Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 137 (2007) 684–696

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Adaptation over parametric families of symmetric linear estimators[☆]

Rudolf Beran[*]

Department of Statistics, University of California, Davis, CA 95616, USA

Available online 28 July 2006

Abstract

This paper treats an abstract parametric family of symmetric linear estimators for the mean vector of a standard linear model. The estimator in this family that has smallest *estimated* quadratic risk is shown to attain, asymptotically, the smallest risk achievable over all candidate estimators in the family. The asymptotic analysis is carried out under a strong Gauss–Markov form of the linear model in which the dimension of the regression space tends to infinity. Leading examples to which the results apply include: (a) penalized least squares fits constrained by multiple, weighted, quadratic penalties; and (b) running, symmetrically weighted, means. In both instances, the weights define a parameter vector whose natural domain is a continuum.

MSC: Primary 62G08; secondary 62G20

Keywords: Biased estimator; Estimated risk; Dimensional asymptotics; Penalized least squares; Running weighted average; Regularization; Linear model

1. Introduction

Consider the standard linear model in which the $n \times 1$ observation vector y satisfies

$$y = \eta + e, \quad \eta = X\beta. \tag{1.1}$$

Here, the design matrix X is $n \times p$ and has rank p, the $p \times 1$ vector β is unknown, and the components of the $n \times 1$ vector e are independent, identically distributed with mean zero, unknown variance σ^2, and finite fourth moment. For brevity, we will call this the *strong Gauss–Markov model*. The problem of estimating η well under this model has motivated major developments in statistical theory and practice.

For any matrix A, including the special case of a vector, let $|A|$ denote the Euclidean (or Frobenius) norm: $|A|^2 = \mathrm{tr}(A'A) = \mathrm{tr}(AA')$. Define the normalized quadratic loss of any estimator $\hat{\eta}$ of η to be

$$L(\hat{\eta}, \eta) = p^{-1}|\hat{\eta} - \eta|^2. \tag{1.2}$$

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

[*] Tel.: +1 530 754 7765; fax: +1 530 750 7099.
E-mail address: beran@wald.ucdavis.edu.

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.06.029

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

$$R(\hat{\eta}, \eta, \sigma^2) = EL(\hat{\eta}, \eta), \tag{1.3}$$

where the expectation is calculated under the strong Gauss–Markov model (1.1).

A *linear estimator* of η has the form $\hat{\eta} = Ay$, where A is a $n \times n$ matrix that does not depend on y. It may be biased or unbiased for η. The least squares estimator, $\hat{\eta}_{\mathrm{LS}} = X(X'X)^{-1}X'y$, is an unbiased linear estimator with risk $R(\hat{\eta}_{\mathrm{LS}}, \eta, \sigma^2) = \sigma^2$. According to the Gauss–Markov theorem, $\hat{\eta}_{\mathrm{LS}}$ has smallest risk among all linear unbiased estimators of η. However, Stein (1956) proved that the least squares estimator is inadmissible for η under quadratic loss whenever $p \geqslant 3$ and the errors are independent, identically normally distributed. In statistical practice, $\hat{\eta}_{\mathrm{LS}}$ is often too variable an estimator unless the number p of regressors is small.

These findings have led to consideration of biased linear estimators for η. The risk function of the linear estimator $\hat{\eta} = Ay$ is

$$R(Ay, \eta, \sigma^2) = p^{-1}[\sigma^2 \operatorname{tr}(A'A) + \eta'(I_n - A)'(I_n - A)\eta]. \tag{1.4}$$

This risk is a convex function of A. Its form suggests the possibility of reducing risk through trade-off, by choice of A, between the variance terms $\sigma^2 \operatorname{tr}(A'A)$ and the bias term $\eta'(I_n - A)'(I_n - A)\eta$.

We find the matrix $\tilde{A}$ that minimizes the risk (1.4). Because

$$R(Ay, \eta, \sigma^2) = p^{-1}[\sigma^2 \operatorname{tr}(A'A) + \eta'\eta - 2\eta'A\eta + \eta'A'A\eta] \tag{1.5}$$

and the matrix derivatives

$$\partial \operatorname{tr}(A'A)/\partial A = 2A, \quad \partial\eta'A\eta/\partial A = \eta\eta', \quad \partial\eta'A'A\eta/\partial A = 2A\eta\eta' \tag{1.6}$$

(cf. Section A.15 in Rao and Toutenberg, 1995), it follows that:

$$\partial R(Ay, \eta, \sigma^2)/\partial A = p^{-1}[2\sigma^2 A - 2\eta\eta' + 2A\eta\eta']. \tag{1.7}$$

Setting this risk derivative equal to zero and simplifying yields

$$\tilde{A} = I_n - (I_n + \sigma^{-2}\eta\eta')^{-1} = (\sigma^2 + |\eta|^2)^{-1}\eta\eta'. \tag{1.8}$$

Let

$$H = X'X, \quad U = XH^{-1/2}. \tag{1.9}$$

Evidently U is $n \times p$ and $U'U = I_p$. It follows from (1.9) that $\eta = U\xi$ with $\xi = H^{1/2}\beta$. Consequently,

$$\tilde{A} = U\tilde{S}U', \quad \tilde{S} = (\sigma^2 + |\xi|^2)^{-1}\xi\xi'. \tag{1.10}$$

Note that $\tilde{S}$ is $p \times p$ symmetric with all eigenvalues between 0 and 1. The oracle linear estimator $\tilde{A}y$ minimizes risk among all linear estimators Ay. It is usually not realizable because we usually lack accurate knowledge of $\eta\eta'$ and σ^2. However, representation (1.10) indicates that, among linear estimators of η, we may reasonably restrict attention to those having the form USU', where S is a $p \times p$ symmetric matrix with all eigenvalues between 0 and 1 and U is given by (1.9). The least squares estimator $\hat{\eta}_{\mathrm{LS}}$ has this form with $S = I_p$.

An extensive literature, reviewed in Buja et al. (1989) and in Kneip (1994), has developed specific examples of symmetric linear estimators that can reduce risk by smoothing or shrinkage, with submodel fitting as a limiting case. In considering this work, it is important for understanding to distinguish between two problems: estimation of the discrete vector η versus estimation of a function that coincides with η at certain design points. The first problem, estimating discrete η, arises in analyzing discrete complete or incomplete multi-way layouts, including regression. Solutions to the discrete problem do not require existence or smoothness of an interpolating function that is to be estimated.

