EXTREMAL FITS IN REACT CONFIDENCE SETS

RUDOLF BERAN[1]

University of California, Berkeley

REACT estimators use ideas from signal processing, model-selection, and shrinkage to achieve much smaller risk in one-way layouts and other linear models than does the classical least squares estimator. The REACT method can generate automatic scatterplot smoothers that compete well on standard data sets with the best fits obtained by other methods. This paper addresses two further questions: Which features in a REACT estimator are not necessarily present in the true mean vector; and which features of the true mean vector might have been smoothed out by the REACT estimator? We answer both questions by constructing extremal members of a confidence set of asymptotic coverage probability α that is centered at the REACT estimator. The methodology is demonstrated on two data-sets from the smoothing literature.

AMS subject classifications: 62J05, 62G07.
Keywords and phrases: economical basis, risk estimation, adaptation, superefficiency, saturated fit, unsaturated fit.

1 Introduction

Consider the Gaussian linear model in which the $n \times 1$ response vector y has a $N(X\beta, \sigma^2 I_n)$ distribution, the regression parameters β and the variance σ^2 being both unknown. Suppose that the $n \times p$ regression matrix X has full rank $p \leq n$. The least squares estimator of the mean $\eta = \mathrm{E}(y) = X\beta$ is then $\hat{\eta}_{LS} = X(X'X)^{-1}X'y$. Under normalized quadratic loss, the risk of an estimator $\hat{\eta}$ of η is $p^{-1}\mathrm{E}|\hat{\eta} - \eta|^2$. This risk is precisely σ^2 for the least squares fit $\hat{\eta}_{LS}$.

Stein (1956) proved the inadmissibility of the least squares fit $\hat{\eta}_{LS}$ to the Gaussian linear model when dimension of the regression space exceeds 2. This defect in least squares becomes intuitively clear in special cases. Consider the one-way layout with one observation per cell (e.g. a digitized signal observed in white noise) or the two-way layout with one observation per cell (e.g. a digitized image observed in white noise). In such examples, the least squares estimator of the signal η is the raw data y. It is not

[1] This research was supported in part by National Science Foundation Grant DMS95-30492.

surprising that electronic devices such as television sets do not rely on the Gauss-Markov theorem to separate signal from noise.

REACT fits combine ideas from signal processing, model-selection and shrinkage to accomplish superefficient estimation of η. The acronym itself is a reminder of the steps in the methodology: Risk Estimation and Adaptation after Coordinate Transformation. The first step in REACT is to devise an orthonormal regression basis for the linear model such that η can be well-approximated by a linear combination of the first few orthonormal basis vectors. It is then reasonable to consider candidate model-selection or shrinkage estimators of η that concentrate on estimating the regression coefficients of these important basis vectors, so as to take advantage of possible bias-variance trade-off. Finally, the candidate estimator that minimizes estimated risk is the REACT estimator of η.

Four decades of development in Statistics preceded REACT fits. James and Stein (1961) and Stein (1966) pioneered shrinkage improvements over least squares estimators. Mallows (1973) examined the use of C_L as a criterion for model-selection, noting that minimum C_L misleads if the class of candidate estimators is too large. Model-selection is a particular form of variable shrinkage. Scenarios in which minimum C_L succeeds in minimizing asymptotic risk over a class of candidate estimators were isolated by Li (1987), Kneip (1994), Beran and Dümbgen (1998). The role of the basis emerged tacitly in Stein's (1966) treatment of multiple shrinkage estimators and in Pinsker's (1980) asymptotic minimax analysis of trend estimation in Gaussian noise. Donoho and Johnstone (1994, 1995) emphasized the importance of a suitable basis and of risk-based adaptation. REACT, the application of signal recovery and shrinkage techniques to superefficient estimation in linear models, was described in Beran (2000). Common to these varied papers are models in which the unknown signal is deterministic.

A parallel literature on random signals has developed estimation for stationary time-series, for Bayes or empirical Bayes formulations, and for hidden Markov models. It is remarkable how often results for random signal models have turned out to have analogs for deterministic signal models with many parameters. Considerable mystery still surrounds this correspondence.

Less explored are ways of representing the likely errors incurred in estimating deterministic signals. Stein (1981) suggested a construction of confidence sets around trend estimators that generalizes the classical confidence set for η centered at $\hat{\eta}_{LS}$. Beran (1996) and Beran and Dümbgen (1998) justified this method asymptotically for confidence sets centered at estimators obtained through nested model-selection or through adaptive monotone shrinkage. For given coverage probability α, better REACT estimators at the center turn out to yield asymptotically smaller confidence sets. How may these high-dimensional confidence sets be put to use? This paper probes the confidence sets for extremal members so as to determine which features of η may have been smoothed out in $\hat{\eta}$ and which features of $\hat{\eta}$ may not be present in η.

