IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 33–46
© Institute of Mathematical Statistics, 2013
DOI: 10.1214/12-IMSCOLL904

Multivariate regression through affinely weighted penalized least squares

Rudolf Beran

University of California, Davis

Abstract: Stein's [In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1** (1956) 197–206 University of California Press] asymptotically superior shrinkage estimator of p univariate means may be rederived through adaptive penalized least squares (PLS) estimation in a regression model with one nominal covariate. This paper treats adaptive PLS estimators for p unknown d-dimensional mean vectors, each of which depends on k_0 scalar covariates that may be nominal or ordinal. The initial focus is on complete regression designs, *not necessarily balanced*: for every k_0-tuple of possible covariate values, at least one observation is made on the corresponding mean vector. The results include definition of suitable candidate classes of PLS estimators in multivariate regression problems, comparison of these candidate estimators through their estimated quadratic risks under a general model that makes no assumptions about the regression function, and supporting asymptotic theory as the number p of covariate-value combinations observed tends to infinity. Empirical process theory establishes that: (a) estimated risks converge to their intended targets *uniformly* over large classes of candidate PLS estimators; (b) a candidate PLS estimator that minimizes estimated risk within such a class has risk that converges to the minimal possible risk within the class. Extension of the results to incomplete regression designs is outlined. The Efron-Morris [*Journal of the American Statistical Association* **68** (1973) 117–130] and Beran [*Annals of the Institute of Statistical Mathematics* **60** (2008) 843–864] estimators for multivariate means in balanced complete designs are seen as special cases.

There arises a problem of finding the reasons for applicability of the mathematical theory of probability to the phenomena of the real world. — A. N. Kolmogorov

1. Introduction

This paper treats estimation in a very general multivariate regression model that relates d-dimensional mean responses to k_0 scalar covariates. To define the model, consider first the multivariate linear model

$$Y = CM + E, \tag{1.1}$$

where M is a $p \times d$ matrix whose rows are unknown mean vectors, Y is a $n \times d$ data matrix, each of whose rows is an observation on one of the mean vectors, and E is an $n \times d$ matrix of random errors with means zero. We assume for now that the design is complete: there is a least one row in Y that constitutes an observation

[1] Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA, e-mail: rjberan@ucdavis.edu

AMS 2000 subject classifications: Primary 62J07, 62F12; secondary 62H12

Keywords and phrases: unbalanced design, quadratic risk, adaptation, multiple penalties

with error on each row in M. Thus $n \geq p$. The $n \times p$ matrix C is the data-incidence matrix of zeroes and ones that suitably maps the rows of M into the rows of Y. Thus, C is of full rank p and $C'C$ is a diagonal matrix that reports the number of observations taken on the successive rows of M.

In fixed-effects regression, each d-dimensional row vector in M depends on k_0 non-random covariates. The k-th covariate, where $1 \leq k \leq k_0$, assumes p_k distinct real values $x_{k,1} < x_{k,2} < \cdots x_{k,p_k}$. 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 $1 \leq k \leq k_0$. These k_0-tuples index all possible combinations of the covariate values. Without loss of generality, we order the $p = \prod_{k=1}^{k_0} p_k$ elements of 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.

Let

$$x_i = (x_{1,i_1}, x_{2,i_2}, \ldots, x_{k_0,i_{k_0}}), \qquad i \in \mathcal{I}. \tag{1.2}$$

The $1 \times d$ row vectors of M have the form

$$M_i = f(x_i), \qquad i \in \mathcal{I}, \tag{1.3}$$

where f is an *unknown* function. This statement puts no restrictions in the row vector M_i beyond asserting that the covariate-value combination x_i determines its value. Without any loss of generality, the $p \times d$ mean matrix M in (1.1) is obtained by stacking the $\{M_i\}$ vertically as i runs through successive values of the ordered index set $\mathcal{I}$.

The *strong Gauss-Markov regression model* for the data Y consists of the preceding three paragraphs plus the assumption that the elements of the error matrix E are independent, identically distributed random variables with variance 1 and finite fourth moment. Because the value of the matrix M is unrestricted by (1.3), it is natural to estimate M by the *least squares estimator* $\hat{M}_{ls} = C^+Y = (C'C)^{-1}C'Y$. On the other hand, to gain insight into the function f in (1.3) and to improve (hopefully) on the risk of $\hat{M}_{ls}$ through unarticulated bias-variance trade-off, applied statisticians have often restricted f to relatively simple function classes before fitting the model.

This paper develops a formal risk improvement strategy that puts *no* restrictions on f or M: devise a large class of candidate penalized least squares estimators for M that tentatively express competing suppositions about f; and let the data choose the candidate estimator that has smallest estimated risk. All risk and estimated risk calculations are done under the *unrestricted* model where f is completely unknown. This point of view—keeping the model as general as is feasible—may be traced back to Stein [12]. The introduction to his paper treats the asymptotic risks of competing shrinkage estimators for p univariate means, as p tends to infinity, without imposing any restrictions on the unknown mean vector.

Section 7 sketches how the assumption of a complete design and the assumptions on the errors E may be weakened. These extensions matter for data analysis. Section 8 develops a data example.

