Ann. Inst. Statist. Math.
Vol. 54, No. 4, 900–917 (2002)

IMPROVING PENALIZED LEAST SQUARES THROUGH ADAPTIVE SELECTION OF PENALTY AND SHRINKAGE

RUDOLF BERAN

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

(Received February 7, 2001; revised July 2, 2001)

Abstract. Estimation of the mean function in nonparametric regression is usefully separated into estimating the means at the observed factor levels—a one-way layout problem—and interpolation between the estimated means at adjacent factor levels. Candidate penalized least squares (PLS) estimators for the mean vector of a one-way layout are expressed as shrinkage estimators relative to an orthogonal regression basis determined by the penalty matrix. The shrinkage representation of PLS suggests a larger class of candidate monotone shrinkage (MS) estimators. Adaptive PLS and MS estimators choose the shrinkage vector and penalty matrix to minimize estimated risk. The actual risks of shrinkage-adaptive estimators depend strongly upon the economy of the penalty basis in representing the unknown mean vector. Local annihilators of polynomials, among them difference operators, generate penalty bases that are economical in a range of examples. Diagnostic techniques for adaptive PLS or MS estimators include basis-economy plots and estimates of loss or risk.

Key words and phrases: Nonparametric regression, one-way layout, adaptation, loss estimator, risk estimator, economical basis, orthogonal polynomial, local annihilator.

1. Introduction

The regression model that motivates statistical procedures studied in this paper is

$$y_i = m(t_i) + e_i, \qquad 1 \le i \le n. \tag{1.1}$$

The nonrandom design points are ordered so that $t_1 \le t_2 \le \cdots \le t_n$. The errors $\{e_i\}$ are independent, identically distributed, each having a $N(0, \sigma^2)$ distribution. Both the function m and the variance σ^2 are unknown. Estimation of m from the observed $\{y_i, t_i\}$ is the task undertaken. This probabilistic formulation serves for the derivation and initial study of estimators for m. Asymptotic theory developed under the model is supplemented with computational experiments on real and artificial data that respect the fundamental distinction between data and probability model and bring out additional aspects of estimator performance. These experiments also explore the use of estimated losses and certain diagnostic plots to assess the performance of competing estimators on particular data.

Let $y = \{y_i\}$, $\mu = \{m(t_i)\}$, and $e = \{e_i\}$ be $n \times 1$ vectors with the stated components. Nonparametric regression as just described can be separated logically into two problems. The first is to estimate the values $\{m(t_i) : 1 \le i \le n\}$. This amounts to estimation of the vector μ in the possibly unbalanced one-way layout

$$y = \mu + e, \tag{1.2}$$

where e has a multivariate $N(0, \sigma^2 I_n)$ distribution. It follows from Stein (1956) that the least squares estimator of μ is inadmissible under quadratic loss whenever the number of factor levels exceeds 2. As will be seen, the least squares estimator can have high quadratic risk when compared with alternative estimators less prone to overfitting the data.

Given an efficient estimator of μ, the second problem is interpolation among its components so as to estimate the function m. This is a problem in approximation theory that is highly sensitive to assumptions on the nature of m. The observed $\{y_i, t_i\}$ will not tell us how many derivatives m has. In the absence of strong prior information about the smoothness of m, we may settle for straightforward linear interpolation or spline interpolation between the estimated components of μ. At a minimum, such interpolation is a convenient visual device for displaying estimators of m at the design points. To consider separately the estimation at design points and the interpolation between design points clarifies what can be done in nonparametric regression. Examples presented in this paper support the claim that efficient estimation of the mean function at the design points is often more important for data analysis than sophisticated interpolation between adjacent estimates.

Suppose that the design points $\{t_i\}$ contain $p \leq n$ distinct values $s_1 < s_2 < \cdots < s_p$, which are the factor levels. Let X denote the $n \times p$ incidence matrix defined as follows: row i contains a 1 in the column j such that $s_j = t_i$ and has zeroes in the other $p - 1$ positions. Let $\beta = (m(s_1), m(s_2), \ldots, m(s_p))'$ denote the mean responses at the factor levels. The mean vector of the one-way layout (1.2) is then

$$\mu = X\beta \tag{1.3}$$

and the least squares estimator of μ is $\hat{\mu}_{LS} = X(X'X)^{-1}X'y$.

Let D be any matrix with p columns, let ν be an element of the extended non-negative reals $[0, \infty]$, and let $|\cdot|$ denote quadratic norm. The candidate *penalized least squares* (PLS) estimator of μ is

$$\hat{\mu}_{PLS}(D, \nu) = X\hat{\beta}_{PLS}(D, \nu) \tag{1.4}$$

where

$$\hat{\beta}_{PLS}(D, \nu) = \operatorname*{argmin}_{\beta \in R^p} [|y - X\beta|^2 + \nu |D\beta|^2]. \tag{1.5}$$

It is understood that $\hat{\beta}_{PLS}(D, \infty) = \lim_{\nu \to \infty} \hat{\beta}_{PLS}(D, \nu)$. Explicitly,

$$\hat{\mu}_{PLS}(D, \nu) = X(X'X + \nu D'D)^{-1}X'y. \tag{1.6}$$

In this form, $\hat{\mu}_{PLS}(D, \nu)$ may be viewed as a generalized ridge estimator.

Effective choice of penalty matrix D and of the non-negative penalty weight ν are central issues. When ν is zero, the candidate PLS estimator reduces to the least squares estimator $\hat{\mu}_{LS}$. For very large ν, the PLS estimator effectively minimizes the residual sum of squares subject to the constraint that $|D\beta|^2$ is approximately zero. To guide the choice of D and ν, we will assess the quality of any estimator $\hat{\mu}$ through normalized quadratic loss and corresponding risk

$$L(\hat{\mu}, \mu) = p^{-1}|\hat{\mu} - \mu|^2, \qquad R(\hat{\mu}, \mu, \sigma^2) = \mathrm{E}L(\hat{\mu}, \mu). \tag{1.7}$$

Let

$$S(D,\nu) = X(X'X + \nu D'D)^{-1}X' \tag{1.8}$$

and let $|\cdot|$ denote Euclidean matrix norm. That is, $|C|^2 = \mathrm{tr}(CC') = \mathrm{tr}(C'C)$ for any matrix C. The risk of the candidate estimator $\hat{\mu}_{PLS}(D,\nu)$ is then

$$R(\hat{\mu}_{PLS}(D,\nu),\mu,\sigma^2) = p^{-1}[\sigma^2|S(D,\nu)|^2 + |\mu - S(D,\nu)\mu|^2]. \tag{1.9}$$

For the least squares estimator $\hat{\mu}_{LS} = \hat{\mu}_{PLS}(D,0)$, this risk reduces to σ^2.

Let $\hat{\sigma}^2$ be a trustworthy estimator of σ^2. Customary when n substantially exceeds p is the variance estimator $\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\mu}_{LS}|^2$. The derivation of the Mallows (1973) C_L criterion yields the risk estimator

$$\hat{R}(D,\nu) = p^{-1}[|y - S(D,\nu)y|^2 + \{2\,\mathrm{tr}[S(D,\nu)] - n\}\hat{\sigma}^2]. \tag{1.10}$$

In particular, when $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$, the estimated risk for the least squares estimator of μ is $\hat{R}(D,0) = \hat{\sigma}^2_{LS}$. We propose to choose both the penalty weight ν and the penalty matrix D so as to minimize the estimated risk $\hat{R}(D,\nu)$.

When represented with respect to the orthogonal penalty basis for the regression space that is defined in the next section, PLS estimators suggest a larger class of candidate monotone shrinkage (MS) estimators for μ. The themes of this paper are: asymptotic theory to support the strategy of choosing the candidate estimator that minimizes estimated risk; the advantages of adaptive MS over adaptive PLS; methods for designing effective penalty matrices; and the use of estimated loss/risk and of diagnostic plots to assess the performance of adaptive PLS or MS estimators on given data.