Section 2 reviews REACT methodology for one-way layouts, drawing particular attention to three orthonormal bases for this design: orthonormal polynomial contrasts, certain trigonometric contrasts that enrich the discrete cosine basis, and the smooth contrasts described in Beran (2000). Section 3 defines most unsaturated and most saturated members of confidence sets centered at REACT estimators of η. These fits at the extremes of credibility exhibit potential differences between the REACT estimator and the true mean vector η. Analyses of two data-sets illustrate the possible economy of orthonormal polynomial contrasts and the methodology for probing REACT confidence sets. Section 4 sketches technical underpinnings.

2 REACT Fits in the One-Way Layout

For a one-way layout with p factor levels, the classical choice of regression matrix X is the incidence matrix. Each row of this X contains a single 1, the remaining $p-1$ entries being 0. Rows are repeated according to number of replications at each factor level. The column index of the 1 indicates factor level. This section reviews REACT fits to one-way layouts. For further background, see Beran (2000).

Scatterplot fits are naturally related to the one-way layout. Given the $\{x_i: 1 \le i \le n\}$, suppose that the $\{y_i: 1 \le i \le n\}$ are conditionally independent and that the conditional distribution of y_i is $N(m(x_i), \sigma^2)$. If m is an unknown function and p denotes the number of distinct values among the $\{x_i\}$, then this conditional model is a one-way layout. By reordering labels as necessary, suppose that x_i is a nondecreasing function of i. Interpolation between successive components of an estimator of η then yields a curve fit to the scatterplot. While this interpolated estimator is poor when based on $\hat{\eta}_{LS}$, it can be very satisfactory when based on more efficient REACT estimators of η.

2.1 Transformation to an economical coordinate system

For any matrix A, let $\mathcal{M}(A)$ denote the subspace spanned by the columns of A. Let U_E denote a suitably chosen $n \times p$ matrix with orthonormal columns such that $\mathcal{M}(U_E) = \mathcal{M}(X)$. Considerations that enter into the choice of U_E will be discussed below. Define

$$z = U_E' y, \qquad \xi = \mathrm{E} z. \tag{2.1}$$

Evidently, z has a normal $N_p(\xi, \sigma^2 I_p)$ distribution. The mapping between ξ, whose range is R^p, and η, whose range is the p-dimensional regression space

$\mathcal{M}(X) \subset R^n$, is one-to-one with

$$\xi = U_E'\eta, \qquad \eta = U_E\xi. \tag{2.2}$$

Similarly, any estimator $\hat{\eta}$ of η corresponds in one-to-one fashion to the estimator $\hat{\xi} = U_E'\hat{\eta}$ of ξ, the inverse relation being $\hat{\eta} = U_E\hat{\xi}$. Quadratic loss is preserved under this correspondence because $U_E'U_E = I_p$ entails

$$p^{-1}|\hat{\eta} - \eta|^2 = p^{-1}|\hat{\xi} - \xi|^2. \tag{2.3}$$

We assume that the orthonormal basis U_E provides an *economical* representation of η in the sense that all but the first few components of ξ are very nearly close to zero. Economy is designated by the subscript E. For an economical basis, we need only identify and estimate the relatively few nonzero components of ξ. The quadratic risk of $\hat{\xi}$ then accumulates many small squared biases from ignoring the nearly zero components of ξ but does not accumulate the many variances that would arise from an attempt to estimate these components from z. Theorem 4 in Beran (2000) gives a precise statement of the bias-variance trade-off in terms of Pinsker's (1980) asymptotic minimax bound on quadratic risk. Note that the concept of economical basis is more restrictive than the concept of sparse basis that is used in treatments of thresholding estimators (cf. Donoho and Johnstone (1994)).

In the author's experiments to-date, three particular orthonormal bases U_E for one-way layouts have proved effective in analyzing scatterplots from the smoothing literature. Let s denote the $p \times 1$ column vector whose i-th component is i and let $u = Xs$, where X is the incidence matrix defined at the start of this section.

a) *Polynomial contrast basis.* The regression space of the one-way layout is spanned by the columns of the matrix $A = (u^0, u, \ldots, u^{p-1})$, where operations on u are performed componentwise. The columns of A are linearly independent because a polynomial of degree $p-1$ has at most $p-1$ distinct roots while $n \geq p$. The polynomial contrast basis is defined as the Gram-Schmidt orthonormalization of the columns of A. Because A is nearly collinear for large p, sophisticated methods are needed to compute this basis. The function poly in S-Plus is one possibility.

b) *Trigonometric contrast basis.* In this case the columns of A are the first p entries in the list $v^0, \{\{\cos(k\pi v), \sin(k\pi v)\}: k \geq 1\}$ where $v = (u - 1/2)/p$. When the rows of the incidence matrix are distinct, A is an enrichment of the discrete cosine basis that avoids the edge artifacts of the discrete Fourier basis. The trigonometric contrast basis is defined as the Gram-Schmidt orthonormalization of the columns of A.

c) *Smooth contrast basis.* This generalization of the discrete cosine transform is described in Beran (2000).