2. Candidate penalized least squares estimators

Let $y = \text{vec}(Y)$, $m = \text{vec}(M)$, $e = \text{vec}(E)$ and $\tilde{C} = I_d \otimes C$, where I_d is the $d \times d$ identity matrix. Equivalent to (1.1) is the equation

$$y = \tilde{C}m + e. \tag{2.1}$$

The least squares estimator of m is $\hat{m}_{ls} = \tilde{C}^{+}y = \mathrm{vec}(\hat{M}_{ls})$.

Let $\mathcal{S}$ be an index set of fixed finite cardinality. Let $\{Q_s \colon s \in \mathcal{S}\}$ be a set of $p \times p$ symmetric, positive semidefinite *penalty matrices*. Useful constructions of these are described later in Section 4. Let $N = \{N_s \colon s \in \mathcal{S}\}$ be a set of $d \times d$ symmetric, positive semidefinite *affine penalty weights*.

Let $|\cdot|$ denote the Frobenius matrix norm: $|A|^2 = \mathrm{tr}(A'A) = \mathrm{tr}(AA')$ for any matrix A. The *penalized least squares criterion* for estimating m is

$$G(m, N) = |y - \tilde{C}m|^2 + m'Q(N)m, \qquad Q(N) = \sum_{s \in \mathcal{S}} (N_s \otimes Q_s). \tag{2.2}$$

Equivalently,

$$G(m, N) = |Y - CM|^2 + \sum_{s \in \mathcal{S}} |Q_s^{1/2} M N_s^{1/2}|^2. \tag{2.3}$$

Factorizations of Q_s and N_s other than matrix square roots may be substituted into (2.3).

The *candidate penalized least squares (PLS) estimator* of m is

$$\hat{m}_{pls}(N) = \operatorname*{argmin}_{m \in R^{pd}} G(n, N) = \big[\tilde{C}'\tilde{C} + Q(N)\big]^{-1} \tilde{C}'y. \tag{2.4}$$

The inverse exists because C and hence $\tilde{C}$ are of full rank. For each choice of $Q(N)$, the candidate $\hat{m}_{pls}(N)$ has the form of a generalized ridge estimator.

Let $\eta = \tilde{C}m = \mathrm{E}(y)$, the expectation being taken under the strong Gauss-Markov model described in Section 1. The *candidate PLS estimator* of η is

$$\hat{\eta}_{pls}(N) = \tilde{C}\hat{m}_{pls}(N) = \tilde{C}\big[\tilde{C}'\tilde{C} + Q(N)\big]^{-1} \tilde{C}'y. \tag{2.5}$$

For each choice of $Q(N)$, the candidate $\hat{\eta}_{pls}(N)$ is a symmetric linear estimator of η in the sense of Buja, Hastie and Tibshirani [5].

For univariate means and observations, when $d = 1$, the affine penalty weights $\{N_s\}$ become non-negative scalars. The foregoing definitions then yield candidate PLS estimators for univariate means using multiple quadratic penalties weighted by non-negative weights. In general, univariate PLS estimators based on one or more quadratic penalty terms are biased. We take the purpose of biased estimation to be trade-off between bias and variance so as to reduce quadratic risk in estimating η. Studies of biased linear estimators for η when $d = 1$ include ridge regression (Hoerl and Kennard [9]), shrinkage estimators for complete balanced multi-way layouts with nominal covariates (Stein [13]), monotone shrinkage estimators for abstract one-way layouts (Beran and Dümbgen [4]), ordered linear smoothers (Kneip [10]), fitting a smooth trend factor in field experiments (Green, Jennison and Seheult [8]), and multiple penalty PLS estimators in Wood [16] and in Beran [1, 2]. The scope of penalized least squares estimation goes well beyond smoothing over ordinal covariates. For example, through suitable construction of the penalty matrices, Beran [1] closely approximated, by an adaptive PLS estimator, Stein's [13] superior shrinkage estimator for a complete balanced multi-way ANOVA layout.

When $d > 1$, non-negative scalar penalty weights are naturally replaced with positive semidefinite penalty matrices, as in (2.2) and (2.3). Results on these multivariate PLS estimators of m and η for $d > 1$ seem sparse. Beran [3] studied, in effect, the special case where the design is complete with an equal number of observations taken on each d-dimensional mean vector and the penalty matrices $\{Q_s\}$ are mutually orthogonal, orthogonal projections. In this scenario, the inverse matrix in (2.4) and (2.5) can be calculated algebraically and the candidate PLS estimators have an elementary form that greatly facilitates their study and use. The present paper considers more general penalty matrices and unbalanced designs. Treated are constructions of penalty matrices, adaptive choice of the affine penalty weights by minimizing estimated risk, and asymptotics that support such adaptation.

