Computational Statistics and Data Analysis 71 (2014) 654–666

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Hypercube estimators: Penalized least squares, submodel selection, and numerical stability

Rudolf Beran [*]

Department of Statistics, University of California, Davis, One Shields Avenue, Davis CA 95616-8705, USA

ARTICLE INFO

Article history:
Received 1 March 2012
Received in revised form 17 May 2013
Accepted 22 May 2013
Available online 27 May 2013

Keywords:
Linear model
Condition number
Estimated risk
Submodel fits
Mean arrays
Multiple shrinkage
Spline fits

ABSTRACT

Hypercube estimators for the mean vector in a general linear model include algebraic equivalents to penalized least squares estimators with quadratic penalties and to submodel least squares estimators. Penalized least squares estimators necessarily break down numerically for certain penalty matrices. Equivalent hypercube estimators resist this source of numerical instability. Under conditions, adaptation over a class of candidate hypercube estimators, so as to minimize the estimated quadratic risk, also minimizes the asymptotic risk under the general linear model. Numerical stability of hypercube estimators assists trustworthy adaptation. Hypercube estimators have broad applicability to any statistical methodology that involves penalized least squares. Notably, they extend to general designs the risk reduction achieved by Stein's multiple shrinkage estimators for balanced observations on an array of means.

1. Introduction

Consider the general linear model

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

Here, y is the $n \times 1$ vector of observations, X is a given $n \times p$ design matrix of rank $p \leq n$, β is an unknown $p \times 1$ vector of regression coefficients, and e is an $n \times 1$ vector of errors. For theoretical developments, the error vector is taken to satisfy the *strong Gauss–Markov condition*: the components of e are independent, identically distributed random variables whose unknown distribution has mean 0, variance σ^2, and finite fourth moment.

The foregoing *general model* is taken to express truth. Credibility of this assumption in a data analysis may require a design matrix X whose p is close to n. A fundamental problem is estimation, under the general model, of

$$\eta = \mathrm{E}(y) = X\beta. \tag{1.2}$$

The regression coefficients may be recovered through $\beta = (X'X)^{-1}X'\eta$. A basic issue is that the *least squares estimator* $\hat{\eta}_{\mathrm{LS}} = X(X'X)^{-1}X'y$ of η usually overfits when p is not small. Stein (1966) and earlier first expressed this phenomenon formally. If the error vector e is Gaussian and $p \geq 3$, then $\hat{\eta}_{\mathrm{LS}}$ is an *inadmissible* estimator of η under the risk function $\mathrm{E}|\hat{\eta}_{\mathrm{LS}} - \eta|^2$.

Regularized estimators of η rely on bias–variance trade-off to achieve smaller quadratic risk than $\hat{\eta}_{\mathrm{LS}}$. The risks of all such competing estimators are calculated under the general model (1.1). Two well-established regularization strategies are submodel least squares estimation, in which the estimator of η is constrained to a subspace of the range space of X, and penalized least squares estimation of η. Introduced and studied here are *hypercube estimators*, a richer class of regularized estimators of η. The key points are as follows.

[*] Tel.: +1 530 746 8284.
E-mail address: rjberan@ucdavis.edu.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.020

- *Definition of hypercube estimators of η.* Let I_p be the $p \times p$ identity matrix. Let V be any $p \times p$ symmetric matrix whose eigenvalues all lie in $[0, 1]$. The associated *hypercube estimator* of η is

$$\hat{\eta}_H(V) = A(V)y \quad \text{with } A(V) = XV(VX'XV + I_p - V^2)^{-1}VX'. \tag{1.3}$$

Evidently the least squares estimator $\hat{\eta}_{\text{LS}}$ coincides with the hypercube estimator $\hat{\eta}_H(I_p)$. Of interest is choosing the matrix V so as to minimize (approximately) the quadratic risk of $\hat{\eta}_H(V)$ under general model (1.1).

- *Penalized least squares estimators are hypercube estimators.* Let W be any $p \times p$ positive semidefinite matrix. The associated *penalized least squares (PLS) estimator* of η is

$$\hat{\eta}_{\text{PLS}}(W) = X(X'X + W)^{-1}X'y = X\hat{\beta}_{\text{PLS}}, \tag{1.4}$$

where

$$\hat{\beta}_{\text{PLS}}(W) = \underset{\beta \in R^p}{\operatorname{argmin}}[|y - X\beta|^2 + \beta'W\beta] = (X'X + W)^{-1}X'y. \tag{1.5}$$

The mapping (1.4) from $\hat{\beta}_{\text{PLS}}(W)$ to $\hat{\eta}_{\text{PLS}}(W)$ is one-to-one, because $\hat{\beta}_{\text{PLS}}(W) = (X'X)^{-1}X'\hat{\eta}_{\text{PLS}}(W)$. The matrix $V = (I_p + W)^{-1/2}$ is symmetric with all eigenvalues in $(0, 1]$. It will be shown in Section 2 that

$$\hat{\eta}_{\text{PLS}}(W) = \hat{\eta}_H((I_p + W)^{-1/2}). \tag{1.6}$$

Penalized least squares estimators thus form a large proper subset of the hypercube estimators (1.3), in which the smallest eigenvalue of V does *not* vanish.

- *Submodel least squares estimators are hypercube estimators.* Let X_0 be any matrix with n rows whose range space lies within the range space of X. Let the superscript $^+$ denote the Moore–Penrose pseudoinverse. Consider the submodel of the general model (1.1) in which η is restricted to the range space of X_0. The *submodel least squares estimator* of η under this constraint is

$$\hat{\eta}_{\text{SUB}}(X_0) = X_0X_0^+y. \tag{1.7}$$

Let $P = (X^+X_0)(X^+X_0)^+$, a symmetric idempotent matrix whose eigenvalues are either 0 or 1. It will be shown in Section 2 that

$$\hat{\eta}_{\text{SUB}}(X_0) = \hat{\eta}_H(P). \tag{1.8}$$

The submodel least squares estimators customarily considered in fitting a general linear model thus form a small proper subset of all hypercube estimators for that general model. Though the submodel least squares estimators described are not penalized least squares estimators, they are limits of certain sequences of penalized least squares estimators. This is seen most readily through their hypercube representations.

- *Numerical instability of penalized least squares estimators.* Let $|\cdot|$ denote the Euclidean norm. The *condition number* of any nonsingular symmetric matrix A is

$$\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A), \tag{1.9}$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are, respectively, the largest and smallest eigenvalues of A. Section 3 shows that, for fixed X, the condition number of the matrix $X'X + W$ to be inverted in (1.4) can be arbitrarily large as the largest eigenvalue of W tends to infinity. In this event, numerical breakdown in the computation of $\hat{\eta}_{\text{PLS}}(W)$ is to be expected. A data example illustrates the point.

- *Numerical stability of hypercube estimators.* In contrast, Section 3 also shows that, for fixed X of full rank, the condition number of the matrix $VX'XV + I_p - V^2$ in Eq. (1.3) stays bounded as V varies over all matrices with eigenvalues in $[0, 1]$. To avoid numerical instability, it is therefore recommended to re-express a penalized least squares estimator as a hypercube estimator, using (1.6), prior to any numerical evaluations.

- *Asymptotic validity of minimizing the estimated risk over V.* The hypercube estimator $\hat{\eta}_H(V) = A(V)y$ defined in (1.3) has, under general model (1.1), the normalized quadratic risk

$$\begin{aligned} R(\hat{\eta}_H(V), \eta, \sigma^2) &= p^{-1}\mathrm{E}|\hat{\eta}_H(V) - \eta|^2 \\ &= p^{-1}\mathrm{tr}[\sigma^2A^2(V) + (I_n - A(V))^2\eta\eta']. \end{aligned} \tag{1.10}$$

The risk depends on the unknown parameters η and σ^2. Let $\hat{\sigma}^2$ be a trustworthy estimator of σ^2. A heuristic argument, as in Mallows (1973), yields the normalized *estimated risk*

$$\begin{aligned} \hat{R}_H(V) &= p^{-1}\mathrm{tr}[\hat{\sigma}^2A^2(V) + (I_n - A(V))^2(yy' - \hat{\sigma}^2I_n)] \\ &= p^{-1}[|y - A(V)y|^2 + \{2\mathrm{tr}(A(V)) - n\}\hat{\sigma}^2]. \end{aligned} \tag{1.11}$$

Let $\mathcal{V}$ denote a closed subset of the set of all $p \times p$ symmetric matrices whose eigenvalues all lie in $[0, 1]$. Suppose that $\hat{V}$ minimizes the estimated risk over all $V \in \mathcal{V}$. When is it true that the *adaptive hypercube estimator* $\hat{\eta}_H(\hat{V})$ has a risk that nearly achieves the minimum risk $\min_{V \in \mathcal{V}} R(\hat{\eta}_H(V), \eta, \sigma^2)$? Section 4 presents answers to this question under asymptotics where p tends to infinity.