The need for asymptotic analysis and for restrictions on the extent of adaptation is indicated by an example. Suppose that S is permitted to vary over all $n \times n$ symmetric matrices that have a specified set of eigenvectors and that σ^2 is known. The symmetric matrix S that minimizes the right side of (1.10) over the class just described then generates an estimator of μ whose risk is dominated by that of the least squares estimator $\hat{\mu}_{LS}$. This may be seen from Remark A on p. 1829 of Beran and Dümbgen (1998).

For fixed penalty matrix D, the shrinkage-adaptive PLS estimator is defined to be $\hat{\mu}_{PLS}(D,\hat{\nu})$, where $\hat{\nu}$ minimizes the estimated risk $\hat{R}(D,\nu)$ over all ν in $[0,\infty]$. We will call this the PLS(D) estimator. Section 2.3 describes how to compute it effectively. Under the probability model described there, the risk of the adaptive estimator PLS(D) converges to the risk of the unrealizable candidate PLS estimator with smallest risk. Thus, the asymptotic risk of the PLS(D) estimator cannot exceed that of the least squares estimator. In practice, it is often far smaller and the shrinkage-adaptive MS(D) estimator to be defined in Subsection 2.2 typically reduces risk further. Subsection 3.2 develops possibilities for adaptation through choice of the penalty matrix D in addition to ν.

Though valuable in exploring the scope of adaptation and the overall behavior of an estimator, ensemble results such as asymptotic minimaxity or rates of convergence do not indicate the adequacy of a particular estimator on particular data. Section 3 addresses the use of estimated loss and of diagnostic basis-economy and shrinkage-vector plots to assess adaptive PLS and MS estimators on given data.

Figure 1 exhibits penalized least squares estimates on three sets of artificial data. The smooth case was suggested by the Canadian earnings data that was analyzed, with further background, in Chu and Marron (1991). The respective mean functions are:

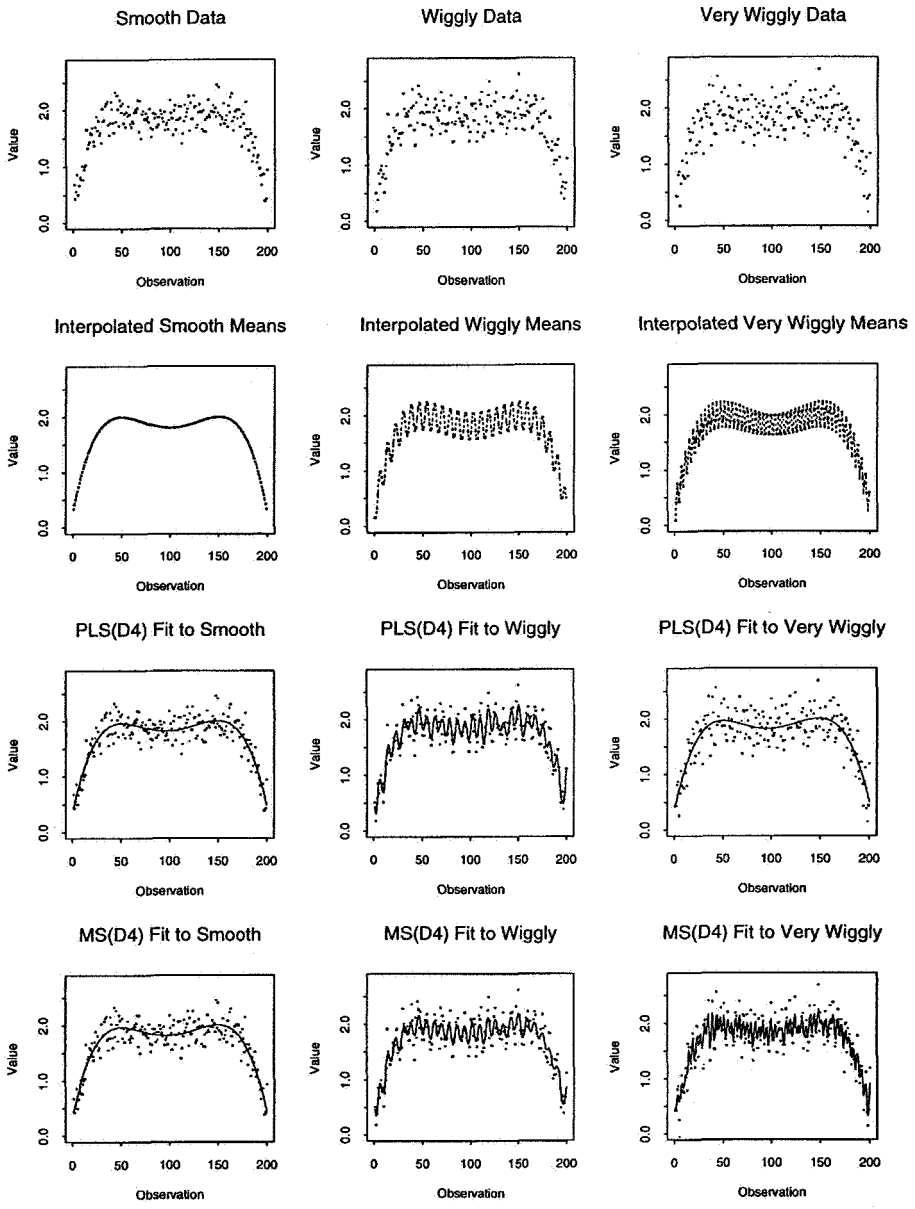

Fig. 1. Each column displays the artificial data, the true mean vector, the PLS(D_4) estimate, and the MS(D_4) estimate.

Smooth: $m_1(t) = 2 - 50((t-25)(t-75))^2$.
Wiggly: $m_2(t) = m_1(t) - .25\sin(50\pi t)$.
Very Wiggly: $m_3(t) = m_1(t) - .25\sin(100\pi t)$.

The design points are $\{t_i = i/(n+1): 1 \le i \le n\}$ with $n = 200$. The j-th artificial data set is $\{m_j(i/201): 1 \le i \le 200\} + e$, where e is a single pseudo-random sample drawn from the $N(0, \sigma^2 I_{200})$ distribution and $\sigma = .2$. In this design, $p = n$. The variance σ^2 is estimated by the high component estimator defined in (2.13), with $q = .75p$.

As penalty matrix we use the $(p-4) \times p$ fourth difference matrix D_4. The first row of D_4 consists of entries 1, −4, 6, −4, 1 followed by zeros. The second row shifts the non-zero entries one place to the right and puts a zero in the first column. Construction of subsequent rows continues the shift of nonzero entries to the right. The d-th difference penalty matrix D_d, defined formally in Section 3.2, is particularly appropriate when the components of β are equally spaced values on a curve whose local behavior mimics a polynomial of degree $d-1$. In the present example, either $d=4$ or $d=5$ works well. Computations for this and other examples in the paper were done with S-Plus 2000 for Windows.

Column j in Fig. 1 plots the j-th artificial sample in the first row and the linearly interpolated (dashed line) components of the mean vector $\mu = \{m_j(i/201)\}$ in the second row. The function `rnorm`, initialized with `set.seed(2)`, produced the pseudo-Gaussian errors that are added to the means in the second row to obtain the artificial samples. Any sinusoidal wiggles present in μ are not apparent to the eye in the scatterplots of this data.

The third row in the figure superposes on the data the linearly interpolated (solid line) components of estimator PLS(D_4). This shrinkage-adaptive PLS estimator recovers the means well from the Smooth sample and detects the sinusoid underlying the Wiggly sample, even though it distorts that sinusoid's amplitude and regularity. However, on the Very Wiggly sample, PLS(D_4) fails utterly, like the eye, to detect the sinusoid and settles for estimating the smooth component of the trend. The fourth row of Fig. 1 plots the adaptive MS(D_4) generalization of PLS(D_4) that is defined in Section 2.2. This estimator succeeds in handling the Very Wiggly sample as well as the other two.

