Ann. Inst. Statist. Math.
Vol. 57, No. 2, 201–220 (2005)

ASP FITS TO MULTI-WAY LAYOUTS*

RUDOLF BERAN

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

(Received November 25, 2002; revised March 29, 2004)

Abstract. The balanced complete multi-way layout with ordinal or nominal factors is a fundamental data-type that arises in medical imaging, agricultural field trials, DNA microassays, and other settings where analysis of variance (ANOVA) is an established tool. ASP algorithms weigh competing biased fits in order to reduce risk through variance-bias tradeoff. The acronym ASP stands for **A**daptive **S**hrinkage on **P**enalty bases. Motivating ASP is a penalized least squares criterion that associates a separate quadratic penalty term with each main effect and each interaction in the general ANOVA decomposition of means. The penalty terms express plausible conjecture about the mean function, respecting the difference between ordinal and nominal factors. Multiparametric asymptotics under a probability model and experiments on data elucidate how ASP dominates least squares, sometimes very substantially. ASP estimators for nominal factors recover Stein's superior shrinkage estimators for one- and two-way layouts. ASP estimators for ordinal factors bring out the merits of smoothed fits to multi-way layouts, a topic broached algorithmically in work by Tukey.

Key words and phrases: Nominal factors, ordinal factors, estimated risk, penalized least squares, annihilator matrix, balanced complete layout, multiparametric asymptotics.

1. Introduction

A fundamental data type arising in the sciences, engineering, and informatics is the balanced complete multi-way layout. Instances include data collected in DNA microassays, in medical imaging, in agricultural field trials, and in other settings where ANOVA is an established tool. The factors in a multi-way layout may be ordinal or nominal. The levels of an ordinal factor are real-values that indicate at least order and possibly more. The levels of a nominal factor are pure labels that convey no ordering information. This paper describes an adaptive approach to fitting balanced complete k_0-way layouts that stems from a penalized least squares (PLS) criterion with $2^{k_0} - 1$ penalty terms. A separate penalty term is associated with each main effect and interaction in the usual ANOVA decomposition of means.

The acronym ASP, which stands for **A**daptive **S**hrinkage on **P**enalty bases, summarizes principal steps in our methodology. The broad approach is: (a) use prior conjecture about the unknown means in the standard Gaussian one-way layout to devise the penalty terms in the PLS criterion; (b) estimate the risk of each candidate PLS estimator *without* making assumptions on the unknown means; (c) define an ASP estimator to be

*This research was supported in part by National Science Foundation Grants DMS 0300806 and 0404547.

the candidate PLS procedure with smallest estimated risk; (d) define additional ASP estimators from certain larger classes of candidate estimators that include the PLS candidates; (e) study the asymptotic risk of ASP estimators under minimal assumptions on the unknown means, using multiparametric asymptotics where the total number of factor-level combinations tend to infinity.

The unrestricted LS estimator tends to overfit the means of a k_0-way layout. Under a homoscedastic independent Gaussian error model, Stein (1956) proved that the LS estimator is inadmissible for quadratic loss whenever the number of factor-level combinations exceeds 2. The drawbacks to LS estimators in both theory and practice have inspired work on competing estimators. Candidate submodel fits, including submodel polynomial fits, and ridge regression, are particular symmetric linear estimators. So are the candidate PLS and other fits used in this paper to construct ASP estimators. Notable studies of symmetric linear estimators include Stein (1981), Li and Hwang (1984), Buja *et al.* (1989), Kneip (1994), and Beran and Dümbgen (1998).

Kimeldorf and Wahba (1970) showed that candidate PLS fits can typically be derived as candidate Bayes estimators. Green *et al.* (1985) studied the use of penalized least squares to fit a smooth trend factor in field experiments. Wood (2000) treated penalized least squares with multiple quadratic penalties. The present paper differs from his work in several ways: (a) the construction of the multiple penalty terms; (b) the use of estimated risk under the full model (with unrestricted means) rather than cross-validation to select penalty weights and terms; (c) consideration of shrinkage strategies more general than PLS; (d) developing multiparametric asymptotics under which the risk of the candidate estimator that minimizes estimated risk converges to that of the unrealizable candidate estimator that minimizes risk.

ASP estimators weigh a large class of competing fits, including smooth fits for ordinal factors, to reduce risk through bias-variance tradeoff. Risks of candidate estimators are evaluated under the full model for the k_0-way layout. The procedures and results in this paper extend to balanced complete k_0-way layouts the adaptive PLS technology described for one-way layouts in Beran (2002). Important for the generalization are a PLS criterion that has a penalty term for each main effect and interaction in the ANOVA decomposition of a k_0-way array; and a canonical representation of candidate PLS estimators with respect to an orthonormal product basis that is determined by the $2^{k_0} - 1$ penalty terms. Multiparametric asymptotics in Section 3 show that the asymptotic risk of an ASP fit never exceeds that of the unrestricted least squares fit whether the k_0 factors are all ordinal, all nominal, or some of each. Both in theory and practice, ASP fits often improve substantially on the risk and visual appearance of least squares fits.

The James-Stein (1961) estimator that shrinks toward the average observation coincides essentially with an ASP estimator for the one-way layout in which the factor is nominal. Stein (1966) studied multiple shrinkage estimators that dominate LS estimators in abstract ANOVA models and gave, as an example, a superior estimator for the two-way layout with nominal factors. Section 3 of this paper shows that ASP estimators for k_0-way layouts in which each factor is nominal are close approximations to Stein's (1966) multiple-shrinkage estimators.

Tukey (1977) proposed and experimented with certain smoothing algorithms for fitting one- and higher-way layouts with ordinal factors. In ordinal one-way layouts where wavelet bases provide a sparse representation of the means, Donoho and Johnstone (1995) used adaptive shrinkage through soft-thresholding. Beran and Dümbgen (1998) proposed and studied adaptive symmetric linear estimators that perform monotone shrinkage relative to a fixed orthonormal basis. Section 2 shows that ASP estimators for k_0-way layouts with either ordinal or nominal factors can be represented as polytone shrinkage estimators.

Modern statistical theory distinguishes among data, probability model, pseudo-random numbers, and algorithms. ASP estimators implicitly fit the probability model that motivates their construction. Using an ASP estimator on data differs from believing that a probability model governs the data. Data is not provably random—even randomized experiments rely on pseudo-random numbers that mimic certifiably only a few properties of random variables. Mathematical study of an estimator under a probability model tests the estimator on virtual data governed by that model. Such mathematical exploration gains pertinence if the probability model can approximate salient relative frequencies in data to be analyzed. Trustworthiness of an estimator in data analysis is ultimately an empirical matter that benefits from interplay between interpretations of mathematical results and computational experiments. Section 4 presents two data analyses where ASP algorithms bring out striking submodel structure or smoothness without detailed intervention by the analyst.

2. Model, representations, and candidate estimators

We consider a balanced complete k_0-way layout of observations on unknown means. Each of the k_0 factors that influence the means may be nominal or ordinal. An algebraic representation for the ANOVA decomposition of the k_0-way array of means induces a penalized least squares criterion that has a separate penalty term for each main effect and interaction. A product basis generated by the penalty terms in the PLS criterion—the penalty basis—yields a canonical representation of candidate PLS estimators. This canonical representation in turn suggests larger classes of candidate estimators for the means.

2.1 *Model*

Let $\mathcal{I}$ denote the set of all k_0-tuples $i = (i_1, i_2, \ldots, i_{k_0})$ such that $1 \leq i_k \leq p_k$ for $1 \leq k \leq k_0$. The means in the k_0-way layout are unknown real values $\{m_i : i \in \mathcal{I}\}$ that depend on k_0 factors as follows. Factor k has p_k levels, denoted by $t_{k1}, t_{k2}, \ldots, t_{k,p_k}$. The factor levels associated with the observations on mean m_i are $t_i = (t_{1i_1}, t_{2i_2}, \ldots, t_{k_0 i_{k_0}})$. In other words,

$$m_i = \mu(t_i) \qquad \text{for} \quad i \in \mathcal{I}, \tag{2.1}$$

where μ is an unknown real-valued function. When factor k is nominal, the factor levels are numerical labels whose distinctness is important. When factor k is ordinal, the values and order of the factor levels are significant. Subscripting of the cells in the k_0-way layout is arranged hereafter so that $t_{k1} < t_{k2} < \cdots < t_{kp_k}$ for $1 \leq k \leq k_0$.

Under the Gaussian model for the balanced complete k_0-way layout, we collect $j_0 \geq 1$ independent noisy observations on each mean m_i:

$$y_{ij} = m_i + \epsilon_{ij} \qquad i \in \mathcal{I}, \;\; 1 \leq j \leq j_0. \tag{2.2}$$

The $\{y_{ij}\}$ are the observations, t_i gives the factor levels associated with mean m_i through (2.1), and the errors $\{\epsilon_{ij}\}$ are independent, identically distributed $N(0, \sigma^2)$ random variables. The variance σ^2 is unknown. The total number of unknown means is $p = \prod_{k=1}^{k_0} p_k$ and the total number of observations is $n = j_0 p$. The normality assumption plays a key role in the proof of Theorem 3.1 through the property that any orthogonal transformation of the data vector has homoscedastic, independent, normally distributed components.

