

# Superefficient Estimation of Multivariate Trend

Rudolf Beran\*

Department of Statistics  
University of California, Berkeley  
Berkeley, CA 94720–3860, USA

Revised June 1999

**Abstract.** The question of recovering a multiband signal from noisy observations motivates a model in which the multivariate data points consist of an unknown *deterministic* trend  $\Xi$  observed with multivariate Gaussian errors. A cognate *random* trend model suggests affine shrinkage estimators  $\hat{\Xi}_A$  and  $\hat{\Xi}_B$  for  $\Xi$ , which are related to an extended Efron-Morris estimator. When represented canonically,  $\hat{\Xi}_A$  performs componentwise James-Stein shrinkage in a coordinate system that is determined by the data. Under the original deterministic trend model,  $\hat{\Xi}_A$  and its relatives are asymptotically minimax in Pinsker’s sense over certain classes of subsets of the parameter space. In such fashion,  $\hat{\Xi}_A$  and its cousins dominate the classically efficient least squares estimator. We illustrate their use to improve on the least squares fit of the multivariate linear model.

*AMS classification:* 62H12, 62J05

*Keywords and phrases:* multivariate linear model, deterministic trend, risk estimator, minimum  $C_L$ , adaptive estimator, Efron-Morris estimator, asymptotic minimax, Pinsker bound.

**1. Introduction.** The least squares fit to a multivariate trend that is observed with error at many points is unsatisfactory because it emphasizes unbiasedness at the expense of risk. This paper develops adaptive affine shrinkage estimators that have two advantages: They asymptotically dominate least squares fits, pointwise in the parameter space; and they are asymptotically minimax over certain classes of subsets in the parameter space.

Consider the multivariate trend model in which we observe the independent  $p \times 1$  random vectors  $\{x_t: 1 \leq t \leq n\}$ , the distribution of  $x_t$  being  $N_p(\xi_t, \Sigma)$  with  $n$  greater than  $p$ . The  $p \times 1$  mean vectors  $\{\xi_t: 1 \leq t \leq n\}$  are unknown *constants*, as is the positive definite  $p \times p$  covariance matrix  $\Sigma$ . The observations  $\{x_t\}$  are organized into the  $n \times p$  data matrix  $X = (x_1, x_2, \dots, x_n)'$  whose expectation is the matrix  $\Xi = EX = (\xi_1, \xi_2, \dots, \xi_n)'$ .

---

\* Research supported in part by National Science Foundation Grant DMS95–30492 and by the Alexander von Humboldt Foundation.

Let  $\hat{\Xi} = (\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_n)'$  denote any estimator of  $\Xi$ . The quality of  $\hat{\Xi}$  is assessed through the quadratic loss

$$\begin{aligned} L_n(\hat{\Xi}, \Xi, \Sigma) &= (np)^{-1} \text{tr}[(\hat{\Xi} - \Xi)\Sigma^{-1}(\hat{\Xi} - \Xi)'] \\ &= (np)^{-1} \sum_{t=1}^n (\hat{\xi}_t - \xi_t)' \Sigma^{-1} (\hat{\xi}_t - \xi_t), \end{aligned} \quad (1.1)$$

where  $\text{tr}$  denotes the trace operator. The risk  $R_n(\hat{\Xi}, \Xi, \Sigma)$  is the expectation of this loss.

The adaptive estimator  $\hat{\Xi}_A$  developed in this paper has asymptotic risk as follows. Let

$$V = n^{-1} \sum_{t=1}^n \xi_t \xi_t' = n^{-1} \Xi' \Xi, \quad W = \Sigma^{-1/2} V \Sigma^{-1/2} \quad (1.2)$$

and denote the eigenvalues of  $W$  by  $\lambda_1(W) \geq \dots \geq \lambda_p(W) \geq 0$ . We will show, among other results, that for every finite positive  $r$ ,

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} |R_n(\hat{\Xi}_A, \Xi, \Sigma) - \tau(W)| = 0 \quad (1.3)$$

where

$$\tau(W) = 1 - p^{-1} \text{tr}[(I_p + W)^{-1}] < 1 \quad \forall \Xi, \Sigma. \quad (1.4)$$

The quantity  $\lambda_1(W)$  that defines the domain of  $\Xi$  in the supremum is a multivariate measure of signal-to-noise ratio. The asymptotic risk in (1.3) dominates the risk of the least squares estimator  $\hat{\Xi}_{LS} = X$ , which is 1 for every value of  $\Xi$  and  $\Sigma$ . Moreover, unlike  $\hat{\Xi}_{LS}$ , the adaptive affine estimator  $\hat{\Xi}_A$  turns out to be asymptotically minimax over certain classes of subsets of  $\Xi$  centered at  $\Xi = 0$ . In this sense,  $\hat{\Xi}_A$  is asymptotically superefficient relative to the classically efficient  $\hat{\Xi}_{LS}$ .

The construction of  $\hat{\Xi}_A$  involves the following steps. Let  $\hat{\Sigma}$  be an independent consistent estimator of  $\Sigma$  and let

$$\hat{V} = n^{-1} \sum_{t=1}^n x_t x_t' - \hat{\Sigma} = n^{-1} X' X - \hat{\Sigma}, \quad \hat{W} = \hat{\Sigma}^{-1/2} \hat{V} \hat{\Sigma}^{-1/2}. \quad (1.5)$$

Suppose that  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \geq -1$  are the eigenvalues of  $\hat{W}$  and  $\{\hat{\gamma}_j\}$  are corresponding eigenvectors. Letting  $[\cdot]_+$  denote the positive-part function, define

$$\hat{A} = \sum_{j=1}^p [\hat{\lambda}_j / (1 + \hat{\lambda}_j)]_+ \hat{\gamma}_j \hat{\gamma}_j'. \quad (1.6)$$

The adaptive estimator of  $\Xi$  is then

$$\hat{\Xi}_A = X \hat{\Sigma}^{-1/2} \hat{A} \hat{\Sigma}^{1/2}. \quad (1.7)$$

Expressions (1.6) and (1.7) reveal that  $\hat{\Xi}_A$  carries out componentwise James-Stein shrinkage in a canonical coordinate system for  $R^p$  that is estimated from the data. Multiple shrinkage in a *fixed* coordinate system was introduced by Stein [17]. Unlike the least squares estimator  $\hat{\Xi}_{LS}$ , the affine shrinkage estimator  $\hat{\Xi}_A$  uses  $\hat{W}$ , which estimates the matrix  $W$ , in order to reduce asymptotic risk.

