

REACT TREND ESTIMATION IN CORRELATED NOISE

RUDOLF BERAN¹

Department of Statistics, University of California, Berkeley,
Berkeley, CA 94720-3860
E-mail: beran@stat.berkeley.edu

ABSTRACT

Suppose that the data is modeled as replicated realizations of a p -dimensional random vector whose mean μ is a trend of interest and whose covariance matrix Σ is unknown, positive definite. REACT estimators for the trend involve transformation of the data to a new basis, estimating the risks of a class of candidate linear shrinkage estimators, and selecting the candidate estimator with smallest estimated risk. For Gaussian samples and quadratic loss, the maximum risks of REACT estimators proposed in this paper undercut that of the classically efficient sample mean vector. The superefficiency of the proposed estimators relative to the sample mean is most pronounced when the new basis provides an economical description of the vector $\Sigma^{-1/2}\mu$, dimension p is not small, and sample size is much larger than p . A case study illustrates how vague prior knowledge may guide choice of a basis that reduces risk substantially.

1. INTRODUCTION

The average of a sample of random vectors drawn from a $N_p(\mu, \Sigma)$ normal distribution is inadmissible, under suitable quadratic loss, as an estimator of the mean vector μ whenever the dimension p of the distribution exceeds two (see Stein [8]). The insistence of the sample mean on unbiasedness can result in over-fitting of μ when p is not small. Recent work on model-selection, shrinkage, and thresholding estimators when $\Sigma = \sigma^2 I_p$ has shown, in that case, that even uncertain prior knowledge about the nature of μ can be translated into major reductions in estimation risk (cf. Donoho and Johnstone [3], Efromovich [4], and Beran [1]). This paper develops REACT shrinkage estimators of μ and their risk properties for situations where the covariance matrix Σ is unknown, though possibly restricted as in spatial or time-series analysis. The superior performance of the proposed estimators is illustrated on a set of multivariate lumber-thickness measurements collected in a study of saw-mill operations.

As data model, suppose that (x_1, x_2, \dots, x_n) are independent random column vectors, each of which has a $N_p(\mu, \Sigma)$ distribution. The components of μ constitute a trend that is observed in correlated noise. The word trend indicates that component order matters. Both μ and the covariance matrix Σ are unknown, though the latter is

¹ This research was supported at Universität Heidelberg by the Alexander von Humboldt Foundation and at Berkeley by National Science Foundation Grant DMS 99-70266. Dean Huber of the U.S. Forest Service in San Francisco provided the lumber-thickness data, both numbers and context.

assumed positive definite and may sometimes have further structure. It is tacitly assumed that observation dimension p is not small and that sample size n is much larger than p , in ways that will be made precise. Let $\hat{\mu}$ denote any estimator of μ . The quality of $\hat{\mu}$ is assessed through the quadratic loss

$$L_{n,p}(\hat{\mu}, \mu, \Sigma) = (n/p)(\hat{\mu} - \mu)' \Sigma^{-1} (\hat{\mu} - \mu). \quad (1)$$

The risk $R_{n,p}(\hat{\mu}, \mu, \Sigma)$ is the expectation of this loss under the model. The normalization factor n/p is convenient for asymptotics in which both n and p tend to infinity. In particular, the risk of the sample mean \bar{x} is 1 for every value of μ and Σ .

The REACT estimator $\hat{\mu}_M$ developed in this paper has asymptotic risk that can be characterized after we introduce some notation. The acronym itself will be explained below. Let U be an orthogonal matrix, to be specified in the description of the REACT method, and let $\xi = n^{1/2}U'\Sigma^{-1/2}\mu$. Define the function $\text{ave}(\cdot)$, applied to any p -dimensional vector, to be the average of its components. For every vector $f \in [0, 1]^p$ and every ξ in R^p , define

$$\rho(f, \xi^2) = \text{ave}[f^2 + (1 - f)^2 \xi^2], \quad (2)$$

which is convex in f . The operations inside the average are performed coordinatewise, as in the S language. Let \mathcal{F}_M denote the convex set of monotone nonincreasing shrinkage vectors $\{f \in [0, 1]^p: f_1 \geq f_2 \geq \dots \geq f_p\}$ and let

$$\tau_M(\xi^2) = \min_{f \in \mathcal{F}_M} \rho(f, \xi^2) < 1 \quad \forall \xi, \Sigma. \quad (3)$$

The quantity $\text{ave}(\xi^2) = (n/p)\mu'\Sigma^{-1}\mu$ measures the signal-to-noise ratio under the model.

We will prove, among other results, that the REACT estimator $\hat{\mu}_M$ satisfies

$$\lim_{n,p \rightarrow \infty} \sup_{(n/p)\mu'\Sigma^{-1}\mu \leq r} |R_{n,p}(\hat{\mu}_M, \mu, \Sigma) - \tau_M(\xi^2)| = 0 \quad (4)$$

for every finite positive r . Here n must tend to infinity faster than p^2 unless Σ is significantly constrained. The asymptotic risk of $\hat{\mu}_M$ is thus strictly less than the risk of the sample mean for every value of μ and of Σ . Moreover, $\hat{\mu}_M$ turns out to be asymptotically minimax over certain subsets of the parameter space. The minimax bound is smallest over subsets where all but the first few components of ξ are very small, or equivalently, when the inner product of $\Sigma^{-1/2}\mu$ with successive columns of U is very small after the first few columns. Prior information can sometimes be used to find such an *economical* basis U . This point is demonstrated in the case study of Section 2. While limit (4) holds for every choice of orthogonal matrix U , we will see that the superefficiency of $\hat{\mu}_M$ over the classically efficient (albeit inadmissible) sample mean is most pronounced when U is most economical.