To facilitate linear algebra, we order the elements $i = (i_1, i_2, \ldots, i_{k_0})$ 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 we assume that $\mathcal{I}$ is so ordered. Taken in order, the indexed means form the $p \times 1$ vector

$$\begin{aligned} m &= \{m_i : i \in \mathcal{I}\} \\ &= \{\cdots\{\{m_{i_1,i_2,\ldots,i_{k_0}} : 1 \le i_1 \le p_1\}, 1 \le i_2 \le p_2\}, \ldots, 1 \le i_{k_0} \le p_{k_0}\}. \end{aligned} \tag{2.3}$$

The observations may be correspondingly ordered in the $n \times 1$ vector

$$y = \{\{y_{ij} : 1 \le j \le j_0\}, i \in \mathcal{I}\}. \tag{2.4}$$

Let $e = (1, 1, \ldots, 1)'$ denote the $j_0 \times 1$ vector of ones. Model (2.2) is equivalent to the assertion

$$y \sim N(Xm, \sigma^2 I_n) \quad \text{with} \quad X = I_p \otimes e. \tag{2.5}$$

We are interested in estimating $\eta = Xm$, the expectation of the observation vector y.

2.2 *ANOVA decomposition*

The orthogonal projections that define the ANOVA decomposition of means into overall mean, main effects, and interactions are given by the following algebra. 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 $J_k = u_k u_k'$ and $H_k = I_{p_k} - u_k u_k'$. 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 $\mathcal{S}$ denote the set of all subsets of $\{1, 2, \ldots, k_0\}$, including the empty set $\emptyset$. The cardinality of $\mathcal{S}$ is 2^{k_0}. For every set $S \in \mathcal{S}$, define the $p_k \times p_k$ matrix

$$P_{S,k} = \begin{cases} J_k & \text{if} \quad k \notin S \\ H_k & \text{if} \quad k \in S \end{cases}. \tag{2.6}$$

Define the $p \times p$ Kronecker product matrix

$$P_S = \bigotimes_{k=1}^{k_0} P_{S,k_0-k+1}. \tag{2.7}$$

The foregoing discussion implies that:

- P_S is symmetric, idempotent for every $S \in \mathcal{S}$.
- If $S \ne \emptyset$, 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 $\mathcal{S}$, then $P_{S_1} P_{S_2} = 0 = P_{S_2} P_{S_1}$.
- $\sum_{S \in \mathcal{S}} P_S = I_p$.

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

The last bulleted point yields for every $m \in R^p$ the identity

$$m = \sum_{S \in \mathcal{S}} P_S m, \tag{2.8}$$

whose right side expresses, in readily computable form, the ANOVA decomposition for the means of a 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 $S \in \mathcal{N}$, where $\mathcal{N}$ is a specified subset of $\mathcal{S}$. The choice of $\mathcal{N}$ identifies the main effects or interaction terms that vanish in the submodel.

2.3 *PLS candidate estimators*

This subsection defines PLS candidate estimators of $\eta = Xm$ and expresses them in canonical form. Let $\mathcal{S}_0 = \mathcal{S} - \emptyset$, the set of all nonempty subsets of $\{1, 2, \ldots, k_0\}$. Introduce the penalty weights $\nu = \{\nu_S : S \in \mathcal{S}_0\}$, where each ν_S lies in $[0, \infty]$. For every $1 \leq k \leq k_0$, let A_k be a matrix with p_k columns such that $A_k u_k = 0$ and each row of A_k has the same Euclidean norm. The rows of A_k are thus contrasts. Examples of such *annihilator* matrices are developed in Subsection 2.4. Let $B_k = A_k' A_k$ and let $A = \{A_k : 1 \leq k \leq k_0\}$ denote the *annihilator string.*

For every set $S \in \mathcal{S}_0$ and $1 \leq k \leq k_0$, define the $p_k \times p_k$ matrix

$$Q_{S,k} = \begin{cases} J_k & \text{if} \quad k \notin S \\ B_k & \text{if} \quad k \in S \end{cases} \tag{2.9}$$

and the $p \times p$ Kronecker product matrix

$$Q_S = \bigotimes_{k=1}^{k_0} Q_{S, k_0 - k + 1}. \tag{2.10}$$

If S_1, S_2 are different subsets of $\mathcal{S}_0$, then there exists k such that $k \in S_1$ and $k \notin S_2$. Then $Q_{S_1,k} = B_k$ by (2.9) while $P_{S_2,k} = J_k$ by (2.6). By the annihilator property of A_k, it follows that $Q_{S_1,k} P_{S_2,k} = 0$. Hence

$$Q_{S_1} P_{S_2} = \bigotimes_{k=1}^{k_0} [Q_{S_1, k_0 - k + 1} P_{S_2, k_0 - k + 1}] = 0. \tag{2.11}$$

Let

$$T(m, \nu, A) = |y - Xm|^2 + j_0 \sum_{S \in \mathcal{S}_0} \nu_S m' Q_S m. \tag{2.12}$$

The *candidate PLS estimator* of $\eta = Xm$ determined by ν and A is defined to be

$$\hat{\eta}_{PLS}(\nu, A) = X \operatorname*{argmin}_{m \in R^p} T(m, \nu, A). \tag{2.13}$$

The factor j_0 in front of the penalty term on the right side of (2.12) is harmless and simplifies subsequent algebra. Let $U_0 = j_0^{-1/2}X$. From the definition of X in (2.5), $U_0'U_0 = I_p$. By calculus,

$$\hat{\eta}_{PLS}(\nu, A) = U_0 \left[I_p + \sum_{S \in \mathcal{S}_0} \nu_S Q_S\right]^{-1} U_0'y. \tag{2.14}$$

If a term $P_S m$ vanishes in the ANOVA decomposition (2.8), then the corresponding penalty term $m'Q_S m$ vanishes in (2.12) because of (2.11). For the annihilator $A_k = H_k$ (to be called the flat annihilator in Subsection 2.4), the penalty term $m'Q_S m$ vanishes if and only if the model term $P_S m$ vanishes in (2.12) because $Q_S = P_S$. Further insight into how A and ν affect the candidate PLS estimator is obtained through the canonical representation derived below.

Suppose that the $p_k \times p_k$ symmetric matrix $B_k = A_k'A_k$ has the spectral decomposition $B_k = U_k \Lambda_k U_k'$, where the eigenvector matrix satisfies $U_k U_k' = U_k'U_k = I_{p_k}$ and the diagonal matrix $\Lambda_k = \text{diag}\{\lambda_{ki}\}$ gives the ordered eigenvalues with $0 = \lambda_{k1} \leq \lambda_{k2} \leq \cdots \leq \lambda_{kp_k}$. This eigenvalue ordering, the reverse of the customary, is adopted here because the eigenvectors associated with the smallest eigenvalues play the greatest role in determining the numerical value and risk of a candidate PLS estimator. Because an annihilator A_k satisfies $A_k u_k = 0$, the eigenvalue λ_{k1} is necessarily zero and has u_k as corresponding eigenvector. Thus, the first column of U_k is u_k or may be chosen to be u_k if the eigenvalue 0 is multiple.

The $p_k \times p_k$ matrix $J_k = u_k u_k'$ is symmetric, idempotent, has eigenvalue 1 associated with the eigenvector u_k, and has eigenvalue 0 repeated $p_k - 1$ times. Let E_k denote the $p_k \times p_k$ diagonal matrix that has 1 in the $(1,1)$ cell and zeroes elsewhere. Because u_k is the first column of U_k, the spectral decomposition $J_k = U_k E_k U_k'$ follows.

For every set $S \in \mathcal{S}_0$ and $1 \leq k \leq k_0$, define the $p_k \times p_k$ diagonal matrix

$$\Gamma_{S,k} = \begin{cases} E_k & \text{if} \quad k \notin S \\ \Lambda_k & \text{if} \quad k \in S \end{cases}. \tag{2.15}$$

From (2.9), the spectral decompositions of B_k, J_k, and (2.15), $Q_{S,k} = U_k \Gamma_{S,k} U_k'$. Consequently, by (2.10), Q_S has spectral decomposition

$$Q_S = U \Gamma_S U', \tag{2.16}$$

where

$$\Gamma_S = \bigotimes_{k=1}^{k_0} \Gamma_{S, k_0 - k + 1}, \qquad U = \bigotimes_{k=1}^{k_0} U_{k_0 - k + 1}. \tag{2.17}$$

The columns of U form an orthonormal product basis for R^p.