The estimator  $\hat{\Xi}_A$  is more easily understood in a multivariate *random* trend model. Instead of the model described above, suppose that the  $\{\xi_t\}$  are independent random vectors, each having a  $N_p(0, V)$  distribution. Given  $\Xi$ , suppose that the  $\{x_t\}$  are conditionally independent, the conditional distribution of  $x_t$  being  $N_p(\xi_t, \Sigma)$ . If  $X$  is observed and  $V, \Sigma$  are known, then the minimum risk predictor of  $\Xi$  under loss (1.1) is  $\tilde{\Xi} = X\Sigma^{-1/2}\tilde{A}\Sigma^{1/2}$ , where

$$\tilde{A} = I_p - (I_p + W)^{-1} = W(I_p + W)^{-1}. \quad (1.8)$$

When  $W$  is nonsingular, it is also true that  $\tilde{A} = (I_p + W^{-1})^{-1}$ . Since  $\hat{A}$  is a consistent, positive semidefinite estimator of  $\tilde{A}$  in the random trend model,  $\hat{\Xi}_A$  is the natural empirical version of the minimum risk predictor for  $\Xi$ . Another consistent estimator of  $\tilde{A}$  is  $\check{A} = I_p - (I_p + \hat{W})^{-1}$  which, unlike  $\hat{A}$ , need not be positive semidefinite. This generates the alternative empirical predictor

$$\hat{\Xi}_B = X\hat{\Sigma}^{-1/2}\check{A}\hat{\Sigma}^{1/2} = X[I_p - n(X'X)^{-1}\hat{\Sigma}]. \quad (1.9)$$

Related predictors, albeit more complex, have been used to analyze multiband satellite image data (see [6]).

Up to second-order refinements,  $\hat{\Xi}_A$  and  $\hat{\Xi}_B$  are multivariate versions of James-Stein [7] estimators for univariate trend. Indeed, when  $p = 1$ ,  $\hat{W} = (n\hat{\sigma}^2)^{-1} \sum_{t=1}^n x_t^2 - 1 = \hat{\lambda}_1$  and  $\hat{\gamma}_1 = 1$ . Then

$$\hat{\Xi}_A = [1 - n\hat{\sigma}^2 / \sum_{t=1}^n x_t^2]_+ X, \quad \hat{\Xi}_B = [1 - n\hat{\sigma}^2 / \sum_{t=1}^n x_t^2] X. \quad (1.10)$$

Statistical folklore, supported by a growing number of results, posits that procedures with good behavior in models with many random parameters may also have desirable properties when those parameters are deterministic. Pfanzagl [13] discussed one aspect of the matter, estimation of a real parameter in the presence of many nuisance parameters, and reviewed earlier contributions to the literature. Another aspect is estimation of the entire high-dimensional parameter  $\Xi$  under the deterministic trend model described above. When  $\Sigma = \hat{\Sigma} = I_p$ , Efron and Morris [4] showed that a refinement of  $\hat{\Xi}_B$  is globally minimax and dominates  $\hat{\Xi}_{LS}$ . Bilodeau and Kariya [2] extended both the Efron-Morris estimator and its global asymptotic minimaxity to the case of unknown  $\Sigma$ . For details, see (3.12).

The aim of this paper is to study the performance of  $\hat{\Xi}_A$  as  $n$  tends to infinity with  $p$  fixed. Section 2 of the paper draws on Pinsker's [14] theorem to give an asymptotic minimax

bound for the estimation of  $\Xi$  over a certain rich class of subsets of the parameter space. The idealized estimator  $\tilde{\Xi}$ , suggested by the random trend model when  $V, \Sigma$  are known, is asymptotically minimax in this deterministic setting, unlike the least squares estimator  $\hat{\Xi}_{LS}$ . The results of Section 3, on the success of adaptation by minimizing estimated risk, entail limit (1.3) and the asymptotic minimaxity, over designated subsets, of  $\hat{\Xi}_A, \hat{\Xi}_B$ , and the extended Efron-Morris estimator. Section 4 describes how these estimators may be used to improve fitting of the multivariate linear model.

**2. Asymptotic Minimax Bound.** This section obtains asymptotic minimax bounds for estimation of  $\Xi$  over certain subsets of the parameter space and constructs two pertinent estimators. The first of these is asymptotically minimax for a specified subset of the parameter space. The second is asymptotically minimax over all subsets of the form considered but still requires knowledge of  $\Sigma$  and  $W$ . The adaptive estimator to be developed in Section 3 depends only on the data.

For the purposes of this section, we will reduce the estimation problem to a canonical form. Let  $\Gamma$  denote a  $p \times p$  matrix whose  $j$ -th column is an eigenvector of  $W$  corresponding to the eigenvalue  $\lambda_j(W)$ . Define

$$y_t = \Gamma' \Sigma^{-1/2} x_t, \quad \eta_t = \Gamma' \Sigma^{-1/2} \xi_t. \quad (2.1)$$

The distribution of  $y_t$  is  $N_p(\eta_t, I_p)$  and the  $\{y_t\}$  are independent random vectors that define the data matrix  $Y = (y_1, y_2, \dots, y_n)'$ . Moreover,

$$n^{-1} \sum_{t=1}^n \eta_t \eta_t' = \text{diag}\{\lambda_j(W)\}. \quad (2.2)$$

Any estimator  $\hat{\Xi}$  of  $\Xi$  induces the estimator

$$\hat{H} = (\hat{\eta}_1, \hat{\eta}_2, \dots, \hat{\eta}_n)' = \hat{\Xi} \Sigma^{-1/2} \Gamma \quad (2.3)$$

of  $H = (\eta_1, \eta_2, \dots, \eta_n)' = \Xi \Sigma^{-1/2} \Gamma$ . The correspondence between  $\hat{\Xi}$  and  $\hat{H}$  is one-to-one as is the correspondence between  $\Xi$  and  $H$ . Risks map through the identity

$$L_n(\hat{\Xi}, \Xi, \Sigma) = L_n(\hat{H}, H, I_p) = (np)^{-1} \sum_{t=1}^n |\hat{\eta}_t - \eta_t|^2. \quad (2.4)$$

The problem of estimating  $\Xi$  under loss (1.1) is therefore equivalent to the simpler problem of estimating  $H$  under quadratic loss, as exhibited in (2.4).