The acronym REACT stands for **r**isk **e**stimation after **c**oordinate **t**ransformation. The construction of $\hat{\mu}_M$ is briefly as follows. Let $\hat{\Sigma}$ denote a suitably consistent estimator of Σ that is independent of \bar{x} . One candidate is the sample covariance matrix. After selecting a tentatively economical orthogonal basis U , define the canonical mean vector

$$\hat{z} = n^{1/2}U'\hat{\Sigma}^{-1/2}\bar{x}. \quad (5)$$

This is the coordinate transformation step. Let $\text{diag}(f)$ denote the diagonal matrix whose diagonal is given by the vector f . The quantity

$$\hat{\rho}(f) = \text{ave}[f^2 + (1 - f)^2(z^2 - 1)]. \quad (6)$$

will be seen to estimate the risk of the candidate estimator

$$\hat{\mu}(f, \hat{\Sigma}) = \hat{\Sigma}^{1/2} U \text{diag}(f) U' \hat{\Sigma}^{-1/2} \bar{x} \quad (7)$$

for μ . This is the risk estimation step. Let $\hat{f}_M = \text{argmin}_{f \in F_M} \hat{\rho}(f)$. This is the adaptation step, which identifies the candidate estimator with smallest estimated risk. Combining these three operations yields the REACT estimator

$$\hat{\mu}_M = \hat{\mu}(\hat{f}_M, \hat{\Sigma}) = n^{-1/2} \hat{\Sigma}^{1/2} U \text{diag}(\hat{f}_M) \hat{z}. \quad (8)$$

This estimator turns out to have the theoretical properties sketched above.

The aims of this paper are to establish the superefficiency of $\hat{\mu}_M$ as n and p tend to infinity at suitable relative rates and to argue that this superefficiency has statistical value. Section 2 illustrates how $\hat{\mu}_M$ improves on the sample mean vector in a case study of lumber-thickness measurements that motivated parts of this paper. Section 3 begins with an asymptotic minimax bound for estimation of the mean vector μ as its dimension p tends to infinity. The success of the adaptation step, the asymptotic minimaxity of $\hat{\mu}_M$, and the remarkable benefits of basis economy are the main topics of that section. Section 4 gives proofs.

2. THE LUMBER-THICKNESS DATA

Softwood lumber mills in the western U.S. typically produce green boards through a series of sawing operations. Initial slicing of the logs by a headrig yields boards that are subsequently resawn one or more times by secondary saws. Variability in each of the sequential sawing operations contributes to irregularities in the thickness of the final green lumber. The data analyzed in this section was collected as part of a larger study by the U.S. Forest Service that investigated how lumber thickness errors are propagated through sequential sawing operations.

Boards selected “at random” as they came off a headrig bandsaw were followed through two horizontal resaws. In a horizontal resaw, the board being divided in two is pressed flat against a horizontal reference plane that is parallel to the saw blade. Thickness errors in the offspring board that touches the reference plane are due entirely to the resaw. However, thickness errors in the other offspring board are the sum of resaw errors and of thickness errors in the parent board. Initially and at each subsequent stage of processing, the thickness of every board produced was measured at eight standardized points, the first four along the “upper” edge, the next four at the opposed points along the “lower” edge. Board orientations were preserved throughout the sequence of resawings and measurements.

The particular sample analyzed in this section arose as follows. Boards of nominal four inch thickness coming off a headrig were resawn horizontally into two inch lumber and then again into one inch lumber. The top and bottom offspring boards from the first resaw were coded, respectively, as samples 1 and 2. The second resaw of these

samples yielded four samples that were coded 11, 12, 21, 22. Here the right digit refers to the position of the offspring board (top or bottom) during the second resaw. The sample 11 that we consider consists of the top offspring from the second resaw of the top offspring from the first resaw.

The thickness measurements for each board are viewed as an 8×1 vector. Components 1 to 4 come from the upper edge of the board while components 5 to 8 come from the lower edge. The measurement sites

1	2	3	4
5	6	7	8

are opposed in pairs and ordered as indicated. In the notation of the Introduction, the dimension p is 8. Figure 1 exhibits the thickness measurements for the 25 boards in sample 11. In most cases, one edge of the board is thicker than the other, but whether the upper or lower edge is thicker varies from board to board. The plot of \bar{x} in cell (1,1) of Figure 2 shows that, on average, the upper edge is thinner than the lower edge, despite considerable board-to-board variation.

Construction of the adaptive estimator $\hat{\mu}_M$ defined in (8) requires estimating the covariance matrix Σ , choosing the orthogonal basis U , and computing $\hat{f}_M = \operatorname{argmin}_{f \in F_M} \hat{\rho}(f)$, where $\hat{\rho}(f)$ is the estimated risk function defined in (6). We consider these matters in turn.

Estimation of Σ . It seems plausible that the sawing errors at different measurement sites are homoscedastic and positively correlated, the amount of correlation depending on distance between the measurement sites. Because board width is very small relative to the distance between measurement sites along either edge, these considerations suggest that

$$\Sigma = \Sigma(A, B, C, D, E) = \begin{pmatrix} A & B & C & D & E & B & C & D \\ B & A & B & C & B & E & B & C \\ C & B & A & B & C & B & E & B \\ D & C & B & A & D & C & B & E \\ E & B & C & D & A & B & C & D \\ B & E & B & C & B & A & B & C \\ C & B & E & B & C & B & A & B \\ D & C & B & E & D & C & B & A \end{pmatrix} \quad (9)$$

with $A \geq E \geq B \geq C \geq D > 0$. By averaging the entries in the sample covariance matrix that correspond to equal entries in (9), we obtain for Σ the estimate

$$\hat{\Sigma} = \Sigma(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}) \quad (10)$$

where $(\hat{A}, \hat{E}, \hat{B}, \hat{C}, \hat{D}) = (.00317, .00209, .00134, .00079, .00044)$.