2. Estimated risk and shrinkage adaptation

A canonical representation assists both theoretical study and numerical computation of the candidate PLS estimators $\hat{\mu}_{PLS}(D,\nu)$. These and the candidate MS estimators defined in Section 2.2 are particular symmetric linear smoothers in the sense of Buja *et al.* (1989) and are candidate REACT estimators in the sense of Beran (2000).

2.1 *The penalty basis*

The replication matrix $R = X'X$ is a $p \times p$ diagonal matrix whose k-th diagonal element indicates the number of $\{t_i\}$ that equal s_k. For any matrix C, let $\mathcal{M}(C)$ denote the subspace spanned by its columns. The columns of the matrix $U_0 = XR^{-1/2}$ provide an orthonormal basis for the regression problem: $U_0'U_0 = I_p$ and $\mathcal{M}(U_0) = \mathcal{M}(X)$. Let $B = R^{-1/2}D'DR^{-1/2}$. Because $X = U_0R^{1/2}$, equation (1.6) is equivalent to

$$\hat{\mu}_{PLS}(D,\nu) = U_0(I_p + \nu B)^{-1}U_0'y. \tag{2.1}$$

The symmetric matrix B has spectral representation $B = \Gamma\Lambda\Gamma'$ where the eigenvector matrix satisfies $\Gamma'\Gamma = \Gamma\Gamma' = I_p$ and the diagonal matrix $\Lambda = \mathrm{diag}\{\lambda_i\}$ gives the ordered eigenvalues with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$. This eigenvalue ordering, the reverse of the customary, is used here because the eigenvectors associated with the smallest eigenvalues largely determine the value and performance of candidate estimator $\hat{\mu}_{PLS}(D,\nu)$. Let $U = U_0\Gamma$. It follows from (2.1) that

$$\hat{\mu}_{PLS}(D,\nu) = U(I_p + \nu\Lambda)^{-1}U'y. \tag{2.2}$$

The columns of the matrix U constitute the orthonormal *penalty basis* for the regression space determined by the penalty matrix D : $U'U = I_p$ and $\mathcal{M}(U) = \mathcal{M}(X)$.

Variational characterization of U. Alternatively, the successive columns $u_1, u_2, \ldots, u_p$ of the penalty basis matrix U may be defined through their variational properties:

- As above, let $U_0 = XR^{-1/2}$ provide an initial orthonormal basis matrix for the regression space $\mathcal{M}(X)$.
- Find a unit vector $u_1 = U_0\gamma$ in $\mathcal{M}(X)$ that minimizes the penalty $|D(X'X)^{-1}X'u_1|^2$. This reduces to finding the $p \times 1$ unit vector γ that minimizes $|DR^{-1/2}\gamma|^2 = \gamma'B\gamma$. The desired minimum penalty vector is thus $u_1 = U_0\gamma_1$, where γ_j is the j-th column of the eigenvector matrix Γ.
- Find a unit vector $u_2 = U_0\gamma$ in $\mathcal{M}(X)$ that minimizes the penalty $|D(X'X)^{-1}X'u_2|^2$ subject to the constraint that u_2 is orthogonal to u_1. This reduces to finding the $p \times 1$ unit vector γ orthogonal to γ_1 that minimizes $|DR^{-1/2}\gamma|^2 = \gamma'B\gamma$. The desired minimum penalty vector is thus $u_2 = U_0\gamma_2$.
- Continue sequential constrained minimization to obtain the penalty basis matrix

$$U = (U_0\gamma_1, U_0\gamma_2, \ldots, U_0\gamma_p) = U_0\Gamma. \tag{2.3}$$

In the one-way layout under consideration, $(X'X)^{-1}X'u_k$ extracts the components of basis vector u_k that are associated with the p factor levels. The penalty for this extracted vector is

$$|D(X'X)^{-1}X'u_k|^2 = |DR^{-1/2}\gamma_k|^2 = \gamma_k'B\gamma_k = \lambda_k. \tag{2.4}$$

When the penalty matrix is a d-th difference operator, the preceding variational characterization of U explains intuitively why its successive column vectors are increasingly wiggly.

2.2 *From PLS to monotone shrinkage estimators*

Fix D so that the penalty basis U is determined. Let $z = U'y$ and let $f(\nu)$ denote the column vector $(1/(1+\nu\lambda_1), 1/(1+\nu\lambda_2), \ldots, 1/(1+\nu\lambda_p))'$, with the understanding that $f(\infty) = \lim_{\nu\to\infty} f(\nu)$. The distribution of z is $N_p(\xi, \sigma^2 I_p)$, where $\xi = U'\mu$. The PLS estimator of ξ implied by expression (2.2) is

$$\hat{\xi}_{PLS}(D,\nu) = U'\hat{\mu}_{PLS}(D,\nu) = f(\nu)z, \tag{2.5}$$

where the multiplication of vectors in the expression to the right is performed componentwise as in the S language. Equivalently,

$$\hat{\mu}_{PLS}(D,\nu) = U\hat{\xi}_{PLS}(D,\nu) = U\,\mathrm{diag}\{f(\nu)\}U'y. \tag{2.6}$$

The structure of representation (2.6) suggests a larger family of candidate estimators for μ. Let

$$\mathcal{F}_{MS} = \{f \in [0,1]^p : f_1 \geq f_2 \geq \cdots f_p\} \tag{2.7}$$

and let

$$\hat{\xi}_{MS}(D,f) = fz \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.8}$$

The candidate *monotone shrinkage* (MS) estimators for μ associated with penalty matrix D are defined by

$$\hat{\mu}_{MS}(D,f) = U\hat{\xi}_{MS}(D,f) = U\,\mathrm{diag}\{f\}U'y \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.9}$$

It follows from (2.6) that the candidate PLS estimators are a proper subset of the MS family in which the shrinkage vector f is restricted to the form $\{f(\nu): \nu \in [0,\infty]\}$.

The next section develops three good reasons for considering monotone shrinkage estimators. First, for every candidate PLS estimator there is an MS estimator whose risk is at least as small. Second, minimizing the *estimated* risk of candidate MS or PLS estimators over all shrinkage vectors permitted by their definitions turns out to minimize asymptotic risk over the respective classes of candidate estimators. Third, computation of adaptive MS estimators is faster than computation of their adaptive PLS counterparts.

2.3 *Estimated risks and shrinkage adaptation*

For any vector h, let $\mathrm{ave}(h)$ denote the average of its components. Define the function

$$\rho(f,\xi^2,\sigma^2) = \mathrm{ave}[f^2\sigma^2 + (1-f)^2\xi^2] \quad \text{for} \quad f \in [0,1]^p. \tag{2.10}$$

Because $|\hat{\mu}_{MS}(D,f) - \mu|^2 = |fz - \xi|^2$, it follows that the normalized quadratic risk of the candidate MS estimator is

$$R(\hat{\mu}_{MS}(D,f),\mu,\sigma^2) = \rho(f,\xi^2,\sigma^2) \quad \text{for} \quad f \in \mathcal{F}_{MS}. \tag{2.11}$$

In particular, the risk $R(\hat{\mu}_{PLS}(D,\nu),\mu,\sigma^2)$ of the candidate PLS estimator, expressed in the original coordinate system by equation (1.9), is simply $\rho(f(\nu),\xi^2,\sigma^2)$.

The risk function $\rho(f,\xi^2,\sigma^2)$ depends on the unknown parameters ξ^2 and σ^2. Having obtained a variance estimator $\hat{\sigma}^2$, we may estimate ξ^2 by $z^2 - \hat{\sigma}^2$ and hence $\rho(f,\xi^2,\sigma^2)$ by

$$\hat{\rho}(D,f) = \mathrm{ave}[f^2\hat{\sigma}^2 + (1-f)^2(z^2 - \hat{\sigma}^2)]. \tag{2.12}$$

Expression (2.12) expresses in canonical form the Mallows risk estimator (1.10).