- *Broad applicability of hypercube estimators*. Hypercube estimators, defined in (1.3), form a rich subclass of the set of symmetric linear estimators considered by Buja et al. (1989). As noted above, hypercube estimators are numerically stable and include algebraic equivalents to both penalized least squares estimators and submodel least squares estimators of η for linear model (1.1). On the other hand, penalized least squares estimators need not be numerically stable near submodel fits. This supports using hypercube estimator representations to improve the numerical stability in any statistical methodology involving penalized least squares.
  Green et al. (1985) studied the use of penalized least squares to fit a smooth trend in field experiments. Wood (2000) treated penalized least squares with multiple quadratic penalties. Penalized least squares plays a major technical role in the spline and semiparametric regression literature. See, for instance, Wahba et al. (1995), Eilers and Marx (1996), Heckman and Ramsay (2000), and Ruppert et al. (2003), especially Chapter 3. By no means, however, do potential applications to spline fitting exhaust the scope of hypercube estimators for linear model (1.1).
  Indeed, any PLS estimator can be hypercubed through transformation (1.6) so as to avoid unnecessary numerical instabilities. Example 1 in Section 3 illustrates this point in smoothing the fit to a one-way layout of means where the covariate is ordinal and discrete. Example 2 in Section 5 provides a different illustration in an unbalanced two-way layout where the two covariates are nominal. Here, the adaptive hypercube estimator searches over ANOVA submodel fits *and* over penalized least squares fits that interpolate among these so as to achieve low estimation risk. Neither example is a curve estimation problem. The point being emphasized is that adaptive hypercube estimators are a very general tool for obtaining low-risk numerically stable fits to linear model (1.1). As such, the range of potential applications is enormous.
- *Adaptive hypercube estimators and Stein multiple shrinkage*. Section 6 treats further the estimation of vectorized arrays of means. When the number of observations taken on each mean is equal, adaptive hypercube estimators are asymptotically identical to the multiple-shrinkage estimators of Stein (1966). Adaptive hypercube estimators thus resolve a longstanding problem: how to extend to general designs the risk reduction achieved by Stein's multiple shrinkage estimators for balanced observations on an array of means.

2. Hypercube estimators embed PLS and submodel estimators

Let V be any $p \times p$ symmetric matrix whose eigenvalues all lie in $[0, 1]$. From (1.3) above, the associated *hypercube estimator* of η is

$$\hat{\eta}_H(V) = XV(VX'XV + I_p - V^2)^{-1}VX'y. \tag{2.1}$$

The inverse exists because $X'X$ is positive definite. For the argument, see the proof of Theorem 2 in Section 3. The present section shows how PLS estimators and submodel least squares estimators may each be expressed as certain hypercube estimators.

2.1. PLS estimators are instances of hypercube estimators

The PLS estimator $\hat{\beta}_{\text{PLS}}(W)$ of β in linear model (1.1) is required to minimize over all $\beta \in R^p$ the PLS criterion

$$T(\beta, W) = |y - X\beta|^2 + \beta' W\beta. \tag{2.2}$$

Here, W is a positive semidefinite penalty matrix. Let

$$\tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \qquad \tilde{X} = \begin{pmatrix} X \\ W^{1/2} \end{pmatrix}, \tag{2.3}$$

where 0 is a $p \times 1$ vector of zeros. Then $T(\beta, W) = |\tilde{y} - \tilde{X}\beta|^2$. Because X has full rank p, so does $\tilde{X}$. Finding $\hat{\beta}_{\text{PLS}}(W)$ thus amounts to doing a least squares minimization, for which the normal equation is $\tilde{X}'\tilde{X}\beta = \tilde{X}'\tilde{y}$. Because this normal equation reduces to $(X'X + W)\beta = X'y$, it follows that $\hat{\beta}_{\text{PLS}}(W) = (X'X + W)^{-1}X'y$ uniquely. The PLS estimator of η is uniquely $\hat{\eta}_{\text{PLS}}(W) = X\hat{\beta}_{\text{PLS}}$, as described in (1.4).

Let $V = (I_p + W)^{-1/2}$, a symmetric matrix whose eigenvalues lie in $(0, 1]$ because W is positive semidefinite. Then $W = V^{-1}(I_p - V^2)V^{-1}$ and

$$\begin{aligned}(X'X + W)^{-1} &= [X'X + V^{-1}(I_p - V^2)V^{-1}]^{-1} \\ &= V[VX'XV + I_p - V^2]^{-1}V.\end{aligned} \tag{2.4}$$

Thus, the PLS estimator $\hat{\eta}_{\text{PLS}}(W)$ is equivalent algebraically to the hypercube estimator $\hat{\eta}_H((I_p + W)^{-1/2})$, as asserted in (1.6).

2.2. Submodel least squares estimators are instances of hypercube estimators

Let X_0 be any matrix with n rows whose range space $\mathcal{R}(X_0)$ lies within the range space $\mathcal{R}(X)$ of X. Consider the submodel of the general model (1.1) in which η is restricted to $\mathcal{R}(X_0)$. The *submodel least squares estimator* of η under this constraint is $\hat{\eta}_{\text{SUB}}(X_0) = X_0X_0^+y$, as noted in (1.7). See also Schott (2005).

Because $\mathcal{R}(X_0) \subset \mathcal{R}(X)$ implies that $X_0 = XA$ for $A = X^+X_0$, it follows that

$$\mathcal{R}(X_0) = \mathcal{R}(XA) = \mathcal{R}(XAA^+A) = \mathcal{R}(XAA^+) = \mathcal{R}(XP), \tag{2.5}$$

where $P = (X^+X_0)(X^+X_0)^+$ is symmetric and idempotent, an orthogonal projection with eigenvalues equal to either 0 or 1. Eq. (2.5) implies that $\hat{\eta}_{\text{SUB}}(X_0) = \hat{\eta}_{\text{SUB}}(XP)$.

Moreover, the submodel least squares estimator $\hat{\eta}_{\text{SUB}}(X_0)$ coincides with the hypercube estimator $\hat{\eta}_H(P)$. If B_1 and B_2 are symmetric matrices such that $B_1B_2 = 0$, then $(B_1 + B_2)^+ = B_1^+ + B_2^+$. Applying this and other standard properties of the Moore–Penrose pseudoinverse (see Schott, 2005) to (2.1) yields

$$\begin{aligned}
\hat{\eta}_H(P) &= XP[PX'XP + (I_p - P)]^+PX'y \\
&= XP[(XP)'(XP)]^+(XP)'y + XP(I_p - P)^+(XP)'y \\
&= XP(XP)^+y = \hat{\eta}_{\text{SUB}}(XP) = \hat{\eta}_{\text{SUB}}(X_0).
\end{aligned} \tag{2.6}$$

3. Condition number for hypercube and PLS estimators

This section analyzes the numerical stability of the matrix inversions in the definitions of the hypercube estimator $\hat{\eta}_H(V)$ and of the PLS estimator $\hat{\eta}_{\text{PLS}}(W)$. For fixed X of full rank, a PLS estimator is numerically unstable for certain choices of the penalty matrix W. Its re-expression (1.6) as a hypercube estimator is stable for all W.

3.1. Condition number and matrix inversion

Inverting a nonsingular symmetric matrix A amounts to solving for β the consistent equation $A\beta = a$. How stable are the solutions when A is perturbed to $A + \epsilon\,\Delta A$ and a is perturbed to $a + \epsilon\,\Delta a$? Here, $\epsilon \geq 0$ is small while ΔA and Δa are fixed. These small perturbations model errors that arise in finite-precision computer arithmetic. The task is to compare the solutions $\beta(\epsilon)$ and $\beta = \beta(0)$ of the respective equations

$$(A + \epsilon\,\Delta A)\beta(\epsilon) = a + \epsilon\,\Delta a \quad \text{and} \quad A\beta = a. \tag{3.1}$$

Let $|\cdot|$ denote the Euclidean norm. The induced matrix norm is defined by $\|A\| = \sup_{x\neq 0}[|Ax|/|x|]$. Following Golub and Van Loan (1996), $\beta(\epsilon)$ is differentiable in a neighborhood of $\epsilon = 0$. Using (3.1), the first derivative $\beta'(0) = A^{-1}(\Delta a - \Delta A\,\beta)$. The Taylor expansion for $\beta(\epsilon)$ is thus

$$\beta(\epsilon) = \beta + \epsilon\beta'(0) + O(\epsilon^2). \tag{3.2}$$

For every $\beta \neq 0$, it follows that the relative solution error satisfies

$$\begin{aligned}
|\beta(\epsilon) - \beta|/|\beta| &= |\epsilon|\,|\beta'(0)|/|\beta| + O(\epsilon^2) \\
&\leq |\epsilon|\,\|A^{-1}\|\,[|\Delta a|/|\beta| + \|\Delta A\|] + O(\epsilon^2).
\end{aligned} \tag{3.3}$$

Because $\|A\| \geq |A\beta|/|\beta| = |a|/|\beta|$, it follows from (3.3) that

$$|\beta(\epsilon) - \beta|/|\beta| \leq |\epsilon|\,\kappa(A)\,[|\Delta a|/|a| + \|\Delta A\|/\|A\|] + O(\epsilon^2), \tag{3.4}$$

where $\kappa(A) = \|A\|\,\|A^{-1}\|$ is called the *condition number* of A. If $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively denote the largest and smallest eigenvalues of A, then

$$\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A). \tag{3.5}$$

According to (3.4), large $\kappa(A)$ points to possible numerical instability when inverting the matrix A. Further analysis of $\kappa(A)$ strengthens this conclusion (see Golub and Van Loan, 1996).