Let  $\mathcal{M}$  consist of all vectors in  $R^p$  whose components  $\{b_j\}$  each satisfy  $1 \leq b_j \leq \infty$  and are nondecreasing in  $j$ . For every  $b \in \mathcal{M}$  and every  $r > 0$ , define

$$D(r, b) = \{\Xi: p^{-1} \sum_{j=1}^p b_j \lambda_j(W) \leq r\}, \quad (2.5)$$

a subset of the original parameter space for  $\Xi$ . Evidently,  $\Xi \in D(r, b)$  if and only if the canonical parameter  $H$  lies in

$$E(r, b) = \{H \in R^{pn}: (np)^{-1} \sum_{j=1}^p b_j \sum_{t=1}^n \eta_{t,j}^2 \leq r\} \quad (2.6)$$

and the sums of squares in this definition are nonincreasing in  $j$ . Application of Pinsker's (1980) theorem to the canonical estimation problem (see Section 5) yields the asymptotic minimax bound

$$\liminf_{n \rightarrow \infty} \sup_{\hat{H}} \sup_{H \in E(r, b)} (np)^{-1} \mathbf{E} \sum_{t=1}^n |\hat{\eta}_t - \eta_t|^2 = \nu_0(r, b), \quad (2.7)$$

where

$$\nu_0(r, b) = p^{-1} \sum_{j=1}^p [(\mu b_j)^{-1/2} - 1]_+ / (1 + [(\mu b_j)^{-1/2} - 1]_+) \quad (2.8)$$

and  $\mu = \mu(r, b)$  is the unique positive real number such that

$$p^{-1} \sum_{j=1}^p [(b_j/\mu)^{1/2} - b_j]_+ = r. \quad (2.9)$$

Furthermore, the bound (2.7) is attained asymptotically by the linear estimator  $\hat{H}^*$  given by  $\hat{\eta}_t^* = \text{diag}\{g_j\}y_t$ , where  $g_j = [1 - (\mu b_j)^{1/2}]_+$  for  $1 \leq j \leq p$ . The least favorable  $H$  satisfies the sums-of-squares restriction stated after (2.6). Let  $X = (x_t, x_2, \dots, x_n)'$ . For any symmetric matrix  $A$  and positive definite matrix  $S$ , both of dimensions  $p \times p$ , let

$$\hat{\Xi}(A, S) = X S^{-1/2} A S^{1/2}. \quad (2.10)$$

The image of estimator  $\hat{H}^*$  in the original parametrization is

$$\hat{H}^* \Gamma' \Sigma^{1/2} = X \Sigma^{-1/2} A^* \Sigma^{1/2} = \hat{\Xi}(A^*, \Sigma), \quad (2.11)$$

where

$$A^* = A^*(r, b, W) = \Gamma \text{diag}\{g_j\} \Gamma'. \quad (2.12)$$

The discussion in this and the preceding paragraph yields the following asymptotic minimax theorem.

**Theorem 2.1.** *For every  $b \in \mathcal{M}$ , every  $r > 0$ , and every positive definite  $\Sigma$ ,*

$$\liminf_{n \rightarrow \infty} \sup_{\hat{\Xi}} \sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}, \Xi, \Sigma) = \nu_0(r, b) \quad (2.13)$$

and  $\hat{\Xi}(A^*, \Sigma)$  is an asymptotically minimax estimator of  $\Xi$  in that

$$\lim_{n \rightarrow \infty} \sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}(A^*, \Sigma), \Xi, \Sigma) = \nu_0(r, b). \quad (2.14)$$

The bound  $\nu_0(r, b)$ , defined in (2.8), is nonincreasing in  $b$  under the usual partial order on  $\mathcal{M}$ .

A major drawback to the estimator  $\hat{\Xi}(A^*, \Sigma)$  is its dependence on  $r$ ,  $b$ ,  $W$ , and  $\Sigma$ . The second estimator to be discussed in this section dispenses with knowledge of  $r$  and  $b$ , though still requiring  $W$  and  $\Sigma$ , and will lead to the fully adaptive estimator  $\hat{\Xi}_A$  that is treated in Section 3.

Let  $\mathcal{A}$  denote all symmetric  $p \times p$  matrices with eigenvalues restricted to  $[0, 1]$ . Evidently  $A^* \in \mathcal{A}$ . Consider the class of *candidate* estimators  $\{\hat{\Xi}(A, \Sigma): A \in \mathcal{A}\}$  defined through (2.10). This class assumes knowledge of  $\Sigma$ . Let

$$\tilde{A} = W(I_p + W)^{-1} = I_p - (I_p + W)^{-1}. \quad (2.15)$$

Using the spectral decomposition  $W = \Gamma\Lambda\Gamma'$  yields the spectral decomposition  $\tilde{A} = \sum_{j=1}^p [\lambda_j / (1 + \lambda_j)] \gamma_j \gamma_j'$ , which shows that  $\tilde{A} \in \mathcal{A}$ . The risk of the candidate estimator  $\hat{\Xi}(A, \Sigma)$  simplifies algebraically to

$$\begin{aligned} R_n(\hat{\Xi}(A, \Sigma), \Xi, \Sigma) &= p^{-1} \text{tr}[A^2 + (I_p - A)^2 W] \\ &= p^{-1} \text{tr}[(A - \tilde{A})^2 (I_p + W)] + p^{-1} \text{tr}[W(I_p + W)^{-1}] \\ &= \rho(A, W), \quad \text{say.} \end{aligned} \quad (2.16)$$

It follows from this display that  $\tilde{A} = \text{argmin}_{A \in \mathcal{A}} \rho(A, W)$  and

$$\min_{A \in \mathcal{A}} \rho(A, W) = \rho(\tilde{A}, W) = \tau(W) \quad (2.17)$$

for  $\tau(W)$  defined by (1.4).

Because  $A^* \in \mathcal{A}$ ,

$$\sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma) \leq \sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}(A^*, \Sigma), \Xi, \Sigma). \quad (2.18)$$

This inequality and the limit (2.14) yield

**Corollary 2.2.** *For every  $b \in \mathcal{M}$ , every  $r > 0$ , and every positive definite  $\Sigma$ ,*

$$\lim_{n \rightarrow \infty} \sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma) = \nu_0(r, b). \quad (2.19)$$

Thus, the estimator

$$\hat{\Xi}(\tilde{A}, \Sigma) = X\Sigma^{-1/2}\tilde{A}\Sigma^{1/2} = X - X(I_p + \Sigma^{-1}V)^{-1}, \quad (2.20)$$

which requires knowledge of  $W$  and  $\Sigma$ , is asymptotically minimax for every choice of  $b \in \mathcal{M}$  and  $r > 0$ . The next section devises a fully adaptive asymptotically minimax estimator that depends only on data.