The following definitions carry out several strategies for estimating the variance σ^2:

• *The least squares variance estimator.* The least squares variance estimator $\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\mu}_{LS}|^2$ is unbiased and is consistent for σ^2 when $n-p$ tends to infinity.

• *The first-difference estimator.* This estimator, $\hat{\sigma}^2_{D1} = [2(n-1)]^{-1}\sum_{i=2}^{n}(y_i - y_{i-1})^2$, was treated by Rice (1984). It is consistent for σ^2 when n tends to infinity and the bias $\lim_{n\to\infty}[2(n-1)]^{-1}\sum_{i=2}^{n}(\mu_i - \mu_{i-1})^2 = 0$. Similar estimators may be constructed from higher-order differences of y.

The next two variance estimators make use of the penalty basis U. Choose $\bar{U}$ so that the concatenated matrix $(U \mid \bar{U})$ is orthogonal. Set $\bar{z} = \bar{U}'y$ in analogy to the earlier $z = U'y$.

• *The high-component variance estimator.* The strategy of pooling sums of squares in analysis of variance suggests

$$\hat{\sigma}^2_H = (n-q)^{-1}\left[\sum_{i=q+1}^{p} z_i^2 + |\bar{z}|^2\right] = (n-q)^{-1}\left[\sum_{i=q+1}^{p} z_i^2 + |y - \hat{\mu}_{LS}|^2\right], \tag{2.13}$$

where $q \leq \min\{p, n-1\}$. The bias of $\hat{\sigma}_H^2$ is $(n-q)^{-1}\sum_{i=q+1}^{p}\xi_i^2$. Consistency of $\hat{\sigma}_H^2$ is assured when this bias tends to zero as $n-q$ tends to infinity. When $q = p < n$, the estimator $\hat{\sigma}_H^2$ reduces to $\hat{\sigma}_{LS}^2$.

• *The robust high-component variance estimator.* Let w denote the vector obtained by concatenating $\{z_i: q+1 \leq i \leq p\}$ with $\bar{z}$. Robustness theory suggests the estimator

$$\hat{\sigma}_{RH} = \text{median}\{|w_j|: 1 \leq j \leq n-q\}/\Phi^{-1}(.75) \tag{2.14}$$

for σ, where Φ^{-1} is the standard normal quantile function. Under model (1.2), $\hat{\sigma}_{RH}^2$ approaches σ^2 in probability when $n-q$ is large and the high order components of ξ are small.

Let $\hat{g} = (z^2 - \hat{\sigma}^2)/z^2$. The risk estimator $\hat{\rho}(D, f)$ in (2.12) can be rewritten in the form

$$\hat{\rho}(D, f) = \text{ave}[(f - \hat{g})^2 z^2] + \hat{\sigma}^2 \,\text{ave}(\hat{g}). \tag{2.15}$$

For fixed penalty matrix D, the *shrinkage-adaptive* PLS(D) estimator is defined to be $\hat{\mu}_{MS}(D, \hat{\nu})$, where

$$\hat{\nu} = \underset{\nu \in [0,\infty]}{\text{argmin}}\, \hat{\rho}(D, f(\nu)) = \underset{\nu \in [0,\infty]}{\text{argmin}}\, \text{ave}[(f(\nu) - \hat{g})^2 z]. \tag{2.16}$$

Computation of $\hat{\nu}$ is thus a weighted least squares problem that can be solved with the S-Plus function `nls` in the manner exhibited on p. 244 of Venables and Ripley (1999). The PLS fits plotted in the third row of Fig. 1 were obtained in this fashion.

Similarly, for fixed penalty matrix D, the *shrinkage-adaptive* MS(D) estimator is defined to be $\hat{\mu}_{MS}(D, \hat{f}_{MS})$, where

$$\hat{f}_{MS} = \underset{f \in \mathcal{F}_{MS}}{\text{argmin}}\, \hat{\rho}(D, f) = \underset{f \in \mathcal{F}_{MS}}{\text{argmin}}\, \text{ave}[(f - \hat{g})^2 z]. \tag{2.17}$$

To facilitate this minimization, let $\mathcal{H} = \{h \in R^p: h_1 \geq h_2 \geq \cdots \geq h_p\}$ and let

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}}\, \text{ave}[(h - \hat{g})^2 z]. \tag{2.18}$$

Then $\hat{f}_{MS} = \hat{h}_+$. That is, each component of $\hat{f}_{MS}$ is the positive part of the corresponding component of $\hat{h}$. For a proof, see Beran and Dümbgen (1998). Computation of $\hat{h}$ is a weighted isotonic least squares problem that can be solved with the pool-adjacent-violators algorithm (cf. Robertson *et al.* (1988)). The MS fits plotted in the last row of Fig. 1 were obtained in this fashion. Computation is faster for MS(D) than for PLS(D). S-Plus code for the examples in this paper is available from the author.

The following theorem shows that adaptation works in the sense that minimizing estimated risk over either the MS or PLS shrinkage class for fixed D succeeds in minimizing risk asymptotically over that class. The result makes no smoothness assumptions on the unknown mean vector μ and follows from Theorems 2.1 and 2.2 in Beran and Dümbgen (1998).

THEOREM 2.1. *Let $\mathcal{F}$ be any subset of $\mathcal{F}_{MS}$ that is closed in $[0,1]^p$. In particular, $\mathcal{F}$ could be either the PLS shrinkage class $\{f(\nu): \nu \in [0,\infty]\}$ or the monotone shrinkage class $\mathcal{F}_{MS}$. Suppose that $\hat{\sigma}^2$ is consistent in that, for every $r > 0$ and $\sigma^2 > 0$,*

$$\lim_{p \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2 r} E|\hat{\sigma}^2 - \sigma^2| = 0. \tag{2.19}$$

Let $V(f)$ denote either the loss $L(\hat{\mu}(D,f),\mu)$ or the estimated risk $\hat{\rho}(D,f)$. Then, for every penalty matrix D, every $r>0$, and every $\sigma^2>0$,

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} E \sup_{f\in\mathcal{F}} |V(f) - \rho(f,\xi^2,\sigma^2)| = 0. \tag{2.20}$$

Moreover, if $\hat{f} = \text{argmin}_{f\in\mathcal{F}}\,\hat{\rho}(D,f)$, then

$$\lim_{p\to\infty} \sup_{\text{ave}(\mu^2)/\sigma^2\le r} |R(\hat{\mu}(D,\hat{f}),\mu,\sigma^2) - \min_{f\in\mathcal{F}} R(\hat{\mu}(D,f),\mu,\sigma^2)| = 0. \tag{2.21}$$

By (2.20), the loss, risk and estimated risk of a candidate estimator converge together asymptotically. Uniformity of this convergence over $\mathcal{F}$ makes the estimated risk of candidate estimators a reasonable surrogate for true risk or loss. By (2.21), the risk of the shrinkage-adaptive estimator $\hat{\mu}(D,\hat{f})$ converges to that of the best candidate estimator. These conclusions break down when the class of shrinkage vectors $\mathcal{F}$ is too large in a covering number sense. In particular, it does not hold if $\mathcal{F} = [0,1]^p$, as shown in Beran and Dümbgen (1998).

Remarks. Condition (2.19) holds for the variance estimator $\hat{\sigma}^2_{LS}$ if $n-p$ tends to infinity with p. Asymptotic results for other variance estimators are given in Beran (1996) and Beran and Dümbgen (1998). The quantity $\text{ave}(\mu^2)/\sigma^2 = \text{ave}(\xi^2)/\sigma^2$ in (2.21) measures the signal to noise ratio. Limits (2.20) and (2.21) both hold without any restrictions on the smoothness of μ. Because the monotone shrinkage class $\mathcal{F}_{MS}$ is strictly larger than the generating PLS shrinkage class $\{f(\nu)\colon \nu\in[0,\infty]\}$, the asymptotic risk of MS(D) cannot exceed that of PLS(D).