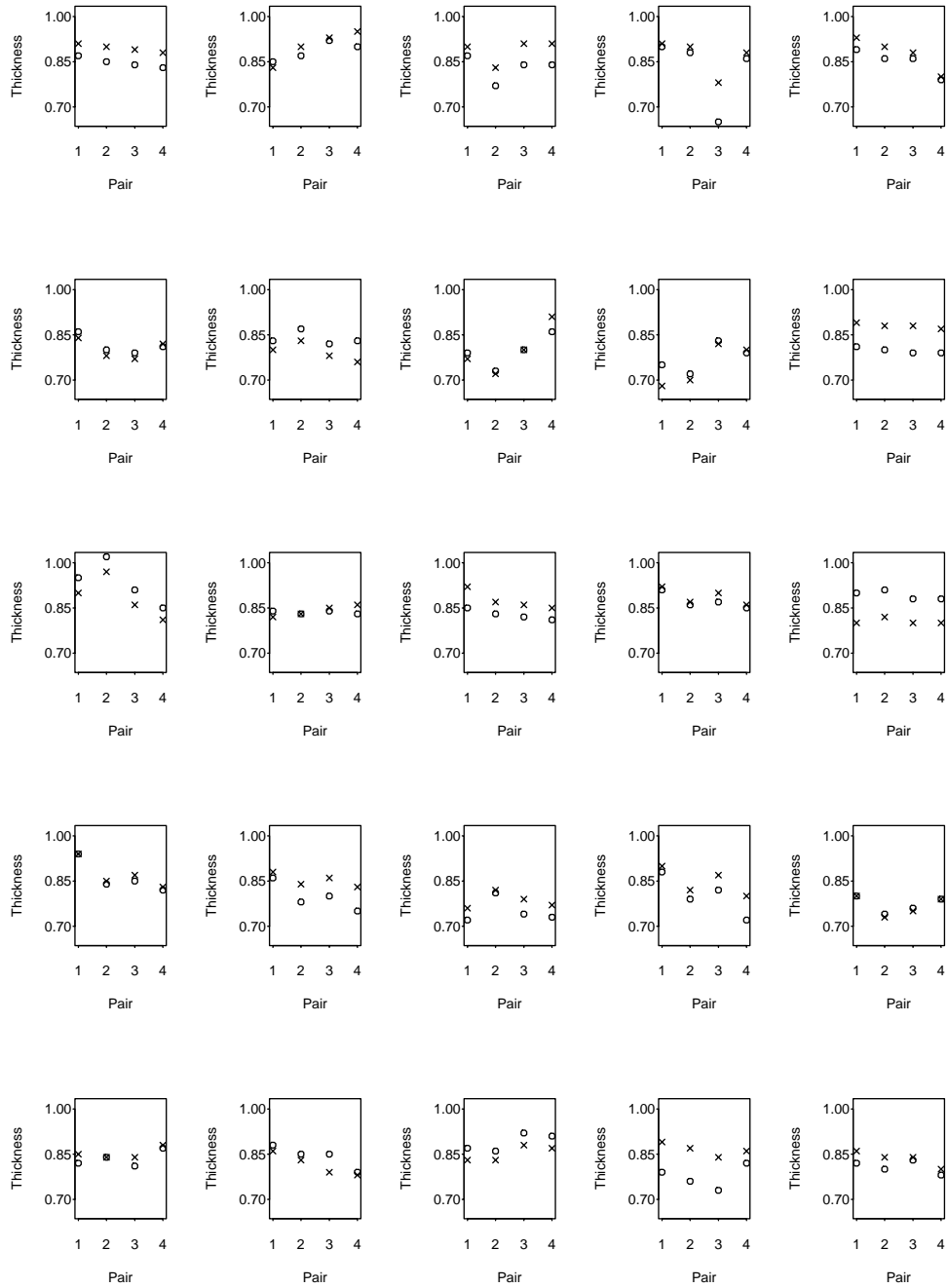


Figure 1. Thickness measurements on a sample of 25 boards. The symbols o and x denote opposed upper and lower edge measurements at the four pairs of sites on each board.