This paper studies parametric families of symmetric linear estimators for discrete η in model (1.1). Examples motivate the following abstract description. Let N denote a closed subset of $[0, 1]^k$, where k is fixed a priori. Let U be the matrix defined in (1.9) and let $\{S(t): t \in N\}$ be a family of symmetric $p \times p$ matrices indexed by the vector parameter t. The eigenvalues of each matrix in the family lie between 0 and 1. The estimators

$$\hat{\eta}(t) = US(t)U'y, \quad t \in N \tag{1.11}$$

constitute a *parametric family* of symmetric linear estimators of η. Representation (1.11) is not unique: replacing U with UO and $S(t)$ with $O'S(t)O$ for any $p \times p$ orthogonal matrix O leaves the estimator unchanged.

Remark. The asymptotic theory to be developed in this paper lets p, the dimension of the regression space, tend to infinity. The sample size n satisfies the condition $n \geqslant p$. Almost every mathematical object considered in the paper depends on p and some depend also on n. For instance, a fuller notation would write $S_p(t)$, $U_{p,n}$ and $\hat{\eta}_{p,n}(t)$ in (1.11). All estimated quantities marked by the caret ˆ depend on p and n, as may their loss and risk. To avoid burdensome typography, we generally omit the subscripts p and n.

Example 1. *Running weighted means*. As an example of structure (1.11), consider a balanced one-way layout in which the factor influencing response is ordinal, with p equally spaced levels, and r observations are made at each factor level. If the vector y records replicated measurements in adjacent components, this one-way layout is a special case of (1.1) with $n = rp$, $X = I_p \otimes u_r$, and u_r an $r \times 1$ vector whose components are each 1. Thus $H = rI_p$ and $U = r^{-1/2}X$.

Let

$$S(t) = \begin{pmatrix} t_1 + t_2 & t_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ t_2 & t_1 & t_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & t_2 & t_1 & t_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & t_2 & t_1 & t_2 \\ 0 & 0 & 0 & 0 & \dots & 0 & t_2 & t_1 + t_2 \end{pmatrix} \tag{1.12}$$

for $t \in [0, 1]^2$. The first and last rows of this matrix are obtained by reflection at the boundary. Let N be $\{t \in [0, 1]^2: t_1 + 2t_2 = 1, t_1 \geqslant .5\}$. For $t \in N$, the eigenvalues of $S(t)$ all lie in $[0, 1]$. The running weighted mean with span 3 is defined to be the positive symmetric linear estimator $\hat{\eta}_{\mathrm{RWM}}(t)$ defined by (1.11) and (1.12) for $t \in N$. The definition is readily extended from span 3 to any other odd span.

Example 2. *Multiply penalized least squares.* For any matrix A, let

$$\rho(A) = \rho[A] = \sup_{x \neq 0} \frac{|Ax|}{|x|}. \tag{1.13}$$

The function ρ is a matrix norm whose further properties are summarized in Proposition 1 of Section 4.

Let $\{t_i: 1 \leqslant i \leqslant k\}$ be relative penalty weights such that $0 \leqslant t_i \leqslant 1$. Let c be a positive constant, possibly large. Let $\{Q_i: 1 \leqslant i \leqslant k\}$ be symmetric positive semi-definite matrices. The corresponding penalized least squares estimator of η is defined to be $\hat{\eta}_{\mathrm{PLS}}(t) = X\hat{\beta}(t)$, where

$$\hat{\beta}(t) = \operatorname*{argmin}_{\beta} \left[|y - X\beta|^2 + c \sum_{i=1}^{k} t_i \beta' Q_i \beta \right] = \left(X'X + c \sum_{i=1}^{k} t_i Q_i \right)^{-1} X'y. \tag{1.14}$$

The penalized least squares construction generates a parametric family of symmetric linear estimators with $N = [0, 1]^k$, U as in (1.9), and

$$S(t) = \left[I_p + c \sum_{i=1}^{k} t_i V_i \right]^{-1}, \quad V_i = H^{-1/2} Q_i H^{-1/2}, \quad t \in [0, 1]^k. \tag{1.15}$$

Note that the matrix $S(t)$ in (1.15) has all eigenvalues between 0 and 1 for every $t \in N$. For the theory to be developed, we will scale each penalty matrix Q_i so that $\rho(V_i) = 1$.

An instance of this structure arises for the one-way layout in which the responses y depend smoothly upon an ordinal factor whose levels are equally spaced. It is natural in this case to consider penalized least squares estimators (1.15) that are constructed from ith difference operators $\{D_i : 1 \leqslant i \leqslant k\}$ as follows: each Q_i is proportional to the symmetric matrix $D_i' D_i$, scaled so that $\rho(V_i) = 1$. This class of penalized least squares estimators includes one-parameter smoothers described by Buja et al. (1989) and by Beran (2000, 2002). With other definitions of the $\{Q_i\}$, the class of candidate estimators (1.15) includes discrete analogs of spline estimators considered by Heckman and Ramsay (2000) that use a single quadratic penalty term based on a linear differential operator.

Example 3. *Submodel selection*. Suppose that the $\{K_j : 1 \leqslant j \leqslant m\}$ are m distinct subsets of $\{1, 2, \dots, k\}$. For $1 \leqslant j \leqslant m$, define $t^{(j)} \in [0, 1]^k$ as follows: the ith component is 1 if $i \in K_j$ and is zero otherwise. Submodel fit $\hat{\eta}_{\mathrm{SM}}(t^{(j)})$ is defined to be the constrained least squares estimator of $\eta = X\beta$ that minimizes residual sum of squares subject to $\beta' Q_i \beta = 0$ for every $i \in K_j$. These quadratic constraints are, respectively, equivalent to the linear constraints $Q_i^{1/2}\beta = 0$ for every $i \in K_j$. As a linearly constrained least squares fit, $\hat{\eta}_{\mathrm{SM}}(t^{(j)})$ is an orthogonal projection of y onto a subspace of the column space of X. It is therefore a symmetric linear estimator of the form (1.11). The finite set N indexing competing submodel fits consists of the elements $\{t^{(j)} \in [0, 1]^k : 1 \leqslant j \leqslant m\}$.

Submodel fit $\hat{\eta}_{\mathrm{SM}}(t^{(j)})$ is also the limit of the penalized least squares fit $\hat{\eta}_{\mathrm{PLS}}(t^{(j)})$ when c tends to infinity. Thus, penalized least squares estimators with $t \in [0, 1]$ and c not small essentially enlarge a certain class of submodel estimators of η.