COROLLARY 2.1. *Under the conditions for Theorem* 2.1,

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} \mathrm{E}\mid \hat{\rho}(D,\hat{f}) - W)\mid = 0 \tag{2.22}$$

for W equal to either $L(\hat{\mu}(D,\hat{f}),\mu)$ or $R(\hat{\mu}(D,\hat{f}),\mu,\sigma^2)$.

PROOF. Equation (2.20) implies that

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} E \sup_{f\in\mathcal{F}} |\hat{\rho}(D,f) - L(\hat{\mu}(D,f),\mu)| = 0, \tag{2.23}$$

which yields (2.22) for the first choice of W. Because $\hat{f}$ minimizes $\hat{\rho}(D,f)$ over $f\in\mathcal{F}$, equation (2.20) also implies that

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} \mathrm{E}|\hat{\rho}(D,\hat{f}) - \min_{f\in\mathcal{F}} \rho(f,\xi^2,\sigma^2)| = 0. \tag{2.24}$$

Combining this with (2.21) yields (2.22) for the second choice of W.

That the *plug-in* loss/risk estimator $\hat{\rho}(D,\hat{f})$ converges asymptotically to the actual loss/risk of PLS(D) or MS(D) is useful when comparing adaptive estimators on specific data. For the examples of Fig. 1, the plug-in loss/risk estimates and actual losses for

	MS loss	MS plug-in	PLS loss	PLS plug-in	LS loss	LS plug-in
Smooth	.0011	−.0068	.0013	−.0066	.0372	.0455
Wiggly	.0111	.0015	.0138	.0072	.0372	.0456
Very Wiggly	.0127	.0092	.0326	.0290	.0372	.0454

PLS(D_4), MS(D_4), and the least squares estimator are shown in Table 1. In scrutinizing this table, we observe that:

• The plug-in estimated losses for the shrinkage-adaptive MS and PLS estimates are noticeably smaller than the true losses.

• The plug-in losses indicate correctly the ordering of the true losses for the MS, PLS and LS estimates.

• The loss of the LS estimator in each of the three examples is .0372, a value reasonably close to the LS risk $\sigma^2 = .04$. The high-component variance estimator used in this experiment overestimates the true variance modestly.

3. Penalty matrix adaptation

Section 3.1 analyzes the manner in which the penalty matrix D affects the asymptotic risks of adaptive estimators MS(D) and PLS(D). The economy of the penalty basis in representing the unknown mean vector μ is a key factor. Section 3.2 develops candidate penalty matrices for equally and unequally spaced factor levels and considers adaptation over both penalty matrix and shrinkage vector. Section 3.3 discusses diagnostic plots that display the empirical economy of candidate penalty bases and considers an alternative to plug-in estimates for the loss or risk of adaptive estimators.

3.1 *Role of an economical penalty basis*

As will be seen, the risk of the shrinkage-adaptive PLS or MS estimator for μ is relatively small if all but the first few components of $\xi = U'\mu$ are very nearly zero. In this event, we say that the columns of the matrix U provide an *economical* basis for the regression space $\mathcal{M}(X)$. The benefit of using an economical regression basis is clear heuristically. In that case, we need only identify and estimate from the data the relatively few non-zero components of ξ, using the naive estimate zero for the remaining components. The quadratic risk then accumulates small squared biases from ignoring the nearly zero components of ξ but does not accumulate the many variances that would arise from an attempt to estimate these unbiasedly.

An idealized formulation of basis economy enables mathematical analysis of how economy affects risk. For every $b \in [0,1]$, every $r > 0$, and every $\sigma^2 > 0$, consider the projected ball

$$B(r,b,\sigma^2) = \{\xi\colon \text{ave}(\xi^2)/\sigma^2 \le r \text{ and } \xi_i = 0 \text{ for } i > bp\}. \tag{3.1}$$

Suppose that the regression basis U associated with penalty matrix D is economical in the formal sense that the transformed mean vector ξ lies in $B(r,b,\sigma^2)$ for some small value of b and some finite positive value of r. Though this description is too simple to serve as a complete definition of basis economy, it yields the following quantitative results about the effect of basis economy on the risk of estimators of μ.

THEOREM 3.1. *Fix the penalty basis U by choice of D. For every $b \in [0,1]$, every $r > 0$, and every $\sigma^2 > 0$, the asymptotic minimax quadratic risk over all estimators of μ is*

$$\lim_{p\to\infty} \inf_{\hat{\mu}} \sup_{\xi \in B(r,b,\sigma^2)} R(\hat{\mu}, \mu, \sigma^2) = \sigma^2 rb/(r+b). \tag{3.2}$$

The shrinkage-adaptive estimator $\hat{\mu}_{MS}(D, \hat{f}_{MS})$ achieves asymptotic minimax bound (3.2) in that

$$\lim_{p\to\infty} \sup_{\xi \in B(r,b,\sigma^2)} R(\hat{\mu}_{MS}(D, \hat{f}_{MS}), \mu, \sigma^2) = \sigma^2 rb/(r+b) \tag{3.3}$$

for every possible b, r, and σ^2.

Limit (3.3) follows from Theorem 4 in Beran (2000). As discussed in that paper, equation (3.2) is a special case of Pinsker's (1980) asymptotic minimax bound. Note that (3.3) establishes more than formal asymptotic minimaxity of shrinkage-adaptive estimator MS(D). When b is small, the right side of (3.3) is much smaller than the risk σ^2 of the least squares estimator $\hat{\mu}_{LS}$. To the extent that estimator PLS(D) approximates estimator MS(D), its performance also benefits strongly from economy of the penalty basis. This phenomenon underlies the very similar appearance of PLS(D_4) and MS(D_4) in the first column of Fig. 1.

3.2 *Candidate penalty matrices and adaptation*

The ideal choice of penalty basis U would have its first column proportional to the unknown mean vector μ so that only the first component of ξ would be nonzero. Though unrealizable, this ideal choice indicates that prior information or conjecture about μ can be exploited in devising the penalty matrix D that generates the penalty basis. The discussion in this section relates prior notions about the local behavior of the mean function m to the construction of reasonable candidate penalty matrices.

Difference operators. Consider the important case when the factor level vector $s = (s_1, s_2, \ldots, s_p)'$ has equally spaced components. To define the d-th difference matrix D_d, consider the $(p-1) \times p$ matrix $\Delta(p) = \{\delta_{i,j}\}$ in which $\delta_{i,i} = 1$, $\delta_{i,i+1} = -1$ for every i and all other entries are zero. Then,

$$D_1 = \Delta(p) \quad \text{and} \quad D_d = \Delta(p-d+1)D_{d-1} \quad \text{for } 2 \le d < p. \tag{3.4}$$

It may be verified that the $(p-d) \times p$ matrix D_d annihilates powers of s up to power $d-1$ in the sense that

$$D_d s^k = 0 \quad \text{for} \quad 0 \le k \le d-1. \tag{3.5}$$

Moreover, in row i of D_d, the elements not in columns $i, i+1, \ldots i+d$ are zero.

The penalty term in (1.5) is proportional to $|D\beta|^2$ where $\beta = m(s)$. When m behaves locally like a polynomial of degree $d-1$, property (3.5) and the subsequent remark about zeros entail that $|D_d\beta|$ is small. We may therefore expect that both PLS(D_d) and MS(D_d) will favor fits with local polynomial behavior of degree $d-1$. This implicit preference is appropriate whenever m has such local polynomial behavior. The success of fits based on penalty matrix D_4 in the first column of Fig. 1 illustrates

the point. We note that normalizing the row vectors of D_d to have unit length does not change the corresponding penalty basis U. However, (3.5) breaks down for $k \geq 1$ when the components of s are not equally spaced.

