AISM (2007) 59: 171–195
DOI 10.1007/s10463-006-0050-z

Rudolf Beran

Multiple penalty regression: fitting and extrapolating a discrete incomplete multi-way layout

Received: 10 September 2004 / Revised: 9 August 2005 / Published online: 11 May 2006

Abstract The discrete multi-way layout is an abstract data type associated with regression, experimental designs, digital images or videos, spatial statistics, gene or protein chips, and more. The factors influencing response can be nominal or ordinal. The observed factor level combinations are finitely discrete and often incomplete or irregularly spaced. This paper develops low risk biased estimators of the means at the observed factor level combinations; and extrapolates the estimated means to larger discrete complete layouts. Candidate penalized least squares (PLS) estimators with multiple quadratic penalties express competing conjectures about each of the main effects and interactions in the analysis of variance decomposition of the means. The candidate PLS estimator with smallest estimated quadratic risk attains, asymptotically, the smallest risk over all candidate PLS estimators. In the theoretical analysis, the dimension of the regression space tends to infinity. No assumptions are made about the unknown means or about replication.

Keywords Nominal factor · Ordinal factor · Estimated risk · Tensor-product penalty · Multiparametric asymptotics · Penalized least squares · Bayes estimator

1 Introduction

The discrete multi-way layout is an abstract data type that is associated with regression, experimental designs, digital images or videos, spatial statistics, gene or protein chips, and other applications. In a discrete k_0-way layout, each of the k_0 factors assumes a finite number of levels. The levels of a factor may be either nominal (i.e., pure labels) or ordinal (i.e., real-values whose order and magnitude bear information). Factors of both types may occur in a multi-way layout.

R. Beran
Department of Statistics, University of California at Davis, Davis, CA 95616, USA
E-mail: beran@wald.ucdavis.edu

Example 1 The *coal ash* data from Cressie (1993, p. 34) records percentage of coal ash in 208 assay samples. The data forms an incomplete two-way layout with one observation per observed cell; the factor pairs are the grid coordinates at which the assay samples were obtained. The factors row coordinate and column coordinate are both ordinal, are both equally spaced, and range over 23 and 16 equally spaced levels, respectively. On the grid of 368 factor level pairs so defined, an assay sample was taken at the 208 points identified in subplot (1, 2) of Fig. 1. Subplot (1,1) is a mesh plot of the observations. Relating mean coal ash percentage to the grid coordinates is the regression problem.

Example 2 The starch data from Freeman (1942, pp. 120–121), reprinted in Scheffé (1959, pp. 216–217), gives the breaking strength and the film thickness in tests on seven types of starch film. The data forms an incomplete unbalanced two-way layout with 94 observations in which breaking strength is the observed response, factor one is the type of starch, and factor two is the thickness of the starch film. In this case, factor one is nominal with seven levels while factor 2 is ordinal with 69 unequally spaced values. On the grid of 483 factor level pairs so defined, one or more observations of breaking strength were taken at the 81 points identified in subplot (1) of Fig. 2. At all but a few of the 81 observed design points, we have only one breaking strength measurement. Relating mean breaking strength to the factors starch type and film thickness is the regression problem.

As these two examples indicate, nominal or ordinal factors, significant incompleteness, and lack of replication are not uncommon in multi-way layout data. Indeed, any regression model with k_0 covariates (or factors) and real-valued responses forms a k_0-way layout of data that is usually incomplete and may be unbalanced.

An incomplete multi-way layout may be modeled in terms of a larger complete layout of means. We will first describe the complete layout and then extract the observed incomplete layout as a subset. Consider k_0 factors, whether nominal or ordinal, in which factor k has p_k distinct levels. Let $\mathcal{I}$ denote the set of all k_0-tuples $i = (i_1, i_2, \ldots, i_{k_0})$ such that $1 \le i_k \le p_k$ for $1 \le k \le k_0$. The component i_k indexes the levels of factor k. These k_0-tuples express all possible combinations of the factor levels. A complete k_0-way layout of means consists of real values $\{m_i : i \in \mathcal{I}\}$. To facilitate algebra, we order the $p = \prod_{k=1}^{k_0} p_k$ elements of the index set $\mathcal{I}$ in mirrored dictionary order: i_{k_0} serves as the first "letter" of the word, i_{k_0-1} as the second "letter", and so forth. Hereafter, the components of $\mathcal{I}$ are always taken in this order. With the index set so ordered, the means for the complete multi-way layout form the $p \times 1$ vector

$$m = \{m_i : i \in \mathcal{I}\} = \big\{\ldots\{\{m_{i_1,i_2,\ldots,i_{k_0}} : 1 \le i_1 \le p_1\}, \\ 1 \le i_2 \le p_2\}, \ldots, 1 \le i_{k_0} \le p_{k_0}\big\}. \qquad (1)$$

Observations are available only on the means $\{m_i : i \in \mathcal{I}_0\}$, where $\mathcal{I}_0$ is a subset of $\mathcal{I}$. When the cardinality q of $\mathcal{I}_0$ is less than p, these observations $y = \big\{\{y_{ij} : 1 \le j \le n_i\}, i \in \mathcal{I}_0\big\}$ form an *incomplete* k_0-way layout, which is unbalanced unless the $\{n_i\}$ are equal. The observation vector y is $n \times 1$ with $n = \sum_{i \in \mathcal{I}_0} n_i$. Define the means-incidence matrix D to be the $q \times p$ matrix of zeroes and ones

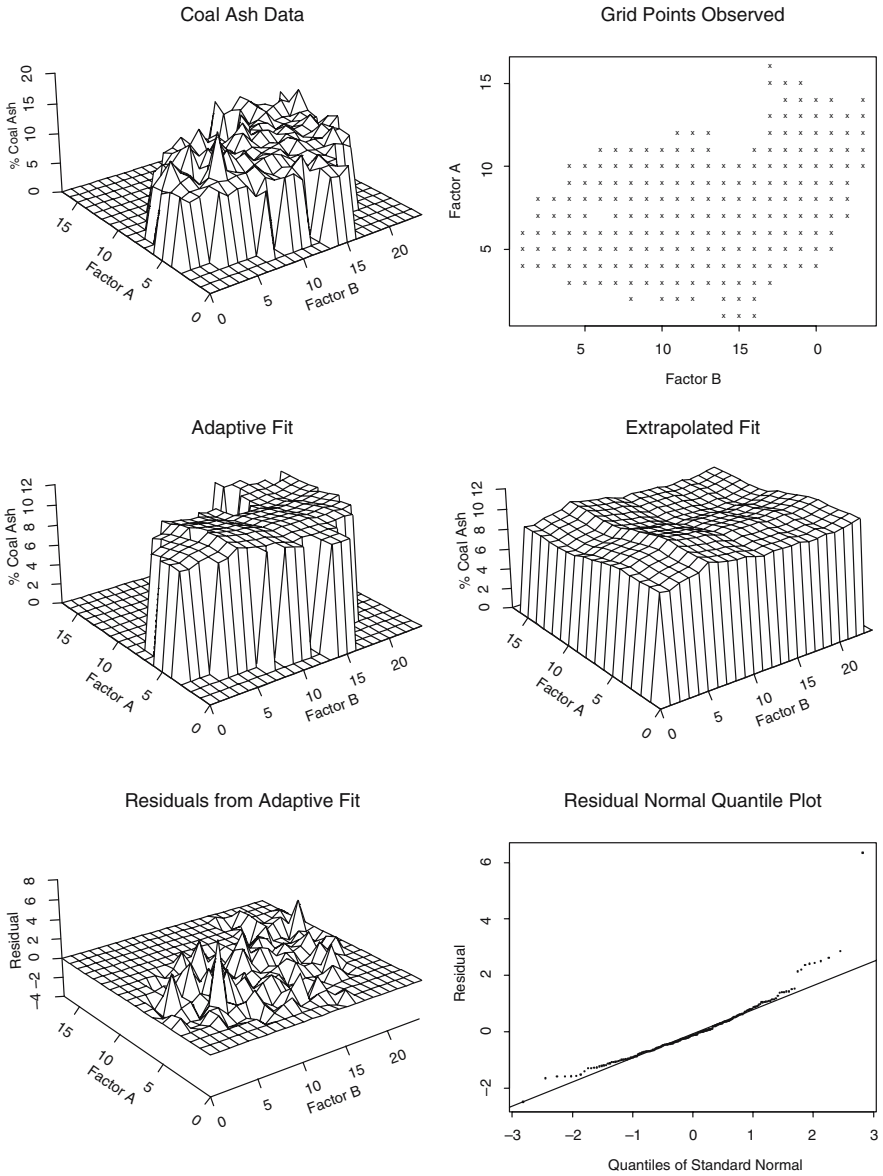

Fig. 1 The coal ash data, its observed factor-level grid, the adaptive PLS estimate $\hat{m}_D(\hat{t})$ of mean coal ash percentages, its extrapolation to the estimated regression function $\hat{m}(\hat{t})$, and residual plots that identify one very large outlier

such that $m_D = Dm$ lists, in vector form, the means $\{m_i : i \in \mathcal{I}_0\}$ for the observed incomplete k_0-way layout. Let C be the $n \times q$ data-incidence matrix of zeroes and ones that suitably replicates components of the vector $m_D = Dm$ into the vector $\eta = \mathrm{E}(y) = Cm_D$. For a complete layout of data, q equals p and D is just the