The estimators $\{\hat{\eta}(t) : t \in N\}$ defined in (1.11) constitute a class of candidate symmetric linear estimators for η. How are we to choose t to obtain an estimator with relatively low risk within the class of candidates? If we knew the risk function of $\hat{\eta}(t)$, we would naturally use the oracle estimator $\hat{\eta}(\tilde{t})$, where $\tilde{t}$ minimizes the risk over all $t \in N$. Because the risk function is usually unknown, we pursue the following modified program:

- Construct an estimator $\hat{r}(t)$ of the risk function of $\hat{\eta}(t)$.
- Construct adaptive symmetric linear estimator $\hat{\eta}(\hat{t})$ such that $\hat{t} = \mathrm{argmin}_{t \in N}\, \hat{r}(t)$.
- Find theoretical conditions in the strong Gauss–Markov model under which the loss and estimated risk functions of $\hat{\eta}(t)$ converge uniformly over $t \in N$ to the true risk function as p tends to infinity.
- Hence, deduce that the risk of adaptive estimator $\hat{\eta}(\hat{t})$ converges to the risk of oracle estimator $\hat{\eta}(\tilde{t})$ as p tends to infinity. In other words, show that the asymptotic risk of $\hat{\eta}(\hat{t})$ converges to the smallest risk achievable over the class of candidate estimators $\{\hat{\eta}(t) : t \in N\}$.

The estimated risk function used in this paper is equivalent to the Mallow's (1973) C_p criterion. The main results provide a firm theoretical basis for the program outlined above.

Pertinent earlier work includes the following papers. For the special cases of nested model selection, ridge regression, and some other examples where $k = 1$, Li (1985, 1987) established the convergence of $\hat{r}(t)$ to the loss of $\hat{\eta}(t)$. Kneip (1994) gave related results for the class of ordered linear smoothers. The latter class abstracts essential properties of Example 1 when $k = 1$ but does not cover Example 2 with $k > 1$. Beran and Dümbgen (1998) carried out the foregoing program for *normally* distributed errors with $S(f) = V\,\mathrm{diag}(f)V'y$, where V is a *fixed* $p \times p$ matrix whose columns are orthonormal and f is a $p \times 1$ vector that is free to range all vectors in R^p with $1 \geqslant f_1 \geqslant f_2 \geqslant \cdots \geqslant f_p \geqslant 0$. Their results may be extended to other examples where the eigenvectors of $US(t)U'$ do not depend on t (cf. Beran, 2000, 2002).

Examples 1 and 2 provide instances of symmetric linear estimators for which *both* the eigenvectors and eigenvalues of $S(t)$ depend on t and the number k of penalty terms may exceed 1. When N is not a finite set, asymptotic analysis of data-based choice of $t \in N$ in such instances falls outside the scope of the earlier work just cited. The theorems established in this paper provide a new template for checking success of adaptation over parametric families of symmetric linear estimators. The theorems hold under the strong Gauss–Markov model, do not restrict the form of the spectral decomposition of $S(t)$, and will be applied to the examples.

2. Adaptive symmetric linear estimators

This section carries out the program outlined in the Introduction. We assume throughout the strong Gauss–Markov linear model (1.1). It follows from (1.9) that $\eta = U\xi$ with $\xi = H^{1/2}\beta$. Hence $\xi = U'\eta$. Let $z = U'y$. Evidently z has mean vector ξ and covariance matrix $\sigma^2 I_p$. The loss function of candidate symmetric linear estimator $\hat{\eta}(t) = US(t)U'y$ is

$$L(\hat{\eta}(t), \eta) = p^{-1}|\hat{\eta}(t) - \eta|^2 = p^{-1}|S(t)z - \xi|^2. \tag{2.1}$$

Let $T(t) = [S(t)]^2$ and $\bar{T}(t) = [I_p - S(t)]^2$. From (2.1), the risk function of the candidate symmetric linear estimator is

$$r(t) = p^{-1}[\sigma^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\{\bar{T}(t)\xi\xi'\}]. \tag{2.2}$$

Let $\hat{\sigma}^2$ be an L_1-consistent estimator of σ^2. An asymptotically unbiased estimator of $\xi\xi'$ is $zz' - \hat{\sigma}^2 I_p$. Mallows (1973) estimator of risk function (2.1), implicit in the derivation of the C_p criterion, is then

$$\hat{r}(t) = p^{-1}[\hat{\sigma}^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\{\bar{T}(t)(zz' - \hat{\sigma}^2 I_p)\}]. \tag{2.3}$$

The (not necessarily unique) adaptive symmetric linear estimator is $\hat{\eta}(\hat{t})$, where

$$\hat{t} = \operatorname*{argmin}_{t \in N} \hat{r}(t). \tag{2.4}$$

Remark. Computation of the matrix U and so of $z = U'y$ is an issue because the matrix H may be ill-conditioned. Let $X = WLV'$ denote the reduced singular value decomposition of X: L is $p \times p$ diagonal with positive elements, W is $n \times p$, V is $p \times p$, and $W'W = V'V = VV' = I_p$. It follows from definition (1.9) that $U = WV'$. Stable algorithms for the singular value decomposition thus enable stable computation of z.

When $n > p$, the *least squares variance estimator* is

$$\hat{\sigma}^2_{\mathrm{LS}} = (n-p)^{-1}|y - \hat{\eta}_{\mathrm{LS}}|^2 = e'(I_n - UU')e. \tag{2.5}$$

It is L_1-consistent in the sense of (2.7) below, under the strong Gauss–Markov model (1.1), if $n - p$ tends to infinity as p tends to infinity.

When $n = p$, the absence of replication requires that an L_1-consistent estimator of σ^2 rely on trustworthy prior information about η. For instance, the *first-difference variance estimator*

$$\hat{\sigma}^2_{\mathrm{FD}} = [2(p-1)]^{-1} \sum_{i=2}^{p} (y_i - y_{i-1})^2 \tag{2.6}$$

satisfies (2.7) under the additional side condition that $\lim_{p\to\infty} p^{-1}\sum_{i=2}^{p} (\eta_i - \eta_{i-1})^2 = 0$ (cf. Rice, 1984). Other consistent variance estimators for the case $n = p$, reviewed in Beran (2002), stem from the pooling idea used in ANOVA.

The following Assumptions support the theorems of this paper:

A1. Either $N = [0, 1]^k$ or the cardinality of N is finite, not depending on p. The family of symmetric matrices $\{S(t): t \in N\}$ is such that $\sup_p \sup_{t\in N} \rho[S(t)] < \infty$. When $N = [0, 1]^k$, $S(t)$ is continuous on N and is differentiable on the interior of N with partial derivatives $\{\nabla_i S(t) = \partial S(t)/\partial t_i : 1 \leqslant i \leqslant k\}$ such that $\sup_{i,p} \sup_{t \in N} \rho[\nabla_i S(t)] < \infty$.

A2. The strong Gauss–Markov model (1.1) holds.

A3. Under the strong Gauss–Markov model, the variance estimator $\hat{\sigma}^2$ is L_1-consistent:

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E|\hat{\sigma}^2 - \sigma^2| = 0 \tag{2.7}$$

for every finite $a > 0$ and $\sigma^2 > 0$.