Local annihilators. To devise useful candidate penalty matrices for arbitrary factor levels $s \in R^p$ and for other notions about m, we draw on the mathematical interpretation of (3.5) as an orthogonality property. Let $g_0, g_1, \ldots, g_{d-1}$ be a given set of real-valued functions defined on the real line. We hypothesize that the mean function m behaves locally like a linear combination of the $\{g_k: 0 \leq k \leq d-1\}$. Local polynomial behavior is the special case where $g_k(s_i) = s_i^k$ for every k.

For each i such that $1 \leq i \leq p-d$, assume that the d vectors $\{(g_k(s_i), \ldots, g_k(s_{i+d}): 0 \leq k \leq d-1\}$ are linearly independent in R^{d+1}. This is a condition on the functions $\{g_k\}$ that is satisfied, for instance, when $g_k(s_i) = s_i^k$. Let $\mathcal{G}_i$ denote the d-dimensional subspace of R^{d+1} that is spanned by these vectors. Define the $(p-d) \times p$ *local annihilator* matrix $A_d = \{a_{i,j}\}$ as follows: In the i-th row of A_d, the subvector $\{a_{i,j}: i \leq j \leq i+d\}$ is the unit vector in R^{d+1}, unique up to sign, that is orthogonal to $\mathcal{G}_i$. The remaining elements of A_d are zero.

THEOREM 3.2. *Let* $g_k(s) = (g_k(s_1), g_k(s_2), \ldots, g_k(s_p))'$. *Each row vector of the local annihilator matrix* A_d *has unit length and*

$$A_d g_k(s) = 0 \quad \text{for} \quad 0 \leq k \leq d-1. \tag{3.6}$$

PROOF. The definition of A_d ensures that its rows have unit length and

$$\sum_{j=1}^{p} a_{i,j} g_k(s_j) = \sum_{j=i}^{i+d} a_{i,j} g_k(s_j) = 0 \quad \text{for} \quad 0 \leq k \leq d-1. \tag{3.7}$$

Of particular utility as the generalization of D_d for unequally spaced factor levels is the *local polynomial annihilator*. This is obtained by setting $g_k(s_i) = s_i^k$ in the definition of A_d. Thus, in the i-th row of the local polynomial annihilator, the subvector $\{a_{i,j}: i \leq j \leq i+d\}$ is the basis vector of degree d in the orthonormal polynomial basis on the factor levels $(s_i, \ldots, s_{i+d})$. All other elements in the row are zero. The S-Plus function `poly` enables computation of the local polynomial annihilator in a numerically stable way for d up to 50 or so. When the components of s are equally spaced, the local polynomial annihilator A_d becomes a scalar multiple of the d-th difference matrix D_d. Of course, local polynomial A_1 is proportional to D_1 for every factor level vector s.

Remark. A referee kindly pointed out that the foregoing discussion of annihilators can be linked to the algorithm for L-splines described at the end of Heckman and Ramsay (2000). Let $m^{(j)}$ denote the j-th derivative of m and let L be a differential operator such that $Lm = \sum_{j=0}^{d-1} a_j m^{(j)}$. The set of all m such that $Lm = 0$ is a linear space of dimension d. Let $g_0, g_1, \ldots, g_{d-1}$ denote a basis for this space. The construction of the sparse matrix A_d in Theorem 3.2 follows from the Heckman and Ramsay algorithm by setting Q', D and U in their notation to A_d, identity matrix and $(g_0(s), g_1(s), \ldots, g_{d-1}(s))$ in the present setting.

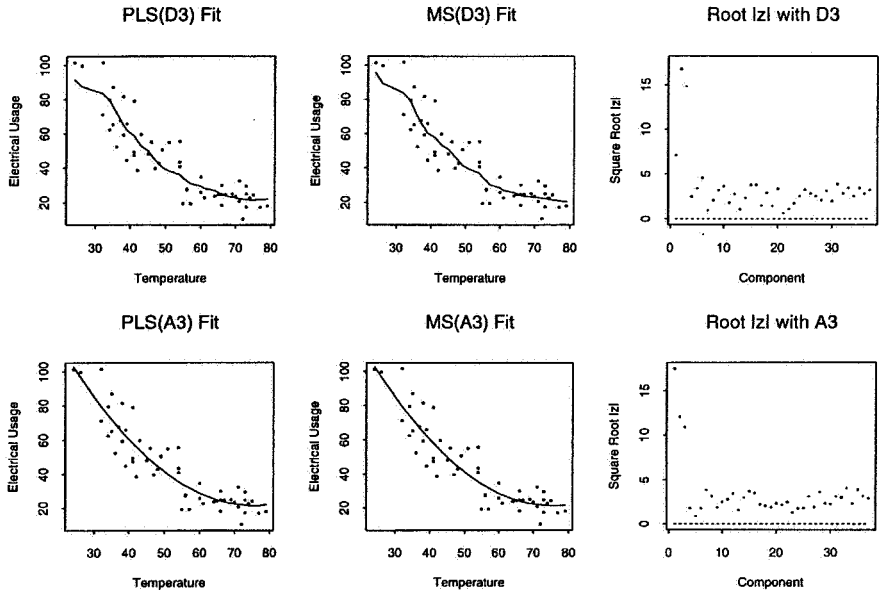

Fig. 2. Using penalty matrices D_3 and local polynomial A_3 respectively, each row displays adaptive PLS and MS estimates for conditional mean electrical usage and the associated basis economy plot.

	D_3 penalty	A_3 penalty
PLS plug-in loss	−33.55	−42.95
MS plug-in loss	−35.57	−42.99

Figure 2 exhibits competing PLS and MS estimates for mean electrical usage as a function of temperature. The data is described in Simonoff (1996). We estimate mean electrical usage conditional on the observed temperatures, whose distinct values are not equally spaced. The variance is estimated by $\hat{\sigma}^2_{LS}$. Because the trend in the data appears to be roughly quadratic, we expect that MS and PLS fits generated with the local polynomial annihilator A_3 as penalty matrix will have relatively low estimated risks. This turns out to be the case. The first row of Fig. 2 gives the PLS and MS fits when the penalty matrix is D_3 while the second row gives the corresponding fits when the penalty matrix is local polynomial A_3. The plug-in loss/risk estimates for these competing fits are shown in Table 2. In sharp contrast, the loss/risk estimate for the least squares estimator of μ is 129.70.

The negativity of the risk estimates in this table is an artifact of the small regression space dimension, $p = 37$. The ordering of the estimated risks matches the visual quality of the competing fits in Fig. 2. In this example, MS does not improve significantly upon PLS. However, choosing the penalty matrix to handle unequal spacing of the design points is clearly beneficial. The basis-economy plots in the third column of Fig. 2 exhibit the superior empirical economy of the local polynomial A_3 penalty basis relative to the D_3 basis.