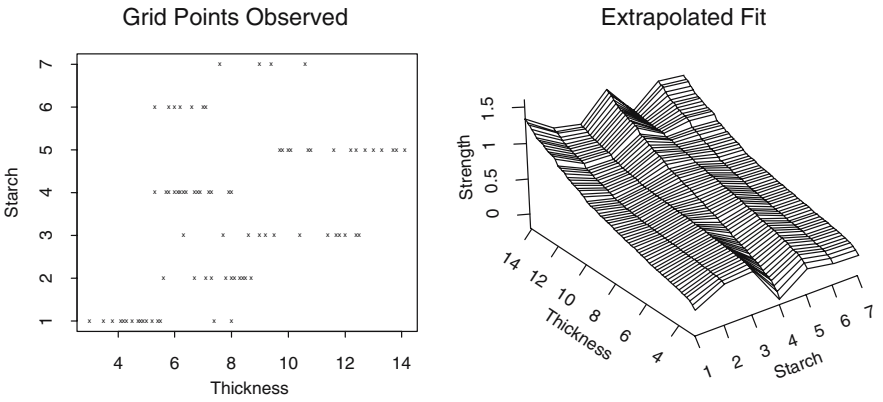

Fig. 2 The observed factor-level grid of the Starch data and the extrapolated adaptive PLS estimate $\hat{m}(\hat{t})$ of mean breaking strengths

identity matrix. In general, $q \leq \min\{p, n\}$, $\operatorname{rank}(C) = q$ and $DD' = I_q$. Consequently, $\operatorname{rank}(D) = q$ and $\operatorname{rank}(CD) = q$. For the coal ash data, $q = n = 208$ and $p = 368$. For the starch data, $q = 81$, $n = 94$, and $p = 483$.

This paper identifies regression with two tasks:

- Estimating efficiently the means $m_D = \{m_i : i \in \mathcal{I}_0\}$ on the factor level combinations where response data is observed. Because $m_D = C^+\eta$, where the superscript $+$ denotes pseudoinverse, this task is equivalent to estimating η.
- Extrapolating the fit to estimate the means $m = \{m_i : i \in \mathcal{I}\}$. The extrapolation defines the estimated regression function.

Note that the product set $\mathcal{I}$ can be chosen larger than the product set of observed factor levels. In defining $\mathcal{I}$, we may include all unobserved factor level combinations that are of interest for extrapolation.

The *strong Gauss-Markov model* for the incomplete layout of observations y asserts

$$y = \eta + e, \quad \eta = Cm_D = CDm, \quad m \in R^p. \tag{2}$$

There are no restrictions on the mean vector m. The components of e are independent, identically distributed with mean 0, unknown variance σ^2, and finite fourth moment. All non-Bayesian risk calculations and asymptotic convergences in this paper use the strong Gauss-Markov model (2) for y. The model expresses the idea that we know little about the means and are not willing to assume away this lack of knowledge through restrictions on the means.

When $q < p$, the parametric function $m_D = C^+\eta$ has an unbiased linear estimator while m does not. The *least squares* estimator of η in model (2) is

$$\hat{\eta}_{\mathrm{LS}} = CD(CD)^+ y = CC^+ y. \tag{3}$$

The corresponding least squares estimator of m_D is $\hat{m}_{D,\mathrm{LS}} = C^+\hat{\eta}_{\mathrm{LS}} = C^+ y$.

For any matrix A, including the special case of a vector, let $|A|$ denote the Euclidean (or Frobenius) norm: $|A|^2 = \operatorname{tr}(A'A) = \operatorname{tr}(AA')$. Define the normalized

quadratic loss of any estimator $\hat{\eta}$ of η to be

$$L(\hat{\eta}, \eta) = q^{-1}|\hat{\eta} - \eta|^2. \tag{4}$$

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

$$R\left(\hat{\eta}, \eta, \sigma^2\right) = EL(\hat{\eta}, \eta), \tag{5}$$

where the expectation is calculated under the strong Gauss-Markov saturated model (2).

The least squares estimator $\hat{\eta}_{\mathrm{LS}}$ in model (2) is an unbiased linear estimator with risk $R\left(\hat{\eta}_{\mathrm{LS}}, \eta, \sigma^2\right) = \sigma^2$. 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 $q \geq 3$ and the errors are independent, identically normally distributed. In statistical practice, $\hat{\eta}_{LS}$ is often too variable an estimator unless the rank q of the regression space is small. The basic message, gradually strengthened by subsequent statistical theory, is the need to consider biased estimators of η.

This paper studies penalized least squares (PLS) *estimators* for η, m_D, and m that rely on multiple quadratic penalties to determine the fit. Let $\mathcal{D}$ be a set of finite cardinality d. Let $\{t_s : s \in \mathcal{D}\}$ be relative penalty weights such that $0 \leq t_s \leq 1$. Let $\{Q_s : s \in \mathcal{D}\}$ be symmetric positive semi-definite $p \times p$ penalty matrices. Section 2 will provide specific useful constructions of $\mathcal{D}$ and of Q_s. Without loss of generality, assume that the elements of $\mathcal{D}$ have been ordered. Let t denote the vector in $[0, 1]^d$ formed by taking the $\{t_s\}$ in the order imposed on the subscript s. For any matrix A, let

$$\rho(A) = \sup_{x \neq 0} \frac{|Ax|}{|x|}. \tag{6}$$

The function ρ is a matrix norm. In the theory to be developed, we scale each penalty matrix Q_s so that $\rho(Q_s) = 1$.

Let c be a large positive constant and let ϵ be a small positive constant, say 10^{-7}. Define

$$Q(t) = \epsilon I_p + c \sum_{s \in \mathcal{D}} t_s Q_s. \tag{7}$$

The ϵ term ensures nonsingularity of $Q(t)$. For $t \in [0, 1]^d$, the PLS estimators of m, η, and m_D are

$$\begin{aligned} \hat{m}(t) &= \underset{m \in R^p}{\operatorname{argmin}} \left[|y - CDm|^2 + m' Q(t) m\right] \\ &= \left[D'C'CD + Q(t)\right]^{-1} D'C'y \end{aligned} \tag{8}$$

and

$$\begin{aligned} \hat{\eta}(t) &= CD\hat{m}(t) = CD\left[D'C'CD + Q(t)\right]^{-1} D'C'y \\ \hat{m}_D(t) &= C^+\hat{\eta}(t) = D\hat{m}(t). \end{aligned} \tag{9}$$

The foregoing expressions define linked families of candidate PLS estimators. The candidate estimator $\hat{m}_D(t)$ estimates the means at factor level combinations where data is observed. The candidate estimator $\hat{m}(t)$ uses prior conjecture expressed through the penalty term $m'Q(t)m$ to estimate m at every factor level k_0-tuple, including those at which no data is observed. Through this extrapolation, $\hat{m}(t)$ defines a candidate regression function on $\mathcal{I}$. Section 2.1 rederives the candidate PLS estimators as Bayes estimators in the restriction of model (2) to normally distributed errors.

Remark Families of candidate least squares fits to submodels of model (2) are limits of a subset of the candidate PLS fits just described. Indeed, suppose that submodel s, for $s \in \mathcal{D}$, asserts that m lies in specified subspace of R^p. Let P_s denote the $p \times p$ matrix that orthogonally projects R^p into that subspace. Let $Q_s = I_p - P_s$. The constraint $Q_s m = 0$ is satisfied if and only if m lies in the subspace that forms the range of P_s. In the penalty weight vector t, set component $t_s = 1$ and set the remaining components equal to 0. Then, as c tends to infinity, the PLS estimator $\hat{\eta}(t)$ converges to the least squares estimator of η for submodel s of (2). In this fashion, the class of candidate PLS estimators effectively extends the class of candidate submodel fits.

The PLS estimators $\left\{\hat{\eta}(t) : t \in [0, 1]^d\right\}$ defined in Eq. (9) constitute a class of candidate symmetric linear estimators for η. How should we construct the penalty matrices $\{Q_s : s \in \mathcal{D}\}$ defining this class? This construction needs to address the nominal or ordinal type of each factor. How should we 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 $t \in [0, 1]^d$. Because the risk function is usually unknown, we pursue the following modified program:

- Construct suitable penalty matrices $\{Q_s : s \in \mathcal{D}\}$ for multi-way layouts with nominal and ordinal factors. The penalty matrices are carefully designed to respect the nature of each factor and to express competing hypothetical notions about the smoothness (if pertinent) and the magnitude of each main effect and interaction in the analysis of variance (ANOVA) decomposition of m. According to their construction, the penalty matrices generate candidate PLS estimators that shrink toward ANOVA submodel fits with or without smoothing (see Sect. 2.3).
- Construct an estimator $\hat{r}(t)$ of the risk function of $\hat{\eta}(t)$. The estimated risk function used in this paper is equivalent to the Mallows (1973) C_p criterion and involves an estimator of σ^2 (see Sect. 2.2 for discussion of both estimators).
- Construct the adaptive estimator $\hat{\eta}(\hat{t})$ such that $\hat{t} = \text{argmin}_{t \in [0,1]^d} \hat{r}(t)$ (see Sect. 2.2).
- Find theoretical conditions in the strong Gauss-Markov model under which the loss and estimated risk functions of $\hat{\eta}(t)$ converge uniformly over $t \in [0, 1]^d$ in the L_1 sense to the true risk function as q tends to infinity (see Theorem 3 in Sect. 3).
- Hence deduce that the risk of adaptive estimator $\hat{\eta}(\hat{t})$ converges to the risk of oracle estimator $\hat{\eta}(\tilde{t})$ as q 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 $\{\hat{\eta}(t) : t \in [0, 1]^d\}$ (see Theorem 4 in Sect. 3).