**3. Adaptive Estimation.** Let  $\hat{\Sigma}$  be a consistent estimator of  $\Sigma$ . The risk function  $\rho(A, W)$  in (2.16) is estimated plausibly by

$$\hat{\rho}(A) = p^{-1}\text{tr}[A^2 + (I_p - A)^2\hat{W}], \quad (3.1)$$

where  $\hat{W}$  defined in (1.5) approximates  $W$ . By analogy with the construction of  $\hat{\Xi}(\tilde{A}, \Sigma)$ , the proposed adaptive estimator of  $\Xi$  is

$$\hat{\Xi}_A = \hat{\Xi}(\hat{A}, \hat{\Sigma}) = X\hat{\Sigma}^{-1/2}\hat{A}\hat{\Sigma}^{1/2} \quad (3.2)$$

with  $\hat{A} = \text{argmin}_{A \in \mathcal{A}} \hat{\rho}(A)$ . Lemma 5.3 in Section 5 verifies that  $\hat{A}$  is given explicitly by (1.6).

The procedure just described is a multivariate version of adaptation by minimizing  $C_L$ , a methodology that Mallows [10] first discussed critically and connected to Stein estimation. Li [9] developed properties of minimum  $C_L$  procedures, relating them to cross-validation methods. Kneip [8] treated the success of minimum  $C_L$  for ordered linear smoothers. On the other hand, Efroimovich and Pinsker [5] and Golubev [6] pioneered adaptive estimators whose maximum risk converges asymptotically to the Pinsker bound for each member of a class of ellipsoids in the parameter space. The extensive univariate literature on such adaptive asymptotically minimax estimators is reviewed by Nussbaum [12].

For every  $b \in \mathcal{M}$  and  $r > 0$ , the set  $D(r, b)$  defined in (2.5) satisfies

$$D(r, b) \subset \{\Xi: \lambda_1(W) \leq pr\}. \quad (3.3)$$

This ordering links the results in the next two theorems with the task of proving that  $\hat{\Xi}$  is asymptotically minimax in the sense of Theorem 1.1.

We will impose the following assumption on the estimator  $\hat{\Sigma}$  of  $\Sigma$ . Note that the condition includes the case when  $\Sigma$  is known and  $\hat{\Sigma} = \Sigma$ . For any matrix argument,  $|\cdot|$  will denote the Frobenius norm, which is defined by  $|A|^2 = \text{tr}[AA'] = \text{tr}[A'A]$ . We note for later use that if  $\{A_i: 1 \leq i \leq k\}$  are  $p \times p$  matrices, then  $|\text{tr}[\prod_{i=1}^k A_i]| \leq \prod_{i=1}^k |A_i|$ .

**Condition C.** The estimator  $\hat{\Sigma}$  and  $X$  are independent. Let  $\hat{Z} = \Sigma^{-1/2}\hat{\Sigma}^{1/2}$ . For every  $r > 0$ ,

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \mathbb{E}J = 0, \quad (3.4)$$

where  $J$  is any one of  $|\hat{Z}^{-1} - I_p|^2$ ,  $|\hat{Z}^{-1}|^2|\hat{Z} - I_p|^2$ , or  $|\hat{Z}^{-1}|^2|\hat{Z}^{-1} - I_p|^2$ .

The next two theorems, proved in Section 5, establish that the estimated risk function  $\hat{\rho}(A)$  and the adaptive estimator  $\hat{\Xi}_A$ , both defined above, serve asymptotically as surrogates for the true risk function  $\rho(A, W)$  and for  $\hat{\Xi}(\tilde{A}, \Sigma)$ .

**Theorem 3.1.** Suppose that Condition C holds. Then, for every  $r > 0$  and every positive definite  $\Sigma$ ,

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \mathbb{E} \sup_{A \in \mathcal{A}} |L_n(\hat{\Xi}(A, \hat{\Sigma}), \Xi, \Sigma) - \rho(A, W)| = 0 \quad (3.5)$$

and

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \mathbb{E} \sup_{A \in \mathcal{A}} |\hat{\rho}(A) - \rho(A, W)| = 0. \quad (3.6)$$

From this result follows

**Theorem 3.2.** Suppose that Condition C holds. Then, for every  $r > 0$  and every positive definite  $\Sigma$ ,

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \mathbb{E}|T - \tau(W)| = 0, \quad (3.7)$$

where  $T$  can be any one of  $L_n(\hat{\Xi}_A, \Xi, \Sigma)$ ,  $L_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma)$  or  $\hat{\rho}(\hat{A})$  and  $\tau(W)$  is defined in (1.4).

The convergence (1.3) of the risk of  $\hat{\Xi}_A$  is immediate from this result. Another consequence is the following corollary, which establishes the asymptotic minimaxity of  $\hat{\Xi}_A$ .

**Corollary 3.3.** Suppose that Condition C holds. For every  $b \in \mathcal{M}$ , every  $r > 0$ , and every positive definite  $\Sigma$ ,

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} |R_n(\hat{\Xi}_A, \Xi, \Sigma) - R_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma)| = 0 \quad (3.8)$$

and

$$\lim_{n \rightarrow \infty} \sup_{\Xi \in D(r, b)} R_n(\hat{\Xi}_A, \Xi, \Sigma) = \nu_0(r, b). \quad (3.9)$$

To verify (3.8), observe that

$$\sup_{\lambda_1(W) \leq r} |R_n(\hat{\Xi}_A, \Xi, \Sigma) - R_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma)| \leq \sup_{\lambda_1(W) \leq r} \mathbb{E}|L_n(\hat{\Xi}_A, \Xi, \Sigma) - L_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma)| \quad (3.10)$$

which tends to zero by Theorem 3.2. Corollary 2.2, (3.3), and (3.10) then imply (3.9).

Related to Corollary 3.3 are the following remarks:

a) A uniform integrability argument yields

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} (np)^{-1} \mathbb{E} |\Sigma^{-1/2} (\hat{\Xi}_B - \hat{\Xi}(\tilde{A}, \Sigma))'|^2 = 0. \quad (3.11)$$

Consequently, by Corollary 2.2, the estimator  $\hat{\Xi}_B$  is asymptotically minimax in the sense (3.9).

b) Suppose that  $\hat{\Sigma}$  is independent of  $X$  and  $(m+p+1)\hat{\Sigma}$  has a Wishart( $\Sigma, m$ ) distribution. Bilodeau and Kariya [2] showed that the extended Efron-Morris estimator

$$\hat{\Xi}_{EM} = X - X[(n-p-1)(X'X)^{-1} + (p-1)I_p/\text{tr}(X'X)]\hat{\Sigma} \quad (3.12)$$

is then globally minimax. Under the hypotheses just stated, this refinement of  $\hat{\Xi}_B$  also has the Pinsker asymptotic minimaxity (3.9), provided  $m$  tends to infinity with  $n$ .

c) Specialized to the case  $p = 1$ , Corollary 3.3 implies that the James-Stein estimator and the positive-part James Stein estimator are asymptotically minimax over every ball centered at the origin in the parameter space. Of course, this result also follows directly from Pinsker's theorem (see Theorem 5.2) or by developing ideas sketched in Stein [16] (see [1]).