Candidate PLS estimator (2.14) thus has the canonical representation

$$\hat{\eta}_{PLS}(\nu, A) = V \,\text{diag}\{f(\nu)\} V'y, \tag{2.18}$$

where

$$\text{diag}\{f(\nu)\} = \left[I_p + \sum_{S \in \mathcal{S}_0} \nu_S \Gamma_S\right]^{-1}, \qquad V = U_0 U \tag{2.19}$$

with $U_0 = j_0^{-1/2}X$. The columns of V form the orthonormal *penalty basis* generated by annihilator string A for the k_0-way layout.

To express the elements of vector $f(\nu) = \{f_i(\nu) : i \in \mathcal{I}\}$ more simply, we partition the index set $\mathcal{I}$ according to the subsets of $\mathcal{S}$. For every $S \in \mathcal{S}$, let

$$\mathcal{I}_S = \{i \in \mathcal{I} : i_k = 1 \text{ if } k \notin S, i_k \geq 2 \text{ if } k \in S\}. \tag{2.20}$$

In other words, $i = (i_1, i_2, \ldots, i_{k_0}) \in \mathcal{I}_S$ if and only if $S = \{k : i_k \geq 2\}$. Evidently, $\mathcal{I} = \bigcup_{S \in \mathcal{S}} \mathcal{I}_S$ and $\mathcal{I}_{S_1} \cap \mathcal{I}_{S_2} = \emptyset$ whenever $S_1 \neq S_2$. For $i \in \mathcal{I}_S$, it follows from (2.19) that

$$f_i(\nu) = \left[1 + \nu_S \prod_{k \in S} \lambda_{k i_k}\right]^{-1}. \tag{2.21}$$

Example. (The three-way layout.) In this case, mirrored dictionary ordering of the factor-level triples yields

$$\mathcal{I} = \{\{\{(i_1, i_2, i_3) : 1 \leq i_1 \leq p_1\}, 1 \leq i_2 \leq p_2\}, 1 \leq i_3 \leq p_3\}. \tag{2.22}$$

The corresponding vectorization of m is obtained by running through the row subscript first, the column subscript second, and the layer subscript third.

Here $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,3\}, \{1,2,3\}\}$. The projections that define the ANOVA decomposition are:

$$\begin{aligned}
&P_\emptyset = J_3 \otimes J_2 \otimes J_1, \quad P_{\{1,2,3\}} = H_3 \otimes H_2 \otimes H_1,\\
&P_{\{1\}} = J_3 \otimes J_2 \otimes H_1, \quad P_{\{2\}} = J_3 \otimes H_2 \otimes J_1, \quad P_{\{3\}} = H_3 \otimes J_2 \otimes J_1,\\
&P_{\{1,2\}} = J_3 \otimes H_2 \otimes H_1, \quad P_{\{2,3\}} = H_3 \otimes H_2 \otimes J_1,\\
&P_{\{1,3\}} = H_3 \otimes J_2 \otimes H_1.
\end{aligned} \tag{2.23}$$

Thus, for instance, if $\mathcal{N} = \{\{1,2\}, \{2,3\}, \{1,3\}, \{1,2,3\}\}$, then the constraints $P_S m = 0$ for every $S \in \mathcal{N}$ define the additive submodel.

The partition of $\mathcal{I}$ induced by $\mathcal{S}$ consists of the subsets

$$\begin{aligned}
&\mathcal{I}_\emptyset = \{(1,1,1)\}, \quad \mathcal{I}_{\{1\}} = \{i \in \mathcal{I} : i_1 \geq 2, i_2 = i_3 = 1\},\\
&\mathcal{I}_{\{2\}} = \{i \in \mathcal{I} : i_2 \geq 2, i_1 = i_3 = 1\}, \quad \mathcal{I}_{\{3\}} = \{i \in \mathcal{I} : i_3 \geq 2, i_1 = i_2 = 1\},\\
&\mathcal{I}_{\{1,2\}} = \{i \in \mathcal{I} : i_1 \geq 2, i_2 \geq 2, i_3 = 1\},\\
&\mathcal{I}_{\{2,3\}} = \{i \in \mathcal{I} : i_2 \geq 2, i_3 \geq 2, i_1 = 1\},\\
&\mathcal{I}_{\{1,3\}} = \{i \in \mathcal{I} : i_1 \geq 2, i_3 \geq 2, i_2 = 1\},\\
&\mathcal{I}_{\{1,2,3\}} = \{i \in \mathcal{I} : i_1 \geq 2, i_2 \geq 2, i_3 \geq 2\}.
\end{aligned} \tag{2.24}$$

The candidate PLS estimator is defined by (2.18), the value of $f_i(\nu)$ being

$$\begin{aligned}
&1 \quad \text{if } i \in \mathcal{I}_\emptyset, \quad [1 + \nu_{\{1\}}\lambda_{1i_1}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{1\}},\\
&[1 + \nu_{\{2\}}\lambda_{2i_2}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{2\}}, \quad [1 + \nu_{\{3\}}\lambda_{3i_2}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{3\}},\\
&[1 + \nu_{\{1,2\}}\lambda_{1i_1}\lambda_{2i_2}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{1,2\}}, \quad [1 + \nu_{\{2,3\}}\lambda_{2i_2}\lambda_{3i_3}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{2,3\}},\\
&[1 + \nu_{\{1,3\}}\lambda_{1i_1}\lambda_{3i_3}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{1,3\}},\\
&[1 + \nu_{\{1,2,3\}}\lambda_{1i_1}\lambda_{2i_2}\lambda_{3i_3}]^{-1} \quad \text{if } i \in \mathcal{I}_{\{1,2,3\}}.
\end{aligned} \tag{2.25}$$

2.4 *More general candidate estimators*

A shrinkage class $\mathcal{F}$ consists of $p \times 1$ vectors $f = \{f_i : i \in \mathcal{I}\}$ such that $0 \le f_i \le 1$. An annihilator class $\mathcal{A}$ is a collection of annihilator strings $A = \{A_k : 1 \le k \le k_0\}$, each of which generates a penalty basis V for the regression space of the k_0-way layout. By generalization of (2.18), we define the *SP candidate estimators* through

$$\hat{\eta}_{SP}(f, A) = V \operatorname{diag}\{f\} V' y \quad \text{for} \quad f \in \mathcal{F},\ A \in \mathcal{A}. \tag{2.26}$$

The acronym SP stands for **S**hrinkage on **P**enalty bases.

Candidate shrinkage classes. We will consider several shrinkage classes. The *Unrestricted* shrinkage class $\mathcal{F}_U$ consists of all $p \times 1$ vectors f with elements in $[0, 1]$. Though too large for successful adaptation, this class is a starting point for defining more useful shrinkage classes.

The *Penalized Least Squares* shrinkage class $\mathcal{F}_{PLS}$ is the subset of shrinkage vectors in $\mathcal{F}_U$ that take the form $f(\nu)$ in (2.19) or (2.21). For fixed A, the candidate estimator (2.26) with $\mathcal{F} = \mathcal{F}_{PLS}$ coincides with PLS candidate estimator (2.18).

The *Polytone Score* shrinkage class $\mathcal{F}_{PS}$ is the subset of $\mathcal{F}_U$ that is restricted as follows: $f_{(1,1,\ldots,1)} = 1$; for every $S \in \mathcal{S}_0$ and $i \in \mathcal{I}_S$,

$$f_i = g_S \left[\prod_{k \in S} \lambda_{k i_k} \right], \tag{2.27}$$

where each g_S is any function nonincreasing in its argument and having range in $[0, 1]$.

The *Polytone* shrinkage class $\mathcal{F}_P$ is the convex subset of $\mathcal{F}_U$ that is restricted as follows: $f_{(1,1,\ldots,1)} = 1$; for every $S \in \mathcal{S}_0$ and $i \in \mathcal{I}_S$, $f_i = f_{(i_1, i_2, \ldots, i_{k_0})}$ is a nonincreasing function in each subscript i_k such that $k \in S$.

Evidently, $\mathcal{F}_{PLS} \subset \mathcal{F}_{PS} \subset \mathcal{F}_P$.

Candidate annihilator classes. In constructing annihilator string $A = (A_k : 1 \le k \le k_0)$, we distinguish between ordinal factors and nominal factors. If factor k is *nominal*, permutation of its levels $\{t_{kj} : 1 \le j \le p_k\}$ should not affect the corresponding candidate SP estimator. This consideration leads to setting $A_k = H_k$ for every k, the latter projection being defined in Subsection 2.2. This choice will be called the flat annihilator, a term suggested by the constant spectrum of the reduced singular value decomposition of H_k.