- On the basis of the foregoing considerations, use $\hat{m}_D = C^+\hat{\eta}$ to estimate the means m_D at factor levels where data is observed; and use the adaptive estimator $\hat{m}(\hat{t})$ to extrapolate the fit to m, that is, to estimate the regression function on $\mathcal{I}$ (see Sect. 4).

A candidate PLS fit based on one or more quadratic penalty terms is a biased linear estimator, constructed in this instance through a mathematical regularization strategy. Under quadratic loss, the aim of biased estimation is to achieve a favorable trade-off between bias and variance that reduces risk in estimating η. Earlier studies of biased linear estimators for η include ridge regression (Hoerl and Kennard, 1970), PLS fits to discrete one-way layouts with nominal or ordinal factors (Beran, 2002), shrinkage estimators for complete balanced multi-way layouts with nominal factors (Stein, 1966), monotone shrinkage estimators for abstract one-way layouts (Beran and Dümbgen, 1998), symmetric linear estimators (Buja et al. 1989), and certain multiple penalty PLS estimators (Wood, 2000; Beran, 2005).

The scope of PLS goes well beyond smoothing if one pays attention to the choice of the penalty matrices. For example, through suitable construction of the penalty matrices, Beran (2005) closely approximated, by an adaptive PLS estimator, Stein's (1966) superior shrinkage estimator for a complete balanced multi-way layout with all factors nominal. The appropriate penalty matrices in this instance, based on flat annihilators defined in Sect. 2.3, encourage shrinkage when fitting each main effect and interaction in the ANOVA decomposition of the mean vector. The present paper handles incomplete unbalanced layouts with nominal or ordinal factors and uses only the strong Gauss-Markov assumption on model errors rather than normality. As a special case, our adaptive PLS methodology yields a superior estimator for an incomplete unbalanced multi-way layout with all factors nominal.

The results in this paper carry out the bulleted program above. Sect. 2 provides systematic constructions of multiple penalty terms for incomplete, unbalanced multi-way layouts that respect the ANOVA structure and the nominal or ordinal character of each factor. The asymptotics in Section 3 justify adaptive choice of the penalty-weight vector t to minimize estimated risk. The theory developed uses asymptotics for loss and risk in which the dimension q of the regression space for linear model (2) tends to infinity while the sample size $n \geq q$. Because the number of parameters being estimated tends to infinity with q, such asymptotics are aptly termed "multiparametric". These asymptotics suit multi-way layouts that provide only a few observations per observed combination of factor-levels—a not uncommon situation.

Remark Under classical asymptotics where n tends to infinity for fixed q, the variance contribution to quadratic risk is small. Thus, trading off variance against bias cannot help much unless the bias is also small. Under these asymptotics, we conjecture that the adaptive estimator $\hat{\eta}(\hat{t})$ behaves asymptotically like the least squares estimator except at or sufficiently near superefficiency points. This happens demonstrably in the Stein special case cited above.

In general, the spectral decomposition of the symmetric linear candidate estimator $\hat{\eta}(t)$ in (9) has eigenvalues and eigenvectors that *both* depend on the penalty-weight vector t. Earlier asymptotics in the literature [cf. Kneip 1994; Beran and Dümbgen 1998; Beran 2002) justified adaptation over t only in cases where the eigenvectors do not depend on t. This simplification of the spectral decomposition

occurs for complete balanced multi-way layouts (Beran 2005) but not for unbalanced or incomplete multi-way layouts. The arguments in the proofs of Sect. 5 do not rely on the spectral decomposition, justify adaptation for PLS estimators in incomplete unbalanced multi-way layouts, and require only the strong Gauss-Markov error assumption rather than the normality assumption that is made in some of the work cited.

For the coal ash data, cell (2,1) in Fig. 1 displays $\hat{m}_D(\hat{t})$, the adaptively estimated means at the factor level pairs where data is observed. Cell (2,2) presents the extrapolated adaptive mean estimator $\hat{m}(\hat{t})$. For the starch data, Fig. 3 and cell (1,2) of Fig. 2 display the extrapolated adaptive mean estimator $\hat{m}(\hat{t})$. Each extrapolation expresses the regression function implicit in the adaptive PLS estimation technique. Section 4 develops the details of these two case studies.

The extensive literature on regularization also treats estimators of a mean function that is deemed to be a function of *continuous* ordinal factors. The observations are made at a discrete grid of factor level combinations, as above. However, a usual aim with continuous-factor approaches is to estimate the smooth (by assumption) mean function on a continuum [cf. Wahba 1990; Wahba et al. 1995; Heckman and Ramsay 2000; Lin 2000]. This objective differs in mathematical formulation and analysis from the discrete estimation problem addressed by this paper. At first glance, estimating the means m on a very fine grid $\mathcal{I}$ of factor level combinations, at most of which we lack observations, might seem to differ little from estimating a smooth mean function. However, our discrete formulation of the estimation problem requires no assumptions on the mean vector m, even for the asymptotics of Sect. 3, in which the cardinality of $\mathcal{I}$ tends to infinity.

2 Candidate estimators, adaptation, and penalty matrices

Section 2.1 rederives the candidate PLS estimators (8) and (9) as Bayes estimators. The risk and estimated risk of the candidate estimator $\hat{\eta}(t)$ and adaptation over t are the subjects of Sect. 2.2. Section 2.3 constructs penalty matrices $\{Q_s : s \in \mathcal{D}\}$ that are suitable for fitting an incomplete multi-way layout with nominal or ordinal factors.

2.1 Bayes estimators and penalized least squares

Candidate Bayes estimators for η and m_D are the subject of this section. For their derivation, we assume temporarily that the errors in model (2) are normally distributed. The competing Gaussian Bayesian models, indexed by the choice of the matrix $Q(t)$, are given by

$$y|m \sim N\left(CDm, \sigma^2 I_n\right), \quad m \sim N\left(0, \sigma^2 Q^{-1}(t)\right). \tag{10}$$

The small constant ϵ in definition (7) of $Q(t)$ makes the prior distribution proper, thereby ensuring admissibility of each candidate Bayes estimator.

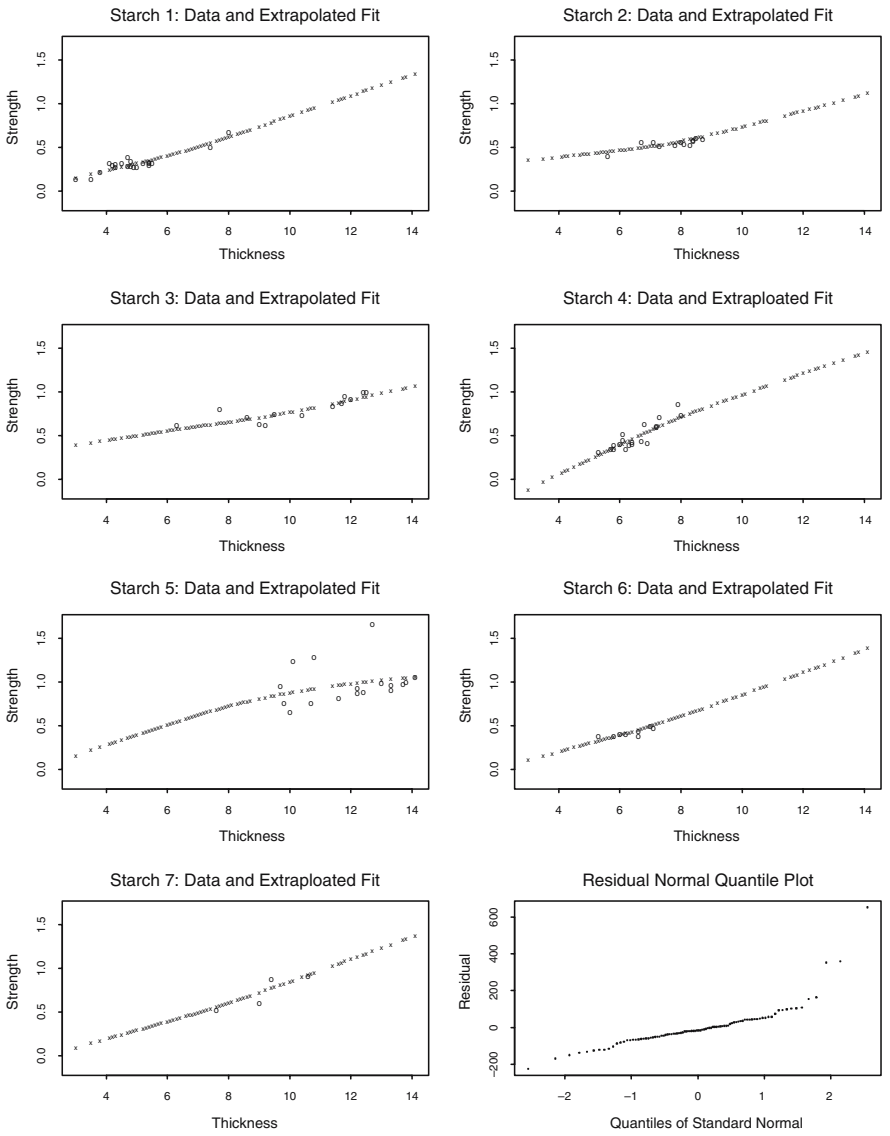

Fig. 3 The Starch data (*small o*), plotted by starch type, and the extrapolated adaptive PLS estimate $\hat{m}(\hat{t})$ of mean breaking strengths (*small x*). The normal quantile residual plot draws attention to the large outliers for starch 5

Theorem 1 *Under Bayesian model (10) and for quadratic loss, the respective Bayes estimators for* m, m_D *and* η *are*

$$\hat{m}(t) = \left[D'C'CD + Q(t)\right]^{-1} D'C'y, \tag{11}$$

$$\begin{aligned}\hat{m}_D(t) &= D[D'C'CD + Q(t)]^{-1} D'C'y \\ &= \left[C'C + (DQ^{-1}(t)D')^{-1}\right]^{-1} C'y,\end{aligned} \tag{12}$$

and

$$\hat{\eta}(t) = CD\left[D'C'CD + Q(t)\right]^{-1} D'C'y$$
$$= C\left[C'C + (DQ^{-1}(t)D')^{-1}\right]^{-1} C'y. \tag{13}$$

Section 5 proves the theorem. The equalities in Eqs. (12) and (14) will be seen to follow both by direct algebra and by probability reasoning.