In general, which orthonormal basis is most economical in fitting a particular one-way layout is an empirical matter that depends on the unknown η. Prior information about the nature of η may help delimit bases to be considered. Diagnostic plots of the components of z and estimated risk calculations assist subsequent selection from a modestly large collection of bases. Such diagnostic techniques illuminate, later in this section, the success of the polynomial contrast basis in fitting two case studies.

2.2 Estimated risks of candidate estimators

Let $\mathcal{F}$ be a closed subset of $[0,1]^p$. In the transformed coordinate system, consider $\{fz: f \in \mathcal{F}\}$ as candidate shrinkage estimators for ξ. For any vector $h \in R^p$, let $\text{ave}(h) = p^{-1}\sum_{i=1}^p h_i$. The risk of fz under normalized quadratic loss is

$$(2.4) \qquad p^{-1}\mathrm{E}|fz - \xi|^2 = \text{ave}[\sigma^2 f^2 + \xi^2(1-f)^2] \equiv \rho(f, \xi^2, \sigma^2).$$

If this risk function were known, we would estimate ξ by the best candidate estimator $\tilde{f}z$, where

$$(2.5) \qquad \tilde{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \rho(f, \xi^2, \sigma^2).$$

This is equivalent to estimating η by $\tilde{\eta}_{\mathcal{F}} = U_E \text{diag}(\tilde{f}) U_E' y$, a symmetric linear smoother.

Usually the risk function in (2.4) is unavailable because both ξ^2 and σ^2 are unknown. Two methods for estimating σ^2 will be considered in this paper:

a) *The least squares variance estimator.* When $n > p$, least squares theory provides the estimator

$$(2.6) \qquad \hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\eta}_{LS}|^2,$$

which is consistent provided $n - p$ tends to infinity.

b) *The high-component variance estimator.* The ANOVA strategy of pooling suggests

$$(2.7) \qquad \hat{\sigma}^2_H = (n-q)^{-1}[\sum_{i=q+1}^{p} z_i^2 + |y - \hat{\eta}_{LS}|^2],$$

where $q < p \leq n$. The bias of $\hat{\sigma}^2_H$ is

$$(2.8) \qquad (n-q)^{-1}\sum_{i=q+1}^{p} \xi_i^2.$$

Consistency of $\hat{\sigma}^2_H$ is assured when this bias tends to zero as n tends to infinity. Economy of U_E makes the bias small when q exceeds the number of basis vectors needed to approximate η well.

The estimator $\hat{\sigma}^2_H$ is particularly useful when $n = p$, the one-way layout with one observation per factor level.

Having devised a consistent estimator $\hat{\sigma}^2$ of σ^2, we estimate ξ^2 by $z^2 - \hat{\sigma}^2$ and the risk $\rho(f, \xi^2, \sigma^2)$ by

$$\hat{\rho}(f) = \text{ave}[\hat{\sigma}^2 f^2 + (z^2 - \hat{\sigma}^2)(1-f)^2]. \tag{2.9}$$

Tacit in the construction of $\hat{\rho}$ is the supposition that the law of large numbers will make $\text{ave}[(1-f)^2(z^2 - \hat{\sigma}^2)]$ consistent for $\text{ave}[(1-f)^2\xi^2)]$. Because $\hat{\rho}(f)$ can sometimes be negative, we will consider as well the risk estimator $\hat{\rho}_+(f) = \max\{\hat{\sigma}^2 \text{ave}(f^2), \hat{\rho}(f)\}$. The uniform consistency of $\hat{\rho}(f)$ and $\hat{\rho}_+(f)$ over suitable $\mathcal{F}$ is treated by Beran and Dümbgen (1998).

2.3 Adaptation

It is natural to use $\hat{\rho}(f)$ as a surrogate for the risk $\rho(f, \xi^2, \sigma^2)$ in identifying the best candidate estimator. This strategy generates the fully data-based estimator $\hat{\eta}_{\mathcal{F}} = U_E \text{diag}(\hat{f}) U'_E$, where

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \hat{\rho}(f). \tag{2.10}$$

Apart from the details of the variance estimator $\hat{\sigma}^2$, this construction of $\hat{f}$ amounts to minimizing the Mallows (1973) C_L criterion or minimizing the Stein (1981) unbiased estimator of risk.