On the other hand, suppose that factor k is *ordinal* with equally spaced levels $t_{(k)} = (t_{k1}, t_{k2}, \ldots, t_{kp})$. To have the SP candidate estimator favor a fit that is locally polynomial of degree $r - 1$ in the levels of factor k, we take A_k proportional to the r-th difference operator of column dimension p_k. Explicitly, consider the $(q-1) \times q$ matrix $\Delta(q) = \{\delta_{vw}\}$ in which $\delta_{v,v} = 1$, $\delta_{v,v+1} = -1$ for every v and all other entries are zero. Define recursively

$$D(1, p_k) = \Delta(p_k), \qquad D(r, p_k) = \Delta(p_k - r + 1) D(r - 1, p_k) \quad \text{for} \quad 2 \le r \le p_k - 1. \tag{2.28}$$

Evidently the $(p_k - r) \times p_k$ matrix $D(r, p_k)$ accomplishes the r-th differencing and annihilates powers of $t_{(k)}$ up to power $r - 1$ in the sense that

$$D(r, p_k) t_{(k)}^u = 0 \quad \text{for} \quad 0 \le u \le r - 1. \tag{2.29}$$

Here $t^u_{(k)}$ denotes the column vector $(t^u_{k1}, \ldots, t^u_{kp_k})'$. Moreover, in row i of $D(r, p_k)$, the elements not in columns $i, i+1, \ldots, i+d$ are zero. The r-th difference annihilator for factor k is defined to be $D(r, p_k)$.

More generally, if μ in (2.1) is expected to behave locally like a polynomial of degree $r-1$ in factor k but the factor levels in $t_{(k)}$ are *not* equally spaced, we define A_k as follows. For every integer $1 \le r \le p_k - 1$, the *local polynomial annihilator* $C(r, p_k)$ is a $(p_k - r) \times p_k$ matrix characterized through three conditions. First, for every possible v, all elements in the v-th row of $C(r, p_k)$ other than $\{c_{vw} : v \le w \le v + r\}$ are zero. Second, $C(r, p_k)$ satisfies the orthogonality conditions

$$(2.30) \qquad C(r, p_k) t^u_{(k)} = 0 \quad \text{for} \quad 0 \le u \le r - 1.$$

Third, each row vector in $C(r, p_k)$ has unit length. These requirements are met by setting the non-zero elements in the v-th row of $C(r, p_k)$ equal to the basis vector of degree r in the orthonormal polynomial basis that is defined on the $r+1$ design points $(t_{kv}, \ldots, t_{k,v+r})$. The S-Plus function poly accomplishes this computation. When the components of $t_{(k)}$ are equally spaced, $C(r, p_k)$ is just a scalar multiple of the r-th difference matrix $D(r, p_k)$.

In the construction of $C(r, p_k)$, the powers of the components of $t_{(k)}$ could be replaced by other linearly independent functions to express prior notions about μ other than local polynomial behavior.

For specified integer $a_k \ge 1$, let $\mathcal{A}_k = \{C(r, p_k) : 1 \le r \le a_k\}$. We will consider SP candidate estimators whose annihilator string $A = (A_k : 1 \le k \le k_0)$ lies in the annihilator class $\mathcal{A} = \prod_{k=1}^{k_0} \mathcal{A}_k$. The asymptotics in Section 3 impose limits on the cardinality of $\mathcal{A}$ relative to p.

3. ASP estimators

For given annihilator class $\mathcal{A}$ and shrinkage class $\mathcal{F}$, the ASP estimator of η is the candidate SP estimator with smallest estimated risk. This section gives details of the construction. Under conditions on $\mathcal{F}$ and $\mathcal{A}$, the asymptotic risk of the ASP estimator coincides with the smallest asymptotic risk achievable by the candidate SP estimators. In this sense, adaptation works. The asymptotics are multiparametric in that the number of means p in the k_0-way layout tends to infinity while the number of replications j_0 is fixed, with possibly $j_0 = 1$.

3.1 *Risks and estimated risks of candidate estimators*

We will assess the performance of any estimator $\hat{\eta}$ of $\eta = Xm$ through its normalized quadratic loss and risk:

$$(3.1) \qquad L(\hat{\eta}, m) = p^{-1} |\hat{\eta} - \eta|^2, \qquad R(\hat{\eta}, m, \sigma^2) = \mathrm{E} L(\hat{\eta}, m).$$

Let $z = V'y$, let $\xi = \mathrm{E}(z) = V'\eta$, and let $\hat{\xi}(f, A) = \mathrm{diag}\{f\} z$. Then, from (2.26),

$$(3.2) \qquad \hat{\eta}_{SP}(f, A) = V\hat{\xi}(f, A), \qquad \eta = V\xi.$$

The normalized quadratic loss of the SP candidate estimator is thus

$$(3.3) \qquad L(\hat{\eta}_{SP}(f, A), m) = p^{-1} |\hat{\eta}_{SP}(f, A) - \eta|^2 = p^{-1} |\hat{\xi}(f, A) - \xi|^2.$$

For any vector x, let ave(x) denote the average of its components. It follows from (3.3) that the risk of the candidate SP estimator is

$$R(\hat{\eta}_{SP}(f, A), m, \sigma^2) = r(f, A, \xi^2, \sigma^2), \tag{3.4}$$

where

$$r(f, A, \xi^2, \sigma^2) = \text{ave}[f^2\sigma^2 + (1-f)^2\xi^2]. \tag{3.5}$$

The multiplication of vectors on the right side of this display is to be done componentwise as in the S language.

Were the risk function (3.5) known, we would use the candidate SP estimator of η that minimizes risk over the class of shrinkage vectors $\mathcal{F}$ and the class of annihilator matrices $\mathcal{A}$ under consideration. This is the oracle candidate estimator. In reality, the risk function contains two quantities, σ^2 and ξ^2, that are usually unknown. The sampling scheme and the ordinal or nominal character of the factors influence methods for estimating σ^2. Basic possibilities include:

- *Replicated layout.* In this setting, where $n > p$, a fundamental choice is the least squares estimator of σ^2, the normalized residual sum of squares in the ANOVA table for the multi-way layout.
- *One observation per combination of factor levels.* Here $n = p$. If the penalty basis is such that the coefficients $\{\xi_i : i \in L\}$ are close to zero, then an appropriate variance estimator is

$$\hat{\sigma}^2 = [\#(L)]^{-1} \sum_{i \in L} z_i^2. \tag{3.6}$$

The least squares estimator of σ^2 converges to σ^2 in mean squared error if and only if $n - p$ tends to infinity. This makes it worth considering when $n - p$ is (say) 25 or more. For variance estimator (3.6), which is designed for the difficult case $n = p$, $\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2$ converges to zero if and only if $\#(L)$ tends to infinity and the sum of squared biases $[\#(L)]^{-1} \sum_{i \in L} \xi_i^2$ tends to zero as p tends to infinity. Because ξ is unknown, practical choice of L relies on prior conjecture about the structure of m checked by scrutiny of $\{z_i : i \in L\}$.

The following procedure expresses an empirical engineering approach to choosing L that the author has found useful. Initially, let $S = \{1, 2, \ldots, k_0\}$. Set $L = \{i \in \mathcal{I}_S : \prod_{k \in S} \lambda_{ki_k} \geq c\}$ for a tentatively selected threshold c. To pick a final c, look for the smallest c such that the value of (3.6) remains stable as c is increased. Look then at the values of the $\{|z_i|^{1/2} : i \in L\}$ for evidence that ξ_i^2 is not excessively large for any $i \in L$. When A_k is equal to H_k, as advocated for nominal factors in Subsection 2.4, then $\lambda_{k2} = \cdots = \lambda_{kp_k} = 1$. In that case, for $c > 0$, $L = \mathcal{I}_S$ and the variance estimator (3.6) coincides with the pooled interaction estimator of σ^2 based on the highest-order interaction term in ANOVA decomposition (2.8).

This informal procedure combines vague prior opinions about m with feedback from the data. Variants of the procedure replace the initial choice of S with a smaller subset of $\{1, 2, \ldots, k_0\}$, then proceed similarly. This yields generalizations of classical pooled lower-order interaction variance estimators and may be useful when the highest order interactions are not negligible. Some readers will be dissatisfied by the subjective character of the proposed method for choosing L and, more generally, of all known methods for estimating σ^2 when $n = p$. They are advised to consider variance estimators and ASP fits based on competing plausible a priori considerations. The data examples in Section 4 illustrate some of the issues. Devising an effective theory for estimating σ^2 in the absence of replication remains an important open problem.

Having devised a variance estimator $\hat{\sigma}^2$, we may estimate ξ^2 by $z^2 - \hat{\sigma}^2$ and hence the risk function $r(f, A, \xi^2, \sigma^2)$ by

$$\text{(3.7)} \qquad \hat{r}(f, A) = \text{ave}[\hat{\sigma}^2 f^2 + (1-f)^2(z^2 - \hat{\sigma}^2)] = \text{ave}[(f - \hat{g})^2 z^2] + \hat{\sigma}^2\, \text{ave}(\hat{g}),$$

where $\hat{g} = (z^2 - \hat{\sigma}^2)/z^2$. Apart from the new considerations entering into the estimation of σ^2, this equation expresses Stein's (1981) unbiased estimator of risk or the risk estimator implicit in Mallow's (1973) discussion of C_p.