The Bayes estimators for η and m coincide with the PLS estimators previously derived and provide alternative algebraic expressions for $\hat{\eta}(t)$ and $\hat{m}_D(t)$. Furthermore, it follows from the second expression in Eq. (12) that $\hat{m}_D(t)$ is itself a PLS estimator:

$$\hat{m}_D(t) = \underset{m_D \in R^q}{\operatorname{argmin}} \left[|y - Cm_D|^2 + m_D' \left(DQ^{-1}(t)D'\right)^{-1} m_D \right]. \tag{14}$$

A connection between Bayes and PLS estimators is not unexpected. For instance, Kimeldorf and Wahba (1970) established a correspondence between smoothing splines and and certain Bayes estimators for function estimation. The PLS and Bayes formulations illuminate, in complementary ways, the role of the matrix $Q(t)$ in expressing prior conjecture about m. It is important not to confuse such conjecture with fact. The estimated frequentist risks considered in Sect. 2.2 have, as their point, enabling the data to comment on the quality of prior conjecture.

2.2 Risk, estimated risk, and adaptation

Following the program outlined in Sect. 1, we consider the $\{\hat{\eta}(t) : t \in [0, 1]^d\}$ as candidate estimators for η in the incomplete unbalanced multi-way layout layout. It is not assumed that the Bayesian model used in the preceding subsection is correct. Instead, the performance of $\hat{\eta}(t)$ as a function of the penalty-weight vector t is scored through its estimated risk under the strong Gauss-Markov model (2). This approach assesses the approximate bias-variance trade-off achieved by each candidate estimator $\hat{\eta}(t)$.

Let

$$R = C'C, \quad U = CR^{-1/2}. \tag{15}$$

Evidently, U is $n \times q$ with $U'U = I_q$ and $R = \operatorname{diag}\{n_1, n_2, \ldots, n_q\}$ records the number of replications at each observed factor level combination. It follows that $\eta = U\xi$ with $\xi = R^{1/2}Dm$. Hence $\xi = U'\eta$. In model (2), let $z = U'y$ and $w = U'e$. From the second expression in Eq. (14),

$$\hat{\eta}(t) = US(t)U'y, \quad S(t) = \left[I_q + V(t)\right]^{-1},$$
$$V(t) = R^{-1/2}\left(DQ^{-1}(t)D'\right)^{-1} R^{-1/2}. \tag{16}$$

The loss function (4) of candidate estimator $\hat{\eta}(t)$ is

$$L(\hat{\eta}(t), \eta) = q^{-1}|\hat{\eta}(t) - \eta|^2 = q^{-1}|S(t)z - \xi|^2. \tag{17}$$

Let $T(t) = S^2(t)$ and $\bar{T}(t) = \left[I_q - S(t)\right]^2$. From Eq. (17), the risk function of the candidate estimator is

$$r(t) = \mathrm{E}L(\hat{\eta}(t), \eta) = q^{-1}\left[\sigma^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\{\bar{T}(t)\xi\xi'\}\right]. \tag{18}$$

Let $\hat{\sigma}^2$ be an L_1-consistent estimator of σ^2. An asymptotically unbiased estimator of $\xi\xi'$ is $zz' - \hat{\sigma}^2 I_q$. The Mallows (1973) estimator of risk function (18), implicit in the derivation of the C_p criterion, is then

$$\hat{r}(t) = q^{-1}\left[\hat{\sigma}^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\left\{\bar{T}(t)(zz' - \hat{\sigma}^2 I_q)\right\}\right]. \tag{19}$$

The (not necessarily unique) adaptive PLS estimator of η is

$$\hat{\eta}(\hat{t}), \quad \text{with } \hat{t} = \operatorname*{argmin}_{t\in[0,1]^d} \hat{r}(t). \tag{20}$$

The quality of $\hat{\eta}(\hat{t})$ will be assessed empirically by the estimated risk $\hat{r}(\hat{t})$. The asymptotics in Sect. 3 provide theoretical support for this approach by establishing suitably uniform convergence of estimated risk to risk. The corresponding adaptive PLS estimator of m_D is $\hat{m}_D(\hat{t}) = C^+\hat{\eta}(\hat{t})$.

To estimate m itself, we use $\hat{m}(\hat{t})$. However, we cannot estimate the risk of $\hat{m}(\hat{t})$ at factor level combinations that lack data. This is a logical consequence of model (2), which makes no assumptions about the means m. The extrapolation of $\hat{m}_D(\hat{t})$ on $\mathcal{I}_0$ to $\hat{m}(\hat{t})$ on $\mathcal{I}$, illustrated in Fig. 1 for the coal-ash data and in Fig. 3 for the starch data, draws strongly on the Bayes prior or penalty structure that is used in constructing the adaptive estimator. This extrapolation is a powerful what-if experiment that reveals the regression function implicit in the adaptive PLS fit $\hat{m}_D$.

Variance estimation. When $n > q$, the *least squares variance estimator* is

$$\hat{\sigma}^2_{LS} = (n-q)^{-1}|y - \hat{\eta}_{LS}|^2 = e'(I_n - UU')e. \tag{21}$$

It is L_1-consistent in the sense of (38) below, under the strong Gauss-Markov model (2), if $n - q$ tends to infinity as q tends to infinity.

When $n = q$, the absence of replication requires that an L_1-consistent estimator of σ^2 be based on additional data or on trustworthy prior information about η. The pooling of high-order interactions, as in the ANOVA, provides one way to do this. See Sect. 3.1 in Beran (2005) for details. Variance estimation based on first differences of adjacent observations is another way when the factors are ordinal and the means are thought to vary smoothly with factor level. Example 1 in Sect. 4 illustrates this technique.

2.3 Penalty matrices

This section considers the ANOVA decomposition for the complete k_0-way layout of means associated with the observed incomplete layout. We design competing classes of tensor-product penalty matrices that express the possible unimportance of certain interactions or main effects among the means and the possible smoothness in the dependence of these means on ordinal factors. These tensor-product penalty matrices serve as the $\{Q_s : s \in \mathcal{D}\}$ in the definition of the candidate estimator $\hat{\eta}(t)$.

2.3.1 ANOVA decomposition

The following algebra gives the orthogonal projections that define the ANOVA decomposition of a complete k_0-way layout of means into overall mean, main effects, and interactions. For $1 \le k \le k_0$, define the $p_k \times 1$ unit vector $u_k = p_k^{1/2}(1, 1, \ldots, 1)'$ and the $p_k \times p_k$ matrices

$$J_k = u_k u_k', \quad H_k = I_{p_k} - u_k u_k'. \tag{22}$$

For each k, the symmetric, idempotent matrices J_k and H_k have rank (or trace) 1 and $p_k - 1$, respectively. They satisfy $J_k H_k = 0 = H_k J_k$ and $J_k + H_k = I_{p_k}$. They are thus orthogonal projections that decompose R^{p_k} into two mutually orthogonal subspaces of dimensions 1 and $p_k - 1$, respectively.

Let $\mathcal{E}$ denote the set of all subsets of $\{1, 2, \ldots, k_0\}$, including the empty set $\emptyset$. The cardinality of $\mathcal{E}$ is 2^{k_0}. For every set $s \in \mathcal{E}$, define the $p_k \times p_k$ matrix

$$P_{s,k} = \begin{cases} J_k & \text{if } k \notin s \\ H_k & \text{if } k \in s \end{cases}. \tag{23}$$

Define the $p \times p$ Kronecker product matrix

$$P_s = \bigotimes_{k=1}^{k_0} P_{s,k_0-k+1}. \tag{24}$$

The foregoing discussion implies that:

- P_s is symmetric, idempotent for every $s \in \mathcal{E}$.
- If $s \neq \emptyset$, the rank (or trace) of P_s is $\prod_{k \in s}(p_k - 1)$. The rank (or trace) of $P_\emptyset$ is 1.
- If s_1 and s_2 are two different sets in $\mathcal{E}$, then $P_{s_1} P_{s_2} = 0 = P_{s_2} P_{s_1}$.
- $\sum_{s \in \mathcal{E}} P_s = I_p$.

Consequently, the $\{P_s : s \in \mathcal{E}\}$ are orthogonal projections that decompose R^p into 2^{k_0} mutually orthogonal subspaces.

The last bulleted point yields the identity

$$m = \sum_{s \in \mathcal{E}} P_s m, \tag{25}$$

whose right side expresses, in readily computable form, the ANOVA decomposition for the means of a complete k_0-way layout. Evidently, $P_\emptyset m$ is the overall mean term. If s is nonempty, $P_s m$ is the main effect or interaction term defined by the factors $k \in s$. The submodels considered in ANOVA are defined by constraining m to satisfy $P_s m = 0$ for every $s \in N$, where N is a specified subset of $\mathcal{E}$. The choice of N identifies the main effects or interaction terms that vanish in the submodel.