3. Loss, risk, estimated risk, and adaptation

Let $\hat{\eta}$ be any estimator of $\eta = \mathrm{E}(y) = \tilde{C}m$. Define the normalized quadratic loss of any estimator $\hat{\eta}$ of η to be

$$L(\hat{\eta}, \eta) = p^{-1}|\hat{\eta} - \eta|^2. \tag{3.1}$$

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

$$R(\hat{\eta}, \eta) = \mathrm{E}L(\hat{\eta}, \eta), \tag{3.2}$$

where the expectation is calculated under the strong Gauss-Markov model of Section 1. The least squares estimator $\hat{\eta}_{ls} = \tilde{C}\hat{m}_{ls} = \tilde{C}\tilde{C}^{+}y$ has risk $R(\hat{\eta}_{ls}, \eta) = d$.

The candidate PLS estimator $\hat{\eta}_{pls}(N)$ can be put into a canonical form that assists subsequent analysis. Recall that $\tilde{C} = I_d \otimes C$ is of full rank pd. Let

$$\tilde{U} = \tilde{C}\big(\tilde{C}'\tilde{C}\big)^{-1/2}. \tag{3.3}$$

Evidently $\tilde{U}$ is $nd \times pd$ with $\tilde{U}'\tilde{U} = I_{pd}$ and $\tilde{C}'\tilde{C}$ is a $pd \times pd$ diagonal matrix whose diagonal elements record systematically the number of observations on each component of m. Because $\tilde{C}$ and $\tilde{U}$ have the same range space, $\eta = \tilde{C}m = \tilde{U}\xi$, with $\xi = \tilde{U}'\eta$. Let $z = \tilde{U}'y$. From (2.5) follows a *canonical form* for $\hat{\eta}_{pls}(N)$:

$$\hat{\eta}_{pls}(N) = \tilde{U}S(N)z, \qquad S(N) = \big[I_{pd} + \big(\tilde{C}'\tilde{C}\big)^{-1/2}Q(N)\big(\tilde{C}'\tilde{C}\big)^{-1/2}\big]^{-1}. \tag{3.4}$$

The matrix $S(N)$ is symmetric with eigenvalues in $[0, 1]$. Its effect in (3.4) is to transform z to a new orthogonal coordinate system, shrink the coefficients in that system toward zero, then transform back to the z coordinate system.

The quadratic *loss* of candidate estimator $\hat{\eta}_{pls}(N)$ is thus

$$L\big(\hat{\eta}_{pls}(N), \eta\big) = p^{-1}|\hat{\eta}_{pls}(N) - \eta|^2 = p^{-1}|S(N)z - \xi|^2. \tag{3.5}$$

Let $T(N) = S^2(N)$ and $\bar{T}(N) = [I_{pd} - S(N)]^2$. From (3.5), the *risk* of candidate estimator $\hat{\eta}_{pls}$ is

$$r(N, \eta) = \mathrm{E}L\big(\hat{\eta}_{pls}(N), \eta\big) = p^{-1}\operatorname{tr}\big[T(N) + \bar{T}(N)\xi\xi'\big]. \tag{3.6}$$

The strong Gauss-Markov assumptions made on the error matrix E entail that $\mathrm{E}(z) = 0$ and $\mathrm{Cov}(z) = I_{pd}$. An unbiased estimator of $\xi\xi'$ is $zz' - I_{pd}$. As in Mallows [11], this circumstance motivates estimating the risk $r(N, \eta)$ by the *estimated risk*

$$\hat{r}(N) = p^{-1}\operatorname{tr}\big[T(N) + \bar{T}(N)\big(zz' - I_{pd}\big)\big]. \tag{3.7}$$

Let $\mathcal{N}$ denote the set of penalty weight matrices $N = \{N_s \colon s \in \mathcal{S}\}$ under consideration. For a rich choice of $\mathcal{N}$, the estimated risk converges, as p tends to infinity, to both the risk and loss of $\hat{\eta}_{pls}(N)$, *uniformly* over all $N \in \mathcal{N}$. This point is developed later in Section 6.

Uniform convergence justifies using estimated risks as surrogates for risks in selecting the penalty weight matrices. The *adaptive PLS estimators* of η and m are defined to be

$$\hat{\eta}_{apls} = \hat{\eta}_{pls}(\hat{N}), \qquad \hat{m}_{apls} = \hat{m}_{pls}(\hat{N}), \qquad \text{where } \hat{N} = \operatorname*{argmin}_{N \in \mathcal{N}} \hat{r}(N). \tag{3.8}$$

4. Penalty matrices

This section describes a construction that penalizes separately the mean effects and interactions in the MANOVA decomposition of M. The $\{Q_s \colon s \in \mathcal{S}\}$ are defined as tensor-product penalty matrices that, suitably weighted, 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 covariates.

4.1. The MANOVA decomposition

The following algebra gives the orthogonal projections that define the multivariate analysis of variance (MANOVA) 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

$$J_k = u_k u_k', \qquad H_k = I_{p_k} - u_k u_k'. \tag{4.1}$$

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

$$P_{s,k} = \begin{cases} J_k & \text{if } k \notin s, \\ H_k & \text{if } k \in s. \end{cases} \tag{4.2}$$

Define the $p \times p$ Kronecker product matrix

$$P_s = \bigotimes_{k=1}^{k_0} P_{s, k_0 - k + 1}. \tag{4.3}$$

The foregoing discussion implies that:

- P_s is symmetric, idempotent for every $s \in \mathcal{S}$.
- 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 $\mathcal{S}$, 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 $\{P_s \colon s \in \mathcal{S}\}$ are orthogonal projections that decompose R^p into 2^{k_0} mutually orthogonal subspaces.