3.2 *ASP estimators and algorithms*

For fixed shrinkage class $\mathcal{F}$ and annihilator class $\mathcal{A}$, the *ASP estimator* is defined to be $\hat{\eta}_{ASP} = \hat{\eta}_{SP}(\hat{f}, \hat{A})$, where

$$\text{(3.8)} \qquad (\hat{f}, \hat{A}) = \underset{f \in \mathcal{F}, A \in \mathcal{A}}{\text{argmin}}\ \hat{r}(f, A) = \underset{f \in \mathcal{F}, A \in \mathcal{A}}{\text{argmin}}\ \text{ave}[(f - \hat{g})^2 z^2].$$

Because the cardinality of $\mathcal{A}$ is finite, the minimization in (3.8) may be accomplished by first minimizing estimated risk over $f \in \mathcal{F}$ for each A and by then minimizing over $A \in \mathcal{A}$. For each fixed A, computation of $\hat{f}$ is a weighted least squares problem whose details depend on the shrinkage class constraints, as described next.

Penalized Least Squares shrinkage. For fixed A, it follows from Subsection 2.3 and (3.7) that the PLS shrinkage vector minimizing estimated risk is $f(\hat{\nu})$, where $f(\nu)$ is defined by (2.21) and

$$\text{(3.9)} \qquad \hat{\nu} = \underset{\nu \in [0,\infty]^{\#(\mathcal{S}_0)}}{\text{argmin}}\ \text{ave}[(f(\nu) - \hat{g})^2 z^2].$$

Because of (2.21), equation (3.9) is equivalent to the system of equations

$$\text{(3.10)} \qquad \hat{\nu}_S = \underset{\nu_S \in [0,\infty]}{\text{argmin}} \sum_{i \in \mathcal{I}_S} \left[\left(1 + \nu_S \prod_{k \in S} \lambda_{ki_k}\right)^{-1} - \hat{g}_i \right]^2 z_i^2, \qquad S \in \mathcal{S}_0.$$

Calculation of $\hat{\nu} = \{\hat{\nu}_S : S \in \mathcal{S}_0\}$ thus amounts to solving $2^{k_0 - 1}$ nonlinear, weighted least squares problems, each of which can be treated with minimization algorithms for a function of a single variable.

Special case. (All factors ordinal.) Let $[\cdot]_+$ be the positive-part function. When each of the k factors is ordinal and the corresponding A_k is the flat annihilator H_k discussed in Subsection 2.4, then $\lambda_{k2} = \cdots = \lambda_{kp_k} = 1$. In this important special case, (3.10) and the identity $\sum_{i \in \mathcal{I}_S} z_i^2 = |P_S U_0' y|^2$ imply that $(1 + \hat{\nu}_S)^{-1} = \hat{c}_S$, where

$$\text{(3.11)} \qquad \begin{aligned} \hat{c}_S &= \underset{c \in [0,1]}{\text{argmin}} \sum_{i \in \mathcal{I}_S} (c - \hat{g}_i)^2 z_i^2 \\ &= \left[1 - \hat{\sigma}^2 \prod_{k \in S} (p_k - 1) / |P_S U_0' y|^2 \right]_+, \qquad S \in \mathcal{S}_0. \end{aligned}$$

The ASP estimator for all factors ordinal thus reduces to

$$\hat{\eta}_{ASP} = U_0 \left[I_p + \sum_{s \in \mathcal{S},} \hat{c}_S P_S \right] U_0' y. \tag{3.12}$$

This expression does not require computing the penalty-basis coordinate system. Stein (1966) studied an abstract multiple shrinkage estimator that, like (3.12), uses data-based shrinkage factors that are constant over subspaces. His paper gave the application to the two-way layout and refined the shrinkage factors slightly so as to reduce the risk of the estimator slightly when p is small. Achieving such risk improvements for more general ASP estimators is an open question.

Polytone Score shrinkage. Fix A. Minimizing estimated risk (3.7) over $f \in \mathcal{F}_{PS}$ is accomplished in several steps. Evidently $\hat{f}_{(1,1,\ldots,1)} = 1$. For every $S \in \mathcal{S}_0$, we compute $\{\hat{f}_i : i \in \mathcal{I}_S\}$ as follows:

Let $w = \{z_i : i \in \mathcal{I}_S\}$, let $\tau = \{\prod_{k \in S} \lambda_{k i_k} : i \in \mathcal{I}_S\}$ denote the vector of corresponding scores, and let $q = \#(S) = \prod_{k \in S}(p_k - 1)$. The q components in w and τ are both arranged according to the mirror dictionary ordering of the indices $i = (i_1, i_2, \ldots, i_{k_0})$ that is defined in Subsection 2.1. Let ρ denote the rank vector of τ and define the q dimensional vector $\tilde{w}$ through $\tilde{w}_{\rho_j} = w_j$. Let $\tilde{h} = (\tilde{w}^2 - \hat{\sigma}^2)/\tilde{w}^2$ and let $\mathcal{K}_S = \{k \in R^q : k_1 \geq k_2 \geq \cdots \geq k_q\}$. Find

$$\tilde{k} = \operatorname*{argmin}_{k \in \mathcal{K}_S} \operatorname{ave}[(k - \tilde{h})^2 \tilde{w}^2], \tag{3.13}$$

using an algorithm for weighted isotonic least squares such as the pool adjacent violators algorithm (cf. Robertson *et al.* (1988)). The PAV algorithm converges in a finite number of steps.

Define the q dimensional vector $\hat{k}$ through $\tilde{k}_{\rho_j} = \hat{k}_j$. The j-th component of $\{\hat{f}_i : i \in \mathcal{I}_S\}$ is then $\max\{\hat{k}_j, 0\}$. The final positive part adjustment is needed because the components $\hat{f}_i$ are restricted to $[0, 1]$. Section 5 of Beran and Dümbgen (1998) provides the supporting mathematical argument.

Polytone shrinkage. Fix A. Minimizing estimated risk (3.7) over $f \in \mathcal{F}_P$ is accomplished in several steps. Let $\mathcal{K}_P$ denote the subset of R^p that is restricted as follows: If $k \in \mathcal{K}_P$ then $k_{(1,1,\ldots,1)} = 1$; for every $S \in \mathcal{S}_0$ and $i \in \mathcal{I}_S$, $k_i = k_{(i_1, i_2, \ldots, i_{k_0})}$ is a nondecreasing function in each subscript i_k such that $k \in S$. Find

$$\hat{k} = \operatorname*{argmin}_{k \in \mathcal{K}_P} \operatorname{ave}[(k - \hat{g})^2 z^2]. \tag{3.14}$$

The j-th component of $\{\hat{f}_i : i \in \mathcal{I}_S\}$ is then $\max\{\hat{k}_j, 0\}$. As above, the positive part adjustment is needed because the components of $\hat{f}_i$ are restricted to $[0, 1]$.

When $k_0 = 1$, the PAV algorithm solves (3.14) in a finite number of steps. When $k_0 = 2$, suitable iterative application of the PAV algorithm converges to the solution of (3.14). Bril *et al.* (1984) provided a FORTRAN implementation. Their idea can be used to define an algorithm for general k_0, but the computational efficiency of this approach is problematic. A faster algorithm would be welcome.

3.3 *Multiparametric asymptotics*

The following theorem gives conditions under which adaptation to minimize estimated risk approximately minimizes true risk as p tends to infinity. The proof draws on abstract results for shrinkage estimators established by Beran and Dümbgen (1998).

THEOREM 3.1. *Fix the annihilator string A and let $\mathcal{F}$ be a subset of $\mathcal{F}_P$ that is closed in $[0,1]^p$. In particular, $\mathcal{F}$ can be any shrinkage class listed in Subsection 2.4 other than $\mathcal{F}_U$. Suppose that $\hat{\sigma}^2$ is consistent in that, for every $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \#(\mathcal{A}) \cdot \text{E}|\hat{\sigma}^2 - \sigma^2| = 0. \tag{3.15}$$

Suppose also that

$$\lim_{p\to\infty} \#(\mathcal{A}) \cdot \min_{1\le k\le k_0} \{p_k^{-1/2}\} = 0. \tag{3.16}$$

a) *Let $V(f,A)$ denote either the loss $L(\hat{\eta}_{SP}(f,A),m)$ or the estimated risk $\hat{r}(f,A)$. Then for every $a > 0$, and every $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \text{E} \sup_{f\in\mathcal{F},A\in\mathcal{A}} |V(f,A) - R(\hat{\eta}_{SP}(f,A),m,\sigma^2)| = 0. \tag{3.17}$$

b) *For the ASP estimator defined in* (3.8), *then*

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \left| R(\hat{\eta}_{ASP},m,\sigma^2) - \min_{f\in\mathcal{F},A\in\mathcal{A}} R(\hat{\eta}_{SP}(f,A),m,\sigma^2)\right| = 0. \tag{3.18}$$

c) *For W equal to either $L(\hat{\eta}_{ASP},m)$ or $R(\hat{\eta}_{ASP},m,\sigma^2)$,*