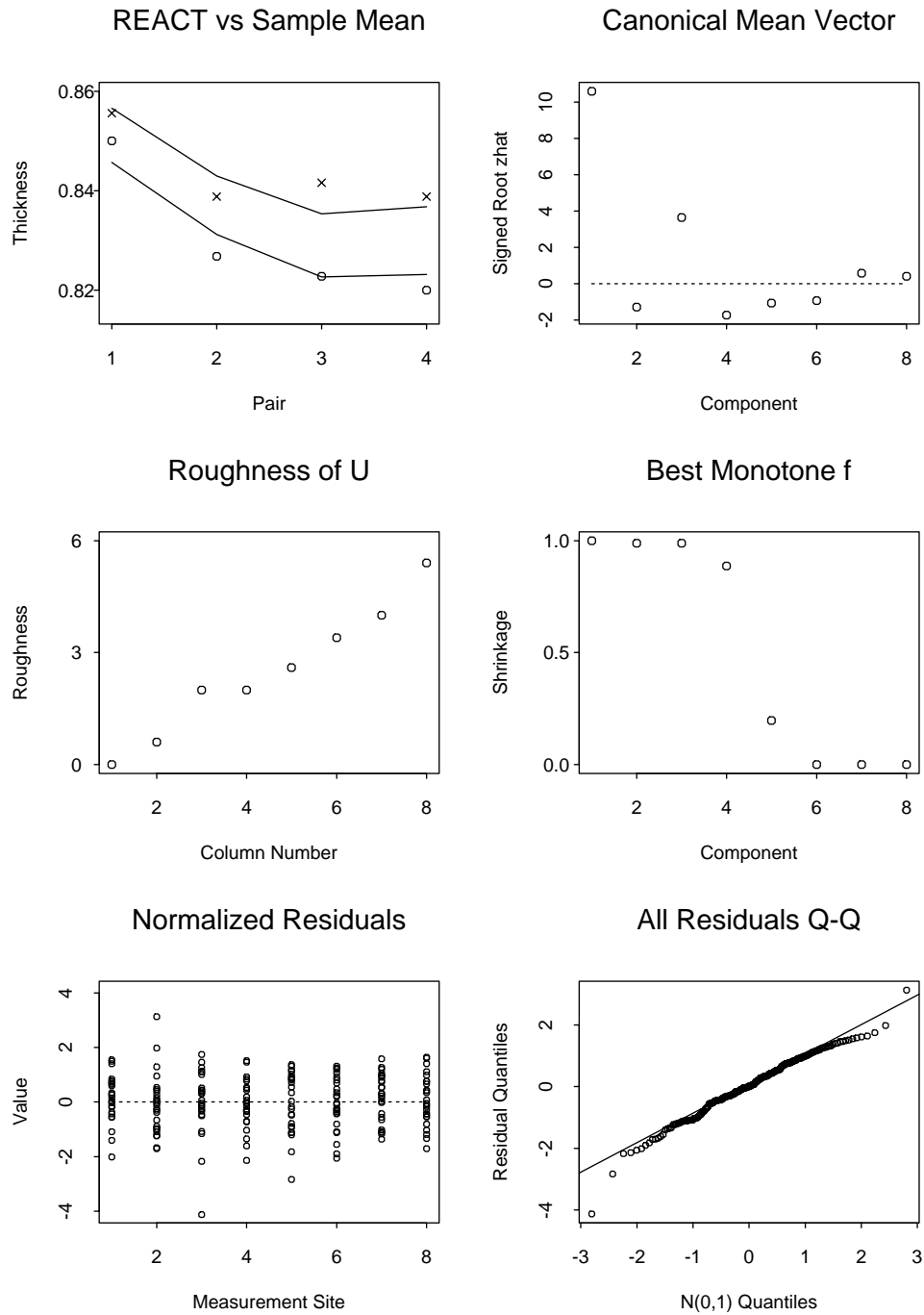


Figure 2. Cell (1,1) displays the REACT estimate $\hat{\mu}_M$ (with interpolated lines) and the sample mean vector (points coded as in Figure 1). The other cells report diagnostic plots discussed in Section 2.

Orthonormal basis U . We construct an ordered tensor-product basis for R^8 as follows. Let $s = (1, 2, 3, 4)$ and let V denote the 4×4 orthogonal matrix whose columns are the orthonormal polynomials in s of degrees 0 to 3. The S-PLUS function `poly()` computes V . Letting v_i denote the i -th column of V , define W to be the partitioned matrix

$$W = 2^{-1/2} \begin{pmatrix} v_1 & v_1 & v_2 & v_2 & v_3 & v_3 & v_4 & v_4 \\ v_1 & -v_1 & v_2 & -v_2 & v_3 & -v_3 & v_4 & -v_4 \end{pmatrix}. \quad (11)$$

The columns of W form an orthonormal basis for R^8 . To obtain a basis that is plausibly economical for expressing the transformed mean thickness vector $\hat{\Sigma}^{-1/2}\mu$, we reorder the columns $\{w_i\}$ of W from least to most rough. Such a reordered basis should be economical if the components of transformed mean thickness vary slowly as we move to adjacent measurement sites. The function

$$\text{Rough}(x) = \sum_{i=2}^4 (x_i - x_{i-1})^2 + \sum_{i=6}^8 (x_i - x_{i-1})^2 + \sum_{i=1}^4 (x_{i+4} - x_i)^2 \quad (12)$$

is taken to measure the roughness of any vector $x \in R^8$. Reordering the columns of W according to their Rough values generates the orthonormal basis matrix

$$U = (w_1, w_3, w_5, w_2, w_4, w_7, w_6, w_8). \quad (13)$$

Cell (2,1) in Figure 2 displays the Rough values for successive columns of U . The corresponding values of the canonical mean vector \hat{z} , defined in (5), are plotted in cell (1,2). The small magnitudes of the higher order components of \hat{z} suggest that the basis U is, in fact, economical in representing the mean vector μ .

Computing $\hat{\mu}_M$. This is straightforward from (8) and the preceding definitions once we have found the empirically best monotone shrinkage vector \hat{f}_M , which minimizes $\hat{\rho}(f)$ over $f \in \mathcal{F}_M$. Let $\hat{g} = 1 - 1/\hat{z}^2$. Then

$$\hat{\rho}(f) = \text{ave}[(f - \hat{g})^2 \hat{z}^2] + \text{ave}(\hat{g}^2). \quad (14)$$

Let $\mathcal{H} = \{h \in R^p: h_1 \geq h_2 \geq \dots \geq h_p\}$. An argument in Beran and Dümbgen [2] deduces from (14) that

$$\hat{f}_M = \check{f}_+ \quad \text{with} \quad \check{f} = \underset{h \in \mathcal{H}}{\text{argmin}} \text{ave}[(h - \hat{g})^2 \hat{z}^2]. \quad (15)$$

The positive-part step arises in (15) because \hat{g} lies in $[-\infty, 1]^p$ rather than in $[0, 1]^p$. The pool-adjacent-violators algorithm, treated by Robertson, Wright and Dykstra [7], provides an effective technique for computing \check{f} and hence \hat{f}_M .