Theorem 1. *Suppose Assumptions* A1 *to* A3 *hold. Let $W(t)$ denote either the loss $L(\hat{\eta}(t), \eta)$ or the estimated risk $\hat{r}(t)$ of candidate estimator $\hat{\eta}(t)$. Then, for every finite $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E\left[\sup_{t\in N} |W(t) - r(t)|\right] = 0. \tag{2.8}$$

This theorem shows that the loss, risk, and estimated risk of candidate estimator $\hat{\eta}(t)$ converge together asymptotically. The uniformity of this convergence over all $t \in N$ makes estimated risk a trustworthy surrogate for its true loss or risk. In the proof, the assumed boundedness of $\rho[S(t)]$ ensures pointwise convergence of $W(t)$ to $r(t)$. The stochastic equicontinuity considerations that strengthen pointwise convergence to the uniform convergence (2.8) draw on the assumed boundedness of the $\{\rho[\nabla_i S(t)]\}$.

Theorem 2. *Suppose Assumptions* A1 *to* A3 *hold. Then, for every finite $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} |R(\hat{\eta}(\hat{t}), \eta, \sigma^2) - r(\tilde{t})| = 0, \quad \tilde{t} = \operatorname*{argmin}_{t\in N} r(t). \tag{2.9}$$

Moreover, for V equal to either the loss $L(\hat{\eta}(\hat{t}), \eta)$ or risk $R(\hat{\eta}(\hat{t}), \eta, \sigma^2)$ of $\hat{\eta}(\hat{t})$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E|\hat{r}(\hat{t}) - V| = 0. \tag{2.10}$$

By (2.9), the risk of the adaptive estimator $\hat{\eta}(\hat{t})$ converges to the risk of the oracle estimator $\hat{\eta}(\tilde{t})$, which achieves minimum risk over the class of symmetric linear estimators $\{\hat{\eta}(t) : t \in N\}$. By (2.10), the plug-in estimator $\hat{r}(\hat{t})$ of the risk of $\hat{\eta}(\hat{t})$ converges to the actual risk or loss of $\hat{\eta}(\hat{t})$. In this manner, we can gauge directly from the data how well we have controlled risk through adaptation.

Example 1 (*continued*). Definition (1.11) of $S(t)$ may be extended to all $t \in [0, 1]^2$. Over the larger domain, $|S(t)x|^2 \leqslant [(t_1 + t_2)^2 + t_2^2]|x|^2$, which implies that $\rho(S(t)) \leqslant 5^{1/2}$ for every $t \in [0, 1]^2$. Moreover, $\nabla_1 S(t) = I_n$ while

$$\nabla_2 S(t) = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{pmatrix} = S(1, 0, 1). \tag{2.11}$$

Consequently, $\sup_{t\in[0,1]^2} \rho[\nabla_1 S(t)] \leqslant 1$ and $\sup_{t\in[0,1]^2} \rho[\nabla_2 S(t)] \leqslant 2^{1/2}$. Theorems 1 and 2 cover this example for $t \in [0, 1]^2$. The theorem conclusions continue to hold when t is restricted to N, a closed subset of $[0, 1]^2$.

Example 2 (*continued*). Let $V(t) = I_p + c\sum_{i=1}^k t_i V_i$. By (1.14), the positive semi-definiteness of each V_i, and Proposition 1 of Section 4,

$$\rho[S(t)] = \rho[V^{-1}(t)] = \lambda_{\max}^{1/2}[V^{-2}(t)] = \lambda_{\max}[V^{-1}(t)] \leqslant 1. \tag{2.12}$$

On the other hand,

$$\nabla_i S(t) = -V^{-1}(t)[\nabla_i V(t)]V^{-1}(t) = -cV^{-1}(t)V_i V^{-1}(t). \tag{2.13}$$

Hence, because Q_i is scaled so that $\rho(V_i) = 1$,

$$\rho[\nabla_i S(t)] \leqslant c\rho[V^{-1}(t)]\rho(V_i)\rho[V^{-1}(t)] \leqslant c \quad \forall t \in C[0, 1]^k. \tag{2.14}$$

The conditions for Theorems 1 and 2 are satisfied.

Example 3 (*continued*). Here the cardinality of N is m and $US(t)U'$ is an orthogonal projection for every $t \in N$. Thus,

$$\rho[S(t)] = \rho[US(t)U'] = 1 \tag{2.15}$$

and the finite cardinality subcase of Theorems 1 and 2 applies.

3. Numerical case study

Data from the Connecticut Tumor Registry, presented on pp. 199–201 of Andrews and Herzberg (1985), reports the age-adjusted incidence of melanoma cases among males in the years 1936–1972. The 37 data-points are plotted in cell (1,1) of both Figs. 1 and 2. There is a sharp upward trend in melanoma incidence that is roughly linear but exhibits substantial ripples. The purpose of the data analysis is to discern further underlying pattern. We apply to the data linear model (1.1), with $p = n = 37$ and matrix $X = I_{37} = U$. The least squares estimate $\hat{\eta}_{\mathrm{LS}}$ of η coincides in this model with the data vector y, so sheds no further light. The first-difference variance estimate (2.6) is $\hat{\sigma}^2_{\mathrm{FD}} = .117$. This value is also the estimated risk of the least squares estimate of η. We compute the adaptive running weighted mean and the adaptive penalized least squares estimates of η.

3.1. *Running weighted mean*

The candidate running weighted mean estimators with span 3 are $\hat{\eta}_{\mathrm{RWM}}(t) = S(t)y$, where $S(t)$ is defined by (1.12) and the range of the vector t is $N = \{t \in [0, 1]^2 : t_1 + 2t_2 = 1, t_1 \geqslant .5\}$. For these t, all eigenvalues of $S(t)$ lie between 0 and 1. Cell (1,2) in Fig. 1 plots the estimated risk of $\hat{\eta}_{\mathrm{RWM}}(t)$ against t_1 when $t \in N$. The smallest estimated risk, .034, is attained at the boundary value $\hat{t} = (.5, .25)$. The adaptive RWM estimate $\hat{\eta}_{\mathrm{RWM}}(\hat{t})$ is plotted in Cell (1,1) of Fig. 1, using linear interpolation only to improve visibility. (Annual mean melanoma incidence is intrinsically a discrete variable.) The estimated risk of the adaptive RWM estimate is .034, about one-quarter of the estimated risk of the least