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \text{E}|\hat{r}(\hat{f},\hat{A}) - W| = 0. \tag{3.19}$$

Because $\max_{1\le k\le k_0} p_k \le p \le [\max_{1\le k\le k_0} p_k]^{k_0}$, the condition $p \to \infty$ is equivalent to $\max_{1\le k\le k_0} p_k \to \infty$. By part a, the loss, risk and estimated risk of a candidate SP estimator converge together asymptotically. Uniformity of this convergence over the shrinkage and annihilator classes makes the estimated risk of a candidate estimators a trustworthy surrogate for its true risk or loss. By part b, the risk of the ASP estimator $\hat{\eta}_{ASP}$ converges to that of the best candidate SP estimator. The theorem covers every shrinkage class defined in Subsection 2.4 except $\mathcal{F}_U$. Because the unrestricted least squares estimator is one of the candidate estimators indexed by these shrinkage classes, its asymptotic risk is at least as large as that of the best-shrinkage adaptive estimator. In practice, the risk of the best shrinkage-adaptive estimator is often much smaller than that of the unrestricted least squares estimator and this is the point. Part c shows that the loss, risk, and plug-in estimated risk of a shrinkage-adaptive estimator converge together asymptotically.

The pleasant properties stated in Theorem 3.1 break down when the shrinkage class is $\mathcal{F}_U$. Then the estimator $\hat{\eta}_{SP}(\hat{f},A)$ is dominated by the least squares estimator (see Beran and Dümbgen (1998), p. 1829). Adaptation works when the class of candidate estimators is not too large, in a sense made precise by the proof below for Theorem 3.1. The richness of a shrinkage class $\mathcal{F} \subset \mathcal{F}_U$ is characterized through the covering number

$J(\mathcal{F})$ that is defined as follows. For any probability measure Q on the set of k_0-tuples $\mathcal{I}$, consider the pseudo-distance $d_Q(f,g) = [\int (f-g)^2 dQ]^{1/2}$ on $[0,1]^{\mathcal{I}}$. For every positive u, let

$$N(u,\mathcal{F},d_Q) = \min\left\{\#\mathcal{F}_0 : \mathcal{F}_0 \subset \mathcal{F}, \inf_{f_0 \in \mathcal{F}_0} d_Q(f_0,f) \le u \;\forall f \in \mathcal{F}\right\}. \tag{3.20}$$

Let

$$N(u,\mathcal{F}) = \sup_Q N(u,\mathcal{F},d_Q), \tag{3.21}$$

where the supremum is taken over all probabilities on T. Define

$$J(\mathcal{F}) = \int_0^1 [\log N(u,\mathcal{F})]^{1/2} du. \tag{3.22}$$

Important in proving Theorem 3.1 is the property $J(\mathcal{F}_P) = O(\min_{1\le k\le k_0}\{\prod_{j\ne k} p_j^{1/2}\})$, which follows from Example 5 on p. 1832 of Beran and Dümbgen (1998) and implies

$$p^{-1/2} J(\mathcal{F}_P) = O\left(\min_{1\le k\le k_0}\{p_k^{-1/2}\}\right). \tag{3.23}$$

The right side of (3.23) tends to zero as $p \to \infty$.

PROOF OF THEOREM 3.1. *Part a.* By Theorem 2.1 in Beran and Dümbgen (1998), there exists a finite constant C such that

$$\begin{aligned} \mathrm{E}\sup_{f\in\mathcal{F}} &|V(f,A) - R(\hat\eta_{SP}(f,A),m,\sigma^2)| \\ &\le C[p^{-1/2}J(\mathcal{F})(\sigma^2 + \sigma\{\mathrm{ave}(\xi^2)\}^{1/2}) + \mathrm{E}|\hat\sigma^2 - \sigma^2|]. \end{aligned} \tag{3.24}$$

Limit (3.17) follows from this, the assumed inclusion of $\mathcal{F}$ in $\mathcal{F}_P$, (3.23), (3.15) and (3.16).

Parts b and c. In analogy to $(\hat f,\hat A) = \mathrm{argmin}_{f\in\mathcal{F},A\in\mathcal{A}}\, \hat r(f,A)$, let

$$(\tilde f,\tilde A) = \operatorname*{argmin}_{f\in\mathcal{F},A\in\mathcal{A}} r(f,A,\xi^2,\sigma^2). \tag{3.25}$$

Then $\min_{f\in\mathcal{F},A\in\mathcal{A}} R(\hat\eta_{SP}(f,A),m,\sigma^2) = r(\tilde f,\tilde A,\xi^2,\sigma^2)$. We show that (3.17) implies

$$\lim_{p\to\infty}\sup_{\mathrm{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|W - r(\tilde f,\tilde A,\xi^2,\sigma^2)| = 0, \tag{3.26}$$

where W can be $L(\hat\eta_{SP}(\hat f,\hat A),m)$ or $L(\hat\eta_{SP}(\tilde f,\tilde A),m)$ or $\hat r(\hat f,\hat A)$. The limits to be proved, (3.18) and (3.19), are immediate consequences of (3.26).

First, (3.17) with $V(f,A) = \hat r(f,A)$ entails

$$\begin{aligned} &\lim_{p\to\infty}\sup_{\mathrm{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|\hat r(\hat f,\hat A) - r(\tilde f,\tilde A,\xi^2,\sigma^2)| = 0 \\ &\lim_{p\to\infty}\sup_{\mathrm{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|\hat r(\hat f,\hat A) - r(\hat f,\hat A,\xi^2,\sigma^2)| = 0. \end{aligned} \tag{3.27}$$

Hence, (3.26) holds for $W = \hat{r}(\hat{f}, \hat{A})$ and

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|r(\hat{f}, \hat{A}, \xi^2, \sigma^2) - r(\tilde{f}, \tilde{A}, \xi^2, \sigma^2)| = 0. \tag{3.28}$$

Second, (3.17) with $V(f, A) = L(\hat{\eta}_{SP}(f, A), m)$ gives

$$\begin{aligned} &\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|L(\hat{\eta}_{SP}(\hat{f}, \hat{A}), m) - r(\hat{f}, \hat{A}, \xi^2, \sigma^2)| = 0 \\ &\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 a} \mathrm{E}|L(\hat{\eta}_{SP}(\tilde{f}, \tilde{A}), m) - r(\tilde{f}, \tilde{A}, \xi^2, \sigma^2)| = 0. \end{aligned} \tag{3.29}$$

These limits together with (3.28) establish the remaining two cases of (3.26).

4. Experiments on data

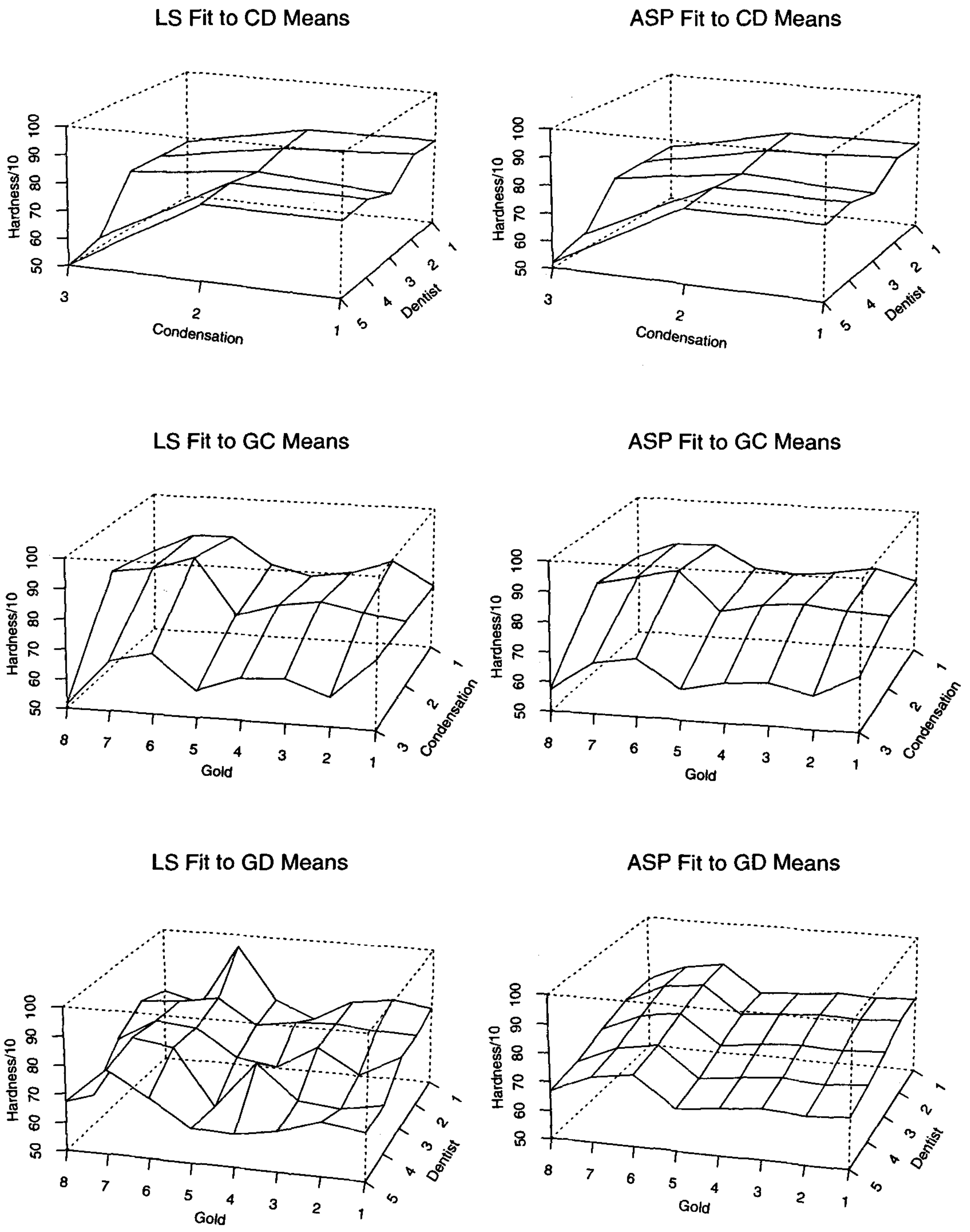
This section describes two data analyses that compare ASP fits with unrestricted LS fits. ASP fits that minimize estimated risk over a pertinent class of candidate estimators reveal striking submodel structure or smoothness in these two data sets. These structural insights are gained algorithmically without detailed intervention by the analyst.