2.3.2 A class of tensor-product penalty matrices

An *annihilator* for factor k is a matrix A_k with p_k columns such that $A_k u_k = 0$. In other words, the rows of A_k are contrasts. The basic idea is that the contrasts in A_k should quantify departures from the conjectured dependence of the means on the levels of factor k. How exactly to do this for ordinal and nominal factors is addressed in Sect. 2.3.3. Here we describe how to build tensor-product penalty matrices $\{Q_s\}$ once the factor annihilators $\{A_k\}$ have been devised. The penalty matrices then generate candidate PLS estimators as discussed in Sects. 1 and 2.1

Let $\mathcal{D} = \mathcal{E} - \emptyset$, the nonempty subsets of $\{1, 2, \ldots k_0\}$. The cardinality of $\mathcal{D}$ is $d = 2^{k_0} - 1$. For every subset $s \in \mathcal{D}$ and for $1 \leq k \leq k_0$, define the $p_k \times p_k$ matrix

$$Q_{s,k} = \begin{cases} J_k & \text{if } k \notin s \\ A_k' A_k & \text{if } k \in s \end{cases} \tag{26}$$

and the $p \times p$ Kronecker product matrix

$$Q_s = \bigotimes_{k=1}^{k_0} Q_{s,k_0-k+1}. \tag{27}$$

Finally, rescale Q_s so that $\rho(Q_s) = 1$. The foregoing definitions entail that $P_s Q_s = Q_s P_s = Q_s$ for every $s \in \mathcal{D}$.

If s_1, s_2 are different subsets of $\mathcal{D}$, then there exists k such that $k \in s_1$ and $k \notin s_2$. Then, $Q_{s_1,k} = A_k' A_k$ by (26) while $P_{s_2,k} = J_k$ by (24). By the annihilator property of A_k, it follows that $Q_{s_1,k} P_{s_2,k} = 0$. Hence

$$P_{s_2} Q_{s_1} = Q_{s_1} P_{s_2} = \bigotimes_{k=1}^{k_0} \left[Q_{s_1,k_0-k+1} P_{s_2,k_0-k+1} \right] = 0. \tag{28}$$

It follows from this and the ANOVA decomposition (25) that

$$m' Q_s m = (P_s m)' Q_s (P_s m) \tag{29}$$

for every $s \in \mathcal{D}$. By Eq. (7),

$$m' Q(t) m = \epsilon I_p + c \sum_{s \in \mathcal{D}} t_s (P_s m)' Q_s (P_s m). \tag{30}$$

Thus, the penalty matrix Q_s in the definition of $Q(t)$ acts solely on the ANOVA component $P_s m$. In this manner, the candidate PLS estimator $\hat{\eta}(t)$ or $\hat{m}_D(t)$ penalizes departures, in the interactions or main effects of m, from attributes determined by the choice of the factor annihilators $\{A_k\}$ that define the $\{Q_s\}$.

2.3.3 Constructing factor annihilators

It remains to devise useful factor annihilators. Let $v_k = \left(v_{k1}, v_{k2}, \ldots, v_{kp_k}\right)$ be the possible levels of factor k, ordered in increasing order if that factor is ordinal. Consider the hypothetical complete k_0-way layout of means associated with the observed incomplete layout. The dependence of m_i on these factor levels is expressed by

$$m_i = f\left(v_{1i_1}, v_{2i_2}, \ldots, v_{k_0 i_{k_0}}\right), \quad i \in \mathcal{I}, \tag{31}$$

where f is an unknown real-valued function whose domain is the set of all factor level combinations.

The complete layout of means forms the array

$$M = \left\{f\left(v_{1i_1}, v_{2i_2}, \ldots, v_{k_0 i_{k_0}}\right) : 1 \le i_1 \le p_1, 1 \le i_2 \le p_2, \ldots, 1 \le i_{k_0} \le p_{k_0}\right\}. \tag{32}$$

Note that the vector m described in Sect. 1 is a systematic vectorization of the array M. Fix k and fix the $\{i_j : j \ne k\}$. Extract from M, as a $p_k \times 1$ vector, the elements that are indexed by $1 \le i_k \le p_k$ and $\{i_j : j \ne k\}$. This vector is

$$m\left(k|\{i_j : j \ne k\}\right) = \left\{f(v_{1i_1}, \ldots, v_{ki_k}, \ldots v_{k_0 i_{k_0}}) : 1 \le i_k \le p_k\right\}. \tag{33}$$

The idea that guides construction of an annihilator A_k for factor k is this: the property that $\left|A_k m\left(k|\{i_j : j \ne k\}\right)\right|$ is relatively small for every choice of the values $\{i_j : j \ne k\}$ should express plausible prior conjecture about the dependence of m on the levels of factor k. It will be necessary to distinguish between nominal factors and ordinal factors.

Factor k is nominal. The levels of a nominal factor are labels that can be permuted freely without loss of information. The corresponding candidate PLS estimators should be therefore be invariant under permutations of nominal levels. This consideration prompts setting $A_k = H_k$ for every k, the latter being defined in (22). This choice of A_k will be called the *flat annihilator* for factor k, a term suggested by the constant spectrum of the reduced singular value decomposition of H_k. With $A_k = H_k$, it follows from Eqs. (23) and (26) that $Q_{s,k} = P_{s,k}$.

Consider the special case where every factor in the layout is nominal. Using the flat annihilator for each factor entails $Q_s = P_s$ for every subset $s \in \mathcal{D}$. Note that $\rho(P_s) = 1$, so no rescaling is needed. In this case, the candidate PLS estimator $\hat{\eta}(t)$ essentially interpolates among the $2^{k_0} - 1$ ANOVA submodel fits to the incomplete layout. The ANOVA submodel fits themselves are limit points of this set of candidate PLS estimators as c tends to infinity (cf. Sect. 1). Consequently, adaptation over the candidate PLS estimators yields an estimator of η whose estimated risk typically undercuts that of the ANOVA submodel fit with smallest estimated risk. The latter is also the ANOVA submodel fit with smallest Mallows C_p.

Factor k is ordinal. Suppose first that the ordered levels of factor k, arranged as the column vector $v_k = (v_{k1}, v_{k2}, \ldots, v_{kp_k})'$, are equally spaced. To have the candidate PLS estimator $\hat{\eta}(t)$ favor a fit that is locally polynomial of degree $h_0 - 1$ in the levels of factor k, we take A_k equal to the h_0-th difference operator of column dimension p_k.

Explicitly, consider the $(g-1)\times g$ matrix $\Delta(g)=\{\delta_{uw}\}$ in which $\delta_{u,u}=1$, $\delta_{u,u+1}=-1$ for every u and all other entries are zero. Define recursively

$$D(1,\, p_k)=\Delta(p_k),$$
$$D(h,\, p_k)=\Delta(p_k-h+1)D(h-1,\, p_k)\ \text{ for } 2\le h\le p_k-1. \tag{34}$$

Evidently the $(p_k-h_0)\times p_k$ matrix $A_k=D(h_0,\, p_k)$ accomplishes h_0th differencing and annihilates powers of v_k up to power h_0-1 in the sense that

$$A_k v_k^h=0\ \text{ for } 0\le h\le h_0-1. \tag{35}$$

The notation v_k^h denotes the vector $\left(v_{k1}^h,\, v_{k2}^h,\, \ldots\, ,\, v_{kp_k}^h\right)'$. Moreover, in row i of A_k, the elements not in columns $i, i+1, \ldots\, , i+h_0$ are zero.

More generally, if the means are expected to behave locally like a polynomial of degree h_0-1 in factor k but the factor levels in v_k are not necessarily equally spaced, we define A_k as follows. The h_0th order *local polynomial annihilator* A_k is a $(p_k-h_0)\times p_k$ matrix characterized through three conditions: First, for every possible i, all elements in the ith row of A_k that are not in columns $i, i+1, \ldots\, , i+h_0$ are zero. Second, A_k satisfies the orthogonality constraints (35). Third, each row vector in A_k has unit length. These requirements are met by setting the non-zero elements in the ith row of A_k equal to the basis vector of degree h_0 in the orthonormal polynomial basis that is defined on the h_0+1 design points $(v_{ki}, \ldots\, , v_{k,i+h_0})$. The S-Plus function `poly` accomplishes this computation.

When the components of v_k are equally spaced, this construction of A_k reduces to a multiple of the h_0th difference annihilator described in the preceding paragraph. In the general construction, the powers of the components of v_k can be replaced by other linearly independent functions of the factor levels to express other prior notions about the dependence of the means on factor k.

3 Asymptotic theory

The main results of this section concern the asymptotic risk and loss of the adaptive estimator $\hat{\eta}(\hat{t})$ that was defined in Eq. (20). For mathematical clarity, we first isolate the properties of the matrix $S(t)$ in Eq. (16) on which the results depend. We ask the reader to recognize that almost every quantity in this paper depends on q. To avoid burdensome notation, we usually omit the subscript q. Section 5 gives all theorem proofs.

Theorem 2 *Let $S(t)$ be the matrix in expression (16) for candidate estimator $\hat{\eta}(t)$. Then, for the matrix norm (16),*

$$\sup_q \sup_{t\in[0,1]^d} \rho[S(t)]<\infty. \tag{36}$$

Moreover, $S(t)$ is continuous on $[0,1]^d$ and is differentiable on the interior of $[0,1]^d$ with partial derivatives $\{\nabla_s S(t)=\partial S(t)/\partial t_s : s\in\mathcal{D}\}$ such that

$$\sup_{s,q} \sup_{t\in[0,1]^d} \rho[\nabla_s S(t)]<\infty. \tag{37}$$

The next two theorems study the asymptotic convergences of the loss, risk, and estimated risk of the candidate estimator $\hat{\eta}(t)$ and of the adaptive estimator $\hat{\eta}(\hat{t})$. The notation and definitions are those of Sect. 2.2. All risks and asymptotics under strong Gauss-Markov model (2). The results proved hold without any assumptions on the unknown means m.