**4. Application to the Multivariate Linear Model.** This section describes some implications of  $\hat{\Xi}_A$  and its cousins for improved fitting of the Gaussian multivariate linear model (see also [2]). For the univariate linear model, Rao and Toutenberg [15] reviewed various biased estimation techniques that have smaller risk than least squares. The multivariate case presents the additional possibility of estimating and using information between response variables.

Consider the multivariate linear model  $Y = CB + E$ , where the observation matrix  $Y$  is  $m \times p$ , the regression matrix  $C$  is  $m \times n$ , the coefficient matrix  $B$  is  $n \times p$ , and the rows of the error matrix  $E$  are independent Gaussian random vectors with mean 0 and covariance matrix  $\Sigma$ . Here  $C$  is a given matrix constant while both  $B$  and  $\Sigma$  are unknown. We will assume that  $\text{rank}(C) = n < m$  and that  $p < n$ . The problem is to estimate  $M = EY = CB$ .

Reducing this linear model to canonical form enables us to apply the preceding results on estimation of multivariate trend. Let  $N$  be an  $m \times n$  matrix whose columns are orthonormal and span the same subspace of  $R^m$  as do the columns of  $C$ . One possible algebraic construction of  $N$  is through the singular value decomposition of  $C$ ,

$$C = NLP' \quad (4.1)$$

where  $P$  is  $n \times n$ ,  $N'N = P'P = PP' = I_n$ , and  $L = \text{diag}\{l_i\}$  with  $l_1 \geq l_2 \geq \dots \geq l_n > 0$ . The columns of  $P$  are eigenvectors of  $C'C$  and  $l_i$  is the positive square root of the  $i$ -th largest eigenvalue.

Having chosen  $N$ , construct the  $m \times (m-n)$  matrix  $\bar{N}$  so that the matrix  $O = \{N|\bar{N}\}$  is orthogonal. If  $N$  comes from the singular value decomposition (4.1), then the columns of  $O$  are eigenvectors of  $CC'$ , ordered in decreasing order of the eigenvalues. Let

$$X = N'Y, \quad \bar{X} = \bar{N}'Y \quad (4.2)$$

and define  $\Xi = EX = LP'B$ , an  $n \times p$  matrix. Because  $(X'|\bar{X}')' = O'Y$ , the rows of  $X$  and  $\bar{X}$  are independent Gaussian random vectors, each having covariance matrix  $\Sigma$ . This structure is a canonical form of the original linear model.

The mapping between  $\Xi$  and  $M = CB$  is one-to-one, because  $M = N\Xi$  and  $\Xi = N'M$ . The columns of the canonical parameter  $\Xi$  can take any value in  $R^n$ ; the columns of the original parameter  $M$  are restricted to the  $n$ -dimensional subspace  $\mathcal{L}(C)$  of  $R^m$  spanned by the columns of  $C$ . The same one-to-one mapping exists between any estimator  $\hat{M} = C\hat{B}$  of  $CB$  and the corresponding estimator  $\hat{\Xi} = N'\hat{M}$  of  $\Xi$ . Because

$$\begin{aligned} L_{m,n}(\hat{M}, M, \Sigma) &= (np)^{-1} \text{tr}[\Sigma^{-1}(\hat{M} - M)'(\hat{M} - M)] \\ &= (np)^{-1} \text{tr}[\Sigma^{-1}(\hat{\Xi} - \Xi)'(\hat{\Xi} - \Xi)] \end{aligned} \quad (4.3)$$

estimation of  $M = CB$  under the loss to the left is equivalent to estimation of the canonical parameter  $\Xi$  under the loss to the right. Denote the corresponding risk by  $R_{m,n}(\hat{M}, M, \Sigma)$ .

Let  $\hat{\Sigma} = (m-n)^{-1}\bar{X}'\bar{X}$  be the usual estimator of  $\Sigma$  based upon the rows of  $\bar{X}$ . In terms of the original parametrization,  $\hat{\Sigma} = (m-n)^{-1}(Y - C\hat{B}_{LS})'(Y - C\hat{B}_{LS})$  where  $\hat{B}_{LS} = (C'C)^{-1}C'Y$  is the least squares estimator of  $B$  (cf. Mardia, Kent and Bibby [11], chapter 6). Define the estimator  $\hat{\Xi}_A$  as in (1.7). Asymptotic minimaxity of  $\hat{\Xi}_A$ , as stated in Corollary 3.3, entails asymptotic minimaxity under loss (4.3) of the estimator

$$\hat{M}_A = N\hat{\Xi}_A = C\hat{B}_A, \quad (4.4)$$

where  $\hat{B}_A = NL^{-1}\hat{\Xi}_A = NL^{-1}X\hat{\Sigma}^{-1/2}\hat{A}\hat{\Sigma}^{1/2}$ .

More precisely, note that  $W$ , defined by (1.2), can be expressed in terms of  $M$  through

$$W = n^{-1}\Sigma^{-1/2}M'NN'M\Sigma^{-1/2} \quad (4.5)$$

and that  $\Xi \in D(r, b)$  if and only if  $M \in C(r, b)$ , where

$$C(r, b) = \{M: M \in \mathcal{L}(C), p^{-1} \sum_{j=1}^p b_j \lambda_j(W) \leq r\}. \quad (4.6)$$

The estimator  $\hat{\Sigma}$  defined above satisfies Condition C with  $n$  replaced by  $m - n$ . Thus, (2.13) and (3.9) imply

**Corollary 4.1.** *Let  $q = \min(n, m - n)$ . For every  $b \in \mathcal{M}$ , every  $r > 0$ , and every positive definite  $\Sigma$ ,*

$$\liminf_{q \rightarrow \infty} \sup_{\hat{M} \in C(r, b)} R_{m, n}(\hat{M}, M, \Sigma) = \nu_0(r, b) \quad (4.7)$$

and

$$\lim_{q \rightarrow \infty} \sup_{M \in C(r, b)} R_{m, n}(\hat{M}_A, M, \Sigma) = \nu_0(r, b). \quad (4.8)$$