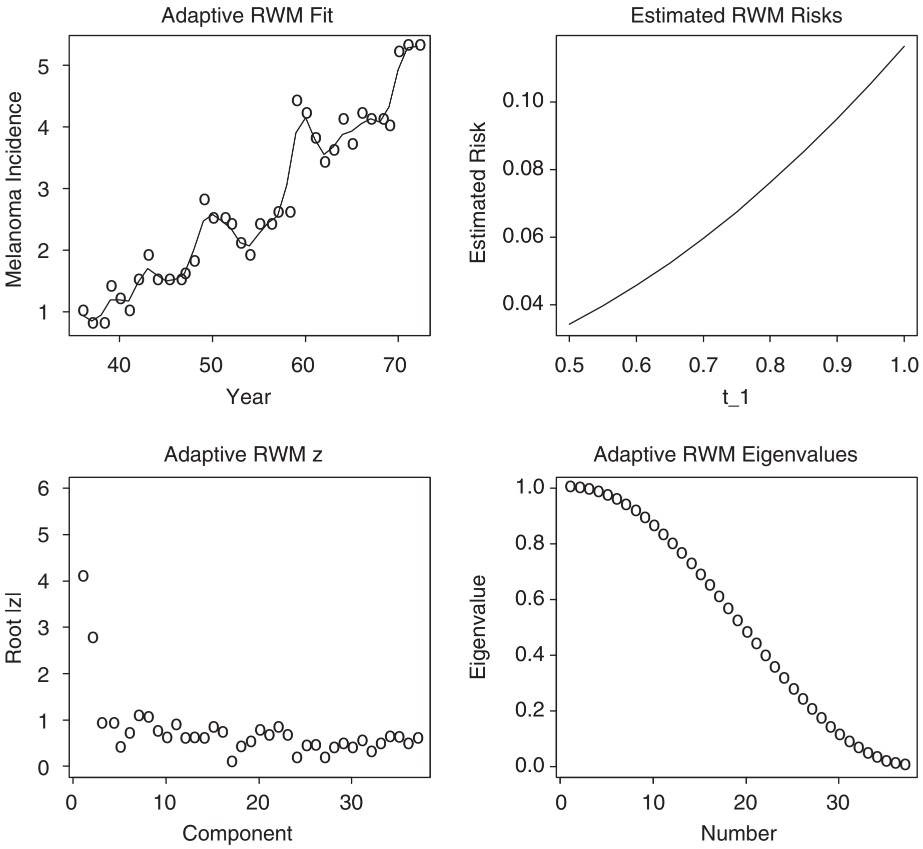

Fig. 1. Adaptive running weighted mean fit to the melanoma incidence data. The span of the running mean is 3. The other plots exhibit inner workings of the fit.

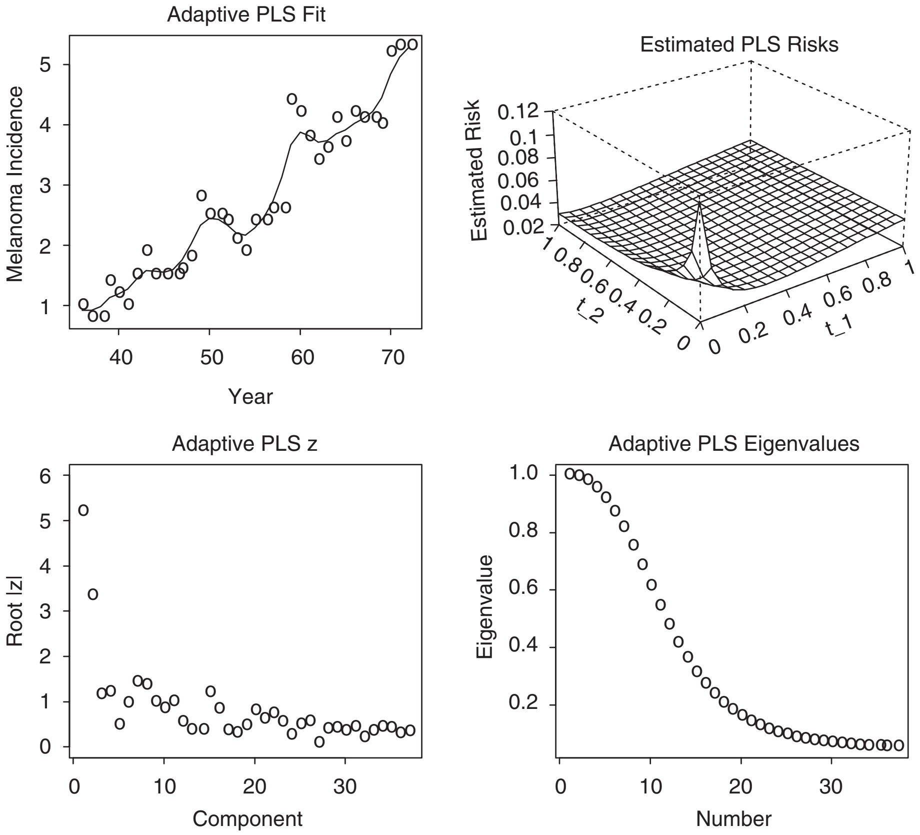

Fig. 2. Adaptive penalized least squares fit to the melanoma incidence data. The two quadratic penalty terms are constructed from the first difference and second difference operators. The other plots exhibit inner workings of the fit.

squares estimate. Theorem 2 provides theoretical support for the credibility of estimated risk as an approximation to risk.

The symmetric matrix $S(t)$ has spectral decomposition $S(t) = O(t)\Lambda(t)O'(t)$, where $\Lambda(t)$ is the diagonal matrix of eigenvalues arranged in nonincreasing order and $O(t)$ is the orthogonal matrix with eigenvectors as columns. Because U is the identity matrix, the adaptive RWM estimate may accordingly be written as

$$\hat{\eta}_{\text{RWM}}(\hat{t}) = O(\hat{t})\Lambda(\hat{t})z(\hat{t}), \quad z(\hat{t}) = O'(\hat{t})y. \tag{3.1}$$

In this expression, the data vector y is first rotated by $O'(\hat{t})$ to obtain $z(\hat{t})$, is then shrunk componentwise by the diagonal elements of $\Lambda(\hat{t})$, and is finally rotated back to the original coordinate system by $O(\hat{t})$ to obtain the adaptive RWM estimate.

Cell (2,1) of Fig. 1 plots the components $\{|z_i(\hat{t})|^{1/2}: 1 \leqslant i \leqslant 37\}$. The square root reduces the dynamic range of the components and reveals more clearly their behavior near the value 0. The first few components are significantly larger in absolute value than the remaining components. It is this feature of the spectral basis $O(\hat{t})$ generated by the adaptive RWM estimate that makes possible variance-bias trade-off through shrinkage. Cell (2,2) of Fig. 1 plots the eigenvalues in $\Lambda(\hat{t})$ that define the shrinkage performed by the adaptive RWM estimate.

