96, 0.96, 0.94, 0.40, 0.76, 0.90, 0.69. For $j > 1$, their values are all very near 0. Thus, $\hat{M}_{\text{sym}}$ essentially replaces each $P_k Y$ in the least squares fit $\hat{M}_{\text{ls}} = \sum_{k=1}^{12} P_k Y$ by its total least squares approximation of rank one and then shrinks these approximations by varying amounts. The shrinkage is greatest when $k = 9$ and remains notable when $k = 10$ or 12.

The 0–1 shrinkage factors $\{I[\hat{w}_{kj} > \hat{\tau}_k] \colon 1 \le k \le 12, 1 \le j \le 3\}$ that define $\hat{M}_{\text{pro}}$ through (11) display a closely related pattern. For $j = 1$, their values as k ranges from 1 to 12 are 1 when $k \ne 9$ and 0 when $k = 9$. For $j > 1$, their values are all 0. Thus, $\hat{M}_{\text{pro}}$ discards $P_9 Y$ from the least squares fit $\sum_{k=1}^{12} P_k Y$ and replaces very other $P_k Y$ by its total least squares approximation of rank one.

Setting $p = 20$ in specifying X was a design choice, made outside the theory of this paper, that mediates between sufficient richness in $\mathcal{R}(X)$ to represent the unknown, possibly irregular, dependence of mean grape yield on row number and sufficient degrees of freedom to estimate σ^2 adequately.

5 Correlated errors

The classical multivariate linear model asserts that the rows of the error matrix E in (1) are independent, identically distributed random vectors, each having a $N(0, \Sigma)$ distribution. Estimation in this model with more general covariance structure can be mapped into the special case treated in the previous sections. The least squares estimator of M is still XX^+Y. Suppose that the covariance matrix Σ is positive definite and known. Consider the *candidate affine shrinkage estimators*

$$\hat{M}(A, \Sigma) = \sum_{k=1}^{s} P_k Y \Sigma^{-1/2} A_k \Sigma^{1/2}, \tag{81}$$

where A lies in $\mathcal{A}^s_{\text{sym}}$ or in $\mathcal{A}^s_{\text{pro}}$. The quadratic risk of $\hat{M}(A, \Sigma)$ is defined to be

$$R(\hat{M}(A, \Sigma), M, \Sigma) = p^{-1}\mathrm{E}\left[\mathrm{tr}\{(\hat{M}(A, \Sigma) - M)\Sigma^{-1}(\hat{M}(A, \Sigma) - M)'\}\right]. \tag{82}$$

Let $Z = Y\Sigma^{-1/2}$, $N = M\Sigma^{-1/2}$, and $\hat{N}(A, \Sigma) = \hat{M}(A, \Sigma)\Sigma^{-1/2}$. Then

$$\hat{N}(A, \Sigma) = \sum_{k=1}^{s} P_k Z A_k \tag{83}$$

and

$$R(\hat{M}(A, \Sigma), M, \Sigma) = \mathrm{E}|\hat{N}(A, \Sigma) - N|^2 = p^{-1}\sum_{k=1}^{s}\mathrm{E}|P_k Z A_k - P_k N|^2. \tag{84}$$

The rows of Z are independent, identically distributed random vectors, each having a $N(0, I_q)$ distribution. Thus, when Σ is known, the foregoing analysis reduces the problem of oracle or adaptive estimation of M, under the linear model and loss function of this section, into the problem already treated in this paper when σ^2 equals 1.

If the covariance matrix Σ is unknown and n is substantially greater than p, we may replace Σ, in the definition of candidate estimator $\hat{M}(A, \Sigma)$ and in its estimated risk, by

$$\hat{\Sigma}_{ls} = (n - p)^{-1}Y'(I_n - XX^+)Y. \tag{85}$$

If both p and $n - p$ tend to infinity, arguments from Sect. 4 show that the loss and estimated risk of $\hat{M}(A, \hat{\Sigma}_{\text{ls}})$ converge together in probability, uniformly over all $A \in \mathcal{A}^s_{\text{sym}}$. Hence, the loss of the adaptive estimator $\hat{M}(\hat{A}_{\text{sym}}, \hat{\Sigma}_{\text{ls}})$, where $\hat{A}_{\text{sym}}$ minimizes the estimated risk of $\hat{M}(A, \hat{\Sigma}_{\text{ls}})$ over all $A \in \mathcal{A}^s_{\text{sym}}$, converges in probability to the minimum loss achievable over the candidate estimator class. Counterparts to Theorems 3 and 4 hold when the estimator of Σ has the L_1-consistency properties used in Beran (1999).

References

Beran, R. (1999). Superefficient estimation of multivariate trend. *Mathematical Methods of Statistics*, *8*, 166–180.

Beran, R. (2007). Adaptive estimators of a mean matrix: total least squares versus total shrinkage. *Econometric Theory*, (in press).

Chatterjee, S., Handcock, M. S., Simonoff, J. S. (1995) *A casebook for a first course in statistics and data analysis*. New York: Wiley

Efron, B., Morris, C. (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika*, *59*, 335–347.

Fuller, W. A. (1987). *Error measurement models*. New York: 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., Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, *17*, 883–893.

Golub, G. H., Van Loan, C. F. (1996). *Matrix computations* (3rd edition). Baltimore: Johns Hopkins University Press.

Lütkepohl, H. (1996). *Handbook of matrices*. Chicester: Wiley.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems of Information Transmission*, *16*, 120–133.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (pp. 197–206). Berkeley: University of California Press.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In F. N. David (Ed.), *Festschrift for Jerzy Neyman* (pp. 351–364). New York: Wiley.

Van Huffel, S. (2004). Total least squares and errors-in-variables modeling: bridging the gap between statistics, computational mathematics and engineering. In J. Antoch (Ed.), *Compstat 2004: Proceedings in computational statistics* (pp. 539–555). Heidelberg, Physica-Verlag.

Van Huffel, S., Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis*. Philadelphia: SIAM.