**Example.** Suppose we observe  $k$  independent replicates of the deterministic trend model described in Section 1. Equivalent is the multivariate linear model in which  $m = kn$ ,  $B$  is  $n \times p$ , and

$$C = (I_n | I_n | \dots | I_n)'. \quad (4.9)$$

Thus  $M = CB = (B' | B' | \dots | B')'$ . The singular value decomposition of  $C$  has  $P = I_n$ ,  $N = k^{-1/2}(I_n | I_n | \dots | I_n)'$ , and  $L = k^{1/2}I_n$ . Let  $Y_1$  denote the first  $n$  rows of  $Y$ ,  $Y_2$  the next  $n$  rows, and so forth until  $Y_k$ . If  $\bar{Y} = k^{-1} \sum_{i=1}^k Y_i$ , then the least squares estimator of  $B$  is  $\hat{B}_{LS} = \bar{Y}$ . Consequently,  $\hat{M}_{LS} = (\bar{Y}' | \bar{Y}' | \dots | \bar{Y}')'$ ,  $X = k^{1/2}\bar{Y}$ ,

$$\hat{\Sigma} = k^{-1} \sum_{i=1}^k (Y_i - \bar{Y})'(Y_i - \bar{Y}), \quad (4.10)$$

and  $\hat{W} = kn^{-1}\hat{\Sigma}^{-1/2}\bar{Y}'\bar{Y}\hat{\Sigma}^{-1/2} - I_p$ . By Corollary 4.1, construction (4.4) yields a superefficient estimator  $\hat{M}_A$  of  $M$  that is asymptotically minimax when  $k$  is fixed and  $n$  tends to infinity. Since  $\Xi = k^{1/2}B$  and  $W = kn^{-1}\Sigma^{-1/2}B'B\Sigma^{-1/2}$ , it follows from (1.3) and (1.4) that  $\hat{M}_A$  improves most significantly on  $\hat{M}_{LS}$  when  $k$  is small.

**5. Argument Details.** This section substantiates various claims made earlier in the paper.

**The Pinsker bound.** Suppose we observe  $u = (u_1, u_2, \dots, u_m)'$ , the  $\{u_i\}$  being independent random variables and the distribution of  $u_i$  being  $N(\theta_i, 1)$ . The problem is to estimate the means  $\theta = (\theta_1, \theta_2, \dots, \theta_m)'$  under normalized quadratic loss. The risk of an estimator  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$  is

$$R_m(\hat{\theta}, \theta) = m^{-1} \mathbb{E} \sum_{i=1}^m (\hat{\theta}_i - \theta_i)^2. \quad (5.1)$$

When specialized to this problem, Pinsker's [14] paper yields two theorems stated below. We emphasize that these two theorems are useful corollaries to Pinsker's more general analysis. Nussbaum's [12] extensive survey reviews other applications of the Pinsker bound.

Let  $\mathcal{N} = \{a \in R^m: a_i \in [1, \infty], 1 \leq i \leq m\}$ . Define addition, subtraction, multiplication and division of  $f$  and  $g$  in  $R^m$  by the specified operation on components, as in coding S-Plus. For instance,  $fg = (f_1g_1, f_2g_2, \dots, f_mg_m)$ . Let  $\text{ave}(f) = m^{-1} \sum_{i=1}^m f_i$ . For every  $a \in \mathcal{N}$  and  $r > 0$ , define the ellipsoid

$$B(r, a) = \{\theta \in R^m: \text{ave}(a\theta^2) \leq r\}. \quad (5.2)$$

Let  $\theta_0^2 = [(\mu a)^{-1/2} - 1]_+$ , where  $\mu$  is the unique positive real number such that  $\text{ave}(a\theta_0^2) = r$ . Define  $\nu_m(r, a) = \text{ave}[\theta_0^2/(1 + \theta_0^2)]$  and  $f_0 = \theta_0^2/(1 + \theta_0^2)$ .

The first theorem drawn from Pinsker's reasoning treats linear estimators for  $\theta$  of the form  $\hat{\theta} = fu$ .

**Theorem 5.1.** *For every  $a \in \mathcal{N}$  and every  $r > 0$ ,*

$$\inf_{f \in R^m} \sup_{\theta \in B(r, a)} R_m(fu, \theta) = \nu_m(r, a) = \sup_{\theta \in B(r, a)} R_m(f_0u, \theta). \quad (5.3)$$

The second theorem from the same source shows that the minimax linear estimator is often asymptotically minimax among all estimators.

**Theorem 5.2.** *If  $\lim_{m \rightarrow \infty} m\nu_m(r, a) = \infty$ , then for every  $a \in \mathcal{N}$  and every  $r > 0$ ,*

$$\lim_{m \rightarrow \infty} \inf_{\hat{\theta}} \sup_{\theta \in B(r, a)} [R_m(\hat{\theta}, \theta)/\nu_m(r, a)] = 1. \quad (5.4)$$

*If  $\lim_{m \rightarrow \infty} \nu_m(r, a) = \nu_0 > 0$ , then also*

$$\lim_{m \rightarrow \infty} \inf_{\hat{\theta}} \sup_{\theta \in B(r, a)} R_m(\hat{\theta}, \theta) = \nu_0. \quad (5.5)$$

**Proof of (2.7).** The canonical estimation problem of Section 2, described in equations (2.1) through (2.7) can be re-expressed in the notation above. Form  $u$  by stacking vertically the columns of  $Y$ . Similarly, form  $\theta$  and  $\hat{\theta}$  by stacking the columns of  $H$  and  $\hat{H}$ . Thus  $m = np$ . Form  $a$  by stacking  $n$  replicates of  $b_1$  atop  $n$  replicates of  $b_2$  and so on through  $n$  replicates of  $b_p$ . With these identifications, equation (5.5) above is equivalent to (2.7).