3.2. *Multiply penalized least squares*

The candidate penalized least squares estimators are $\hat{\eta}_{\text{PLS}}(t) = S(t)y$, where $S(t)$ is defined by (1.15) for $t \in [0, 1]^2$, $k = 2$, and $c = 20$. The two penalty matrices Q_1 and Q_2 are taken proportional to $D_1'D_1$ and $D_2'D_2$, respectively, where D_i is the ith difference operator. All eigenvalues of $S(t)$ are between 0 and 1. Cell (1,2) in Fig. 2 plots the estimated risk of $\hat{\eta}_{\mathrm{PLS}}(t)$ against $t \in [0, 1]^2$. Outside the immediate vicinity of $t = (0, 0)$, which labels the least squares estimate $\hat{\eta}_{\mathrm{LS}}$, the estimated risk function is relatively flat. The smallest estimated risk, .028, is attained at $\hat{t} = (.138, .745)$. The adaptive PLS estimate $\hat{\eta}_{\mathrm{PLS}}(\hat{t})$ is plotted, with linear interpolation, in Cell (1,1) of Fig. 2. The estimated risk of the adaptive PLS estimate is .028, smaller than that of the adaptive RWM estimate. Again, Theorem 2 supports this risk estimate theoretically.

The adaptive PLS estimate has a spectral representation akin to (3.1). Cell (2,1) of Fig. 2 plots the components $\{|z_i(\hat{t})|^{1/2}\colon 1 \leqslant i \leqslant 37\}$. The first few components are again significantly larger in absolute value than the remaining components. Cell (2,2) of Fig. 2 plots the eigenvalues in $\Lambda(\hat{t})$ that define the shrinkage performed by the adaptive PLS estimate.

3.3. Discussion

The estimated risk of the adaptive PLS estimate is smaller than that of the adaptive RWM estimate and both are substantially smaller than the estimated risk of the least squares estimate. Comparing the plots in the lower row of Fig. 1 with their counterparts in Fig. 2 provides an intuitive explanation. In both the PLS and RWM cases, the components $\{|z_i|^{1/2}\}$ decline sharply toward zero as i increases. However, the PLS eigenvalues shrink the smaller $\{z_i\}$ toward zero more aggressively. Thereby PLS achieves, on the present data, a greater reduction in estimated risk through trade-off of estimated variance with estimated bias.

Theorem 2 provides a theoretical scenario under which the ordering of estimated risks approximates, asymptotically, the ordering of the corresponding true risks. On those grounds, the adaptive PLS fit to the melanoma data, with estimated risk .028, is preferable to the adaptive RWM fit, with estimated risk .034. The local peaks in the PLS fit are seen to follow observed peaks in the sunspot cycle. While the association with periods of increased solar emissions might also be detected from the adaptive RWM fit or even from the raw data, the adaptive PLS fit makes it easier to do so, especially in the variable first part of the data.

The statistical logic in this case study deserves comment. An adaptive procedure implicitly fits the probability model that motivates it. However, using the procedure on data is *not* the same as believing that the motivating model generated the data. Indeed, the melanoma incidence data is not certifiably random. In the absence of randomness, the first-difference variance estimate quantifies the level of detail in the data that is deemed to be noise (that is, not predictable). The adaptive PLS or RWM estimate then separates the trend in η from this possibly deterministic noise.

A probability model seeks to describe hypothetical data that is similar in selected relative frequencies to what was observed. Procedures that work well in the world of a realistic probability model may also prove satisfactory in the world of data and computational experiments. This is a matter open to empirical testing. Advances in statistical computing and in empirical process theory have reformulated statistics as an experimentally supported theory and practice of data analysis.

4. Proofs

We first record three results that will be used to prove Theorem 1. Let the $\{\lambda_i(S)\}$ denote the eigenvalues of the generic symmetric matrix S and let $\lambda_{\max}(S)$ denote the largest of these eigenvalues. Recall definition (1.13) of a function ρ that maps matrices into real numbers.

Proposition 1.

(a) *ρ is a matrix norm.*
(b) *If matrices A and B have compatible dimensions, $\rho(AB) \leqslant \rho(A)\rho(B)$.*
(c) *If a is a row or column vector, $\rho(a) = |a|$.*
(d) *If the vectors a, b and the matrix A have compatible dimensions, $|a'Ab| \leqslant |a||b|\rho(A)$.*
(e) $\rho(A) = \lambda_{\max}^{1/2}(A'A) = \lambda_{\max}^{1/2}(AA') = \rho(A')$.
(f) *If S is symmetric, then $\rho(S) = \lambda_{\max}^{1/2}(S^2) = \max_i |\lambda_i(S)|$ and $\rho(S^2) = \rho^2(S)$.*
(g) *If S is $p \times p$ symmetric, then $p^{-1}|\mathrm{tr}(S)| \leqslant \rho(S)$.*

The definitions cited readily yield the properties listed in Proposition 1.

Proposition 2. *Let $\{Y_p\colon p \geqslant 1\}$ be random elements of $C([0,1]^k)$. Let $\operatorname{plim}_{p\to\infty}$ denote the limit in probability as $p \to \infty$. Suppose that*

$$\operatorname*{plim}_{p\to\infty} Y_p(t) = 0 \quad \forall t \in [0,1]^k \tag{4.1}$$

and that

$$\lim_{\delta\to 0} \limsup_{p\to\infty} \mathrm{P}\left[\sup_{|t-s|\leqslant\delta} |Y_p(t) - Y_p(s)| \geqslant \varepsilon\right] = 0. \tag{4.2}$$

Then

$$\operatorname*{plim}_{p\to\infty} \sup_{t\in[0,1]^k} |Y_p(t)| = 0. \tag{4.3}$$

The assumptions of Proposition 2 imply the weak convergence in $C([0,1]^k)$ of $\{Y_p\colon p \geqslant 1\}$ to the zero element. From this, (4.3) follows. See Wichura (1971) for a short proof of the weak convergence or argue directly from the Arzelà–Ascoli theorem.