The MANOVA decomposition of M is the identity

$$M = \sum_{s \in \mathcal{S}} P_s M. \tag{4.4}$$

Here $P_\emptyset M$ is the overall mean term in the decomposition. 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 classical MANOVA are defined by constraining M to satisfy $P_s M = 0$ for every s in a specified subset of $\mathcal{S}$. The choice of the subset identifies the main effects or interaction terms that vanish in the submodel.

4.2. A class of tensor-product penalty matrices

An *annihilator* for factor k is a matrix A_k with p_k columns such that $A_k u_k = 0$. The rows of A_k are contrasts, selected to quantify departures from hypothetical dependence of the means on the levels of covariate k. How to devise the $\{A_k\}$ suitably for ordinal and nominal covariates is addressed in a Section 5. Here we describe how to build tensor-product penalty matrices $\{Q_s\}$ once the covariate annihilators $\{A_k\}$ have been chosen.

Let

$$Q_{s,k} = \begin{cases} J_k & \text{if } k \notin s, \\ A_k' A_k & \text{if } k \in s, \end{cases} \tag{4.5}$$

a $p_k \times p_k$ matrix. Define the $p \times p$ Kronecker product matrix

$$Q_s = \bigotimes_{k=1}^{k_0} Q_{s,k_0-k+1}. \tag{4.6}$$

Note that the MANOVA projection P_s defined in (4.3) arises as the special case of Q_s when each $A_k = H_k$.

Suppose that the spectral representation of $A_k' A_k$ is

$$A_k' A_k = \sum_{j=1}^{p_k} \lambda_{k,j} \pi_{k,j}, \tag{4.7}$$

where the $\{\pi_{k,j}\}$ are rank 1 eigenprojections corresponding to the eigenvalues $0 = \lambda_{k,1} \le \cdots \le \lambda_{k,p_k}$. Note that $\pi_{k,1} = u_k u_k'$. From (4.7) and (4.5),

$$Q_{s,k} = \sum_{j=1}^{p_k} \gamma_{s,k,j} \pi_{k,j}, \tag{4.8}$$

where

$$\gamma_{s,k,j} = \begin{cases} 1 & \text{if } j = 1 \text{ and } k \notin s, \\ 0 & \text{if } j \ge 2 \text{ and } k \notin s, \\ \lambda_{k,j} & \text{if } k \in s. \end{cases} \tag{4.9}$$

Recall from Section 1 the index set $\mathcal{I}$ of ordered k_0-tuples

$$\mathcal{I} = \{\ldots\{\{(i_1, i_2, \ldots, i_{k_0})\colon 1 \le i_1 \le p_1\}, 1 \le i_2 \le p_2\}, \ldots, 1 \le i_{k_0} \le p_{k_0}\}. \tag{4.10}$$

For every $s \in \mathcal{S}$, let $\mathcal{I}_s = \{i \in \mathcal{I}\colon i_k = 1 \text{ if } k \notin s \text{ and } i_k \geq 2 \text{ if } k \in s\}$. Thus, $i = (i_1, i_2, \ldots, i_{k_0}) \in \mathcal{I}_s$ if and only if $s = \{k\colon i_k \geq 2\}$. 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 every $i \in \mathcal{I}$, let

$$\pi_i = \bigotimes_{k=1}^{k_0} \pi_{k_0-k+1,i_k}, \qquad \gamma_{s,i} = \prod_{k=1}^{k_0} \gamma_{s,k_0-k+1,i_k}. \tag{4.11}$$

Then,

$$Q_s = \bigotimes_{k=1}^{k_0} Q_{s,k_0-k+1} = \sum_{i\in\mathcal{I}_s} \gamma_{s,i}\pi_i, \tag{4.12}$$

where, for every $i \in \mathcal{I}_s$,

$$\gamma_{s,i} = \begin{cases} 1 & \text{if } s = \emptyset, \\ \prod_{k\in s} \lambda_{s,i_k} & \text{if } s \neq \emptyset. \end{cases} \tag{4.13}$$

This is a spectral representation of Q_s.

Note that the $\{\pi_i\colon i \in \mathcal{I}\}$ are mutually orthogonal rank 1 projections with $\sum_{i\in\mathcal{I}} \pi_i = I_p$. Moreover, the MANOVA projection P_s defined in (4.3) has the property $P_s = \sum_{i\in\mathcal{I}_s} \pi_i$, a spectral representation. From these facts and the spectral representation of Q_s in (4.12) follows

$$P_s Q_s^{1/2} = Q_s^{1/2} P_s = Q_s^{1/2} \quad \text{and} \quad P_{s_1} Q_{s_2} = Q_{s_2} P_{s_1} = 0 \quad \text{if } s_1 \neq s_2. \tag{4.14}$$

From this and the MANOVA decomposition (4.4) of M, it follows that the quadratic penalty in the PLS criterion (4.4) satisfies

$$\sum_{s\in\mathcal{S}} |Q_s^{1/2} M N_s^{1/2}|^2 = \sum_{s\in\mathcal{S}} |Q_s^{1/2}(P_s M) N_s^{1/2}|^2. \tag{4.15}$$

Thus, the tensor-product penalty matrix Q_s defined in this section acts only on the MANOVA component $P_s M$.