3.2. Condition number for PLS estimators

The following theorem gives a lower bound on $\kappa(X'X + W)$, the condition number that governs the numerical stability of the PLS estimator $\hat{\eta}_{\text{PLS}}(W)$.

Theorem 1. *In definition (1.4) of $\hat{\eta}_{\text{PLS}}(W)$, recall that X is of full rank p and that W is positive semidefinite. Then,*

$$\kappa(X'X + W) \geq [\lambda_{\min}(X'X) + \lambda_{\max}(W)]/[\lambda_{\max}(X'X) + \lambda_{\min}(W)]. \tag{3.6}$$

The matrices $X'X$ and W may be interchanged in (3.6).

Proof. Let B_1 and B_2 be any symmetric positive semidefinite matrices of the same dimensions. By Weyl's theorem (see Schott, 2005),

$$\begin{aligned}
&\lambda_{\max}(B_1) + \lambda_{\min}(B_2) \leq \lambda_{\max}(B_1 + B_2) \leq \lambda_{\max}(B_1) + \lambda_{\max}(B_2) \\
&\lambda_{\min}(B_1) + \lambda_{\min}(B_2) \leq \lambda_{\min}(B_1 + B_2) \leq \lambda_{\min}(B_1) + \lambda_{\max}(B_2).
\end{aligned} \tag{3.7}$$

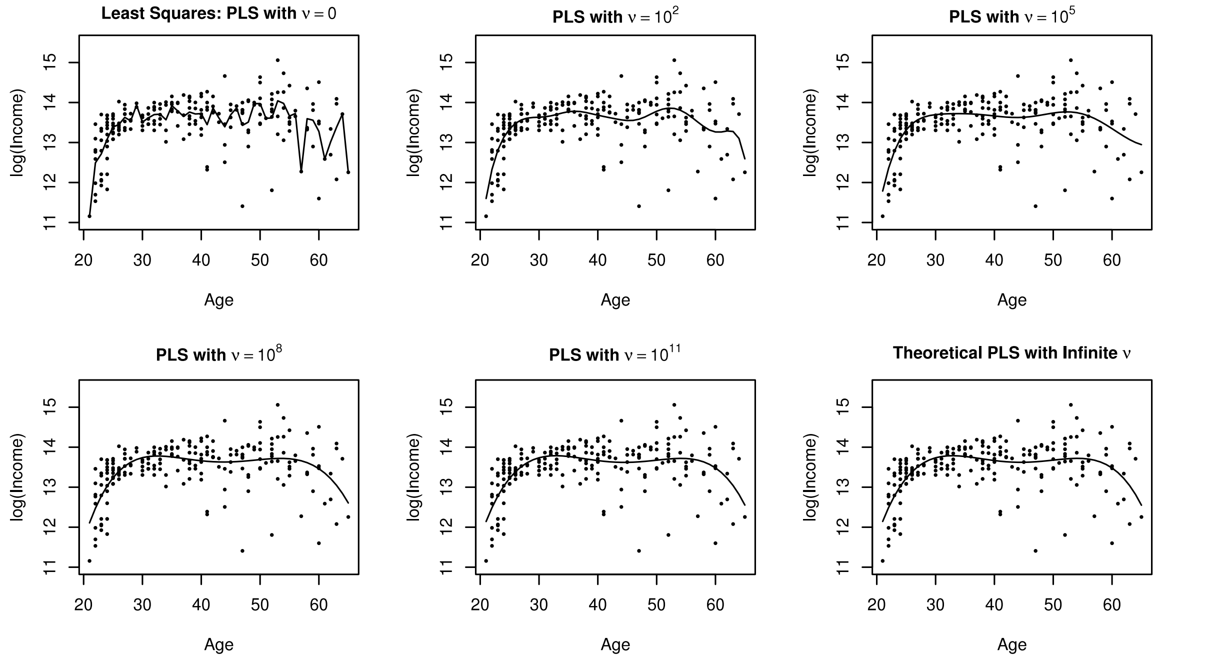

Fig. 1. Linearly interpolated PLS fits to the Canadian earnings data for various values of penalty weight ν. Despite apparent convergence as ν increases, numerical breakdown occurs at $\nu = 10^{13}$.

Taking $B_1 = W$ and $B_2 = X'X$, this yields that

$$\begin{aligned}&\lambda_{\max}(X'X + W) \geq \lambda_{\max}(W) + \lambda_{\min}(X'X)\\ &\lambda_{\min}(X'X + W) \leq \lambda_{\min}(W) + \lambda_{\max}(X'X).\end{aligned} \tag{3.8}$$

Lower bound (3.6) follows from expression (3.5) for $\kappa(X'X + W)$. □

In applications of penalized least squares, it is not uncommon that the penalty matrix W has $\lambda_{\min}(W) = 0$ while $\lambda_{\max}(W)$ can be very large. Then, by (3.6), the condition number that governs the numerical stability of $\hat{\eta}_{\text{PLS}}(W)$ can be large. The following example indicates how easily this can happen.

Example 1. The (1, 1) plot in Fig. 1 displays a scatterplot of the Canadian earnings data considered by Ullah (1985) and by Chu and Marron (1991). The plot displays log(incomes) for $n = 205$ individuals versus their $p = 45$ ages. The design is an unbalanced one-way layout, in which the number of log(income) observations vary from 1 to 12, according to age. The dark lines in the (1, 1) plot, *added only to guide the eye*, join the average log(incomes) observed at each age. These are, in fact, discrete annual quantities. Fitting these data is not a curve estimation problem.

A model for the data is

$$y = Cm + e. \tag{3.9}$$

Here, y is the $n \times 1$ vector of observed log(incomes), m is the $p \times 1$ vector of mean log(incomes), C is the $n \times p$ data-incidence matrix, with elements 0 or 1, that links observations to pertinent means, and e is the $n \times 1$ error vector. This is a special case of linear model (1.1) in which $X = C$ and $\beta = m$.

Consider the $(g - 1) \times g$ difference matrix $\Delta(g) = \{\delta_{u,w}\}$ in which $\delta_{u,u} = 1$, $\delta_{u,u+1} = -1$ for every u and all other entries are zero. Define D_5, the fifth difference matrix with $p = 45$ columns, by the matrix product

$$D_5 = \Delta(p-4)\Delta(p-3)\Delta(p-2)\Delta(p-1)\Delta(p). \tag{3.10}$$

Let $|\cdot|$ denote the Euclidean norm, and let $W(\nu) = \nu D_5' D_5$. Consider the PLS estimator of m defined for every $\nu \geq 0$ by

$$\hat{m}_{\text{PLS}}(W(\nu)) = \underset{m \in R^p}{\operatorname{argmin}}[|y - Cm|^2 + \nu|D_5 m|^2] = (C'C + W(\nu))^{-1}C'y. \tag{3.11}$$

The PLS estimator of $\eta = Cm$ is then $\hat{\eta}_{\text{PLS}}(W(\nu)) = C\hat{m}_{\text{PLS}}(W(\nu))$, the specialization of (1.4) and (1.5) when $X = C$.

The condition number $\kappa(C'C + W(\nu))$ governs the numerical stability of $\hat{\eta}_{\text{PLS}}(W(\nu))$. The eigenvalues of $W(\nu)$ range from zero (with multiplicity 5) to slightly more that 1016ν. The matrix $C'C$ is the diagonal matrix of replication counts at each age, ranging from 1 to 12 for this data set. Hence, by Theorem 1,

$$\kappa(C'C + W(\nu)) \geq (1 + 1016\nu)/12. \tag{3.12}$$

On the one hand, the PLS estimator (3.11) is well defined mathematically for every positive ν. On the other hand, because of the lower bound (3.12) on the condition number, numerical breakdown of $\hat{\eta}_{\text{PLS}}(W(\nu))$ and $\hat{m}_{\text{PLS}}(W(\nu))$ is to be expected in finite-precision computer arithmetic when the penalty weight ν is sufficiently large.