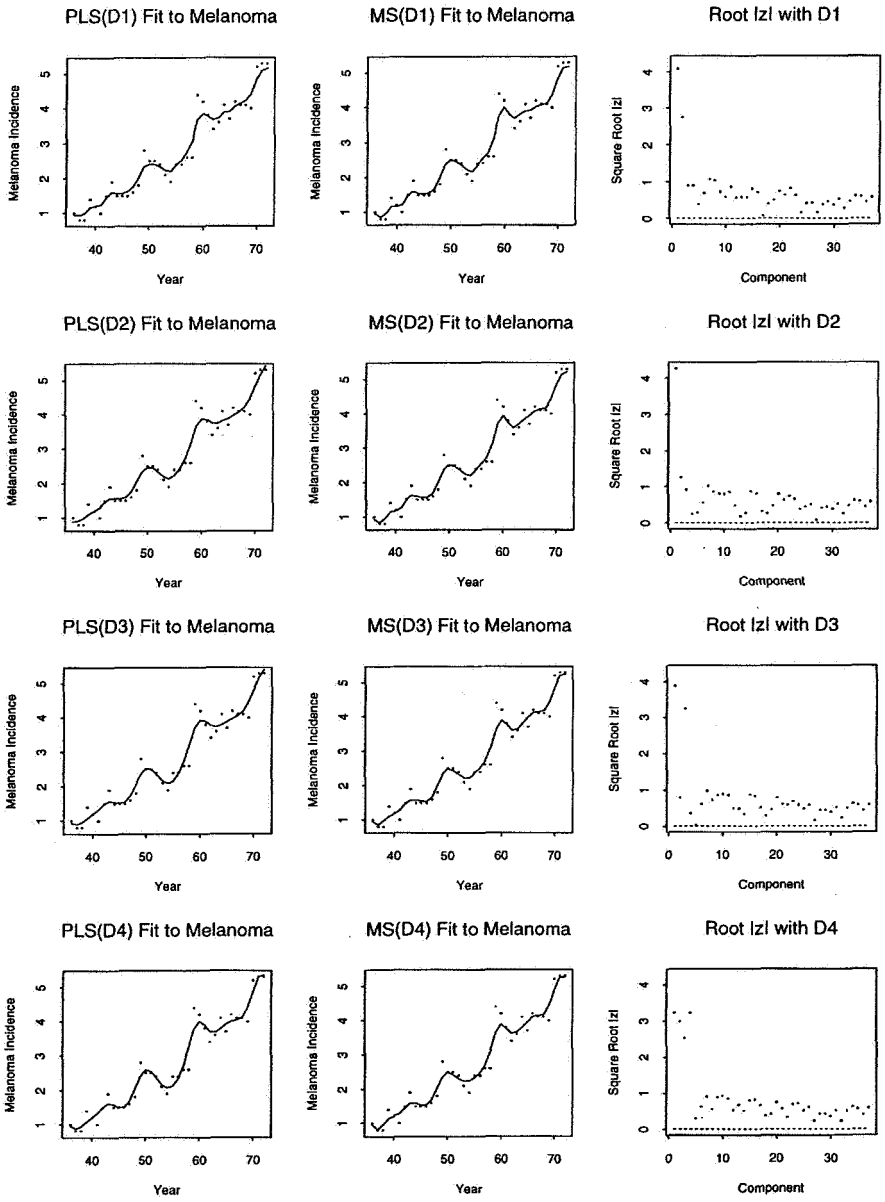

Fig. 3. Row d displays the shrinkage-adaptive PLS(D_d) and MS(D_d) estimates for mean melanoma incidence and, in the third column, the associated basis-economy plot.

Adaptation over penalty bases. Having devised a set $\mathcal{D}$ of candidate penalty matrices, we may use estimated risk to select an empirically best PLS or MS estimator by extending the adaptation method described in Section 2. Over shrinkage class $\mathcal{F}$ and penalty matrix class $\mathcal{D}$, the fully adaptive estimator of μ is defined to be $\hat{\mu}_{\mathcal{D},\mathcal{F}} = \hat{\mu}(\hat{D}, \hat{f})$, where

$$(\hat{D}, \hat{f}) = \underset{D \in \mathcal{D}, f \in \mathcal{F}}{\operatorname{argmin}} \hat{\rho}(D, f). \tag{3.8}$$

	$d=1$	$d=2$	$d=3$	$d=4$
PLS(D_d) plug-in loss	.0310	.0294	.0326	.0349
MS(D_d) plug-in loss	.0165	.0166	.0194	.0230

If the cardinality of $\mathcal{D}$ is $o(p^{1/2})$, $\mathcal{F}$ is a closed subset of $\mathcal{F}_{MS}$, and $E|\hat{\sigma}^2 - \sigma^2| = O(p^{1/2})$, then Theorem 2.1 and Corollary 2.1 may be extended to justify the simultaneous adaptation in (3.8) over both f and D. The extension follows from the error bounds established in Theorems 2.1 and 2.2 of Beran and Dümbgen (1998). Justifying adaptation over larger classes of penalty matrices is an open question. Because local polynomials of degree up to 6 or so approximate a wide range of smooth mean vectors, adaptation over large $\mathcal{D}$ need not be advantageous.

Figure 3 exhibits competing adaptive PLS and MS estimates for mean melanoma incidence in males based on measurements for the years 1936 to 1972 and using the first-difference variance estimator $\hat{\sigma}^2_{D1}$. The data is given on pp. 199–201 of Andrews and Herzberg (1985). The first two columns in Fig. 3 display linearly interpolated PLS(D_d) and MS(D_d) fits to the data, the candidate penalty matrices being $\{D_d: 1 \leq d \leq 4\}$. The plug-in loss/risk estimates for these competing fits are showen in Table 3.

The loss/risk estimate for the least squares estimator of μ, which coincides here with the raw data, is .1165. It is not too surprising that the PLS(D_2) and MS(D_2) estimators have relatively low estimated risk among this group of competing shrinkage-adaptive estimators because the underlying trend in the melanoma data is roughly linear. The plotted shrinkage-adaptive estimators capture ripples in melanoma incidence that are associated with the sunspot cycle. It is notable that the competing adaptive fits in Fig. 3 are visually similar, even though their estimated risks differ. Heckman and Ramsay (2000) obtained similar fits to this data with continuous-spline penalized least squares, using differential penalty operators analogous to D_d and choosing penalty weight by generalized cross-validation or by equivalent degrees-of-freedom. Their treatment also considered a penalty differential operator that annihilates sinusoids of specified frequency.

The third column in Fig. 3 plots the components $\{|z_i|^{1/2}\}$ against i for each of the four penalty bases considered. Such diagnostic plots will be called *basis-economy plots.* The square root transformation reduces the vertical range and makes more visible the values near zero. The purpose of a basis-economy plot is to approximate the unobservable ideal plot of the $\{|\xi_i|^{1/2}\}$ against i so as to assess the economy of the penalty basis. For the melanoma data, the penalty basis generated by D_2 is empirically the most economical in Fig. 3. This finding is consistent with the ranking of estimated risks described above. At the same time, all four penalty matrices $\{D_d: 1 \leq d \leq 4\}$ yield similar looking fits.

3.3 *Diagnostic tools*

The foregoing theory and examples have identified two key factors that govern the risk of PLS and MS estimators. The first and more important factor is the economy of the basis U generated by the penalty matrix D. The second factor is the extent to which adaptive monotone shrinkage or penalized least squares shrinkage is able to exploit whatever economy exists in the chosen basis U. Flexibility in the shrinkage strategy becomes particularly important when, as columns two and three of Fig. 1, high-frequency details in the unknown mean entail that strict economy does not hold.