4.3. Simplification for balanced complete designs

In a balanced complete design, $n_0 \geq 1$ observations are taken for each of the covariate-value combinations $\{x_i\colon i \in \mathcal{I}\}$. Consequently, $C'C = n_0 I_p$. The least squares estimator of m is then

$$\hat{m}_{ls} = \left(\tilde{C}'\tilde{C}\right)^{-1} \tilde{C}' y = n_0^{-1} \tilde{C}' y. \tag{4.16}$$

Using (2.4),

$$\hat{m}_{pls}(N) = \left[\tilde{C}'\tilde{C} + Q(N)\right]^{-1} \tilde{C}' y = \left[I_{pd} + n_0^{-1} Q(N)\right]^{-1} \hat{m}_{ls}. \tag{4.17}$$

The spectral representation (4.12) of Q_s, the definition of $Q(N)$ in (2.2), and the previously noted fact $\sum_{i\in\mathcal{I}} \pi_i = I_p$ yield

$$I_{pd} + n_0^{-1} Q(N) = \sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{I}_s} \left[\left(I_d + n_0^{-1}\gamma_{s,i} N_s\right) \otimes \pi_i\right]. \tag{4.18}$$

Hence, for a balanced complete design,

$$\hat{m}_{pls}(N) = \sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{I}_s} \big[\big(I_d + n_0^{-1}\gamma_{s,i} N_s\big)^{-1} \otimes \pi_i\big]\hat{m}_{ls}. \tag{4.19}$$

In matrix form,

$$\hat{M}_{pls}(N) = \sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{I}_s} \pi_i \hat{M}_{ls}\big(I_d + n_0^{-1}\gamma_{s,i} N_s\big)^{-1}. \tag{4.20}$$

This class of candidate PLS estimators enriches significantly the class of multiple affine shrinkage estimators studied in Beran [3] because (4.20) has p summands, where p goes to infinity in the asymptotics to be described in Section 6. Nevertheless, the risk and estimated risk of $\hat{M}_{pls}(N)$ still retain the additive structure seen in that paper, thereby greatly simplifying construction of the adaptive estimator $\hat{M}_{apls}$. For large p, the Efron-Morris [7] estimator coincides with a special case of $\hat{M}_{apls}$ in a balanced complete one-way layout: Q_1 is set to be the identity matrix. Beran ([3], p. 854) gives details.

5. Constructing covariate annihilators

To complete the construction of candidate PLS estimators, it remains to devise annihilators that work in the world of data. The guiding idea behind construction of annihilator A_k is this: nearly zero values of $A_k M$ should express plausible restrictions on how M depends on covariate k. It will be necessary to distinguish between nominal covariates and ordinal covariates.

Covariate k is nominal. The values of a nominal covariate are labels that can be permuted freely without loss of information. The corresponding candidate PLS estimators should therefore be invariant under permutations of nominal levels. This consideration prompts setting $A_k = H_k$ for every k, the latter being defined in (4.1). This choice of A_k will be called the *flat annihilator* for covariate 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 (4.5) 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 $s \in \mathcal{S}$. In this case, the candidate PLS estimator $\hat{\eta}_{pls}(N)$ interpolates among all possible MANOVA submodel fits to the complete layout. The MANOVA submodel fits themselves are limit points of this set of candidate PLS estimators.

Covariate k is ordinal. Suppose first that the ordered values of covariate k, arranged as the column vector $c_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,p_k})'$, are equally spaced. To have the candidate PLS estimator $\hat{\eta}_{pls}(N)$ favor a fit that is locally polynomial of degree $h_0 - 1$ in the values of covariate 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

$$D(1, p_k) = \Delta(p_k), \quad D(h, p_k) = \Delta(p_k - h + 1) D(h-1, p_k) \quad \text{for } 2 \le h \le p_k - 1. \tag{5.1}$$

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

$$A_k c_k^h = 0 \qquad \text{for } 0 \le h \le h_0 - 1. \tag{5.2}$$

The notation c_k^h denotes the vector $(x_{k,1}^h, x_{k,2}^h, \ldots, x_{k,p_k}^h)'$. 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 c_k are not necessarily equally spaced, we define A_k as follows. The h_0-th 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 i-th 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 (5.2). Third, each row vector in A_k has unit length. These requirements are met by setting the non-zero elements in the i-th 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 $(x_{k,i}, \ldots, x_{k,i+h_0})$. The R function `poly` accomplishes this computation.

When the components of c_k are equally spaced, this construction of A_k reduces to a multiple of the h_0-th difference annihilator described in the preceding paragraph.