The orthogonal polynomials of degrees 0–4 span the eigenspace of the eigenvalue zero. Fig. 1 displays $\hat{m}_{\text{PLS}}(W(\nu))$ for increasing values of ν. These competing PLS fits are computed in R 2.15.2 using the function `solve()` to do the matrix inversion. Visually, the fits interpolate between the least squares polynomial fits of degree 44 (i.e., the sample means at each age) and degree 4 (i.e., the limit of $\hat{m}_{\text{PLS}}(W(\nu))$ as ν tends to infinity). The PLS estimator (3.11) is a perturbation of the fourth degree fit that, as ν decreases, moves the PLS fit closer to the sample means. Which value of ν gives the best fit in the sense of (approximately) minimizing risk under model (3.9)?

Answers to this question require, first of all, the ability to compute the PLS estimator $\hat{m}_{\text{PLS}}(W(\nu))$ reliably for many values of ν. Direct calculation as described in the previous paragraph breaks down totally at $\nu = 10^{13}$. If, instead, the matrix inversion is done with the R function `ginv()` from the library MASS, the calculation of the PLS estimator breaks down even earlier, at $\nu = 10^5$. The lower bound (3.12) on the condition number of the PLS estimator guarantees eventual breakdown in finite-precision arithmetic as ν increases. Numerical stability in this example is achieved through the hypercube representation of the PLS estimator, treated in the next subsection.

3.3. Condition number for hypercube estimators

The following theorem gives an upper bound on $\kappa(VX'XV + I_p - V^2)$, the condition number that governs the numerical stability of the hypercube estimator $\hat{\eta}_H(V)$.

Theorem 2. *In definition (1.3) of $\hat{\eta}_H(V)$, recall that X is of full rank p and V is positive semidefinite with all eigenvalues in* $[0, 1]$. *Then,*

$$\kappa(VX'XV + I_p - V^2) \le \begin{cases} \dfrac{\lambda_{\max}(X'X) + 1}{\lambda_{\min}(X'X)} - 1 & \text{if } \lambda_{\min}(X'X) < 1 \\ \lambda_{\max}(X'X) & \text{if } \lambda_{\min}(X'X) \ge 1. \end{cases} \tag{3.13}$$

This upper bound does not depend on V.

Proof. For brevity, write $\epsilon = \lambda_{\min}(X'X) > 0$. Let $B_1 = V(X'X - \epsilon I_p)V$ and $B_2 = I_p + (\epsilon - 1)V^2$. Then

$$VX'XV + I_p - V^2 = B_1 + B_2. \tag{3.14}$$

On the one hand, $\lambda_{\max}(B_1) \le \lambda_{\max}(X'X) - \epsilon$; and $\lambda_{\max}(B_2) \le \max\{\epsilon, 1\}$ because the eigenvalues of B_2 lie between ϵ and 1. Thus, by (3.7),

$$\lambda_{\max}(B_1 + B_2) \le \lambda_{\max}(X'X) - \epsilon + \max\{\epsilon, 1\}. \tag{3.15}$$

On the other hand, $\lambda_{\min}(B_1) \ge 0$ because $X'X - \epsilon I_p$ is positive semidefinite; and $\lambda_{\min}(B_2) \ge \min\{\epsilon, 1\}$, again because the eigenvalues of B_2 lie between ϵ and 1. Thus, by (3.7),

$$\lambda_{\min}(B_1 + B_2) \ge \min\{\epsilon, 1\}. \tag{3.16}$$

Upper bound (3.13) follows from (3.5) and the foregoing displayed equations. □

Theorem 2 shows that, for fixed X of full rank, the condition number that determines the numerical stability of $\hat{\eta}_H(V)$ is uniformly bounded over all V having eigenvalues in $[0, 1]$. Properties of X alone determine the bound (3.13). By Section 2.1, any PLS estimator $\hat{\eta}_{\text{PLS}}(W)$ equals, algebraically, the hypercube estimator $\hat{\eta}_H(V)$ with $V = (I_p + W)^{-1/2}$. When, as is not uncommon, V can be computed stably from the spectral decomposition of W, the hypercube form avoids unnecessary instability in computing the algebraically equivalent PLS estimator.

Example 1 (*Continued*). Let $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{41}$ denote the distinct eigenvalues of $D_5'D_5$, and let $\{P_k: 1 \le k \le 41\}$ denote the corresponding eigenprojections. The eigenvalue $\lambda_1 = 0$ has multiplicity 5. P_1 is the orthogonal projection into the subspace spanned by the orthogonal polynomials of degrees 0 to 4. The spectral representation $D_5'D_5 = \sum_{k=1}^{41} \lambda_k P_k$ implies that $W(\nu) = \sum_{k=2}^{41} \nu\lambda_k P_k$. By (1.6) and Section 2.1, the PLS estimator $\hat{\eta}_{\text{PLS}}(W(\nu))$ has an equivalent hypercube representation:

$$\hat{\eta}_{\text{PLS}}(W(\nu)) = \hat{\eta}_H[(I_p + W(\nu))^{-1/2}] = \hat{\eta}_H\left[P_1 + \sum_{k=2}^{41}(1 + \nu\lambda_k)^{-1/2}P_k\right]. \tag{3.17}$$

Because $\lambda_{\max}(C'C) = 12$ while $\lambda_{\min}(C'C) = 1$, the upper bound (3.13) on the condition number for the hypercube estimator on the right-hand side of (3.17) is 12 for every $\nu \ge 0$. Section 4 develops the implications of such numerical stability for choosing a best value of ν in this example.

By (3.17) and algebraic reasoning akin to (2.6),

$$\lim_{\nu\to\infty} \hat{\eta}_{\text{PLS}}(W(\nu)) = \hat{\eta}_H(P_1) = CP_1(CP_1)^+ y. \tag{3.18}$$

In other words, the hypercube form of the PLS estimator converges, as ν increases, to a hypercube estimator that coincides with the least squares fit to the fourth-degree polynomial submodel of model (3.9). The foregoing discussion illustrates how hypercube representations of PLS estimators control the condition number and include limits of PLS estimators as a bonus.

4. Risk, estimated risk, and adaptation

The question arises of how to chose the matrix V that defines the "best" hypercube estimator. Risk-adaptive selection methods assume that the general model (1.1) holds and seek to minimize the risk asymptotically over the class of candidate estimators as p increases. Consistent selection methods assume that some components of β may vanish, and then seek to estimate the non-zero components and identify the zero components consistently as p increases. The classical Hodges estimator illustrates that the two goals are antagonistic: as a rule, a consistent selection fit has poor risk in neighborhoods (within the general model) of the submodels it considers. The methodology outlined in this section for selecting V seeks to control the risk, not to achieve consistent submodel selection. For the latter purpose, lasso methods using the Bayesian information criterion (BIC) to select a penalty weight may be attractive (see Wang and Leng, 2008 and references cited therein).

Under the general linear model (1.1), with the strong Gauss–Markov assumption on the error vector e, the normalized quadratic risk of hypercube estimator $\hat{\eta}_H(V)$ is given in (1.10). It is proposed to estimate this risk by the estimated risk (1.11) or by an algebraically equivalent criterion. Doing so is in the tradition of the Mallows C_p criterion and of statistical decision theory, when the estimated risk is a trustworthy approximation to the risk. This section develops conditions under which the estimated risk converges to the risk of $\hat{\eta}_H(V)$ *uniformly* over a usefully large class of matrices V as p tends to infinity. It is noted below that uniform convergence of estimated risks does *not* occur when the class of V is the set of *all* $p \times p$ symmetric matrices with eigenvalues in [0, 1].

4.1. Canonical expressions for risk and estimated risk

Both the risk (1.10) and the estimated risk (1.11) admit canonical re-expressions in terms of matrices that are p dimensional rather than n dimensional. The canonical forms assist computation of the estimated risk when p is much smaller than n, and they are fundamental for the asymptotic theory of the next subsection in which p tends to infinity.

Let $N = X'X$ and $U = XN^{-1/2}$. Then $U'U = I_p$ and

$$\begin{aligned} \hat{\eta}_H(V) &= US(V)U'y \\ S(V) &= N^{1/2}V(VNV + I_p - V^2)^{-1}VN^{1/2}. \end{aligned} \tag{4.1}$$

The eigenvalues of the $p \times p$ symmetric matrix $S(V)$ all lie in [0, 1] whenever V is symmetric with eigenvalues in [0, 1]. In particular, $S(I_p) = I_p$. The unconstrained least squares estimator of η is $\hat{\eta}_H(I_p) = UU'y = X(X'X)^{-1}X'y$.

Model (1.1) states that $y = \eta + e$ for $\eta = X\beta$. Let $\xi = N^{1/2}\beta = U'\eta$. Then $\eta = U\xi$. Let $z = U'y$. It follows that $\mathrm{E}(z) = \xi$ and $\mathrm{Cov}(z) = \sigma^2 I_p$. The normalized quadratic *loss* of the hypercube estimator $\hat{\eta}_H(V) = US(V)U'y$ for η is then

$$L(\hat{\eta}_H(V), \eta) = p^{-1}|\hat{\eta}_H(V) - \eta|^2 = p^{-1}|S(V)z - \xi|^2. \tag{4.2}$$

The corresponding *risk* function is therefore

$$R(\hat{\eta}_H(V), \eta, \sigma^2) = p^{-1}\mathrm{tr}[\sigma^2 S^2(V) + (I_p - S(V))^2 \xi\xi']. \tag{4.3}$$

This expression is algebraically equivalent to (1.10).