Successful adaptation, meaning that the risks of $\hat{\eta}_{\mathcal{F}}$ and $\tilde{\eta}_{\mathcal{F}}$ converge, requires restrictions on the richness of $\mathcal{F}$. Beran and Dümbgen (1998) developed sufficient conditions on the covering number of $\mathcal{F}$ to ensure success of adaptation. The global class $\mathcal{F} = [0,1]^p$ is too large for adaptation. Two smaller but very useful shrinkage classes for which adaptation works are:

a) *Monotone class* $\mathcal{F}_M$. This is the closed convex set $\{f \in [0,1]^p \colon f_1 \geq f_2 \geq \ldots \geq f_p\}$. It makes sense to damp down the higher order components of z in constructing fz precisely because U_E is an economical basis. The value $\hat{f}_M$ that minimizes estimated risk $\hat{\rho}(f)$ over all $f \in \mathcal{F}_M$ is unique and can be computed by algorithms for weighted isotonic regression, as detailed in Beran and Dümbgen (1998). The corresponding estimator of η is $\hat{\eta}_M = U_E \text{diag}(\hat{f}_M) U'_E y$.

b) *Nested selection class* $\mathcal{F}_{NS}$. This subset of $\mathcal{F}_M$ is defined as follows. For $0 \leq k \leq p$, let $e(k)$ denote the p dimensional column vector whose i-th component is 1 if $1 \leq i \leq k$ and is 0 otherwise. Then $\mathcal{F}_{NS}$ is the union of the vectors $\{e(k) \colon 0 \leq k \leq p\}$. Nested model-selection is the idea behind $\mathcal{F}_{NS}$, whose convex hull is $\mathcal{F}_M$. Computation of $\hat{f}_{NS}$, the value that minimizes $\hat{\rho}(f)$ over all $f \in \mathcal{F}_{NS}$, is straightforward. The corresponding estimator of η is $\hat{\eta}_{NS} = U_E \text{diag}(\hat{f}_{NS}) U_E' y$.

In the original parametrization, the candidate estimators have the form $U_E \text{diag}(f) U_E' y$. They are thus symmetric linear smoothers, in the sense of Buja, Hastie and Tibshirani (1989), whose eigenvectors are given by the columns of U_E. This observation explains why REACT estimators can act like adaptive locally linear smoothers.

2.4 Case studies with polynomial contrasts

The underlying scatterplot in Figure 1 exhibits log-income versus age of the individual sampled. This Canadian earnings data was introduced by Ullah (1985) and treated further by Chu and Marron (1991). Conditioning on the observed ages, we fit an unbalanced one-way layout to the $n = 205$ observed log-incomes, the factor levels being the $p = 45$ distinct ages from 21 to 65, taken in numerical order. The top row of Figure 1 exhibits the polynomial contrast estimators $\hat{\eta}_{NS}$ and $\hat{\eta}_M$, using the least squares estimator of variance to compute risks of candidate REACT estimators. In both plots, the components of the estimator have been interpolated linearly. The visual impression created by cubic spline interpolation is similar. Such interpolation is more than a visual device if we consider mean log(income) to be a continuous function of age.

Using $\hat{\rho}(\cdot)$ to estimate the risks of $\hat{\eta}_{LS}$, $\hat{\eta}_{NS}$ and $\hat{\eta}_M$ yields, respectively, $\hat{\rho}_{LS} = \hat{\sigma}^2_{LS} = .295$, $\hat{\rho}_{NS} = -.029$ and $\hat{\rho}_M = -.037$. The negative values are inconvenient here. Using instead $\hat{\rho}_+(\cdot)$ yields the risk estimates $\hat{\rho}_{+,NS} = .039$ and $\hat{\rho}_{+,M} = .036$. Both of the REACT estimators have far smaller estimated risk than the least squares estimator $\hat{\eta}_{LS}$. If the latter is plotted with linear interpolation, the resulting curve is jagged, especially at the higher ages. The two interpolated REACT fits are fully data-based once the basis U_E is selected, with no tuning parameter requiring attention, and resemble fits obtained for this data by locally linear smoothers. Fits to this data by Nadaraya-Watson kernel smoothers are biased upwards near ages 21 and 65 but otherwise resemble the REACT fits (see Chu and Marron (1991)).

Because ages are equally spaced, the two polynomial contrast REACT estimators actually fit polynomials at the earnings data-points, though not in between. The estimator $\hat{\eta}_{NS}$ fits a polynomial of degree 5 (which coincides with the degree 5 least squares fit) while $\hat{\eta}_M$ fits a polynomial of degree 14 (which differs considerably from the degree 14 least squares fit). All coefficients of the fitted REACT polynomials beyond the term of degree 2 are very small. Indeed, a classical F-test at level .10 does not find evidence of nonzero coefficients beyond degree 2. Because rejection, not acceptance, is important in testing, this result only indicates that we should not use a fit of degree less than 2. The parabolic least squares fit to the earnings data completely misses the econometrically interesting dip between ages 40 and 50 and has notably larger estimated risk (namely .106) than either of the two REACT fits. Estimating the mean vector with small risk is an enterprise that differs from seeking overwhelming test evidence that certain regression coefficients are non-zero.