**Lemma 5.3.** *The matrix  $\hat{A} = \text{argmin}_{A \in \mathcal{A}} \hat{\rho}(A)$  is given explicitly by (1.6).*

**Proof.** Let  $\check{A} = I_p - (I_p + \hat{W})^{-1}$ . As in the second line in (2.16),

$$\hat{\rho}(A) = p^{-1} \text{tr}[(A - \check{A})^2(I_p + \hat{W})] + p^{-1} \text{tr}[\hat{W}(I_p + \hat{W})^{-1}]. \quad (5.6)$$

Let  $\mathcal{S}$  denote the set of all  $p \times p$  symmetric matrices. From (5.6),  $\check{A} = \text{argmin}_{A \in \mathcal{S}} \hat{\rho}(A)$ . Write  $\hat{\pi}_j = \hat{\lambda}_j/(1 + \hat{\lambda}_j)$ . If  $\hat{\Lambda} = \text{diag}\{\hat{\lambda}_j\}$  and  $\hat{\Gamma} = \{\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_p\}$ , then  $\hat{W}$  has the spectral

representation  $\hat{W} = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}'$ . Consequently,  $\check{A} = \sum_{j=1}^p \hat{\pi}_j \hat{\gamma}_j \hat{\gamma}_j'$ . Because  $1 + \hat{\lambda}_j \geq 0$ , it follows that  $\hat{\pi}_j \leq 1$  but need not be positive. Consequently  $\check{A}$  is not, in general, an element of  $\mathcal{A}$ .

Define

$$\check{A}_+ = \sum_{\hat{\pi}_j \geq 0} \hat{\pi}_j \hat{\gamma}_j \hat{\gamma}_j', \quad \check{A}_- = \sum_{\hat{\pi}_j < 0} \hat{\pi}_j \hat{\gamma}_j \hat{\gamma}_j', \quad (5.7)$$

noting that  $\check{A}_+ \in \mathcal{A}$ ,  $\check{A} = \check{A}_+ + \check{A}_-$ , and  $\check{A}_+ \check{A}_- = 0$ . For brevity, put

$$\hat{K} = I_p + \hat{W} = \sum_{j=1}^p (1 + \hat{\lambda}_j) \hat{\gamma}_j \hat{\gamma}_j'. \quad (5.8)$$

Then

$$\begin{aligned} \text{tr}[(A - \check{A})^2 \hat{K}] &= \text{tr}[\{(A - \check{A}_+) - \check{A}_-\}^2 \hat{K}] \\ &= \text{tr}[(A - \check{A}_+)^2 \hat{K}] - \text{tr}[A \check{A}_- \hat{K}] - \text{tr}[\hat{K} \check{A}_- A] + \text{tr}[\check{A}_-^2 \hat{K}]. \end{aligned} \quad (5.9)$$

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

$$\begin{aligned} -\text{tr}[A \check{A}_- \hat{K}] &= -\text{tr}[A \sum_{\hat{\pi}_j < 0} \hat{\pi}_j (1 + \hat{\lambda}_j) \hat{\gamma}_j \hat{\gamma}_j'] \\ &= -\sum_{\hat{\pi}_j < 0} \hat{\pi}_j (1 + \hat{\lambda}_j) \hat{\gamma}_j' A \hat{\gamma}_j \geq 0 \end{aligned} \quad (5.10)$$

because  $A$  is positive semidefinite and  $1 + \hat{\lambda}_j \geq 0$ . Similarly,  $-\text{tr}[\hat{K} \check{A}_- A] \geq 0$ . It now follows from (5.10) that

$$\text{tr}[(A - \check{A})^2 \hat{K}] \geq \text{tr}[(A - \check{A}_+)^2 \hat{K}] + \text{tr}[\check{A}_-^2 \hat{K}] \quad (5.11)$$

for every  $A \in \mathcal{A}$ . This implies that  $\hat{A} = \check{A}_+$ , as was to be shown.

**Proof of Theorem 3.1** We first prove (3.6). Let  $z_t = \Sigma^{-1/2} x_t$  and  $\zeta_t = \Sigma^{-1/2} \xi_t$ . The  $\{z_t: 1 \leq t \leq n\}$  are independent random vectors, the distribution of  $z_t$  being  $N_p(\zeta_t, I_p)$ . If  $\hat{U} = n^{-1} \sum_{t=1}^n z_t z_t'$  and  $\hat{Z}$  is the matrix defined in Condition C, then

$$\hat{W} = \hat{Z}^{-1} \hat{U} (\hat{Z}^{-1})' - I_p \quad (5.12)$$

and

$$\hat{U} = W + I_p + \hat{F} + \hat{F}' + \hat{G}, \quad (5.13)$$

where

$$\hat{F} = n^{-1} \sum_{t=1}^n (z_t - \zeta_t) \zeta_t', \quad \hat{G} = n^{-1} \sum_{t=1}^n (z_t - \zeta_t)(z_t - \zeta_t)' - I_p. \quad (5.14)$$

By direct calculations,  $\sup_{\lambda_1(W) \leq r} \text{E}|\hat{F}|^2 = O(n^{-1})$  and  $\sup_{\lambda_1(W) \leq r} \text{E}|\hat{G}|^2 = O(n^{-1})$ ; consequently

$$\sup_{\lambda_1(W) \leq r} \text{E}|\hat{U} - W - I_p| = O(n^{-1/2}). \quad (5.15)$$

Evidently

$$\begin{aligned}
\hat{\rho}(A) - \rho(A, W) &= p^{-1} \text{tr}[(I_p - A)^2(\hat{W} - W)] \\
&= p^{-1} \text{tr}[(I_p - A)^2\{\hat{Z}^{-1}\hat{U}(\hat{Z}^{-1})' - W - I_p\}] \\
&\leq p^{-1} \sum_{j=1}^3 T_j,
\end{aligned} \tag{5.16}$$

where

$$\begin{aligned}
|T_1| &= |\text{tr}[(I_p - A)^2\hat{Z}^{-1}\hat{U}\{(\hat{Z}^{-1})' - I_p\}]| \leq |I_p - A|^2|\hat{Z}^{-1}||\hat{U}||\hat{Z}^{-1} - I_p| \\
|T_2| &= |\text{tr}[(I_p - A)^2(\hat{Z}^{-1} - I_p)\hat{U}]| \leq |I_p - A|^2|\hat{Z}^{-1} - I_p||\hat{U}| \\
|T_3| &= |\text{tr}[(I_p - A)^2(\hat{U} - W - I_p)]| \leq |I_p - A|^2|\hat{U} - W - I_p|.
\end{aligned} \tag{5.17}$$

For every  $A \in \mathcal{A}$ ,  $|I_p - A|^2 \leq p$ . Combining the last three displays with Condition C yields (3.6).