The risk depends on two unknown quantities: $\xi\xi'$ and σ^2. An unbiased estimator of the former is $zz' - \sigma^2$. Let $\hat{\sigma}^2$ be an L_1-consistent estimator of σ^2, possibly but not necessarily the least squares estimator of σ^2. Heuristic reasoning akin to that in Mallows (1973) leads to the *estimated risk* function

$$\begin{aligned} \hat{R}_H(V) &= p^{-1}\mathrm{tr}[\hat{\sigma}^2 S^2(V) + (I_p - S(V))^2(zz' - \hat{\sigma}^2 I_p)] \\ &= p^{-1}[|z - S(V)z|^2 + \{2\mathrm{tr}(S(V)) - p\}\hat{\sigma}^2]. \end{aligned} \tag{4.4}$$

This expression is algebraically equal to (1.11), but involves matrices of dimension p rather than n.

4.2. Adaptation and its asymptotic justification

The following question arises: over what sets $\mathcal{V}$ of $p \times p$ symmetric matrices V with eigenvalues in [0, 1] does $\hat{R}_H(V)$ converge to the risk $R(\hat{\eta}_H(V), \eta, \sigma^2)$, *uniformly* over $V \in \mathcal{V}$? When uniform convergence holds, the estimated risk provides a trustworthy approximation to the true risk of hypercube estimator $\hat{\eta}_H(V)$ as V ranges over $\mathcal{V}$. This in turn enables trustworthy selection of a good V.

A bracing negative result: The desired uniform convergence does *not* occur when $\mathcal{V}$ is the class of all $p \times p$ symmetric matrices with eigenvalues in $[0, 1]$. Indeed, risk adaptation already fails over the subclass of all $V \in \mathcal{V}$ that have a specified orthonormal eigenbasis and unordered eigenvalues that vary freely in $[0, 1]$. For the argument in the Gaussian error case, see Remark A on p. 1829 of Beran and Dümbgen (1998). In that setting, greedy minimization of the estimated risk fails to minimize the risk asymptotically and yields inadmissible estimators.

A useful positive result: Uniform convergence of the estimated risk to the risk does hold under the following restrictions on $\mathcal{V}$. Let $\{P_k\colon 1 \le k \le s\}$, with $1 \le s \le p$, be symmetric, idempotent, mutually orthogonal $p \times p$ matrices such that $\sum_{k=1}^{s} P_k = I_p$. Define the matrices

$$V(d) = \sum_{k=1}^{s} d_k P_k, \quad d = (d_1, \ldots, d_s) \in [0, 1]^s. \tag{4.5}$$

The right-hand side of (4.5) is a spectral representation of $V(d)$. Let $\mathcal{D}$ denote any closed subset of $[0, 1]^s$. The following two theorems treat risk adaptation over all V in $\mathcal{V} = \{V(d)\colon d \in \mathcal{D}\}$. Key is the assumption that both s and the projections $\{P_k\colon 1 \le k \le s\}$ are fixed as p tends to infinity.

Theorem 3. *Fix s and the projections $\{P_k\colon 1 \le k \le s\}$ in (4.5). Suppose that model (1.1) holds with the strong Gauss–Markov assumption on the errors. Suppose that, for every finite $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0. \tag{4.6}$$

Let the quantity $L(d)$ denote either the loss $L(\hat{\eta}_H(V(d)), \eta)$ or the estimated risk $\hat{R}_H(V(d))$. Then, for every finite $a > 0$ and $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathrm{E}\left[\sup_{d\in[0,1]^s} |L(d) - R(\hat{\eta}_H(V(d)), \eta, \sigma^2)|\right] = 0. \tag{4.7}$$

The loss, risk, and estimated risk of the hypercube estimator $\hat{\eta}_H(V(d))$ thus converge together as p tends to infinity, uniformly over all $d \in [0, 1]^s$. A modification of the arguments in Beran (2007) proves Theorem 3.

For $\mathcal{D}$ as defined above, let

$$\hat{d} = \operatorname*{argmin}_{d\in\mathcal{D}} \hat{R}_H(V(d)), \qquad \tilde{d} = \operatorname*{argmin}_{d\in\mathcal{D}} R(\hat{\eta}_H(V(d)), \eta, \sigma^2). \tag{4.8}$$

When d is restricted to $\mathcal{D}$, the *adaptive hypercube estimator* $\hat{\eta}_H(V(\hat{d}))$ is intended to achieve a risk close to that of the unrealizable *oracle hypercube estimator* $\hat{\eta}_H(V(\tilde{d}))$. The following result, a consequence of Theorem 3, implies that this happens asymptotically (see Beran, 2007 for the details of the argument).

Theorem 4. *Suppose that the assumptions for Theorem 3 hold. Let the quantity L denote either the loss or risk of the adaptive hypercube estimator $\hat{\eta}_H(V(\hat{d}))$. Then, for every finite $a > 0$ and $\sigma^2 > 0$,*

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} |L - R(\hat{\eta}_H(V(\tilde{d})), \eta, \sigma^2)| = 0 \tag{4.9}$$

and

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} |L - \hat{R}_H(V(\hat{d}))| = 0. \tag{4.10}$$

Thus, as p tends to infinity, the loss and risk of the adaptive hypercube estimator $\hat{\eta}_H(V(\hat{d}))$ converge to the loss and risk of the oracle hypercube estimator $\hat{\eta}_H(V(\tilde{d}))$. Moreover, the plug-in risk estimator $\hat{R}_H(V(\hat{d}))$ converges to the actual loss or risk of $\hat{\eta}_H(V(\hat{d}))$. This result motivates adaptation in Example 1 below and justifies adaptation in Example 2 of Section 5.

Remarks. For finite p, the estimated risk function $\hat{R}_H(V(d))$ can be negative for some values of d. This does *not* diminish the functionality of definition (4.8) of $\hat{d}$: smaller estimated risk still means better according to the asymptotic theory. Adding a sufficiently large constant to the loss function reduces the likelihood of encountering negative estimated risks without changing anything essential.

As noted at the start of this subsection, minimizing the estimated hypercube risk $\hat{R}_H(V)$ over *all* $p \times p$ symmetric matrices V with eigenvalues in $[0, 1]$ does not minimize the risk asymptotically. However, if the eigenprojections of V are specified and ordered, and the corresponding eigenvalues are required to be monotone decreasing, then it seems very likely that the desired uniform asymptotic convergence of estimated risks to risk occurs as p increases. Beran and Dümbgen (1998) proved this (and slightly stronger variants) for the case when $X'X$ is a multiple of the identity matrix and the error vector e is Gaussian. In that setting, efficient algorithms exist for minimizing the estimated risk over all choices of monotone decreasing eigenvalues.

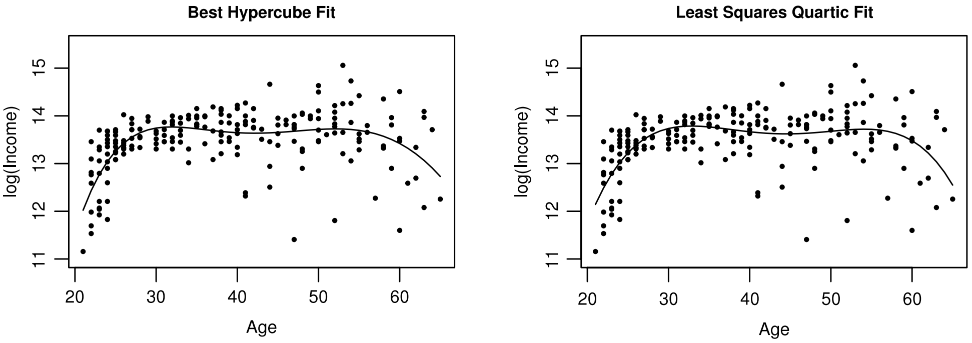

Fig. 2. Linearly interpolated best hypercube fit and least squares quartic fit to the Canadian earnings data.

Example 1 (*Continued*)**.** In definition (4.5) of $V(d)$, let $s = 41$, and choose the $\{P_k: 1 \leq k \leq 41\}$ to be the eigenprojections in the spectral representation of $D_5'D_5$, as in Eq. (3.17). Let $\mathcal{D}_0 = \{1, (1+\nu\lambda_2)^{-1/2}, \ldots, (1+\nu\lambda_{41})^{-1/2}: \nu \geq 0\}$, and let $\mathcal{D} = \mathcal{D}_0 \cup \{(1, 0, \ldots, 0)\}$, a closed subset of $[0, 1]^{41}$. Because of (3.17), the class of hypercube estimators $\{\hat{\eta}_H(V(d)): d \in \mathcal{D}\}$ embeds the PLS estimators $\{\hat{\eta}_{\text{PLS}}(W(\nu)): \nu \geq 0\}$. In view of (3.18), the estimator $\hat{\eta}_H[V((1, 0, \ldots, 0))]$ is the least squares polynomial fit of degree 4, approached theoretically by the PLS estimator as ν tends to infinity. The hypercube embedding of the PLS estimators has bounded condition number for all ν. The numerical instability of the PLS estimators (3.11) for large ν vanishes when computing the algebraically equivalent hypercube estimators.