Cell (2,2) of Figure 2 displays the components of \hat{f}_M for the lumber thickness case study. The first three components are very close to 1, the fourth is .89, the fifth is .20, and the last three components are zero. The estimated risk of $\hat{\mu}_M$ is $\hat{\rho}(\hat{f}_M) = .24$, sharply lower than the risk or estimated risk of the sample mean \bar{x} , which is 1.

Cell (1,1) in Figure 2 plots the components of $\hat{\mu}_M$ (with linear interpolation between adjacent sites along each edge) and the corresponding components of \bar{x} . The plot of $\hat{\mu}_M$ suggests that mean thickness decreases as we move down the length of a board;

that upper edge means are consistently smaller than corresponding lower edge means; and that the difference in cross-board mean thickness grows only slowly down the length of the board. The impression left by the plot of \bar{x} is more confused and does not bring out the last feature. In this particular case study, $\hat{\mu}_M$ smooths \bar{x} through shrinkage and choice of the basis U , even though the primary goal is to reduce risk. As an incidental but useful consequence, $\hat{\mu}_M$ is more intelligible than \bar{x} .

Cell (3,1) of Figure 2 displays, component by component, the normalized residual vectors $n^{1/2}\hat{\Sigma}^{-1/2}(x_i - \hat{\mu}_M)$, where $1 \leq i \leq 25$. The Q-Q plot in cell (3,2) compares all 200 residuals against the standard normal distribution. There is no evidence of serious departures from marginal normality of the lumber thickness measurements, from the postulated covariance structure (9), and from the fitted mean vector $\hat{\mu}_M$.

3. ASYMPTOTICALLY MINIMAX ESTIMATORS

This section begins with asymptotic minimax bounds for estimation of μ over certain subsets of the parameter space. Subsection 3.1 gives an oracle estimator that achieves these bounds. The oracle estimator is usually not realizable because its definition requires knowledge of $\mu'\Sigma^{-1}\mu$ and of Σ . However, the form of the oracle estimator motivates, in Subsection 3.2, the definition of the fully adaptive estimator $\hat{\mu}_M$ and provides a path to establishing asymptotic minimaxity of the latter. The choice of the orthogonal basis U is discussed theoretically after Theorems 1 and 4 and is carried out in Section 2 for the lumber-thickness data.

3.1. Minimax Oracle Estimation. We begin by reparametrizing the estimation problem in the oracle world where Σ and $\mu'\Sigma^{-1}\mu$ are known. Let

$$z = n^{1/2}U'\Sigma^{-1/2}\bar{x} \quad \xi = \mathbb{E}z = n^{1/2}U'\Sigma^{-1/2}\mu. \quad (16)$$

Any estimator $\hat{\mu}$ of μ induces the estimator $\hat{\xi} = n^{1/2}U'\Sigma^{-1/2}\hat{\mu}$ of ξ . The mapping between $\hat{\mu}$ and $\hat{\xi}$ is one-to-one as is the mapping between μ and ξ . Risks are placed into correspondence through the loss identity

$$L_{n,p}(\hat{\mu}, \mu, \Sigma) = p^{-1}|\hat{\xi} - \xi|^2. \quad (17)$$

In the oracle world, the problem of estimating μ under loss (1) is equivalent to estimating ξ under quadratic loss (17).

To formulate the notion of basis economy, consider for every $b \in [0, 1]$ and every $r > 0$ the ball

$$B(r, b) = \{\xi: \text{ave}(\xi^2) \leq r \text{ and } \xi_i = 0 \text{ for } i > bp\}. \quad (18)$$

Let u_i denote the i -th column of U . In the original parametrization, $B(r, b)$ corresponds to the ellipsoid

$$D(r, b) = \{\mu: (n/p)\mu'\Sigma^{-1}\mu \leq r \text{ and } u'_i\Sigma^{-1/2}\mu = 0 \text{ for } i > bp\}. \quad (19)$$

If μ lies in $D(r, b)$, then $\Sigma^{-1/2}\mu$ lies in the subspace spanned by the first $[bp]$ columns of U . Regression coefficients with respect to these orthonormal vectors provide a description of $\Sigma^{-1/2}\mu$ which is highly compressed when b is small. We then say that the

basis is economical for estimating μ . Though overly idealized, this definition of economy leads to explicit results that link the economy of the basis with the superefficiency of $\hat{\mu}_M$.

Consider candidate estimators for ξ of the form $\hat{\xi}(f) = fz$, where $f \in \mathcal{F}_M$. These correspond to the candidate estimators

$$\hat{\mu}(f, \Sigma) = \Sigma^{1/2}U \text{diag}(f)U'\Sigma^{-1/2}\bar{x} = n^{-1/2}\Sigma^{1/2}U \text{diag}(f)z \quad (20)$$

for μ . Because of (17), the risk of $\hat{\mu}(f, \Sigma)$ is

$$R_{n,p}(\hat{\mu}(f, \Sigma), \mu, \Sigma) = \rho(f, \xi^2), \quad (21)$$

the function ρ being defined in (2). Let $\tilde{f}_M = \text{argmin}_{f \in \mathcal{F}_M} \rho(f, \xi^2)$. The oracle estimator is $\hat{\mu}(\tilde{f}_M, \Sigma)$, the candidate estimator that minimizes risk. The restriction to candidate estimators indexed by $f \in \mathcal{F}_M$ makes possible successful adaptation (see remarks preceding Theorem 2) as well as fine performance when the basis U is economical (see remarks following Theorems 1 and 4).