In Figure 2, the $(1,1)$ cell plots the signed square root of z_i versus i. The square-root transformation makes more visible the values of z_i that are close to zero in value. This diagnostic plot supports the hypothesis that the polynomial contrast basis is economical for the earnings data, a plausible finding because the levels of the age factor are ordered in time. The $(1,2)$ cell exhibits the components of first five orthonormal polynomial contrasts, with linear interpolation to aid visibility.

The underlying scatterplot in Figure 3 exhibits, for each row in a vineyard near Lake Erie, the total grape harvest over three years. Simonoff (1996) used this vineyard data as a case study for smoothing methods and provided further background. Conditioning on the $p = 52$ row numbers, we fit a balanced one-way layout to the $n = 52$ observed three-year harvests. The top row of Figure 3 exhibits, with linear interpolation, the polynomial contrast estimators $\hat{\eta}_{NS}$ and $\hat{\eta}_M$. The high-component variance estimator with $q = 15$ served to compute estimated risks. The estimated risks of $\hat{\eta}_{LS}$, $\hat{\eta}_{NS}$ and $\hat{\eta}_M$ are, respectively, $\hat{\rho}_{LS} = \hat{\sigma}^2_H = 6.08$, $\hat{\rho}_{NS} = 1.24$ and $\hat{\rho}_M = 1.02$.

Row numbers being equally spaced, the estimator $\hat{\eta}_{NS}$ fits a polynomial of degree 10 to the points in the vineyard data scatterplot (but not in between). This fit coincides with the degree 10 least squares fit to the same points. The estimator $\hat{\eta}_M$ fits a polynomial of degree 17 to the points in the scatterplot (but not in between). This fit differs substantially from the degree 17 least squares fit to these same points. Linear interpolation between fitted points avoids the wiggliness inherent in polynomial curves. Both REACT fits resemble a locally linear nonparametric regression fit to this data exhibited in Fig. 5.13 of Simonoff (1996). The diagnostic plots in the top row of Figure 4 and the relatively small estimated risks of both REACT fits support the supposition that the polynomial contrast basis is economical for the vineyard data. This conclusion is plausible because the levels of the row-number factor are spatially ordered.

3 Confidence Sets and Saturation

We begin by applying the confidence set idea sketched at the end of Stein (1981). For $\mathcal{F}$ equal to either $\mathcal{F}_M$ or $\mathcal{F}_{NS}$, consider the root

$$\hat{t}_{\mathcal{F}} = p^{1/2}[p^{-1}|\hat{\eta}_{\mathcal{F}} - \eta|^2 - \hat{\rho}(\hat{f})]. \tag{3.1}$$

The right side of (3.1) compares the normalized quadratic loss of $\hat{\eta}_{\mathcal{F}}$ with an estimate of its expectation, which is the estimated risk. A confidence