Proposition 3. *Suppose that $\{V_p\colon p\geqslant 1\}$, V, $\{W_p\colon p \geqslant 1\}$, and W are nonnegative random variables such that $\operatorname{plim}_{p\to\infty} V_p = V$, $V_p \leqslant W_p$ a.s. for every p, $EW < \infty$, and $\lim_{p\to\infty} E|W_p - W| = 0$. Then $\lim_{p\to\infty} E|V_p - V| = 0$.*

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

Proof of Theorem 1. The strategy is to show that $W(t) - r(t)$ converges in probability to zero for every $t \in [0,1]$, then use Proposition 2 to show that $\sup_{t\in[0,1]} |W(t) - r(t)|$ converges in probability to zero, and finally invoke Proposition 3 to establish (2.8). Repeatedly used are the constraint $p^{-1}|\eta|^2 \leqslant a$ and the following properties of $T(t) = S^2(t)$ and $\bar{T}(t) = [I - S(t)]^2$, which follow from the theorem assumptions and Proposition 1:

$$\begin{gathered}\sup_p \sup_{t\in N} \rho[T(t)] < \infty, \quad \sup_p \sup_{t\in N} \rho[\bar{T}(t)] < \infty, \\ \sup_{i,p} \sup_{t\in N} \rho[\nabla_i T(t)] < \infty, \quad \sup_{i,p} \sup_{t\in N} \rho[\nabla_i \bar{T}(t)] < \infty.\end{gathered} \tag{4.4}$$

The bounds on derivatives pertain only to the case $N = [0,1]^k$ and use the following identity: $\nabla_i T(t) = \nabla_i S(t) \cdot S(t) + S(t) \cdot \nabla_i S(t)$.

We first prove the case $W(t) = \hat{r}(t)$ of (2.8). Define $\check{r}(t)$ by replacing $\hat{\sigma}^2$ with σ^2 in Definition (2.3) of $\hat{r}(t)$. Hereafter, we generally omit the argument t, writing $\bar{T}$ in place of $\bar{T}(t)$, for instance. Because of the inequality

$$|\hat{r}(t) - \check{r}(t)| \leqslant |\hat{\sigma}^2 - \sigma^2| p^{-1}[|\mathrm{tr}(T)| + |\mathrm{tr}(\bar{T})|] \leqslant |\hat{\sigma}^2 - \sigma^2|[\rho(T) + \rho(\bar{T})] \tag{4.5}$$

and the L_1 consistency (2.7) of $\hat{\sigma}^2$, we may replace $\hat{r}(t)$ with $\check{r}(t)$ in the subsequent argument.

4.1. Pointwise consistency

Let

$$Y_p(t) = \check{r}(t) - r(t), \quad B(t) = U\bar{T}(t)U' \tag{4.6}$$

and note that $\mathrm{tr}(B) = \mathrm{tr}(\bar{T})$ and $\rho(B) = \rho(\bar{T})$. Recall the definitions at the start of Section 2 and the foregoing definition of $\check{r}(t)$. Then $z = \xi + w$, where $w = U'e$, and

$$\begin{aligned} Y_p(t) &= p^{-1}\mathrm{tr}[\bar{T}(zz' - \sigma^2 I_p - \xi\xi')] \\ &= p^{-1}[2\xi'\bar{T}w + \{w'\bar{T}w - \sigma^2\,\mathrm{tr}(\bar{T})\}] = p^{-1}[2\eta' Be + \{e'Be - \sigma^2\,\mathrm{tr}(B)\}] \end{aligned} \tag{4.7}$$

Evidently, $E(\eta' Be) = 0 = E[e'Be - \sigma^2 \operatorname{tr}(B)]$ and

$$\operatorname{Var}(p^{-1}\eta' Be) = p^{-2}\sigma^2\eta' B^2\eta \leqslant p^{-2}\sigma^2|\eta|^2\rho(B^2) \leqslant p^{-1}\sigma^2 a\rho^2(\bar{T}). \tag{4.8}$$

Moreover, if $B = \{b_{ij}\}$ and $e = \{e_i\}$, then $e'Be = \sum_i b_{ii}e_i^2 + 2\sum_{i<j} b_{ij}e_ie_j$. Let γ denote the kurtosis of e_i, so that $E(e_i^4) = (3+\gamma)\sigma^4$ and $\operatorname{Var}(e_i^2) = (2+\gamma)\sigma^4$. Then, using $|B| = |\bar{T}|$,

$$\begin{aligned}\operatorname{Var}(p^{-1}e'Be) &= p^{-2}\sigma^4\left(2|B|^2 + \gamma\sum_i b_{ii}^2\right) \leqslant p^{-2}\sigma^4(2+\gamma)|B|^2 \\ &= p^{-2}\sigma^4(2+\gamma)|\bar{T}|^2 \leqslant p^{-1}\sigma^4(2+\gamma)\rho^2(\bar{T}).\end{aligned} \tag{4.9}$$

Thus, for every $t \in N$,

$$\operatorname*{plim}_{p\to\infty} Y_p(t) = 0. \tag{4.10}$$

4.2. Uniform consistency

Consider the case when $N = [0, 1]^k$. For any $s, t \in [0, 1]^k$,

$$Y_p(s) - Y_p(t) = p^{-1}\sum_{i=1}^{k}(s_i - t_i)[2\eta'\nabla_i Be + \{e'\nabla_i Be - \sigma^2 \operatorname{tr}(\nabla_i B)\}], \tag{4.11}$$

where $\nabla_i B = \nabla_i B(\bar{s})$ for some $\bar{s}$ on the line segment that joins s and t. Thus,

$$\sup_{|s-t|\leqslant\delta}|Y_p(s) - Y_p(t)| \leqslant \delta p^{-1}\sum_{i=1}^{k}[2|\eta'\nabla_i Be| + |e'\nabla_i Be| + \sigma^2|\operatorname{tr}(\nabla_i B)|]. \tag{4.12}$$

Moreover, using Proposition 1,

$$\begin{aligned}&p^{-1}|\operatorname{tr}(\nabla_i B)| = p^{-1}|\operatorname{tr}(\nabla_i\bar{T})| \leqslant \rho(\nabla_i\bar{T}),\\ &p^{-1}E|\eta'\nabla_i Be| \leqslant p^{-1}|\eta|E|e|\rho(\nabla_i B) \leqslant a^{1/2}\sigma\rho(\nabla_i\bar{T}),\\ &p^{-1}E|e'\nabla_i Be| \leqslant p^{-1}E|e|^2\,\rho(\nabla_i B) = \sigma^2\rho(\nabla_i\bar{T}).\end{aligned} \tag{4.13}$$

Applying Markov's inequality to the right-hand side of (4.13) establishes existence of a finite constant C, not depending on p, such that

$$\mathrm{P}\left[\sup_{|s-t|\leqslant\delta}|Y_p(s) - Y_p(t)| \geqslant \varepsilon\right] \leqslant C\delta. \tag{4.14}$$

Hence,

$$\lim_{\delta\to 0}\limsup_{p\to\infty}\mathrm{P}\left[\sup_{|s-t|\leqslant\delta}|Y_p(s) - Y_p(t)| \geqslant \varepsilon\right] = 0. \tag{4.15}$$

Limits (4.10) and (4.15) plus Proposition 2 yield

$$\operatorname*{plim}_{p\to\infty}\sup_{t\in[0,1]^k}|Y_p(t)| = 0. \tag{4.16}$$

In the situation when the cardinality of N is finite and does not depend on p, this result is immediate from the pointwise convergence (4.10).