4.1 *Hardness of dental fillings*

Brown (1975) and Seheult and Tukey (2001) analyzed a three-factor layout described by Xhonga (1971). The response variable is a measure of the hardness of fillings obtained by 5 Dentists (D) using 8 Gold alloys (G) and 3 Condensation methods (C). According to Xhonga (1971), the objective of the experiment was to find a dental gold filling with greater hardness. Condensation, properly carried out, was known to increase the hardness of a filling. The three condensation techniques used in the experiment were: (1) electromalleting, in which blows are delivered mechanically at a steady frequency; (2) hand malleting, in which a small mallet is used to deliver blows; and (3) hand condensation. The reported hardness observations are each averages of ten measurements that are not available. It was reported anecdotally that dentist 5 appeared to be physically tired before the experiment.

The first two papers cited above give the data and standard ANOVA table for this three-way layout with one observation per combination of factor levels. Analysis of the data is difficult because there is only one observation per cell, because possibly all of the main effects and interactions matter, and because outliers complicate the estimation of variance. Performance of semiautomatic ASP fits on this data thus provides an extreme test case.

The GD mean square in the ANOVA, the smallest of the interaction mean squares, yields the variance estimator $\hat{\sigma}^2 = 7458$. Treating the three factors as nominal, we compute the ASP estimator (3.12) that uses for each factor the flat annihilator (cf. Subsection 2.4). The estimated risk of the ASP estimator is 2428, about one-third the estimated risk 7458 of the unrestricted LS estimator, which here coincides with the data. With subscripts appropriate to the context, the shrinkage constants used by the ASP estimator are: $\hat{c}_G = .76$, $\hat{c}_C = .98$, $\hat{c}_D = .86$, $\hat{c}_{GC} = .50$, $\hat{c}_{GD} = .00$, $\hat{c}_{CD} = .77$, $\hat{c}_{GCD} = .25$. Because of equation (3.11), the present choice of $\hat{\sigma}^2$ forces the vanishing of $\hat{c}_{GD}$, thereby removing the gold-dentist interaction from the ASP fit while retaining varying portions of the other least squares interactions and main effects.


LS Fit to CD Means

ASP Fit to CD Means

LS Fit to GC Means

ASP Fit to GC Means

LS Fit to GD Means

ASP Fit to GD Means

Fig. 1. ASP and unrestricted least squares fits to the two-way marginal means in the three-way layout of filling hardness data, using the GD mean square to estimate variance. Each factor is nominal. All three annihilators are flat.

Averaging a three-way fit over all levels of one factor yields a two-way marginal fit that estimates, under model (2.2), the corresponding two-way array of averaged means. The second column of Fig. 1 displays, as perspective plots, the two-way marginal fits obtained from the ASP fit. In this ASP column, the CD plot indicates that hand condensation by dentists 4 and 5 was less effective in making a filling hard than other condensation-dentist pairings, if we average over gold alloys. The GC plot reveals that electromalleting and hand malletting hardened a filling more than hand condensation, regardless of the gold alloy used, if we average over dentists. The GD plot suggests that gold alloys 6, 7 and 8 (in that order) produced harder fillings than the other alloys, regardless of dentist, if we average over condensation methods. The removal of the GD interaction term by the ASP fit explains the additive structure in this last plot.

The two-way marginal fits obtained from the LS fit are displayed in the first column of Fig. 1. The CD and GC plots tell the same story as their ASP counterparts. However, the GD plot obtained from the LS fit conveys no clear message, unlike its ASP counterpart. Designed to reduce risk under model (2.2) by learning from the data, the ASP fit reveals interesting structure in the results of the dental hardness experiment. Figure 2 plots the residuals after the ASP fit just described. The Q-norm plot reveals outliers in both tails. The residual versus fit plot indicates that some larger residuals of both signs are associated with the smaller fitted values and some larger positive residuals are associated with the larger fitted values.

The value of $\hat{\sigma}^2$ quantifies the level of detail in the ASP fit that is deemed indistinguishable from noise. Because it is difficult to estimate σ^2 in this example, we consider two other plausible values of $\hat{\sigma}^2$ that bracket the value used in the foregoing analysis. The GCD mean square in the ANOVA yields the larger variance estimator $\hat{\sigma}^2 = 9969$, which yields an ASP estimator with estimated risk 1179. The shrinkage constants are now: $\hat{c}_G = .68$, $\hat{c}_C = .97$, $\hat{c}_D = .82$, $\hat{c}_{GC} = .33$, $\hat{c}_{GD} = .00$, $\hat{c}_{CD} = .70$, $\hat{c}_{GCD} = .00$. The present choice of $\hat{\sigma}^2$ forces the vanishing of $\hat{c}_{GCD}$ through (3.11). The vanishing of $\hat{c}_{GD}$ also occurs through (3.11) because the GD mean square is smaller here than the GCD mean square. The main features in the analog (not shown) of Fig. 1 change only slightly for this ASP estimator and the conclusions stated above still hold. Though somewhat larger, the residuals from this ASP fit exhibit the same patterns as in Fig. 2.

A referee drew to attention the much smaller robustified variance estimator $\hat{\sigma}^2 = 2398$ that is suggested by Table 10 in Seheult and Tukey (2001). This yields an ASP estimator with estimated risk 1878 that is closer to the raw data than either of the