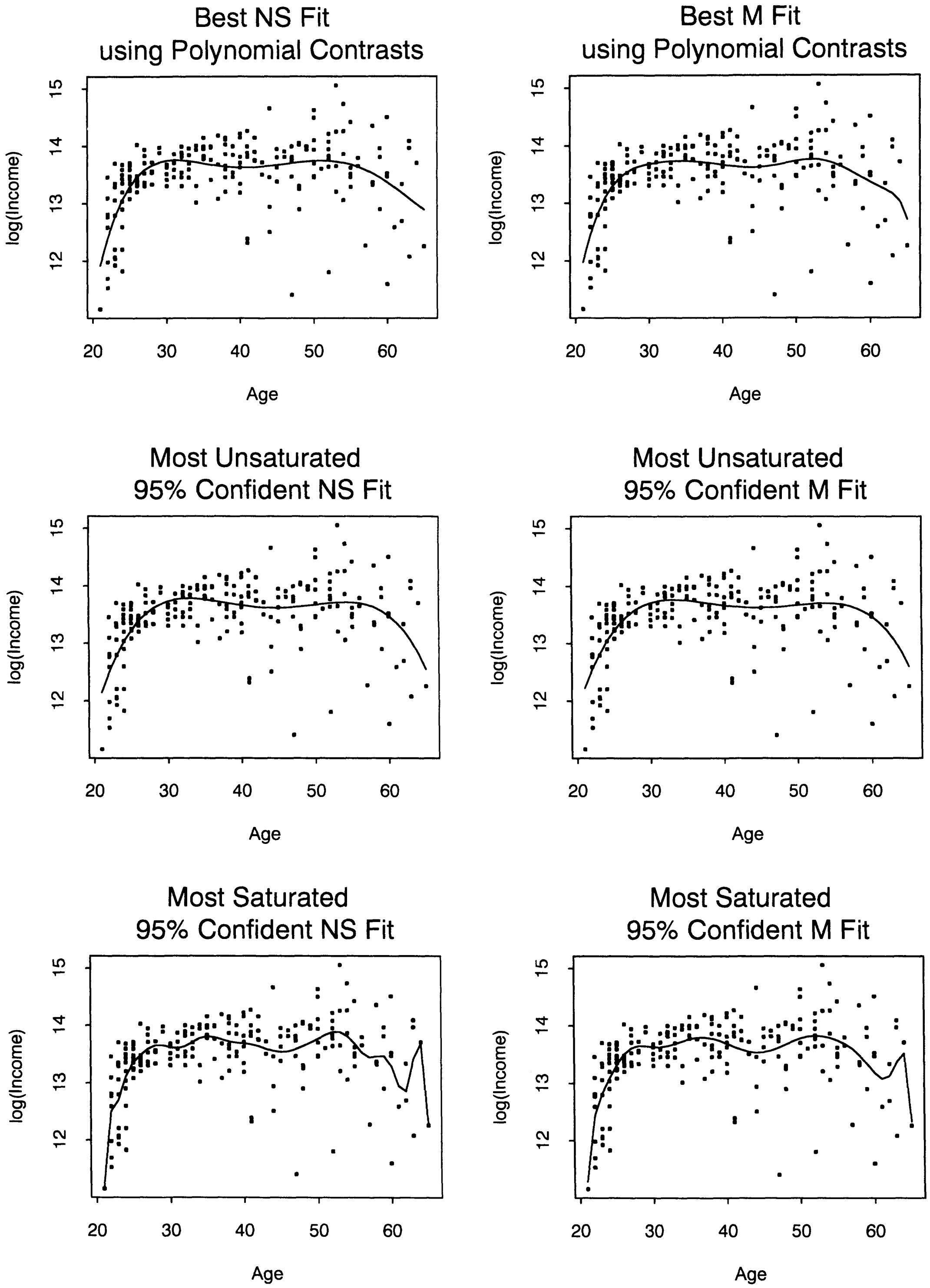

Figure 1. REACT fits and extremal confident fits to the Canadian earnings data.

set for η is obtained by referring $\hat{t}_{\mathcal{F}}$ to the α-th quantile of its estimated distribution.

For reasons detailed in Section 4, the variance estimator $\hat{\sigma}^2$ that enters into the definition of $\hat{t}_{\mathcal{F}}$ strongly affects the next step in the construction. We therefore write $\hat{t}_{\mathcal{F},1}$ and $\hat{t}_{\mathcal{F},2}$ to distinguish the two cases.

Least squares variance estimator. When $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$, the distribution of $\hat{t}_{\mathcal{F},1}$ for large p and $n-p$ is approximately $N(0, \hat{\tau}^2_{\mathcal{F},1})$ with

$$\begin{aligned}\hat{\tau}^2_{\mathcal{F},1} = 2\hat{\sigma}^4_{LS}\text{ave}[(2\hat{f}-1)^2] + 2[p/(n-p)]\hat{\sigma}^4_{LS}[\text{ave}(2\hat{f}-1)]^2 \\ + [4\hat{\sigma}^2_{LS}\text{ave}[(z^2-\hat{\sigma}^2_{LS})(1-\hat{f})^2]]_+.\end{aligned} \tag{3.2}$$

Accordingly, a confidence set of approximate coverage probability α for η is

$$\hat{C}_{\mathcal{F},1}(\alpha) = \{\theta \in \mathcal{M}(X)\colon |\hat{\eta}_{\mathcal{F}} - \theta|^2 \le p\hat{\rho}(\hat{f}) + p^{1/2}\hat{\tau}_{\mathcal{F},1}\Phi^{-1}(\alpha)\}. \tag{3.3}$$

In this expression, Φ^{-1} is the quantile function of the standard normal distribution.

When $\mathcal{F}$ consists of the single vector $e(p)$, defined in Section 2, the corresponding REACT estimator $\hat{\eta}_F$ is just $\hat{\eta}_{LS}$. The classical confidence set for η based on the F-distribution is $\{\theta \in \mathcal{M}(X)\colon |\hat{\eta}_{\mathcal{F}} - \theta|^2 \le p\hat{\sigma}^2_{LS}F^{-1}_{p,n-p}(\alpha)\}$. If p and $n-p$ both tend to infinity in such a way that $p/(n-p)$ converges to a finite constant, then the classical confidence set is asymptotically equivalent to the confidence set given by (3.3) when $\mathcal{F} = \{e(p)\}$.