Theorem 3 *Suppose that $S(t)$ has the properties stated in Theorem 2. Assume that the strong Gauss-Markov model (2) holds and that, for every finite $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0. \tag{38}$$

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 $c > 0, a > 0$, and $\sigma^2 > 0$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} E\left[\sup_{t\in[0,1]^d} |W(t) - r(t)|\right] = 0. \tag{39}$$

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 $t \in [0, 1]^d$ makes estimated risk a trustworthy surrogate for its true loss or risk. In the proof, the boundedness of $\rho[S(t)]$ ensures pointwise convergence of $W(t)$ to $r(t)$. Strengthening this pointwise convergence to the uniform convergence (39) draws on the boundedness of the $\{\rho[\nabla_s S(t)]\}$.

Theorem 4 *Suppose that the conditions for Theorem 3 hold. Then, for every finite $c > 0, a > 0$, and $\sigma^2 > 0$,*

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} |R(\hat{\eta}(\hat{t}), \eta, \sigma^2) - r(\tilde{t})| = 0, \quad \text{with } \tilde{t} = \operatorname*{argmin}_{t\in[0,1]^d} r(t). \tag{40}$$

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 $\hat{\eta}(\hat{t})$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{r}(\hat{t}) - V| = 0. \tag{41}$$

By Eq. (40), the risk of the adaptive PLS 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 candidate estimators $\{\hat{\eta}(t) : t \in [0, 1]^d\}$. By Eq. (41), the plug-in risk estimator $\hat{r}(\hat{t})$ converges to the actual risk or loss of $\hat{\eta}(\hat{t})$. Through $\hat{r}(\hat{t})$, we learn approximately how much adaptation over the class of candidate PLS estimators reduces risk for the data at hand.

4 Application to the examples

This section details the construction of adaptive PLS estimators for the coal ash data and the starch data described in Sect. 1. The statistical logic in these case studies goes beyond the preceding mathematical developments. 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, neither the coal ash data nor the starch data is certifiably random. In the absence of randomness, the variance estimate quantifies the level of detail that is deemed to be noise (i.e., not predictable) and thereby determines the shrinkage transformations that yield the adaptive PLS estimators of m_D, η, and m.

A probability model describes, at best, hypothetical data that is similar in selected relative frequencies to what was observed; and several very different types of model may serve this purpose. Procedures that work well in the world of an insightful probability model may also prove satisfactory in the world of data and computational experiments. Ultimately, this is a matter open to empirical testing. Advances in both statistical computing and empirical process theory have induced a dynamic that encourages experimentation with algorithms for data analysis; pushes into the background notions of statistics as normative mathematical philosophy; returns to prominence the fundamental distinctions among data, probability model, and procedure; and reformulates statistics as an experimentally supported theory and technology for data analysis.

4.1 Coal ash data

These data form an incomplete two-way layout with one observation at each observed grid point. The factor levels are the grid coordinates, which encode the geographical coordinates at which each assay sample was taken. Both factors are ordinal, with $p_1 = 23$ and $p_2 = 16$ equally spaced levels, respectively. We tentatively conjecture that mean coal ash percentage varies slowly as we move from one grid-point to the next. On this supposition, we set the annihilator A_k to be the first-difference operator of dimension $(p_k - 1) \times p_k$ for $k = 1, 2$. The tensor-product penalty terms generated by these annihilators penalize, according to their weights, departures in candidate PLS means from locally constant dependence on each of the two geographical factors (see Sect. 2.3).

Conjectured local constancy in the means motivates the following variance estimator. Form all first differences between adjacent observations in rows and between adjacent observations in columns. Square these differences and then average them to obtain the first-difference variance estimator $\hat{\sigma}^2$. This is an obvious bivariate extension of the first-difference variance estimator for a one-way layout (cf. Rice 1984). It is not difficult to provide conditions on $\hat{m}_D$ that express local constancy and ensure that this variance estimator has L_1-consistency in the sense (38). For the coal ash data, $\hat{\sigma}^2 = 1.038$.

To completely define the candidate PLS (or Bayes) estimators in Eq. (8), (9) and Theorem 1, we set $c = 10^4$ and $\epsilon = 10^{-7}$. For the coal ash data, $p = 368$ while $q = n = 208$. The matrix $C = I_q$. The estimated risk function $\hat{r}(t)$ in (19) is minimized numerically at $\hat{t} = (.000349, .000217, 1)$. Calculations were done in S-Plus 6.2.1 for Linux with the aid of the function `nlmin`. Cell (2,1) in Fig. 1 displays the adaptive estimate $\hat{m}_D(\hat{t}) = \hat{\eta}(\hat{t})$ of the means at observed grid points. Cell (2,2) presents the regression function estimate $\hat{m}(\hat{t})$ that extrapolates the fit to all grid points. The residual plots draw attention to one very large outlier at a grid-point next to that of the interior missing observation. In the three mesh plots, linear interpolation is merely a graphical device to guide the eye.

The estimated risk of the adaptive PLS estimate $\hat{\eta}(\hat{t})$ under the strong Gauss-Markov model (2) is $\hat{r}(\hat{t}) = 0.117$. The estimated risk of the full model least squares estimate, which coincides in this example with the raw data, is $\hat{\sigma}^2 = 1.038$, almost nine times as large. It is to be emphasized that neither the model nor the risk estimator make any assumptions about the unknown means. The adaptive PLS estimate constructed here is nearly additive in the two factors and is considerably smoother than the raw data. Through smoothing and shrinkage to a submodel, this adaptive PLS estimate has reduced estimated risk very substantially, thereby validating empirically the use of first-difference annihilators to construct the tensor-product penalty matrices. Unlike kriging, which relies on random effects modeling, the adaptive PLS methodology in this example uses a general fixed effects model—the strong Gauss-Markov model (2).

4.2 Starch data

These data form a highly incomplete two-way layout with observations at 81 sparsely distributed grid points plotted in cell (1,1) of Fig. 2. The first factor, starch type, is nominal with $p_1 = 7$ levels. The second factor, thickness of the starch film, is ordinal with $p_2 = 69$ distinct levels that are not equally spaced. Having plotted the data, we conjecture that breaking strength of a starch film varies locally linearly with the thickness of the film. On this supposition, we set the annihilator A_2 to be the generalized second-difference operator of dimension $(p_2 - 1) \times p_2$ for $k = 1, 2$. The annihilator A_1 is set equal the projection H_1 of dimension $p_1 \times p_1$ that is appropriate for a nominal factor. The tensor-product penalty terms generated by these annihilators penalize, according to their weights, differences between starch types and departures from locally linear dependence of candidate PLS means on starch film thickness.

Conjectured local linearity in how the means depend on starch-film thickness motivates the following variance estimator. Fit a separate least squares line to the data for each starch. Pool the residual sums of squares from these seven fits to obtain $\hat{\sigma}^2$. It is possible to provide conditions on $\hat{m}_D$ that express local linearity mathematically and ensure that this variance estimator has L_1-consistency in the sense (38). For the starch data, $\hat{\sigma}^2 = 13,976$.

To completely define the candidate PLS (or Bayes) estimators, we set $c = 10^5$ and $\epsilon = 10^{-7}$. For the starch data, $p = 493$ while $q = 81$ and $n = 94$. The matrix C is not the identity because of a few replicated observations at a few grip points. The estimated risk function $\hat{r}(t)$ in Eq. (19) is minimized numerically at $\hat{t} = (1, .7707 \times 10^{-6}, 1)$. Cell (1,2) in Fig. 2 displays the regression function estimate $\hat{m}(\hat{t})$ that extrapolates the adaptive fit to all grid points. The linear interpolation is merely a graphical device to guide the eye. Indeed, starch type is a nominal factor, not ordinal. The plots in Fig. 3 compare $\hat{m}$ with the raw data. The fits appear satisfactory except for starch 5, where the data points for larger film thicknesses exhibit a dual personality. An unrecorded factor may have influenced these measurements. The plotted values of $\hat{m}$ at observed data points define $\hat{m}_D$ in Fig. 3.

The estimated risk of the adaptive PLS estimate $\hat{\eta}(\hat{t})$ under the strong Gauss-Markov model (2) is $\hat{r}(\hat{t}) = 4,116$. The estimated risk of the full model least squares estimate, which coincides in this example with the raw data, is

$\hat{\sigma}^2 = 13,976$, more than three times as large. In this example too, neither the model nor the risk estimator make any assumptions about the unknown means. The adaptive PLS estimate just constructed reduces estimated risk by strongly penalizing the main effects of starch type and the interactions between starch type and starch film thickness. Its success in trading off estimated bias against estimated risk empirically validates the efficacy of the chosen annihilators.

5 Proofs

This section proves Theorems 1–4 and provides algebraic insight into the various expressions for the Bayes or PLS estimators.

Lemma 1 *Let C be the data-incidence matrix and D be the means incidence matrix defined in Sect. 1. Let K be any $p \times p$ positive definite matrix. Then $DK^{-1}D'$ is also positive definite and*

$$D\left(D'C'CD + K\right)^{-1} D' = \left[C'C + \left(DK^{-1}D'\right)^{-1}\right]^{-1}. \tag{42}$$

Proof Because $DD' = I_q$, it follows that $D'a = 0$ for $q \times 1$ vector a if and only if $a = 0$. Consequently, $DK^{-1}D'$ is positive definite.