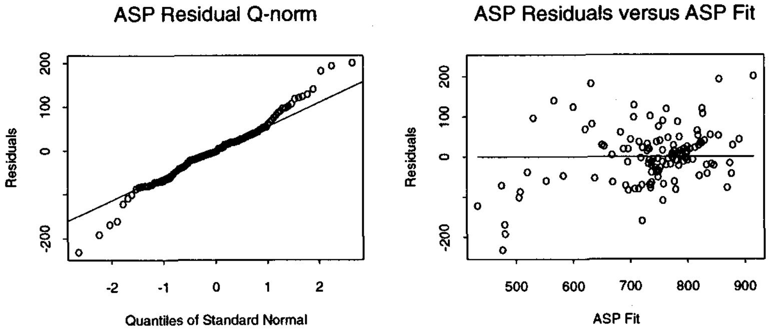
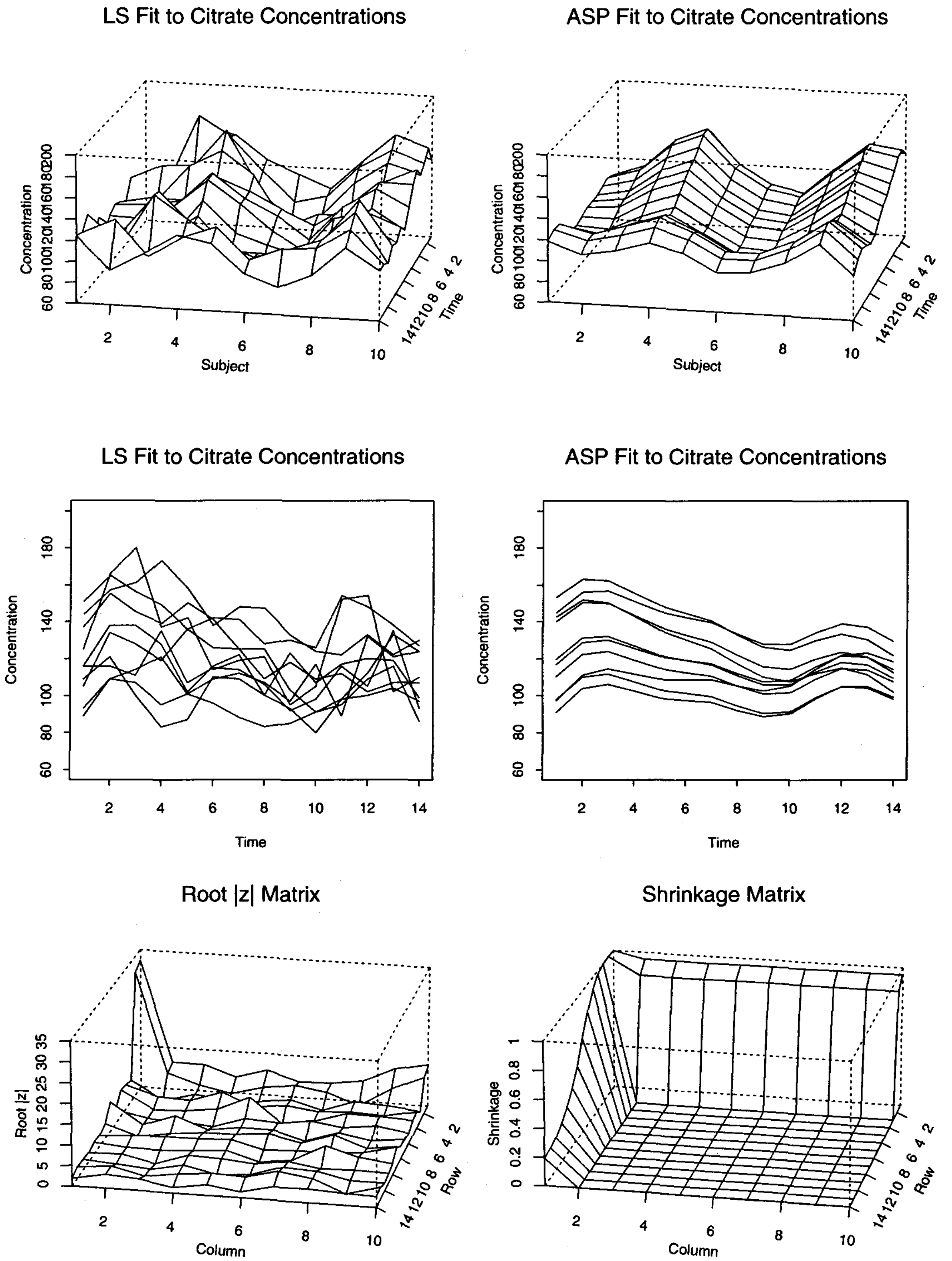


Fig. 2. Residual plots for the ASP fit described in Fig. 1.



LS Fit to Citrate Concentrations

ASP Fit to Citrate Concentrations

LS Fit to Citrate Concentrations

ASP Fit to Citrate Concentrations

Root |z| Matrix

Shrinkage Matrix

Fig. 3. ASP and unrestricted least squares fits to the two-way layout of plasma citrate concentrations. The factor subject is nominal, the factor time is ordinal, and the respective annihilators are flat and second-difference. The diagnostic plots reveal what ASP does.

preceding fits. Indeed, the shrinkage constants are now: $\hat{c}_G = .92$, $\hat{c}_C = .99$, $\hat{c}_D = .96$, $\hat{c}_{GC} = .84$, $\hat{c}_{GD} = .68$, $\hat{c}_{CD} = .93$, $\hat{c}_{GCD} = .76$. For this ASP estimator, the first two entries in the second column of the analog (not shown) of Fig. 1 do not change qualitatively but the third entry now shows moderate departures from additivity in the ASP fitted GD means. Though somewhat smaller, the residuals from this ASP fit exhibit the same patterns as in Fig. 2. Note that smaller residuals do not indicate merit in a fit. The residuals of the LS fit, which coincides with the raw data, vanish entirely.

For all three variance estimators considered, the computed ASP shrinkage constants indicate that the G, C, D main effects and the CD, GC interactions are important. The third fit retains more of the other interactions. Nevertheless, major features of the fitted CD, GC, and GD plots that were described above for the first ASP fit are exhibited by all three ASP fits.

4.2 *Concentration of plasma citrate*

Andersen *et al.* (1981) analyzed a two-factor experiment that is also recorded as data-set 41 in Andrews and Herzberg (1985). For each of 10 subjects, the concentration of citrate in plasma was measured hourly (in μmol per liter) at 14 times from 8 to 21 hours Meals were given at 8, 12, and 17 hours. We treat this data as a two-way layout in which the first factor subject is nominal and the second factor time is ordinal, using the flat annihilator for the nominal factor and the r-th difference annihilator for the ordinal factor (cf. Subsection 2.4). Note that the r-th difference annihilator generates a penalty term that favors a fit of local polynomial degree $r - 1$ to the ordinal factor.

To obtain a variance estimate, we take $r = 2$, which favors a locally linear fit to the ordinal factor, a plausible a priori choice for the plasma data. The variance estimate $\hat{\sigma}^2 = 131.9$ is then obtained from the high-component variance estimator (3.6) with $L = \{(i, j) : 5 \leq i \leq 10, 10 \leq j \leq 14\}$. This choice of L is informal, motivated by the plot of $\{|z_{i,j}|^{1/2}\}$ in Fig. 3.

The candidate PLS estimators considered let the penalty weights range freely and let the differencing order r of the annihilator for the ordinal factor range from 1 to 4. For each r, the estimated risk of the shrinkage adaptive PLS estimator (which chooses penalty weights to minimize estimated risk) is: 22.6 for $r = 1$, 21.2 for $r = 2$; 29.2 for $r = 3$, and 29.4 for $r = 4$. The foregoing value of $\hat{\sigma}^2$ is used to calculate these estimated risks. The ASP estimator that minimizes estimated risk over the candidate PLS estimators thus takes $r = 2$. Note that the estimated risk 21.1 of this ASP estimator is about one sixth the estimated risk 131.9 of the unrestricted least squares estimator but is only slightly smaller than the estimated risk 22.6 of the adaptive PLS estimator induced by $r = 1$.

As the interpolated plots of Fig. 3 reveal, the ASP fit achieves its relatively small estimated risk in two ways: through near additivity in the two factors and through greater smoothness than the least squares fit. Unlike the highly irregular data, the ASP fit plainly shows a common diurnal pattern in mean plasma citrate concentration as a function of time. For each subject, the highest mean citrate concentration is at 9 and 10 hours; the lowest concentration is at 16 and 17 hours. The meal times do not explain the fitted pattern. The analog (not shown) of Fig. 3 for the shrinkage adaptive PLS estimator with $r = 1$ is very similar.

Acknowledgements

The author wishes to thank an anonymous referee for stimulating comments and two references.

References

Andersen, A. H., Jensen, E. B. and Schou, G. (1981). Two-way analysis of variance with correlated errors, *International Statistical Review*, **49**, 153–167.

Andrews, D. F. and Herzberg, A. M. (1985). *Data*, Springer, New York.

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

Bril, G., Dykstra, R., Pillers, C. and Robertson, T. (1984). Isotonic regression in two variables, *Journal of the Royal Statistical Society (C)*, **33**, 352–357.

Brown, M. B. (1975). Exploring interaction effects in the ANOVA, *Applied Statistics*, **24**, 288–298.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown wavelet shrinkage, *Journal of the American Statistical Association*, **90**, 1200–1224.

Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing, *Journal of the Royal Statistical Society (B)*, 299–315.

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proceedings Fourth Berkeley Symposium Mathematical Statistics and Probability*, **1**, 361–380, University of California Press, Berkeley.

Kimeldorf, G. S. and 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.

Li, K.-C. and Hwang, J. T. (1984). The data-smoothing aspect of Stein estimates, *Annals of Statistics*, **12**, 887–897.

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 of Information Transmission*, **16**, 120–133.

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

Seheult, A. H. and Tukey, J. W. (2001). Towards robust analysis of variance, *Data Analysis from Statistical Foundations* (eds. A. K. Mohammed and E. Saleh), 217–244, Nova Science Publishers, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probabability* (ed. J. Neyman), **1**, 197–206, University of California Press, Berkeley.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs, *Festschrift for Jerzy Neyman* (ed. F. N. David), 351–364, Wiley, New York.

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

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society (B)*, **62**, 413–428.

Xhonga, F. (1971). Direct gold alloys—Part II, *Journal of the American Academy of Gold Foil Operators*, **14**, 5–15.