6. Asymptotic theory

The two theorems in this section concern the loss, risk, and estimated risk of the estimators $\hat{\eta}_{pls}(N)$ and $\hat{\eta}_{apls}$ defined in (3.4) and (3.8) respectively. We ask the reader to recognize that almost every quantity in this paper depends on p. To avoid burdensome notation, we generally omit the subscript p.

The notation for loss, risk and estimated risk is that in Section 3. The affine penalty weights $N = \{N_s \colon s \in \mathcal{S}\}$ are restricted to positive semidefinite symmetric matrices with bounded spectral norm. To this end, for every finite $b > 0$, define

$$(6.1) \qquad N \in \mathcal{N}(b) \qquad \text{if and only if} \qquad \max_{s \in \mathcal{S}} |N_s|_{spec} \le b.$$

Theorem 6.1. *Assume that the strong Gauss-Markov model holds. Let $W(N)$ denote either the loss $L(\hat{\eta}_{pls}(N), \eta)$ or the estimated risk $\hat{r}(N)$ of $\hat{\eta}_{pls}(N)$. For every finite $a > 0$ and $b > 0$,*

$$(6.2) \qquad \lim_{p \to \infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E} \sup_{N \in \mathcal{N}(b)} |W(N) - r(N, \eta)| = 0.$$

Proof idea. The argument has three steps: First is to show that, for every $N \in \mathcal{N}(b)$, the difference $W(N) - r(N, \eta)$ converges in probability to zero as p tends to infinity. Second, using empirical process theory, is to show that $\sup_{N \in \mathcal{N}(b)} |W(N) - r(N, \eta)|$ converges in probability to zero. Third is to use uniform integrability to establish (6.2). The condition that $|N_s|_{spec} \le b$ ensures that every element of N_s lies in the compact interval $[-b, -b]$. The full proof is an algebraically detailed extension of the argument given for the univariate case in Theorem 3 of Beran [2]. □

Theorem 6.1 shows that the loss, risk, and estimated risk of candidate estimator $\hat{\eta}_{pls}(N)$ converge together as p tends to infinity. The uniformity of this convergence over all $N \in \mathcal{N}(b)$ makes estimated risk a trustworthy surrogate for either loss or risk when consulting the data regarding a good choice of the affine penalty weights N. The next theorem expresses this point formally.

Refining definition (3.8) for the theorem statement, let $\hat{N} = \operatorname{argmin}_{N \in \mathcal{N}(b)} \hat{r}(N)$ and the adaptive estimator $\hat{\eta}_{apls} = \hat{\eta}_{pls}(\hat{N})$. Let $\tilde{N} = \operatorname{argmin}_{N \in \mathcal{N}(b)} r(N, \eta)$ denote the corresponding oracle choice of $N \in \mathcal{N}(b)$.

Theorem 6.2. *Assume that the strong Gauss-Markov model holds. For every finite $a > 0$ and $b > 0$,*

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} |R(\hat{\eta}_{apls}, \eta) - r(\tilde{N}, \eta)| = 0. \tag{6.3}$$

Moreover, if V denotes either the loss $L(\hat{\eta}_{apls}, \eta)$ or risk $R(\hat{\eta}_{apls}, \eta)$ of $\hat{\eta}_{apls}$, then

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{r}(\hat{N}) - V| = 0. \tag{6.4}$$

Proof. We show that (6.2) in Theorem 6.1 implies

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|Z - r(\tilde{N}, \eta)| = 0, \tag{6.5}$$

where Z can be $L(\hat{\eta}_{apls}, \eta)$ or $\hat{r}(\hat{N})$. The three limits to be proved in (6.3) and (6.4) follow immediately from (6.5).

First, (6.2) with $W(N) = \hat{r}(N)$ entails

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{r}(\hat{N}) - r(\tilde{N}, \eta)| = 0, \quad \lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{r}(\hat{N}) - r(\hat{N}, \eta)| = 0. \tag{6.6}$$

Hence, (6.5) holds for $Z = \hat{r}(\hat{N})$ and

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|r(\hat{N}, \eta) - r(\tilde{N}, \eta)| = 0. \tag{6.7}$$

Second, (6.2) with $W(N) = L(\hat{\eta}_{pls}(N), \eta)$ gives

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|L\big(\hat{\eta}_{pls}(\hat{N}), \eta\big) - r(\hat{N}, \eta)| = 0. \tag{6.8}$$

Because $\hat{\eta}_{apls} = \hat{\eta}_{pls}(\hat{N})$, this and (6.7) establish the remaining case in (6.5). □

By (6.3), the risk of the adaptive PLS estimator $\hat{\eta}_{apls}$ converges to the risk of the oracle estimator $\hat{\eta}_{pls}(\tilde{N})$, which achieves minimum risk over the class of candidate estimators $\{\hat{\eta}_{pls}(N) \colon N \in \mathcal{N}_b\}$. By (6.4), the plug-in risk estimator $\hat{r}(\hat{N})$ converges to the actual risk or loss of $\hat{\eta}_{apls}$. Through $\hat{r}(\hat{N})$, we estimate the extent to which adaptation over the class of candidate PLS estimators reduces risk for the data on hand.