The candidate hypercube fits interpolate, along a curve in $[0, 1]^{41}$ parameterized by ν, between the least squares polynomial fits of degree 4 (when $d = (1, 0, \ldots, 0)$) and degree 44 (when $d = (1, 1, \ldots, 1)$ and the sample means at each age are the fit). The hypercube form enables trustworthy numerical minimization of estimated risk over all $d \in \mathcal{D}$. The least squares estimate of the variance σ^2 is $\hat{\sigma}^2 = 0.295$. The adaptive hypercube estimator, defined by (4.8) to minimize the estimated risk over all $d \in \mathcal{D}$, is $\hat{\eta}(V(\hat{d}))$ with $\hat{d} = (1, (1+\hat{\nu}\lambda_2)^{-1/2}, \ldots, (1+\hat{\nu}\lambda_{41})^{-1/2})$ and $\hat{\nu} = 16, 074, 617$. The estimated risk of this adaptive hypercube estimator is -0.0296, far smaller than the estimated risk 0.295 of the unconstrained least squares fit to the data. The estimated risk of the least squares fourth-degree polynomial fit is -0.0226, only slightly greater than that of the best hypercube fit. Fig. 2 compares these two fits with small estimated risk.

The statistical model for the data is (3.9), a one-way layout of observations on completely unconstrained means. Under model (3.9), with estimated risk as the criterion, the analysis establishes that (a) the best hypercube fit to the Canadian income data is only slightly better than the fourth-degree polynomial fit, and (b) the small differences between the two fits are visible only near age 65, where data is sparse.

5. Application to an array of means

Given an array of means, let m denote the $p \times 1$ vector obtained by taking the means in array order. Given observations made with error on these means, let y denote the $n \times 1$ vector obtained by clumping replications together and taking the clumps in the same order as the corresponding means. The model is linear model (1.1) with β replaced by m and X replaced by the $n \times p$ data-incidence matrix C:

$$y = Cm + e. \tag{5.1}$$

The elements of C are either 0 or 1, chosen to link each component of the observation vector y to the appropriate mean.

The columns of C are orthogonal and $C'C = \text{diag}\{n_i\}$, where n_i is the number of observations made on m_i. It is assumed that the design is complete but not necessarily balanced: the replication numbers $\{n_i\}$ are each at least 1 and may be unequal. Thus, $\text{rank}(C) = p \leq n$. The formulation covers regression models where each m_i is an *unknown* function of one or more covariates. Multi-way ANOVA models are a special case where the covariates are factors.

Define orthogonal projections $\{P_k: 1 \leq k \leq s\}$ and the symmetric matrix $V(d)$ as in (4.5). For $\eta = \text{E}(y) = Cm$, consider the hypercube estimator $\hat{\eta}_H(V(d))$, defined by (1.3) with X specialized to C. Because $V^2(d) = V(d^2)$, the hypercube estimator of $m = (C'C)^{-1}C'\eta$ is

$$\hat{m}_H(V(d)) = V(d)[V(d)C'CV(d) + I_p - V(d^2)]^{-1}V(d)C'y. \tag{5.2}$$

The mapping between $\hat{m}_H(V(d))$ and $\hat{\eta}_H(V(d)) = C\hat{m}_H(V(d))$ is one-to-one. Both estimators are assessed by quadratic risk (4.3) and estimated risk (4.4).

Let $d_k = 1$ if $k \in \mathcal{K} \subset \{1, 2, \ldots, s\}$, and let $d_k = 0$ otherwise. Let $P = \sum_{k\in\mathcal{K}} P_k$. Then $V(d) = P$, an orthogonal projection, and

$$\hat{m}_H(V(d)) = \hat{m}_H(P) = P(CP)^+y = (CP)^+y. \tag{5.3}$$

This is the least squares estimator of m in the submodel of (5.1) where $m = P\beta$ for some $\beta \in R^p$, that is, $\eta \in \mathcal{R}(CP)$. Eq. (5.3) is a special case of the analysis in Section 2.2, which showed how hypercube estimators embed submodel least squares estimators.

The following example illustrates how minimizing the estimated risk over hypercube estimators – the methodology supported by the theorems of Section 4 – can improve on classical ANOVA submodel fits.

Example 2. The rat litter data treated by Scheffé (1959) form an unbalanced two-way layout. Each response recorded is the average weight-gain of a rat litter when the infants in the litter are nursed by a rat foster-mother. Factor 1, with four levels, is the genotype of the foster-mother. Factor 2, with the same levels, is the genotype of the infant litter. Fig. 3 presents the data, giving separate scatterplots for each of the four foster-mother genotypes. The lines interpolate, purely to guide the eye, the unconstrained least squares (LS) fit to the two-way layout. This discrete LS fit estimates the mean average weight-gain in each cell of the two-way layout by averaging the observations in that cell. According to the unconstrained LS fit, foster-mothers of genotype 3 are more successful than foster-mothers of genotype 4 in feeding litters of all genotypes. As the (1, 1) plot in Fig. 4 makes clear, further comparisons among the effects of foster-mother or litter genotypes do not stand out in the unconstrained LS fit.

Two-way ANOVA considers competing least squares submodel fits to the rat litter data. Suppose that the $\{m_{ij}\colon 1 \le i \le 4,\ 1 \le j \le 4\}$ are the means in this two-way array. Vectorize the $\{m_{ij}\}$ in array order (i.e., by stacking successive columns of the matrix) to get the 16×1 mean vector m. Vectorize the observations similarly, clumping replications together, to obtain the observation vector y. With suitable data-incidence matrix C, model (5.1) expresses the observed two-way layout.

To express two-way ANOVA submodels, let $u_r = (1/2, 1/2, 1/2, 1/2)'$ for $r = 1, 2$. Set $J_r = u_r u_r'$ and $H_r = I_4 - J_r$. The standard ANOVA projections are

$$P_1 = J_2 \otimes J_1, \qquad P_2 = J_2 \otimes H_1, \qquad P_3 = H_2 \otimes J_1, \qquad P_4 = H_2 \otimes H_1. \tag{5.4}$$

The $\{P_k\}$ are symmetric, idempotent, mutually orthogonal matrices such that $\sum_{k=1}^{4} P_k = I_{16}$. The two-way ANOVA representation of the means is $m = \sum_{k=1}^{4} P_k m$. Submodels are specified by setting one or more of the summands equal to zero. For instance, requiring $P_4 m = 0$ specifies the ANOVA submodel with no interactions, and requiring $P_3 m = P_4 m = 0$ specifies the ANOVA submodel in which the Factor 2 main effects and the interactions both vanish. There are 2^4 such ANOVA submodels.

Consider the particular hypercube estimators of m defined by (5.2) when the $s = 4$ projections are those given in (5.4). From the discussion surrounding (5.3), the hypercube estimators $\{\hat{m}_H(V(d))\colon d \in \{0, 1\}^4\}$ are precisely the least squares fits to the 16 submodels generated by keeping or dropping summands in the ANOVA decomposition $m = \sum_{k=1}^{4} P_k$. The least squares estimate of variance under model (5.1) is $\hat{\sigma}^2 = 54.2$. The estimated risks of the various submodel least squares fits, computed as in Section 4.1, are as follows.

d	(0, 0, 0, 0)	(1, 0, 0, 0)	(0, 1, 0, 0)	(1, 1, 0, 0)	(0, 0, 1, 0)	(1, 0, 1, 0)
$\hat{R}_H(d)$	11154.6	56.2	11119.6	28.4	11094.0	72.8

d	(0, 1, 1, 0)	(1, 1, 1, 0)	(0, 0, 0, 1)	(1, 0, 0, 1)	(0, 1, 0, 1)	(1, 1, 0, 1)
$\hat{R}_H(d)$	11054.2	44.7	10449.4	57.2	10354.5	35.6

d	(0, 0, 1, 1)	(1, 0, 1, 1)	(0, 1, 1, 1)	(1, 1, 1, 1)
$\hat{R}_H(d)$	10396.3	75.9	10284.0	54.2

Not surprisingly, the least squares fits to the eight ANOVA submodels that omit the overall mean have very large estimated risk. The foster-mother main effects submodel with $d = (1, 1, 0, 0)$ has the smallest estimated risk, 28.4. Next smallest is the estimated risk 35.6 for the no litter main effects submodel $d = (1, 1, 0, 1)$. The additive submodel with $d = (1, 1, 1, 0)$ has estimated risk 44.7. The estimated risk of the full least squares fit to the data is 54.2. On the basis of the ANOVA submodel fit with smallest estimated risk, displayed in the (1, 2) plot in Fig. 4, one might conclude that foster-mothers 2 (dashed line) dominate foster-mothers 1 (solid line), which dominate foster-mothers 3 (dotted line), which dominate foster-mothers 4. Such a simple ordering seems questionable when compared with the complexity of the unconstrained least squares fit exhibited in the (1, 1) plot.