Let $R = C'C$, a non-singular diagonal matrix, and let $E = R^{1/2}D$. Then,

$$\left(K + E'E\right)^{-1} = K^{-1} - K^{-1}E'\left(I + EK^{-1}E'\right)^{-1} EK^{-1} \tag{43}$$

by a standard identity (cf. Sen and Srivastava 1990, p. 275). Note that for any nonsingular matrix W,

$$\begin{aligned} I &= (I+W)^{-1} + W(I+W)^{-1} = (I+W)^{-1} + (I+W)^{-1}W \\ &= (I+W)^{-1} + (I+W^{-1})^{-1}. \end{aligned} \tag{44}$$

Let $W = EK^{-1}E'$, a symmetric matrix. Equations (43) and (44) imply

$$\begin{aligned} E(K+E'E)^{-1}E' &= W - W(I+W)^{-1}W = W - W\left[I - (I+W)^{-1}\right] \\ &= W(I+W)^{-1} = (I+W^{-1})^{-1}. \end{aligned} \tag{45}$$

Substituting the definitions of W and E into Eq. (45) yields the desired identity Eq. (42).

Proof of Theorem 1 The prior distribution on m entails that the posterior distribution is $m|y \sim N\left(VD'C', \sigma^2 V\right)$ with

$$V = \left[D'C'CD + Q(t)\right]^{-1}. \tag{46}$$

This yields the Bayes estimator (11) for m and the first expression in (12) for the Bayes estimator of $m_D = Dm$ and the first expression in (13) for the Bayes estimator of $\eta = CDm$.

On the other hand, the prior distribution on m implies that $m_D \sim N(0, \sigma^2 DQ^{-1}(t)D')$. By Lemma 1, the covariance matrix here is nonsingular. Hence the posterior distribution is $m_D|y \sim N\left(VC'y, \sigma^2 V\right)$ with

$$V = \left[C'C + (DQ^{-1}(t)D')^{-1}\right]^{-1}. \tag{47}$$

This implies the second expression in (12) for the Bayes estimator of m_D and the second expression in (13) for the Bayes estimator of $\eta = Cm_D$. The equivalence of the dual expressions for the Bayes (or PLS) estimators of m_D and of η, just derived by probability reasoning, is also established through the algebraic identity (42).

Recall definition (6): $\rho(A) = \sup_{x\neq 0}[|Ax|/|x|]$. The next lemma summarizes properties of ρ that are used in proving Theorems 2 and 3. The notation $\{\lambda_i(S)\}$ denotes the eigenvalues of the symmetric matrix S and $\lambda_{\max}(S)$ is the largest of these eigenvalues. This lemma is used without further comment in proving Theorems 2 and 3.

Lemma 2

a. ρ *is a matrix norm.*
b. If matrices A *and* B *have compatible dimensions,* $\rho(AB) \leq \rho(A)\rho(B)$.
c. If a *is a vector,* $\rho(a) = |a|$.
d. If the vectors a, b *and the matrix* A *have compatible dimensions,* $|a'Ab| \leq |a||b|\rho(A)$.
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* $q \times q$ *symmetric, then* $q^{-1}|\operatorname{tr}(S)| \leq \rho(S)$.

Proof Parts a and c are immediate from the definition of ρ. Part b holds because

$$\begin{aligned}\rho(AB) &= \sup_{x\neq 0} \frac{|ABx|}{|x|} = \sup_{x:\, Bx\neq 0} \frac{|ABx|}{|x|} \\ &\leq \sup_{x:\, Bx\neq 0} \frac{|ABx|}{|Bx|} \cdot \sup_{x:\, Bx\neq 0} \frac{|Bx|}{|x|} \leq \rho(A)\rho(B).\end{aligned} \tag{48}$$

Part d follows from b and c. Part e holds because

$$\begin{aligned}\rho^2(A) &= \sup_{x\neq 0} \frac{|Ax|^2}{|x|^2} \\ &= \sup_{x\neq 0} \frac{x'A'Ax}{|x|^2} = \lambda_{\max}(A'A) = \lambda_{\max}(AA') = \rho^2(A').\end{aligned} \tag{49}$$

If S is symmetric, then by part e, $\rho^2(S) = \lambda_{\max}(S^2) = [\max_i |\lambda_i(S)|]^2$. Moreover, $\rho(S^2) = \lambda_{\max}^{1/2}(S^4) = \lambda_{\max}(S^2) = \rho^2(S)$. These two calculations establish part f. Part g follows from $q^{-1}|\operatorname{tr}(S)| \leq q^{-1}\sum_{i=1}^{q} |\lambda_i(S)| \leq \max_i |\lambda_i(S)|$ and part f.

Proof of theorem 2 In view of (16) and the positive definiteness of the symmetric matrix $V(t)$,

$$\rho[S(t)] = \rho[V^{-1}(t)] = \lambda_{\max}[V^{-1}(t)] < 1 \tag{50}$$

for every q and t. This establishes (36).

On the other hand,

$$\begin{aligned}
\nabla_s S(t) &= -S(t)[\nabla_s V(t)]S(t) \\
\nabla_s V(t) &= R^{-1/2}[\nabla_s (DQ^{-1}(t)D')^{-1}]R^{-1/2} \\
\nabla_s \left(DQ^{-1}(t)D'\right)^{-1} &= -\left(DQ^{-1}(t)D'\right)^{-1} D[\nabla_s Q^{-1}(t)]D' \left(DQ^{-1}(t)D'\right)^{-1} \\
\nabla_s Q^{-1}(t) &= -cQ^{-1}(t)Q_s Q^{-1}(t). \qquad (51)
\end{aligned}$$

Evidently, $\rho(R^{-1/2}) \leq 1$, $\rho(D) = \lambda_{\max}^{1/2}(DD') = 1$, $\rho(Q_s) = 1$ by the normalization of the penalty matrices, and

$$\rho[Q^{-1}(t)] = \lambda_{\max}\left[Q^{-1}(t)\right] = 1/\lambda_{\min}[Q(t)] \leq \epsilon^{-1}. \qquad (52)$$

Moreover,

$$\begin{aligned}
\rho\left[\left(DQ^{-1}(t)D'\right)^{-1}\right] &= \lambda_{\max}\left[\left(DQ^{-1}(t)D'\right)^{-1}\right] \\
&= 1/\lambda_{\min}\left[DQ^{-1}(t)D'\right] \leq \epsilon + cd \qquad (53)
\end{aligned}$$

because $DD' = I_q$ and

$$\begin{aligned}
\lambda_{\min}\left[DQ^{-1}(t)D'\right] = \inf_x \frac{x'DQ^{-1}(t)D'x}{x'DD'x} &\geq \lambda_{\min}\left[Q^{-1}(t)\right] = 1/\lambda_{\max}[Q(t)] \\
&\geq (\epsilon + cd)^{-1}. \qquad (54)
\end{aligned}$$

It follows from Eqs. (50)–(53), using part b of Lemma 2, that $\rho[\nabla_s S(t)] < c(\epsilon + cd)^2/\epsilon^2$ for every q, s and t. This establishes Eq. (37).

Proof of theorem 3 The strategy is to show that $W(t) - r(t)$ converges in probability to zero for every $t \in [0, 1]$, then show that $\sup_{t\in[0,1]} |W(t) - r(t)|$ converges in probability to zero, and finally use uniform integrability to establish (39). Repeatedly used are the constraint $q^{-1}|\eta|^2 \leq a$ 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 Lemma 2.

$$\begin{aligned}
&\sup_q \sup_{t\in[0,1]^d} \rho[T(t)] < \infty, \quad \sup_q \sup_{t\in[0,1]^d} \rho[\bar{T}(t)] < \infty \\
&\sup_{s,q} \sup_{t\in[0,1]^d} \rho[\nabla_s T(t)] < \infty, \quad \sup_{s,q} \sup_{t\in[0,1]} \rho[\nabla_s \bar{T}(t)] < \infty. \qquad (55)
\end{aligned}$$

The bounds on derivatives use the following identity: $\nabla_s T(t) = \nabla_s S(t) \cdot S(t) + S(t) \cdot \nabla_s S(t)$.

We first prove the case $W(t) = \hat{r}(t)$ of (39). Define $\check{r}(t)$ by replacing $\hat{\sigma}^2$ with σ^2 in the definition (19) 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

$$\begin{aligned}
|\hat{r}(t) - \check{r}(t)| &\leq |\hat{\sigma}^2 - \sigma^2| q^{-1}[|\operatorname{tr}(T)| + |\operatorname{tr}(\bar{T})|] \\
&\leq |\hat{\sigma}^2 - \sigma^2|[\rho(T) + \rho(\bar{T})] \qquad (56)
\end{aligned}$$

and the L_1 consistency (38) of $\hat{\sigma}^2$, we may replace $\hat{r}(t)$ with $\check{r}(t)$ in the subsequent argument.