7. Extensions

We outline two extensions of the theory in this paper that are desirable to make adaptive PLS estimators useful for data analysis. The first addresses correlation within observation vectors, which is seen in the data treated in Section 8. The second treats multivariate regression in *incomplete* layouts, when observations are not available for all combinations of the covariate values.

7.1. Correlation and heteroscedasticity within observation vectors

More realistic for data analysis than model (1.1) is

$$Y = CM + E\Sigma^{1/2}, \tag{7.1}$$

where Σ is a $d \times d$ positive definite covariance matrix and the error matrix E is as before. In this model, the rows of Y each have covariance matrix Σ. In the vectorized notation of Section 2, model (7.1) is equivalent to

$$y_\Sigma = \eta_\Sigma + e, \tag{7.2}$$

where

$$y_\Sigma = \big(\Sigma^{-1/2} \otimes I_p\big)y, \qquad \eta_\Sigma = \big(\Sigma^{-1/2} \otimes I_p\big)\eta. \tag{7.3}$$

If we treat y_Σ as the observation vector, this is the model already considered in this paper. The estimation procedure is then:

- Estimate η_Σ by the adaptive PLS estimator $\hat{\eta}_{\Sigma,apls}$ constructed from y_Σ;
- Estimate η by $\hat{\eta}_{apls} = (\Sigma^{1/2} \otimes I_p)\hat{\eta}_{\Sigma,apls}$; and m by $\hat{m}_{apls} = \tilde{C}^{+}\hat{\eta}_{pls}$.

The asymptotic theory in Section 6 maps over to this procedure when the loss function for any estimator $\hat{\eta}$ of η is

$$p^{-1}(\hat{\eta} - \eta)'\big(\Sigma^{-1} \otimes I_p\big)(\hat{\eta} - \eta) = p^{-1}|\hat{\eta}_\Sigma - \eta_\Sigma|^2, \tag{7.4}$$

where $\hat{\eta}_\Sigma = (\Sigma^{-1/2} \otimes I_p)\hat{\eta}$.

In practice, Σ is unknown and must be replaced by an estimator $\hat{\Sigma}$ to mimic the foregoing construction of $\hat{\eta}_{apls}$. Some remarks on this extension:

- If $\hat{\Sigma}$ is consistent for Σ, the asymptotics in Section 6 can be extended to show that loss and estimated risk converge together in probability. Stronger conditions on $\hat{\Sigma}$, not easily verified for practical constructions, seem necessary to obtain strict analogs of Theorems 6.1 and 6.2 that also address convergence of risk.
- The least squares estimator $\hat{\Sigma}_{ls} = (n-p)Y'(I_n - CC^{+})Y$, available when $n > p$, is consistent for Σ when $n - p$ tends to infinity.
- In the absence of adequate replication, pooling may provide a useful estimator of Σ: fit a plausible linear submodel for M by least squares and construct the least squares estimator of Σ associated with this submodel fit. This estimator will be consistent if its bias tends to zero in the asymptotics. Obviously replication is more trustworthy than pooling in estimating Σ.

7.2. Incomplete designs

Typical in multivariate regression problems with k_0 covariates is multivariate response data collected on an incomplete array, a proper subset of a complete k_0-way of covariate-value combinations. The data incidence matrix C is then not of full rank. The matrix inverse may not exist in definitions (2.4) and (2.5) of candidate PLS estimators. To handle this, we may replace the penalty matrix $Q(N)$ by the necessarily full rank matrix

$$Q_\epsilon(N) = Q(N) + \epsilon I_p, \qquad \epsilon > 0. \tag{7.5}$$

In practice, a choice such as $\epsilon = 10^{-6}$ seems to work well, shrinking the fit slightly towards zero and stabilizing the numerical computations.

The construction of adaptive PLS estimators using $Q_\epsilon(N)$ and their risk asymptotics in incomplete designs parallel those already given for complete designs. The methodology yields adaptive PLS estimators $\hat{m}_{apls}$ and $\hat{\eta}$ over the complete k_0 array—a fitted regression function that extrapolates to missing values. The risk of this estimator is assessed only at design points where observations are available. For a full discussion of the univariate case $d = 1$, see Beran [2].

8. Data example

The data matrix Y in this case study is 52×3. Row i of Y reports the grape yields harvested in three different years from row i of a vineyard with 52 rows of grapevines. The data is taken from Chatterjee, Handcock and Simonoff [6]. The grape yields, measured in lugs of grapes harvested from each row, are plotted in Figure 1, using a different plotting character for each of the three years. Both year-to-year and row-to-row changes in viticulture affect the observed yields. The analysis seeks to find possible patterns in the row harvest yields that persist across the three years. The notation used is that of the preceding theoretical sections.