For a given penalty matrix, a comparative shrinkage-vector plot displays, with linear

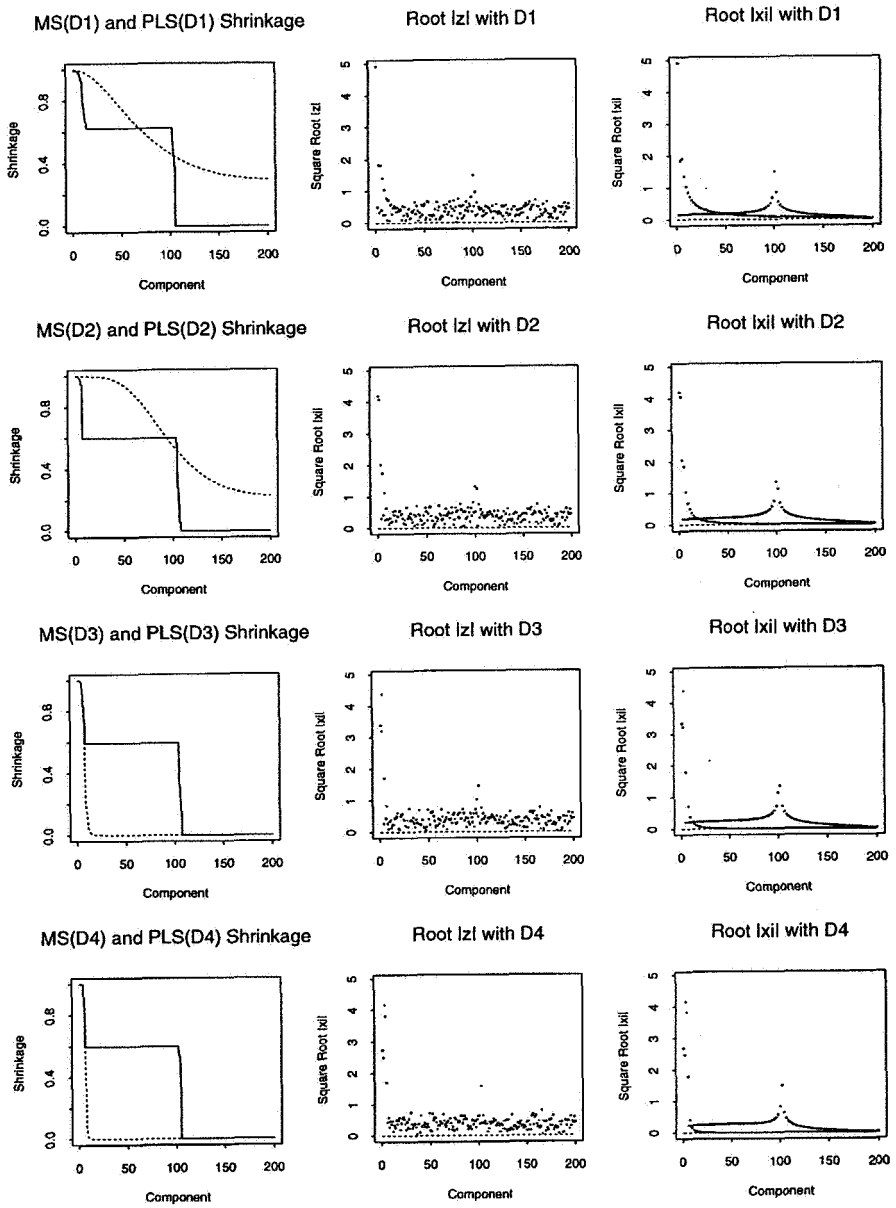

Fig. 4. On the Very Wiggly data, row d displays the shrinkage vectors for estimators PLS(D_d) (solid interpolation) and MS(D_d) (dashed interpolation), the basis economy plot, and the ideal basis economy plot.

interpolation for visibility, the components of the adaptively chosen shrinkage vectors $\hat{f}_{PLS}$ and $\hat{f}_{MS}$. Setting such a plot next to the basis-economy plot enables one to assess how well adaptive MS or PLS estimation exploits the degree of economy present in the penalty basis. For the Very Wiggly data described in the Introduction, the first two columns in Fig. 4 display the shrinkage-vectors and basis-economy plots generated by penalty matrices D_1 through D_4. The D_4 basis appears more economical than the other

three penalty bases, but not much more than the D_3 basis. However, for either the D_3 or D_4 basis, adaptive PLS does a poor job of mimicking adaptive MS. This phenomenon underlies the inability of estimate PLS(D_4) in Fig. 1 to recover the sinusoidal component of trend. The third column of Fig. 4 plots the actual components of ξ, which are available here because the data is artificial and μ is known. It is gratifying that the empirical basis-economy plots in the middle column capture the essential features found in the ideal plots of the third column.

Feedback about which nonparametric regression procedure to use in a particular data analysis can come from estimated performance summaries as well as from diagnostic plots. A broadband diagnostic approach is surely more effective than any single tool. Plug-in estimated losses sharpen our scrutiny of the fits and diagnostic plots in Figs. 1 to 4. However, the discussion accompanying Fig. 1 indicated that plug-in estimated loss/risk for an adaptive PLS or MS estimate tends to underestimate true loss. We therefore consider another approach to estimating the loss or risk of a general estimator $\hat{\mu} = \hat{\mu}(y)$. Let $g(y) = \hat{\mu}(y) - y$. If the function g satisfies assumptions detailed in Stein (1981), then the risk of $\hat{\mu}$ under the Gaussian model described in the Introduction is

$$R(\hat{\mu}, \mu, \sigma^2) = \sigma^2 + E\left[2\sigma^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2\right]. \tag{3.9}$$

The implied estimator of loss or risk is

$$\hat{L}(\hat{\mu}) = \hat{\sigma}^2 + 2\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} \partial g_i(y)/\partial y_i + n^{-1}|g(y)|^2. \tag{3.10}$$

When $\hat{\mu}(y)$ lacks a tractable closed form, the partial derivatives needed in (3.10) may be approximated numerically. Let v_i denote the vector in R^n whose i-th component is 1 and whose other components are 0. Then, for small real values of δ,

$$\partial g_i(y)/\partial y_i \approx \delta^{-1}[g_i(y + \delta v_i) - g_i(y)], \qquad 1 \le i \le n. \tag{3.11}$$

Computing these difference quotients requires computing $\hat{\mu}(y) = y + g(y)$ and the n perturbed estimators $\{\hat{\mu}(y + \delta v_i): 1 \le i \le n\}$.

Sometimes the Stein loss/risk estimator in (3.10) has a closed form expression. For the candidate estimators $\hat{\mu}_{PLS}(D, \nu)$ or $\hat{\mu}_{MS}(D, f)$, the estimator (3.10) reduces to $\hat{\rho}(D, f(\nu))$ or $\hat{\rho}(D, f)$ respectively. For either PLS(D) or MS(D), the loss, the risk, and the plug-in loss/risk estimator converge together as p tends to infinity; Theorem 2.1 and Corollary 2.1 give the details. However, the experiment reported in Section 2.3 indicates that the rate of convergence may not be swift and that plug-in loss/risk estimators may underestimate true loss.

Alternatively, we can construct by numerical approximation the Stein loss/risk estimator (3.11) for the shrinkage-adaptive estimators PLS(D) and MS(D). Does this approach produce better estimates of loss than the plug-in method? For the examples of Fig. 1, where the penalty matrix is D_4, the approximate Stein loss/risk estimate obtained from (3.11) with $\delta = .0001$ may be compared with their plug-in counterparts and the true losses (Table 4). In this table, the Stein and the plug-in estimates for the loss of MS(D_4) and PLS(D_4) are close; their ranking is the same; and the former is only slightly closer to the true loss in most cases. There is no compelling reason in this experiment to prefer the Stein loss/risk estimates over their computationally simpler plug-in counterparts.

	MS loss	MS Stein	MS plug-in	PLS loss	PLS Stein	PLS plug-in
Smooth	.0011	−.0061	−.0068	.0013	−.0061	−.0066
Wiggly	.0111	.0031	.0015	.0138	.0076	.0072
Very Wiggly	.0127	.0100	.0092	.0326	.0294	.0290

Acknowledgements

This research was supported in part by National Science Foundation Grant DMS99–70266.

References

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

Beran, R. (1996). Confidence sets centered at C_p estimators, *Ann. Inst. Statist. Math.*, **48**, 1–15.

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

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

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

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator, *Statist. Sci.*, **6**, 404–436.

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

Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **15**, 661–676.

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

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

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proc. Third Berkeley Symp. on Math. Statist. Prob.*, Vol. 1 (ed. J. Neyman), 197–206, University of California Press, Berkeley.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, **9**, 1135–1151.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*, 3rd ed., Springer, New York.