Theorem 1. *For every $r > 0$ and $b \in [0, 1]$,*

$$\lim_{p \rightarrow \infty} \sup_{\mu \in D(r,b)} R_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma) = rb/(r + b). \quad (22)$$

The asymptotic minimax risk over all estimators of μ is

$$\lim_{p \rightarrow \infty} \inf_{\hat{\mu}} \sup_{\mu \in D(r,b)} R_{n,p}(\hat{\mu}, \mu, \Sigma) = rb/(r + b). \quad (23)$$

The asymptotic minimax bound in (23) is thus achieved by the oracle estimator. For fixed b , the asymptotic maximum risk of $\hat{\mu}(\tilde{f}_M, \Sigma)$ increases monotonically in r but never exceeds b . In sharp contrast, the risk of \bar{x} is always 1 whatever the value of μ . The first message of Theorem 1 is that we can only gain, when p is not small, by using the oracle estimator in place of the sample mean \bar{x} . The second message is that the reduction in maximum risk achieved by the oracle estimator can be remarkable if b is close to zero. This occurs when the basis U used to define the oracle estimator is highly economical. We note that the minimax asymptotics are uniform over subsets of μ and thus are considerably more trustworthy than risk limits computed pointwise in μ .

3.2. Successful Adaptation. The oracle estimator depends on ξ^2 and Σ , both of which are typically unknown. To devise a realizable estimator that does not depend on unknown parameters, we proceed as follows. Let $\hat{\Sigma}$ be a consistent estimator of Σ . Then

$$\hat{z} = n^{1/2}U'\hat{\Sigma}^{-1/2}\bar{x} \quad (24)$$

plausibly estimates $z = n^{1/2}U'\Sigma^{-1/2}\bar{x}$. Consider the realizable candidate estimators $\hat{\mu}(f, \hat{\Sigma})$ where f ranges over \mathcal{F}_M . In view of (21), the function $\hat{\rho}(f)$ defined in (6) estimates the risk of these candidate estimators. This risk estimator is suggested by the Mallows [5] C_L criterion or the Stein [9] unbiased risk estimator, with plug-in estimation of the unknown covariance matrix. By analogy with the construction of

the oracle estimator, we minimize estimated risk over the candidate estimators to obtain the estimator $\hat{\mu}_M$ defined in (8). We will show in Theorem 4 that $\hat{\mu}_M$ shares the asymptotic minimaxity of the oracle estimator.

Let $|\cdot|$ denote the Euclidean matrix norm, which is defined by $|A|^2 = \text{tr}[AA']$. If A_1 and A_2 are both $p \times p$ matrices, then the Cauchy-Schwarz inequality for this norm asserts that $|A_1 A_2| \leq |A_1| |A_2|$. The following consistency condition will be imposed upon the estimator $\hat{\Sigma}$.

Condition C. *The estimators $\hat{\Sigma}$ and \bar{x} are independent. Let $\hat{V} = \Sigma^{-1/2} \hat{\Sigma}^{1/2}$. For every $r > 0$,*

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} \text{E} |\hat{V} - I_p|^2 = 0 \quad \lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} \text{E} |\hat{V}^{-1} - I_p|^2 = 0. \quad (25)$$

In this statement, the relative rates at which n and p tend to infinity will depend on the covariance estimator $\hat{\Sigma}$. For instance, if $\hat{\Sigma}$ is the sample covariance matrix based on the observed (x_1, x_2, \dots, x_n) , then Condition C holds provided p and n tend to infinity in such a way that p^2/n tends to zero. In the lumber data example or in time-series contexts, restrictions may be imposed on the form of Σ . Condition C may then hold for suitably constructed $\hat{\Sigma}$ under less severe limitations on the rate at which p increases with n .

The next two theorems, proved in Section 4, show that the estimated risk function $\hat{\rho}(f)$ and the adaptive estimator $\hat{\mu}_M$ serve asymptotically as valid surrogates for $\rho(f, \xi^2)$ and the oracle estimator $\hat{\mu}(\tilde{f}_M, \Sigma)$. It is important to note that similar results do not hold if the class of monotone shrinkage vectors \mathcal{F}_M , defined before display (3), is replaced by a much larger class of shrinkage vectors such as the global class $\mathcal{F}_G = [0, 1]^p$. Adaptation over \mathcal{F}_G produces an inadmissible estimator of μ , as shown in [2].

Theorem 2. *Suppose that $\hat{\Sigma}$ satisfies Condition C. For every $r > 0$ and every positive definite Σ ,*

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} \text{E} \sup_{f \in \mathcal{F}_M} |L_{n,p}(\hat{\mu}(f, \hat{\Sigma}), \mu, \Sigma) - \rho(f, \xi^2)| = 0 \quad (26)$$

and

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} \text{E} \sup_{f \in \mathcal{F}_M} |\hat{\rho}(f) - \rho(f, \xi^2)| = 0. \quad (27)$$

Because $\tau_M(\xi^2) = \rho(\tilde{f}_M, \xi^2)$, a consequence of Theorem 2 is

Theorem 3. *Suppose that $\hat{\Sigma}$ satisfies Condition C. For every $r > 0$ and every positive definite Σ ,*

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} \text{E} |T - \tau_M(\xi^2)| = 0, \quad (28)$$

where T can be any one of $L_{n,p}(\hat{\mu}_M, \mu, \Sigma)$, $L_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma)$, or $\hat{\rho}(\tilde{f}_M)$.