It is fundamental to recognize that probability models serve, in this paper and elsewhere, as a mathematical device for testing statistical procedures on interesting fake data. The theoretical understanding of a statistical procedure so gained is akin to the role of basic science in developing medical procedures—useful and important but by no means the whole story. Experiments on lab animals and clinical trials on humans are essential in medicine. Trials of statistical procedures on data in the world around us are equally necessary in statistics. The vineyard data are not certifiably random in the sense of axiomatic probability theory. In applying adaptive PLS estimators to this data, we recall Tukey's [15] remark: "In practice, methodologies have no assumptions and deliver no certainties."

Note that the harvest data is intrinsically discrete, the vineyard rows having physical existence. The data is treated as a one-way layout with trivariate responses that depend on a single ordinal covariate (vineyard row number). Here, $p = n = 52$,

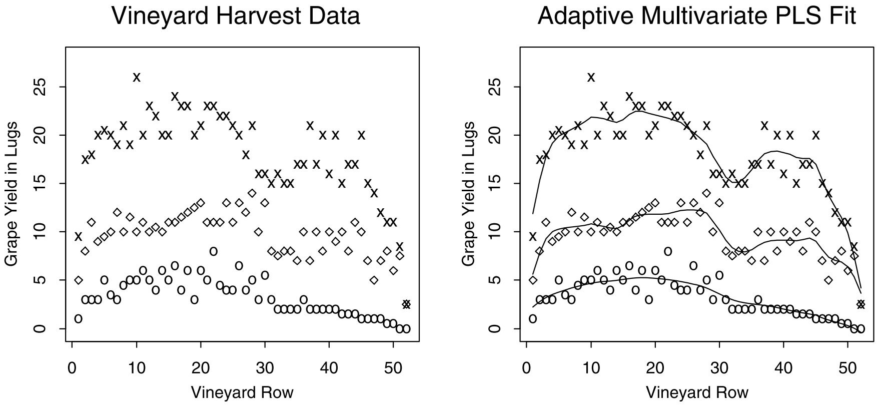

Fig 1. *Vineyard data and linearly interpolated adaptive PLS fit to the trivariate grape yields.*

$d = 3$, and $k_0 = 1$. For a one-way layout, the index set $\mathcal{S}$ defined in Section 4 is

$$\mathcal{S} = \{\emptyset, \{1\}\} \tag{8.1}$$

and

$$\mathcal{I} = \{i\colon 1 \le i \le p\}, \qquad \mathcal{I}_{\emptyset} = \{1\}, \qquad \mathcal{I}_{\{1\}} = \{i\colon 2 \le i \le p\}. \tag{8.2}$$

On the hypothetical conjecture that mean responses may vary smoothly in locally linear manner, we set the annihilator A_1 to be the second-difference matrix. The eigenvectors of $A_1'A_1$, ordered from smallest to largest eigenvalue, give the orthogonal basis O that generates the spectral representations of the two penalty matrices $\{Q_s\colon s \in \mathcal{S}\}$.

In the absence of replication, we estimate Σ from the residuals after the least squares fit of Y to the first 20 columns of O—a pooling strategy. This yields

$$\hat{\Sigma} = \begin{pmatrix} 0.994 & 0.191 & 0.160 \\ 0.191 & 1.782 & -.268 \\ 0.160 & -.268 & 3.054 \end{pmatrix}, \tag{8.3}$$

which indicates slightly correlated heteroscedasticity across the three years.

For numerical simplicity, we restrict the class of candidate PLS estimators by forcing $N_{\emptyset} = 0$. Then the candidate PLS estimators do not shrink the mean response vector. Adaptation is accomplished by minimizing estimated risk over all positive semidefinite affine penalty weights $N_{\{1\}}$. The Cholesky factorization of $N_{\{1\}}$ is a convenient parametrization in performing the minimization.

The estimated risks of $\hat{\eta}_{apls}$ and of $\hat{\eta}_{ls}$ are, respectively, 0.364 and 3. The reduction in estimated risk achieved over the least squares fit by adaptive PLS is more than eightfold. The points in Figure 1 are the harvest yields, which coincide with the least squares estimator $\hat{M}_{ls}$ for M. The solid lines in the figure join the points that represent $\hat{M}_{apls}$. They are purely a visual device to guide the eye. Large dips in $\hat{M}_{apls}$ occur in the outer vineyard rows and near row 33; smaller fluctuations occur elsewhere. These point to possible variations in soil fertility or irrigation and to climate stress on the outer vineyard rows.

References

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

[2] Beran, R. (2007). Multiple penalty regression: Fitting and extrapolating a discrete incomplete multi-way layout. *Annals of the Institute of Statistical Mathematics* **59** 171–195.

[3] Beran, R. (2008). Estimating a mean matrix: Boosting efficiency by multiple affine shrinkage. *Annals of the Institute of Statistical Mathematics* **60** 843–864.

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

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

[6] Chatterjee, S., Handcock, M. S. and Simonoff, J. S. (1995). *A Casebook for a First Course in Statistics and Data Analysis*. Wiley, New York.

[7] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130.

[8] Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society, Series B* **47** 299–315.

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

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

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

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

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

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

[15] Tukey, J. W. (1980). Methodological comments focused on opportunities. In *Multivariate Techniques in Communication Research* (P. R. Monge and J. Cappella, eds.) 489–528. Academic Press, New York.

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