This motivates considering the much larger class of hypercube estimators $\{\hat{m}_H(V(d))\colon d \in [0, 1]^4\}$, which includes the ANOVA submodel least squares fits as well as PLS fits that interpolate among these. By definition (4.8), the adaptive hypercube estimate of m is $\hat{m}_H(V(\hat{d}))$, where $\hat{d}$ minimizes the estimated risk $\hat{R}(V(d))$ over all $d \in [0, 1]^4$. Applying the estimated risk formula (4.4) to this example yields $\hat{d} = (0.997, 0.693, 0.000, 0.415)$. The estimated risk of the adaptive hypercube estimator is $\hat{R}(V(\hat{d})) = 16.1$, which undercuts substantially the estimated risk 28.4 of the best ANOVA submodel fit.

As is seen in the (1, 3) plot in Fig. 4, the adaptive best hypercube fit tells a more nuanced story than the best ANOVA submodel fit. It recognizes some degree of interaction between foster-mother genotype and litter genotype. Foster-mothers 2 still dominate foster-mothers 3, which still dominate foster-mothers 4. However, foster-mothers 1 are now seen to dominate all others in feeding litters of genotype 1 (not entirely surprising) and are otherwise comparable with foster-mothers 3.

The residual plots Fig. 4 also hint that the best ANOVA submodel fit is an underfit relative to the best hypercube fit. However, estimated risk comparisons among competing fits are more sophisticated than residual plots because they are not fooled by overfitting.

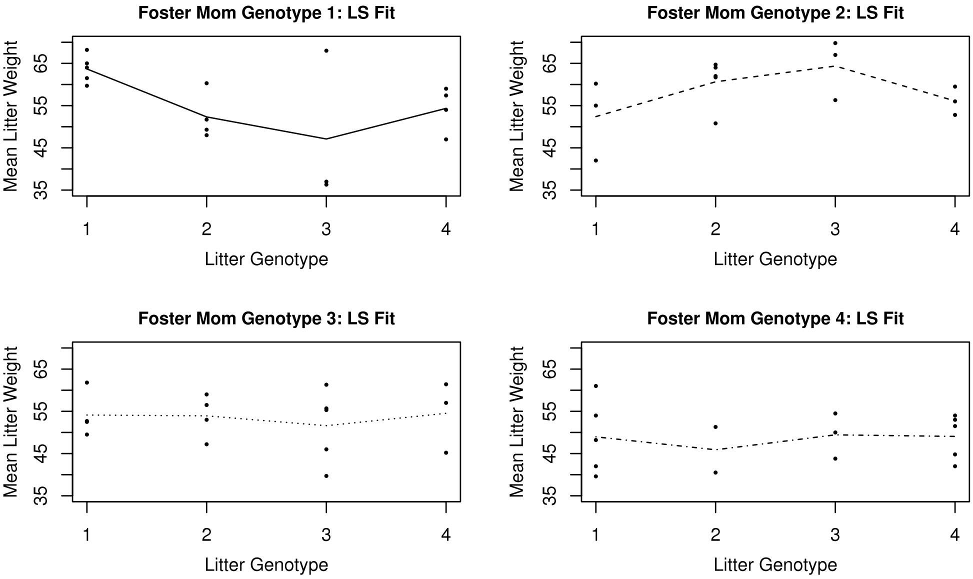

Fig. 3. For each foster-mother genotype, a scatterplot of the average weight-gains of litters nursed against litter genotype. The lines join the unconstrained least squares fit across litter genotypes.

Remarks. Data differ essentially from pseudo-data described by probability models or generated by pseudo-random simulations or constructed from experience. The theorems in Section 4 reveal good performance of certain adaptive hypercube estimators on pseudo-data generated by the general linear model (1.1). Such theory alone does not ensure success in any specific data analysis. As Tukey (1980) observed, "In practice, methodologies have no assumptions and deliver no certainties". This being said, the insight gained through the adaptive hypercube fit to the rat litter data seems valuable.

6. Adaptive hypercube estimators and Stein multiple shrinkage

Section 5 treated adaptive hypercube estimators for a vectorized array of means when the number of observations made may vary for each mean. This section shows that, when the number of observations on each mean is equal, adaptive hypercube estimators simplify greatly: they are then almost identical, for all but small p, to the multiple-shrinkage estimators of Stein (1966). Adaptive hypercube estimators thereby solve a longstanding problem: how to extend Stein multiple shrinkage to unbalanced designs.

Consider the case when the number of observations on each component of the vectorized means m in model (5.1) is $n_0 \geq 1$. Then $C'C = n_0 I_p$. The least squares estimator of m reduces to

$$\hat{m}_{\mathrm{LS}} = (C'C)^{-1}C'y = n_0^{-1}C'y. \tag{6.1}$$

As in (4.5), let $V(d) = \sum_{k=1}^{s} d_k P_k$, where $\sum_{k=1}^{s} P_k = I_p$, and the $\{P_k\}$ are mutually orthogonal, symmetric, and idempotent. The hypercube estimator (5.2) reduces to

$$\begin{aligned}\hat{m}_H(V(d)) &= V(d)[n_0 V(d^2) + I_p - V(d^2)]^{-1} V(d)C'y \\ &= V(d)[(n_0 - 1)V(d^2) + I_p]^{-1} V(d)C'y.\end{aligned} \tag{6.2}$$

The explicit matrix inversion

$$[(n_0 - 1)V(d^2) + I_p]^{-1} = \sum_{k=1}^{s} [(n_0 - 1)d_k^2 + 1]^{-1} P_k \tag{6.3}$$

in (6.2) implies that

$$\hat{m}_H(V(d)) = \sum_{k=1}^{s} c_k P_k \hat{m}_{\mathrm{LS}}, \quad c_k = \frac{n_0 d_k^2}{(n_0 - 1)d_k^2 + 1}. \tag{6.4}$$

Theorem 5. *Suppose that model* (5.1) *holds with* $C'C = n_0 I_p$. *Let* $I(\cdot)$ *denote the indicator function. Let* $\mathcal{D}$ *denote a closed subset of* $[0, 1]^s$.

$$\hat{\tau}_k = p^{-1}\hat{\sigma}^2 \mathrm{tr}(P_k), \qquad \hat{w}_k = p^{-1} n_0 |P_k \hat{m}_{\mathrm{LS}}|^2 - \hat{\tau}_k, \qquad \hat{w}_{k+} = \max\{\hat{w}_k, 0\}. \tag{6.5}$$

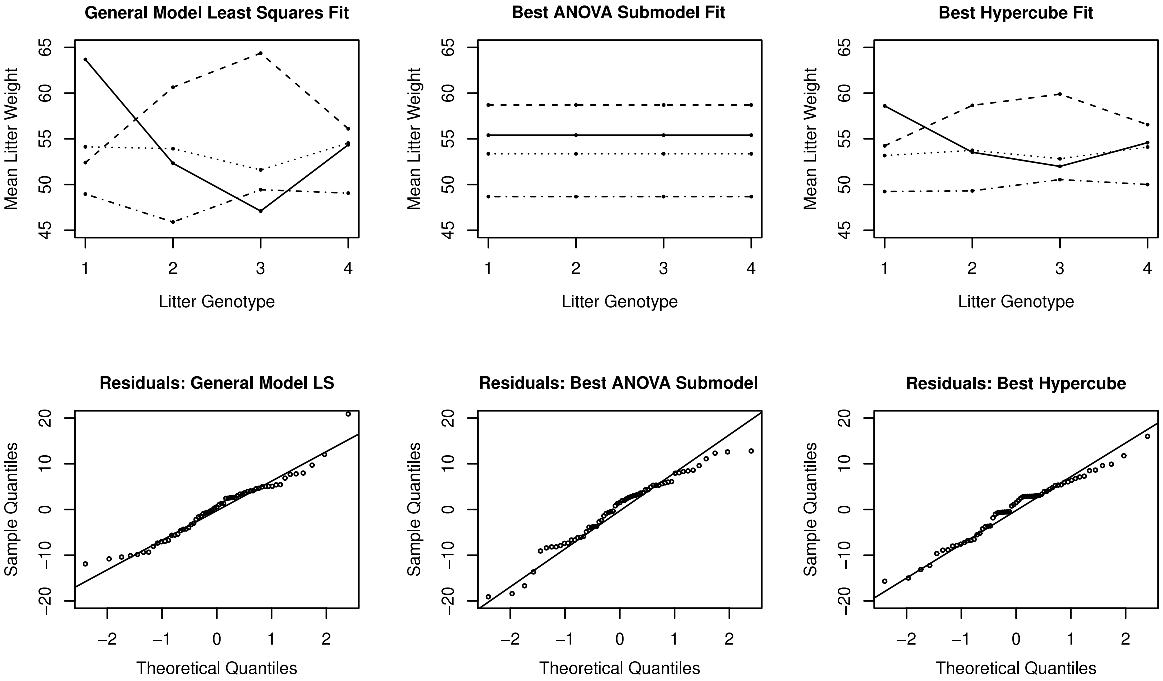

Fig. 4. Unconstrained least squares, adaptive ANOVA submodel least squares, and adaptive hypercube fits to the rat litter data. The line type in the first row identifies the foster-mother. Each line joins the fitted values across litter genotypes.