Theorem 3 implies the risk convergence (4) and

Theorem 4. *Suppose that $\hat{\Sigma}$ satisfies Condition C. For every $r > 0$, every $b \in [0, 1]$, and every positive definite Σ ,*

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,1)} |R_{n,p}(\hat{\mu}_M, \mu, \Sigma) - R_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma)| = 0 \quad (29)$$

and

$$\lim_{n,p \rightarrow \infty} \sup_{\mu \in D(r,b)} R_{n,p}(\hat{\mu}_M, \mu, \Sigma) = rb/(r+b). \quad (30)$$

By comparing (30) with (23), we see that the adaptive estimator $\hat{\mu}_M$ is asymptotically minimax over $D(r,b)$ and has small maximum risk when b is small, in which event the basis U represents $\Sigma^{-1/2}\mu$ economically. Moreover, (29) shows that the risk of $\hat{\mu}_M$ mimics that of the oracle estimator $\hat{\mu}(\tilde{f}_M, \Sigma)$, uniformly over ellipsoids in the parameter space that correspond to bounds on the signal-to-noise ratio. Theorem 4 thus establishes the success of the adaptation strategy over shrinkage vectors $f \in \mathcal{F}_M$ that is expressed in the definition of $\hat{\mu}_M$.

4. PROOFS

Pinsker's paper [6] yields two minimax theorems for the estimation of ξ from z in the oracle world. Let $\mathcal{E} = \{a \in R^p: a_i \in [1, \infty], 1 \leq i \leq p\}$. For every $a \in \mathcal{E}$, define the ellipsoid

$$E(r, a) = \{\xi \in R^p: \text{ave}(a\xi^2) \leq r\}. \quad (31)$$

When $\xi \in E(r, a)$ and $a_i = \infty$, it is to be understood that $\xi_i = 0$ and $a_i^{-1} = 0$. Let

$$\xi_0^2 = [(\delta/a)^{1/2} - 1]_+ \quad g_0 = \xi_0^2/(1 + \xi_0^2) = [1 - (a/\delta)^{1/2}]_+, \quad (32)$$

where δ is the unique positive number such that $\text{ave}(a\xi_0^2) = r$. Define

$$\nu_p(r, a) = \rho(g_0, \xi_0^2) = \text{ave}[\xi_0^2/(1 + \xi_0^2)]. \quad (33)$$

Evidently, $\nu_p(r, a) \in [0, 1]$ for every $r > 0$ and every $a \in \mathcal{E}$.

The first theorem that can be specialized from Pinsker's reasoning identifies the linear estimator that is minimax among all linear estimators of ξ and finds the minimax risk for this class.

Theorem 5. *For every $a \in \mathcal{E}$ and every $r > 0$,*

$$\inf_{f \in R^p} \sup_{\xi \in E(r,a)} E|fz - \xi|^2 = \nu_p(r, a) = \sup_{\xi \in E(r,a)} E|g_0z - \xi|^2. \quad (34)$$

The second theorem gives conditions under which the minimax linear estimator g_0z is asymptotically minimax among all estimators of ξ .

Theorem 6. *For every $a \in \mathcal{E}$ and every $r > 0$ such that $\lim_{p \rightarrow \infty} p\nu_p(r, a) = \infty$,*

$$\lim_{p \rightarrow \infty} [\inf_{\hat{\xi}} \sup_{\xi \in E(r,a)} E|\hat{\xi} - \xi|^2 / \nu_p(r, a)] = 1. \quad (35)$$

If $\liminf_{p \rightarrow \infty} \nu_p(r, a) > 0$, then also

$$\lim_{p \rightarrow \infty} [\inf_{\hat{\xi}} \sup_{\xi \in E(r,a)} E|\hat{\xi} - \xi|^2 - \nu_p(r, a)] = 0. \quad (36)$$

Because g_0 depends on r and a , the asymptotic minimaxity of $g_0 z$ is assured only over the one ellipsoid $E(r, a)$. The following construction yields an oracle estimator that is asymptotically minimax over a class of such ellipsoids. Let $\mathcal{E}_0 \subset \mathcal{E}$ and \mathcal{F} be such that $g_0(r, a) \in \mathcal{F}$ for every $a \in \mathcal{E}_0$ and every $r > 0$. To enable successful adaptation, we will require that the shrinkage class \mathcal{F} be not too large. This requirement limits the choice of \mathcal{E}_0 . Let $\tilde{f} = \operatorname{argmin}_{f \in \mathcal{F}} \rho(f, \xi^2)$. Because both \tilde{f} and g_0 lie in \mathcal{F} , it follows that

$$\sup_{\xi \in E(r, a)} \mathbb{E} |\tilde{f} z - \xi^2| \leq \sup_{\xi \in E(r, a)} \mathbb{E} |g_0 z - \xi|^2 = \nu_p(r, a) \quad (37)$$

for every $a \in \mathcal{E}_0$ and every $r > 0$. This implies the asymptotic minimaxity of $\tilde{f} z$ over the class of ellipsoids $E(r, a)$ that is generated as a ranges over \mathcal{E}_0 and r ranges over the positive reals.

Proof of Theorem 1. In the transformed problem, candidate estimator $\hat{\mu}(f, \Sigma) = fz$. The ball $B(r, b)$ defined in (18) is the specialization of $E(r, a)$ when $a_i = 1$ for $1 \leq i \leq bp$ and $= \infty$ otherwise. In this case, (32) and (33) imply that $\lim_{p \rightarrow \infty} \nu_p(r, a) = rb/(r + b)$ and that g_0 has coefficients $g_{0,i} = [1 - \delta^{-1/2}]_+$ for $1 \leq i \leq bp$ and $= 0$ otherwise. Consequently, $g_0 \in \mathcal{F}_M$. The asymptotic minimax bound (23) is the specialization of (36) while (22) follows from (37) with $\mathcal{F} = \mathcal{F}_M$.