Pointwise consistency. Let

$$Y_q(t) = \check{r}(t) - r(t), \quad B(t) = U\bar{T}(t)U' \tag{57}$$

and note that $\mathrm{tr}(B) = \mathrm{tr}(\bar{T})$ and $\rho(B) = \rho(\bar{T})$. Recall that $z = U'y$, $w = U'e$, $\xi = U'\eta$, and $y = \eta + e$. From Eqs. (18), (19) and the foregoing definition of $\check{r}(t)$,

$$\begin{aligned} Y_q(t) &= q^{-1}\,\mathrm{tr}\left[\bar{T}(zz' - \sigma^2 I_q - \xi\xi')\right] \\ &= q^{-1}\left[2\xi'\bar{T}w + \left\{w'\bar{T}w - \sigma^2\,\mathrm{tr}(\bar{T})\right\}\right] \\ &= q^{-1}\left[2\eta'Be + \left\{e'Be - \sigma^2\,\mathrm{tr}(B)\right\}\right]. \end{aligned} \tag{58}$$

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

$$\mathrm{Var}\left(q^{-1}\eta'Be\right) = q^{-2}\sigma^2\eta'B^2\eta \le q^{-2}\sigma^2|\eta|^2\rho(B^2) \le q^{-1}\sigma^2 a\rho^2(\bar{T}). \tag{59}$$

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 $\mathrm{E}(e_i^4) = (3+\gamma)\sigma^4$ and $\mathrm{Var}(e_i^2) = (2+\gamma)\sigma^4$. Then, using $|B| = |\bar{T}|$,

$$\begin{aligned} \mathrm{Var}(q^{-1}e'Be) &= q^{-2}\sigma^4\left(2|B|^2 + \gamma\sum_i b_{ii}^2\right) \le q^{-2}\sigma^4(2+\gamma)|B|^2 \\ &= q^{-2}\sigma^4(2+\gamma)|\bar{T}|^2 \le q^{-1}\sigma^4(2+\gamma)\rho^2(\bar{T}). \end{aligned} \tag{60}$$

Thus, for every $t \in [0,1]^d$,

$$\operatorname*{plim}_{q\to\infty} Y_q(t) = 0. \tag{61}$$

Uniform consistency. For any $u, t \in [0,1]^d$,

$$Y_q(u) - Y_q(t) = q^{-1}\sum_{s\in\mathcal{D}}(u_s - t_s)\left[2\eta'\nabla_s Be + \{e'\nabla_s Be - \sigma^2\,\mathrm{tr}(\nabla_s B)\}\right], \tag{62}$$

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

$$\sup_{|u-t|\le\delta}|Y_q(u) - Y_q(t)| \le \delta q^{-1}\sum_{s\in\mathcal{D}}\left[2|\eta'\nabla_s Be| + |e'\nabla_s Be| + \sigma^2|\,\mathrm{tr}(\nabla_s B)|\right]. \tag{63}$$

Moreover, using Lemma 2,

$$\begin{aligned} &q^{-1}|\,\mathrm{tr}(\nabla_s B)| = q^{-1}|\,\mathrm{tr}(\nabla_s\bar{T})| \le \rho(\nabla_s\bar{T}) \\ &q^{-1}\mathrm{E}|\eta'\nabla_s Be| \le q^{-1}|\eta|\,\mathrm{E}|e|\,\rho(\nabla_s B) \le a^{1/2}\sigma\rho(\nabla_s\bar{T}) \\ &q^{-1}\mathrm{E}|e'\nabla_s Be| \le q^{-1}\mathrm{E}|e|^2\,\rho(\nabla_s B) = \sigma^2\rho(\nabla_s\bar{T}). \end{aligned} \tag{64}$$

Applying Markov's inequality to Eq. (63) establishes existence of a finite constant C, not depending on q, such that

$$\mathrm{P}\left[\sup_{|u-t|\le\delta}|Y_q(u) - Y_q(t)| \ge \epsilon\right] \le C\delta. \tag{65}$$

Hence,

$$\lim_{\delta\to 0}\limsup_{q\to\infty} \mathrm{P}\left[\sup_{|u-t|\le\delta} |Y_q(u) - Y_q(t)| \ge \epsilon\right] = 0. \tag{66}$$

It follows from (61), (66) and Wichura (1971) that Y_q converges weakly in $C[0,1]^d$ to the zero element of $C[0,1]^d$. Hence,

$$\operatorname*{plim}_{q\to\infty} \sup_{t\in[0,1]^d} |Y_q(t)| = 0. \tag{67}$$

L_1 uniform consistency. Let $V_q = \sup_{t\in[0,1]^d} |Y_q(t)|$ and $\rho_{\max} = \sup_q \sup_{t\in[0,1]^d} \rho(\bar{T}(t))$. Using (58) and $|\eta| = |\xi|$,

$$\begin{aligned} V_q &\le q^{-1} \sup_{t\in[0,1]^d} \left[2|\xi'\bar{T}(t)w| + |w'\bar{T}(t)w| + \sigma^2|\operatorname{tr}(\bar{T}(t))|\right] \\ &\le \left[2a^{1/2}\{q^{-1/2}|w|\} + q^{-1}|w|^2 + \sigma^2\right]\rho_{\max} = W_q \text{ (say)}. \end{aligned} \tag{68}$$

Let $W = \left[2a^{1/2}\sigma + 2\sigma^2\right]\rho_{\max}$. Because $|w|^2 = e'UU'e$, a calculation akin to Eq. (60) shows that

$$\operatorname{Var}\left(q^{-1}|w|^2\right) \le q^{-1}\sigma^4(2+\gamma). \tag{69}$$

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

$$\begin{aligned} &\lim_{q\to\infty} \mathrm{E}|q^{-1}|w|^2 - \sigma^2| = 0 \\ &\lim_{q\to\infty} \mathrm{E}|q^{-1/2}|w| - \sigma| \le \lim_{q\to\infty} \mathrm{E}^{1/2}[q^{-1/2}|w| - \sigma]^2 = 0. \end{aligned} \tag{70}$$

Consequently, in view of Eq. (55), $\lim_{q\to\infty} \mathrm{E}|W_q - W| = 0$. This convergence, inequality (68), and a uniform integrability argument (cf. Neveu 1965, p. 52) imply that (67) can be strengthened to

$$\lim_{q\to\infty} \mathrm{E}\left[\sup_{t\in[0,1]^d} |Y_q(t)|\right] = 0. \tag{71}$$

This completes the proof of (39) when $W(t) = \hat{r}(t)$.

The argument for the case $W(t) = L(\hat{\eta}(t), \eta)$ of Eq. (17) is similar. The loss and risk of $\hat{\eta}(t)$ are given in Eqs. (17) and (18). Note that

$$L(\hat{\eta}(t), \eta) = q^{-1}\left[2\xi'(S - I_q)Sw + w'Tw + \xi'\bar{T}\xi\right]. \tag{72}$$

Let

$$Y_q(t) = L(\hat{\eta}(t), \eta) - r(t) = q^{-1}\left[2\xi'Vw + \left\{w'Tw - \sigma^2\operatorname{tr}(T)\right\}\right], \tag{73}$$

where $V(t) = T(t) - S(t)$. Because the right-hand side of (73) has the same structure as the middle expression in Eq. (58), an argument parallel to the one that follows Eq. (58) completes the proof. Indeed, Eq. (55) and the theorem assumptions on $S(t)$ imply that

$$\sup_q \sup_{t\in[0,1]^d} \rho[V(t)] < \infty, \quad \sup_{s,q} \sup_{t\in[0,1]^d} \rho[\nabla_s V(t)] < \infty. \tag{74}$$

Proof of theorem 4 We show that Eq. (39) implies

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|Z - r(\tilde{t})| = 0, \tag{75}$$

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 Eqs. (40) and (41) are immediate consequences of Eq. (75).

First, (39) with $W(t) = \hat{r}(t)$ entails

$$\begin{aligned} \lim_{q\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|\hat{r}(\hat{t}) - r(\tilde{t})| &= 0 \\ \lim_{p\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|\hat{r}(\hat{t}) - r(\hat{t})| &= 0. \end{aligned} \tag{76}$$

Hence, Eq. (75) holds for $Z = \hat{r}(\hat{t})$ and

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|r(\hat{t}) - r(\tilde{t})| = 0. \tag{77}$$

Second, Eq. (39) with $W(t) = L(\hat{\eta}(t), \eta)$ gives

$$\begin{aligned} \lim_{q\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|L(\hat{\eta}(\hat{t}), \eta) - r(\hat{t})| &= 0 \\ \lim_{q\to\infty} \sup_{q^{-1}|\eta|^2\le a} \mathrm{E}|L(\hat{\eta}(\tilde{t}), \eta) - r(\tilde{t})| &= 0. \end{aligned} \tag{78}$$

These limits together with Eq. (77) establish the remaining two cases of Eq. (75).

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

References

Beran, R. (2002). Improving penalized least squares through adaptive selection of penalty and shrinkage. *Annals of the Institute of Statistical Mathematics*, *54*, 900–917.

Beran, R. (2005). ASP fits to multi-way layouts. *Annals of the Institute of Statistical Mathematics*, *57*, 201–220.

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

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

Freeman, H. A. (1942). *Industrial statistics*. New York: Wiley.

Heckman, N. E., Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics*, *28*, 241–258.

Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Kimeldorf, G. S., Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, *41*, 495–502.

Kneip, A. (1994). Ordered linear smoothers. *Annals of Statistics*, *22*, 835–866.

Lin, Yi (2000). Tensor product space ANOVA Fits. *Annals of Statistics*, *28*, 734–755.

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

Neveu, J. (1965). *Mathematical foundations of the calculus of probability.* San Francisco: Holden-Day.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, *12*, 1215–1230.
Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
Sen, A., Srivastava, M. (1990). *Regression analysis*. Berlin Heidelberg New York: Springer.
Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 197–206. Berkeley: University of California Press.
Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In: F. N. David (Ed.), *Festschrift for Jerzy Neyman*, (pp. 351–364). New York: Wiley.
Wahba, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.
Wahba, G., Wang Y., Gu, C., Klein, R., Klein, B. (1995) Smoothing spline ANOVA for exponential families with application to the Wisconsin epidemiological study of diabetic retinopathy. *Annals of Statistics*, *23*, 1868–1895.
Wichura, M. J. (1971). A note on the weak convergence of stochastic processes. *Annals of Mathematical Statistics*, *42*, 1769–1772.
Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, *62*, 413–428.