To verify (3.5), write  $A_Z = \hat{Z}A\hat{Z}^{-1}$  and observe that

$$\begin{aligned}
L_n(\hat{\Xi}(A, \hat{\Sigma}), \Xi, \Sigma) &= (np)^{-1} \sum_{t=1}^n |A_Z z_t - \zeta_t|^2 \\
&= (np)^{-1} \sum_{t=1}^n |A_Z(z_t - \zeta_t) - (I_p - A_Z)\zeta_t|^2 \\
&= p^{-1} \text{tr}[A_Z' A_Z (I_p + \hat{G}) + (I_p - A_Z)'(I_p - A_Z)W - 2(I_p - A_Z)' A_Z \hat{F}].
\end{aligned} \tag{5.18}$$

Since

$$|A_Z - A| \leq |A||\hat{Z}^{-1}||\hat{Z} - I_p| + |A||\hat{Z}^{-1} - I_p|, \tag{5.19}$$

the limit (3.5) follows from the preceding two displays, the statement after (5.14), and Condition C.

**Proof of Theorem 3.2.** Limit (3.6) implies that

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \text{E}|\hat{\rho}(\hat{A}) - \rho(\hat{A}, W)| = 0 \tag{5.20}$$

and

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \text{E}|\hat{\rho}(\hat{A}) - \rho(\tilde{A}, W)| = 0. \tag{5.21}$$

Since  $\rho(\tilde{A}, W) = \tau(W)$ , limit (3.7) holds for  $T = \hat{\rho}(\hat{A})$  and

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \text{E}|\rho(\hat{A}, W) - \tau(W)| = 0. \tag{5.22}$$

On the other hand, limit (3.5) gives

$$\lim_{n \rightarrow \infty} \sup_{\lambda_1(W) \leq r} \mathbb{E} |L_n(\hat{\Xi}_A, \Xi, \Sigma) - \rho(\hat{A}, W)| = 0. \quad (5.23)$$

Combining this with (5.22) entails (3.7) for  $T = L_n(\hat{\Xi}_A, \Xi, \Sigma)$ . Finally, taking  $\hat{\Sigma} = \Sigma$  yields (3.7) for  $T = L_n(\hat{\Xi}(\tilde{A}, \Sigma), \Xi, \Sigma)$ .

**6. Discussion.** This paper approaches from several directions the affine shrinkage estimator  $\hat{\Xi}_A$  for the multivariate trend  $\Xi$ . In a Gaussian random trend model,  $\hat{\Xi}_A$  is an estimated minimum risk predictor of  $\Xi$ . For the deterministic trend model used in our analysis,  $\hat{\Xi}_A$  is that member of a certain class of candidate affine shrinkage estimators that minimizes estimated risk, or equivalently, minimizes the  $C_L$  criterion. Analysis shows that  $\hat{\Xi}_A$  is asymptotically minimax in Pinsker's sense over certain subsets of trends centered at  $\Xi = 0$ . The asymptotic maximum risk of  $\hat{\Xi}_A$  over such subsets strictly dominates that of the least squares trend estimator. Unlike  $\hat{\Xi}(A^*, \Sigma)$  and  $\hat{\Xi}(\tilde{A}, \Sigma)$ , the other asymptotically minimax estimators studied in Section 2, the estimator  $\hat{\Xi}_A$  is fully adaptive, depending only on data. As exhibited in the Introduction,  $\hat{\Xi}_A$  achieves superefficiency relative to the least squares estimator by performing componentwise James-Stein shrinkage in a coordinate system that is estimated from the data. The construction of  $\hat{\Xi}_A$ , applied to the Gaussian multivariate linear model in canonical form, yields improved regression fits. These main results carry over to cousins of  $\hat{\Xi}_A$  such as  $\hat{\Xi}_B$  and the extended Efron-Morris estimator. The historically distinct ideas of Stein, Mallows, and Pinsker on estimation of a high-dimensional parameter form the background to this paper.

## References

- [1] R. Beran, *Stein estimation in high dimensions: a retrospective*, In: Madan Puri Festschrift, (E. Brunner and M. Denker, eds.), VSP, Zeist (1996), pp. 91–110.
- [2] M. Bilodeau and T. Kariya, *Minimax estimators in the normal MANOVA model*, J. Multivariate Anal., 28 (1989), 260–270.
- [3] S. Yu. Efrimovich and M. S. Pinsker, *Learning algorithm for nonparametric filtering*, Automat. Remote Control, 45 (1984), pp. 1434–1440.
- [4] B. Efron and C. Morris, *Multivariate empirical Bayes and estimation of covariance matrices*, Ann. Statist., 4 (1976), pp. 22–32.
- [5] G. K. Golubev, *Adaptive asymptotically minimax estimators of smooth signals*, Problems Inform. Transmission, 23 (1987), pp. 57–67.
- [6] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, *A transformation for ordering multispectral data in terms of image quality with implications for noise removal*, IEEE

- Trans. Geosci. Remote Sensing, 26 (1985), pp. 65–74.
- [7] W. James and C. Stein, *Estimation with quadratic loss*, In: Proc. Fourth Berkeley Symp. Math. Statist. Prob., 1, (J. Neyman, ed.), University of California Press, Berkeley (1961), pp. 361–380.
- [8] A. Kneip, *Ordered linear smoothers*, Ann. Statist., 22 (1994), pp. 835–866.
- [9] K.-C. Li, *Asymptotic optimality for  $C_p$ ,  $C_L$ , cross-validation and generalized cross-validation: discrete index set*, Ann. Statist., 15 (1989), pp. 958–976.
- [10] C. L. Mallows, *Some comments on  $C_p$* , Technometrics, 15 (1973), pp. 661–676.
- [11] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
- [12] M. Nussbaum, *The Pinsker bound: a review*, In: Encyclopedia of Statistical Sciences: Update Volume 3, (S. Kotz, C. B. Read, eds.), Wiley, New York, to appear.
- [13] J. Pfanzagl, *Incidental versus random nuisance parameters*, Ann. Statist., 21 (1993), pp. 1663–1691.
- [14] M. S. Pinsker, *Optimal filtration of square-integrable signals in Gaussian noise*, Problems Inform. Transmission, 16 (1980), pp. 120–133.
- [15] C. R. Rao and H. Toutenberg, *Linear Models. Least Squares and Alternatives*, Springer, New York, 1995.
- [16] C. Stein, *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*, In: Proc. Third Berkeley Symposium Symp. Math. Statist. Prob., 1, (J. Neyman, ed.), University of California Press, Berkeley (1956), pp. 197–206.
- [17] C. Stein, *An approach to recovery of inter-block information in balanced incomplete block designs*, In: Research Papers in Statistics. Festschrift for Jerzy Neyman (F. N. David, ed.), Wiley, London (1966), pp. 351–366.