(i) *When* $\mathcal{D} = [0, 1]^s$, *the adaptive hypercube estimator of* m, *defined by* (4.8) *and* (6.4), *is*

$$\hat{m}_H(V(\hat{d})) = \sum_{\hat{w}_k \geq 0}^{s} \left(\frac{\hat{w}_k}{\hat{\tau}_k + \hat{w}_k} \right) P_k \hat{m}_{\mathrm{LS}}, \tag{6.6}$$

and has estimated risk

$$\hat{R}_H(V(\hat{d})) = \sum_{\hat{w}_k \geq 0} \left(\frac{\hat{\tau}_k \hat{w}_k}{\hat{\tau}_k + \hat{w}_k} \right) + \sum_{\hat{w}_k < 0} \hat{w}_k. \tag{6.7}$$

(ii) *When* $\mathcal{D} = \{0, 1\}^s$, *the adaptive hypercube estimator of* m, *defined by* (4.8) *and* (6.4), *is*

$$\hat{m}_H(V(\hat{d})) = \sum_{k=1}^{s} I(\hat{w}_k > \hat{\tau}_k) P_k \hat{m}_{\mathrm{LS}}, \tag{6.8}$$

and has estimated risk

$$\hat{R}_H(V(\hat{d})) = \sum_{k=1}^{s} \min\{\hat{w}_k, \hat{\tau}_k\}. \tag{6.9}$$

Proof. Let $\mathcal{D} = [0, 1]^s$. The mapping in (6.4) of d_k into c_k is a homeomorphism that maps $[0, 1]^s$ onto itself. For balanced designs, the class of hypercube estimators $\{\hat{m}_H(V(d)) : d \in [0, 1]^s\}$ thus coincides with the class of *shrinkage* estimators

$$\hat{m}_S(c) = \sum_{k=1}^{s} c_k P_k \hat{m}_{\mathrm{LS}}, \quad c \in [0, 1]^s. \tag{6.10}$$

Let $\hat{\eta}_S(c) = C\hat{m}_S(c)$. Let $\tau_k = p^{-1}\sigma^2 \mathrm{tr}(P_k)$ and $w_k = p^{-1} n_0 |P_k m|^2$. By direct calculation, the risk of $\hat{\eta}_S(c)$ is

$$R(\hat{\eta}_S(c), \eta, \sigma^2) = p^{-1}\mathrm{E}|\hat{\eta}_S(c) - \eta|^2 = \sum_{k=1}^{s} [c_k^2 \tau_k + (1 - c_k)^2 w_k]. \tag{6.11}$$

By the reasoning for (4.4), the estimated risk of $\hat{\eta}_S(c)$ is

$$\hat{R}_S(c) = \sum_{k=1}^{s} [c_k^2 \hat{\tau}_k + (1 - c_k)^2 \hat{w}_k]. \tag{6.12}$$

Let

$$\hat{c}_k = \begin{cases} \dfrac{\hat{w}_k}{\hat{\tau}_k + \hat{w}_k} & \text{if } \hat{w}_k \geq 0 \\ 0 & \text{if } \hat{w}_k < 0. \end{cases} \tag{6.13}$$

When $\hat{w}_k \geq 0$, the summand

$$c_k^2 \hat{\tau}_k + (1 - c_k)^2 \hat{w}_k = (c_k - \hat{c}_k)^2 (\hat{\tau}_k + \hat{w}_k) + \hat{\tau}_k \hat{c}_k. \tag{6.14}$$

Thus, $\hat{c} = \text{argmin}_{c \in [0,1]^s} \hat{R}_S(c) = (\hat{c}_1, \ldots, \hat{c}_s)$. Eq. (6.6) follows because $\hat{m}_H(V(\hat{d}))$ must coincide with $\hat{m}_S(\hat{c})$. The estimated risk formula (6.7) follows by plugging $\hat{c}$ into (6.12).

Let $\mathcal{D} = \{0, 1\}^s$. The mapping in (6.4) of d_k into c_k is a homeomorphism that maps $\{0, 1\}^s$ onto itself. Let

$$c_k^* = I(\hat{c}_k > 1/2) = I(\hat{w}_k > \hat{\tau}_k). \tag{6.15}$$

From (6.12) and (6.14), $c^* = \text{argmin}_{c \in \{0,1\}^s} \hat{R}_S(c) = (c_1^*, \ldots, c_s^*)$. Eq. (6.8) follows because $\hat{m}_H(\hat{d})$ must now coincide with $\hat{m}_S(c^*)$. The estimated risk formula (6.9) follows by plugging c^* into (6.12). □

Remarks. By its definition, the minimal estimated risk (6.7) cannot exceed the minimal estimated risk (6.9). Neither of these can exceed the estimated risk $\hat{\sigma}^2$ of the full model least squares estimator $\hat{m}_{\text{LS}}$.

The adaptive hypercube estimator (6.6) for $\mathcal{D} = [0, 1]^s$ is equivalent algebraically to

$$\hat{m}_H(V(\hat{d})) = \sum_{k=1}^{s} \left[1 - \frac{\hat{\sigma}^2 \text{tr}(P_k)}{n_0 |P_k \hat{m}_{\text{LS}}|^2} \right]_+ P_k \hat{m}_{\text{LS}}, \tag{6.16}$$

where $[\cdot]_+$ denotes the positive-part function. Apart from small p refinements, expression (6.16) applies separate positive-part James–Stein shrinkage to each projection $P_k \hat{m}_{\text{LS}}$. Stein (1966) gave an exact treatment of such multiple shrinkage estimators under the Gaussian error model. Adaptive hypercube estimators thus extend to general designs the risk reduction achieved by Stein's multiple shrinkage estimators for balanced observations on an array of means.

In addition, the adaptive hypercube estimator (6.8) for $\mathcal{D} = \{0, 1\}^s$ is equivalent algebraically to

$$\hat{m}_H(V(\hat{d})) = \sum_{k=1}^{s} I \left\{ \left[1 - \frac{\hat{\sigma}^2 \text{tr}(P_k)}{n_0 |P_k \hat{m}_{\text{LS}}|^2} \right]_+ > \frac{1}{2} \right\} P_k \hat{m}_{\text{LS}}. \tag{6.17}$$

Let $\mathcal{K} \subset \{1, 2, \ldots, s\}$, and let $P = \sum_{k \in \mathcal{K}} P_k$. The condition that $m = P\beta$ for some β defines a submodel of (5.1). Within the class of least squares fits to these submodels, expression (6.17) describes the simplest that minimizes the estimated risk and provides an efficient algorithm for finding it.

References

Beran, R., 2007. Adaptation over parametric families of symmetric linear estimators. Journal of Statistical Planning and Inference 137, 684–696.

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

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

Chu, C.-K., Marron, J.S., 1991. Choosing a kernel regression estimator. Statistical Science 6, 404–436.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–121.

Golub, G.H., Van Loan, C.F., 1996. Matrix Computations, third ed. Johns Hopkins University Press, Baltimore.

Green, P., Jennison, C., Seheult, A., 1985. Analysis of field experiments by least squares smoothing. Journal of the Royal Statistical Society: Series B 47, 299–315.

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

Mallows, C., 1973. Some comments on C_p. Technometrics 15, 661–676.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge.

Scheffé, H., 1959. The Analysis of Variance. Wiley.

Schott, J.R., 2005. Matrix Analysis for Statistics, second ed. Wiley.

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

Tukey, J., 1980. Methodological comments focused on opportunities. In: Monge, P.R., Cappella, J.N. (Eds.), Multivariate Techniques in Human Communication Research. Academic Press, New York, pp. 490–528.

Ullah, A., 1985. Specification analysis of econometric models. Journal of Quantitative Economics 2, 187–209.

Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B., 1995. Smoothing spline ANOVA for exponential families with application to the Wisconsin epidemiological study of diabetic retinopathy. Annals of Statistics 23, 1868–1895.

Wang, H., Leng, C., 2008. A note on adaptive group lasso. Computational Statistics and Data Analysis 52, 5277–5286.

Wood, S.N., 2000. Modeling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society: Series B 62, 413–428.