Proof of Theorem 2. If X and Y are non-negative random variables, then

$$\mathbb{E} |X^2 - Y^2| \leq \mathbb{E} |X - Y|^2 + 2\mathbb{E}^{1/2} Y^2 \cdot \mathbb{E}^{1/2} |X - Y|^2. \quad (38)$$

We first prove (27). The definitions (16) and (24) of z and \hat{z} entail that

$$\hat{z} - z = n^{1/2} U' (\hat{V}^{-1} - I_p) \Sigma^{-1/2} \bar{x}. \quad (39)$$

From this, Condition C, and the Cauchy-Schwarz inequality for the matrix norm,

$$\mathbb{E} |\hat{z} - z|^2 \leq p[1 + \operatorname{ave}(\xi^2)] \mathbb{E} |\hat{V}^{-1} - I_p|^2. \quad (40)$$

Let

$$\check{\rho}(f) = \operatorname{ave}[f^2 + (1 - f)^2(z^2 - 1)]. \quad (41)$$

It follows from the definition (6) of $\hat{\rho}(f)$, (38), (40) and Condition C that

$$\lim_{n, p \rightarrow \infty} \sup_{\xi \in B(r, 1)} \mathbb{E} \sup_{f \in [0, 1]^p} |\hat{\rho}(f) - \check{\rho}(f)|^2 = 0. \quad (42)$$

On the other hand, Lemmas 6.3 (first part) and 6.4 in [2] imply

$$\lim_{p \rightarrow \infty} \sup_{\xi \in B(r, 1)} \mathbb{E} \sup_{f \in \mathcal{F}_M} |\check{\rho}(f) - \rho(f, \xi^2)|^2 = 0. \quad (43)$$

In (43), the distribution of the difference does not depend on n ; and it is not possible to replace $f \in \mathcal{F}_M$ with $f \in [0, 1]^p$ for reasons discussed in [2]. Limit (27) is immediate from (42) and (43).

Next, observe that

$$L_{n, p}(\hat{\mu}(f, \hat{\Sigma}), \mu, \Sigma) = p^{-1} |\hat{V} U \operatorname{diag}(f) \hat{z} - U \xi|^2. \quad (44)$$

and that $|U \operatorname{diag}(f)z - U\xi|^2 = |fz - \xi|^2$. From these facts plus (38), (40) and Condition C follows

$$\lim_{n,p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} \sup_{f \in [0,1]^p} |L_{n,p}(\hat{\mu}(f, \hat{\Sigma}), \mu, \Sigma) - p^{-1}|fz - \xi|^2| = 0. \quad (45)$$

On the other hand, Lemmas 6.3 (second part) and 6.4 in [2] entail

$$\lim_{p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} \sup_{f \in \mathcal{F}_M} |p^{-1}|fz - \xi|^2 - \rho(f, \xi^2)| = 0. \quad (46)$$

Limit (26) is the consequence of (45) and (46).

Proof of Theorem 3. Limit (27) implies that

$$\lim_{n,p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} |\hat{\rho}(\hat{f}_M) - \rho(\hat{f}_M, \xi^2)| = 0 \quad (47)$$

and

$$\lim_{n,p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} |\hat{\rho}(\tilde{f}_M) - \rho(\tilde{f}_M, \xi^2)| = 0. \quad (48)$$

In view of (3), $\tau_M(\xi^2) = \rho(\tilde{f}_M, \xi^2)$. Consequently, limit (28) holds for $T = \hat{\rho}(\hat{f}_M)$ and, in addition,

$$\lim_{n,p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} |\rho(\hat{f}_M, \xi^2) - \tau_M(\xi^2)| = 0. \quad (49)$$

On the other hand, limit (26) implies that

$$\lim_{n,p \rightarrow \infty} \sup_{\xi \in B(r,1)} \mathbb{E} |L_{n,p}(\hat{\mu}_M, \mu, \Sigma) - \rho(\hat{f}_M, \xi^2)| = 0. \quad (50)$$

Combining this result with (49) yields (28) for $T = L_{n,p}(\hat{\mu}_M, \mu, \Sigma)$. Because $\hat{\Sigma} = \Sigma$ satisfies Condition C, it is also true that (28) holds for $T = L_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma)$.

Proof of Theorem 4. Note that

$$|R_{n,p}(\hat{\mu}_M, \mu, \Sigma) - R_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma)| \leq \mathbb{E} |L_{n,p}(\hat{\mu}_M, \mu, \Sigma) - L_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma)|. \quad (51)$$

Limit (29) follows from this inequality and Theorem 3.

Because $D(r, b)$ is a subset of $D(r, 1)$, limit (29) entails

$$\lim_{n,p \rightarrow \infty} \left| \sup_{\mu \in D(r,b)} R_{n,p}(\hat{\mu}_M, \mu, \Sigma) - \sup_{\mu \in D(r,b)} R_{n,p}(\hat{\mu}(\tilde{f}_M, \Sigma), \mu, \Sigma) \right| = 0. \quad (52)$$

This together with (22) implies (30).

REFERENCES

1. R. Beran, REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Soc.* (2000) **95**, in press.
2. R. Beran and L. Dümbgen, Modulation of estimators and confidence sets. *Ann. Statist.* (1998) **26**, 1826–1856.
3. D. L. Donoho and I. M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* (1995) **90**, 1200–1224.
4. S. Efromovich, Quasi-linear wavelet estimation. *J. Amer. Statist. Soc.* (1999) **94**, 189–204.
5. C. L. Mallows, Some comments on C_p . *Technometrics* (1973) **15**, 661–676.
6. M. S. Pinsker, Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* (1980) **16**, 120–133.
7. T. Robertson, F. T. Wright, R. L. Dykstra, *Order Restricted Statistical Inference*. Wiley, New York, 1988.
8. C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: *Proc. Third Berkeley Symp. Math. Statist. Prob.* (ed. J. Neyman). Univ. Calif. Press, Berkeley, 1956, 197–206.
9. C. Stein, Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* (1981) **9**, 1135–1151.