4.3. L_1 uniform consistency

Let $V_p = \sup_{t\in[0,1]^k} |Y_p(t)|$ and let $\rho_{\max} = \sup_p \sup_{t\in[0,1]^k} \rho(\bar{T}(t))$.
Using (4.7) and $|\eta| = |\xi|$,

$$\begin{aligned} V_p &\leqslant p^{-1} \sup_{t\in[0,1]^k} [2|\xi'\bar{T}(t)w| + |w'\bar{T}(t)w| + \sigma^2|\mathrm{tr}(\bar{T}(t))|] \\ &\leqslant [2a^{1/2}\{p^{-1/2}|w|\} + p^{-1}|w|^2 + \sigma^2]\rho_{\max}. \end{aligned} \tag{4.17}$$

Let W_p denote the right-hand side of (4.17) and let $W = [2a^{1/2}\sigma + 2\sigma^2]\rho_{\max}$. Because $|w|^2 = e'UU'e$, a calculation akin to (4.9) shows that

$$\mathrm{Var}(p^{-1}|w|^2) \leqslant p^{-1}\sigma^4(2+\gamma). \tag{4.18}$$

Hence, $p^{-1}|w|^2 = \{p^{-1/2}|w|\}^2$ converges in probability to its expectation σ^2. By Vitali's theorem,

$$\begin{aligned} &\lim_{p\to\infty} E|p^{-1}|w|^2 - \sigma^2| = 0, \\ &\lim_{p\to\infty} E|p^{-1/2}|w| - \sigma| \leqslant \lim_{p\to\infty} E^{1/2}[p^{-1/2}|w| - \sigma]^2 = 0. \end{aligned} \tag{4.19}$$

Consequently, $\lim_{p\to\infty} E|W_p - W| = 0$. This convergence, inequality (4.17), and Proposition 3 imply that (4.16) can be strengthened to

$$\lim_{p\to\infty} E\left[\sup_{t\in[0,1]^k} |Y_p(t)|\right] = 0. \tag{4.20}$$

This completes the proof of (2.8) when $W(t) = \hat{r}(t)$. □

The argument for the case $W(t) = L(\hat{\eta}(t), \eta)$ of (2.8) is similar. The loss and risk of $\hat{\eta}(t)$ are given in (2.1) and (2.2). Note that

$$L(\hat{\eta}(t), \eta) = p^{-1}[2\xi'(S - I_p)Sw + w'Tw + \xi'\bar{T}\xi]. \tag{4.21}$$

Let

$$Y_p(t) = L(\hat{\eta}(t), \eta) - r(t) = p^{-1}[2\xi'Vw + \{w'Tw - \sigma^2\,\mathrm{tr}(T)\}], \tag{4.22}$$

where $V(t) = T(t) - S(t)$. Because the right-hand side of (4.22) has the same structure as the middle expression in (4.7), an argument parallel to the one that follows (4.7) completes the proof. Indeed, (4.4) and the theorem assumptions on $S(t)$ imply that

$$\sup_p \sup_{t\in N} \rho[V(t)] < \infty, \quad \sup_{i,p} \sup_{t\in N} \rho[\nabla_i V(t)] < \infty. \tag{4.23}$$

Proof of Theorem 2. We show that (2.8) implies

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2\leqslant a} E|Z - r(\tilde{t})| = 0, \tag{4.24}$$

where Z can be $L(\hat{\eta}(\hat{t}), \eta)$ or $L(\hat{\eta}(\tilde{t}), \eta)$ or $\hat{r}(\hat{t})$. The three limits to be proved in (2.9) and (2.10) are immediate consequences of (4.24).

First, (2.8) with $W(t) = \hat{r}(t)$ entails

$$\begin{aligned} &\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2\leqslant a} E|\hat{r}(\hat{t}) - r(\tilde{t})| = 0, \\ &\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2\leqslant a} E|\hat{r}(\hat{t}) - r(\hat{t})| = 0. \end{aligned} \tag{4.25}$$

Hence, (4.24) holds for $Z = \hat{r}(\hat{t})$ and

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E|r(\hat{t}) - r(\tilde{t})| = 0. \tag{4.26}$$

Second, (2.8) with $W(t) = L(\hat{\eta}(t), \eta)$ gives

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E|L(\hat{\eta}(\hat{t}), \eta) - r(\hat{t})| = 0,$$

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leqslant a} E|L(\hat{\eta}(\tilde{t}), \eta) - r(\tilde{t})| = 0. \tag{4.27}$$

These limits together with (4.26) establish the remaining two cases of (4.24).

References

Andrews, D.F., Herzberg, A.M., 1985. Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer, New York.

Beran, R., 2000. REACT scatterplot smoothers: superefficiency through basis economy. J. Amer. Statist. Assoc. 63, 155–171.

Beran, R., 2002. Improving penalized least squares through adaptive selection of penalty and shrinkage. Ann. Inst. Statist. Math. 54, 900–917.

Beran, R., Dümbgen, L., 1998. Modulation of estimators and confidence sets. Ann. Statist. 26, 1826–1856.

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models (with discussion). Ann. Statist. 17, 453–555.

Heckman, N.E., Ramsay, J.O., 2000. Penalized regression with model-based penalties. Canad. J. Statist. 28, 241–258.

Kneip, A., 1994. Ordered linear smoothers. Ann. Statist. 22, 835–866.

Li, K.-C., 1985. From Stein's unbiased risk estimates to the method of generalized cross-validation. Ann. Statist. 13, 1352–1377.

Li, K.-C., 1987. Asymptotic optimality for C_{p}, C_{L}, and generalized cross-validations. Ann. Statist. 15, 958–976.

Mallows, C.L., 1973. Some comments on C_{p}. Technometrics 15, 661–676.

Neveu, J., 1965. Mathematical Foundations of the Calculus of Probability. Holden-Day, San Francisco.

Rao, C.R., Toutenberg, H., 1995. Linear Models. Least Squares and Alternatives. Springer, New York.

Rice, J., 1984. Bandwidth choice for nonparametric regression. Ann. Statist. 12, 1215–1230.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Neyman, J. (Ed.), Proceedings of Third Berkeley Symposium on Mathematical Statistics Probability. University of California Press, Berkeley, pp. 197–206.

Wichura, M.J., 1971. A note on the weak convergence of stochastic processes. Ann. Math. Statist. 42, 1769–1772.