High-component variance estimator. The preceding confidence set is not available for one-way layouts with one observation per factor level. Suppose that $n = p$ and $\hat{\sigma}^2 = \hat{\sigma}^2_H$. Let

$$\begin{aligned}\hat{h}_1 &= 2\hat{f} - 1 + [p/(p-q)][\text{ave}(1-2\hat{f})](1-e(q)) \\ \hat{h}_2 &= \hat{f} - 1 + [p/(p-q)][\text{ave}(1-2\hat{f})](1-e(q)).\end{aligned} \tag{3.4}$$

The distribution of $\hat{t}_{\mathcal{F},2}$ for large p and $p-q$ and small $p^{-1/2}\sum_{i=q+1}^{p}\xi_i^2$ is approximately $N(0, \hat{\tau}^2_{\mathcal{F},2})$ with

$$\hat{\tau}^2_{\mathcal{F},2} = 2\hat{\sigma}^4_H\text{ave}(\hat{h}_1^2) + [4\hat{\sigma}^2_H\text{ave}[(z^2-\hat{\sigma}^2_H)\hat{h}_2^2]_+. \tag{3.5}$$

The corresponding confidence of approximate coverage probability α for η is

$$\hat{C}_{\mathcal{F},2}(\alpha) = \{\theta \in \mathcal{M}(X)\colon |\hat{\eta}_{\mathcal{F}} - \theta|^2 \le p\hat{\rho}(\hat{f}) + p^{1/2}\hat{\tau}_{\mathcal{F},2}\Phi^{-1}(\alpha)\}. \tag{3.6}$$

3.1 Saturated and unsaturated fits

Visualizing either of the confidence sets $\hat{C}_{\mathcal{F},1}(\alpha)$ or $\hat{C}_{\mathcal{F},2}(\alpha)$ as subsets of the regression space $\mathcal{M}(X) \subset R^n$ is difficult at best. One useful way of interpreting such confidence sets centered at $\hat{\eta}_{\mathcal{F}}$ is to ask:

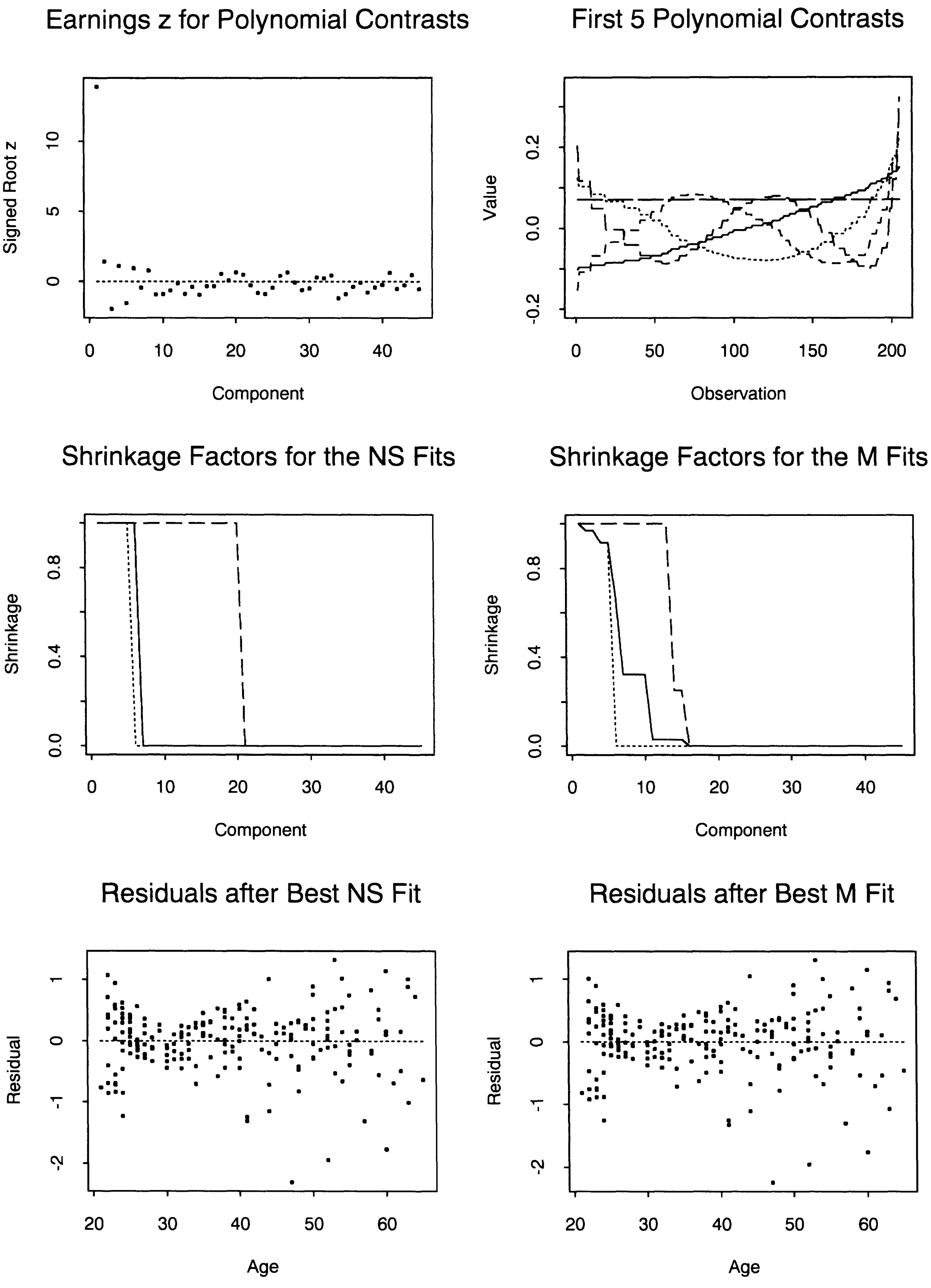

Figure 2. Diagnostic plots for REACT fits to the Canadian earnings data.

a) Which features in the estimator $\hat{\eta}_{\mathcal{F}}$ are not necessarily present in η once sampling error is taken into account?

b) Which features of η might have been smoothed out by the estimator $\hat{\eta}_{\mathcal{F}}$ because of sampling error?

We will construct extremal members of the confidence sets $\hat{C}_{\mathcal{F},i}(\alpha)$ that throw light on both of these questions. For the bases considered in this paper, these extremal elements amount to "smoothest" and "roughest" perturbations of the REACT fit that lie on the boundary of the confidence set.

A shrinkage vector $f \in [0,1]^p$ is said to be *saturated up to order* k if $f_1 = \ldots = f_k = 1$. It is said to be *unsaturated down to order* k if $f_k = \ldots = f_p = 0$. Let

$$\begin{aligned} \mathcal{F}_{M,u}(k) &= \{f \in \mathcal{F}_M \colon f \text{ is unsaturated down to order } k\} \\ \mathcal{F}_{M,s}(k) &= \{f \in \mathcal{F}_M \colon f \text{ is saturated up to order } k\}. \end{aligned} \tag{3.7}$$

Define

$$\hat{f}_{M,u}(k) = \operatorname*{argmin}_{f \in \mathcal{F}_{M,u}(k)} \hat{\rho}(f), \qquad \hat{f}_{M,s}(k) = \operatorname*{argmin}_{f \in \mathcal{F}_{M,s}(k)} \hat{\rho}(f). \tag{3.8}$$

Among the shrinkage vectors $\{\hat{f}_{M,u}(k)\colon k \geq 1\}$ such that the vector $U_E \text{diag}(\hat{f}_{M,u}(k))U_E' y$ lies in $\hat{C}_{\mathcal{F},i}(\alpha)$, let $\hat{f}_{M,u}$ be the one for which k is smallest. We say that $\hat{\eta}_{M,u} = U_E \text{diag}(\hat{f}_{M,u})U_E' y$ is the *most unsaturated α-confident M fit* for η. In the other direction, among the shrinkage vectors $\{\hat{f}_{M,s}(k)\colon k \geq 1\}$ such that $U_E \text{diag}(\hat{f}_{M,s}(k))U_E' y$ lies in $\hat{C}_{\mathcal{F},i}(\alpha)$, let $\hat{f}_{M,s}$ be the one for which k is largest. We say that $\hat{\eta}_{M,s} = U_E \text{diag}(\hat{f}_{M,s})U_E' y$ is the *most saturated α-confident M fit* for η.

Suppose that α is close to 1. Comparing $\hat{\eta}_{M,u}$ with $\hat{\eta}_M$ indicates which features of the the estimator $\hat{\eta}_M$ need not be present in η once allowance is made sampling error. This addresses question (a) above. On the other hand, comparing $\hat{\eta}_{M,s}$ with $\hat{\eta}_M$ indicates which features of η may have been smoothed out by the estimator $\hat{\eta}_M$ because of sampling error. This addresses question (b) above. The strategy just described for identifying interesting extremal members of confidence sets centered at $\hat{\eta}_M$ extends readily to confidence sets centered at $\hat{\eta}_{NS}$.

In the preceding discussion, the word "smooth" tacitly assumes that the column vectors in U_E are are of decreasing smoothness as column index increases. This is the case for the bases mentioned in Section 2.4. More generally, "smooth" should be be replaced by a description of the key feature that is increasingly captured as we move through the basis.

3.2 Further analysis of the case studies

The second row of Figure 1 exhibits the most unsaturated NS and M fits that lie within the 95% confidence balls for η centered, respectively, at $\hat{\eta}_{NS}$

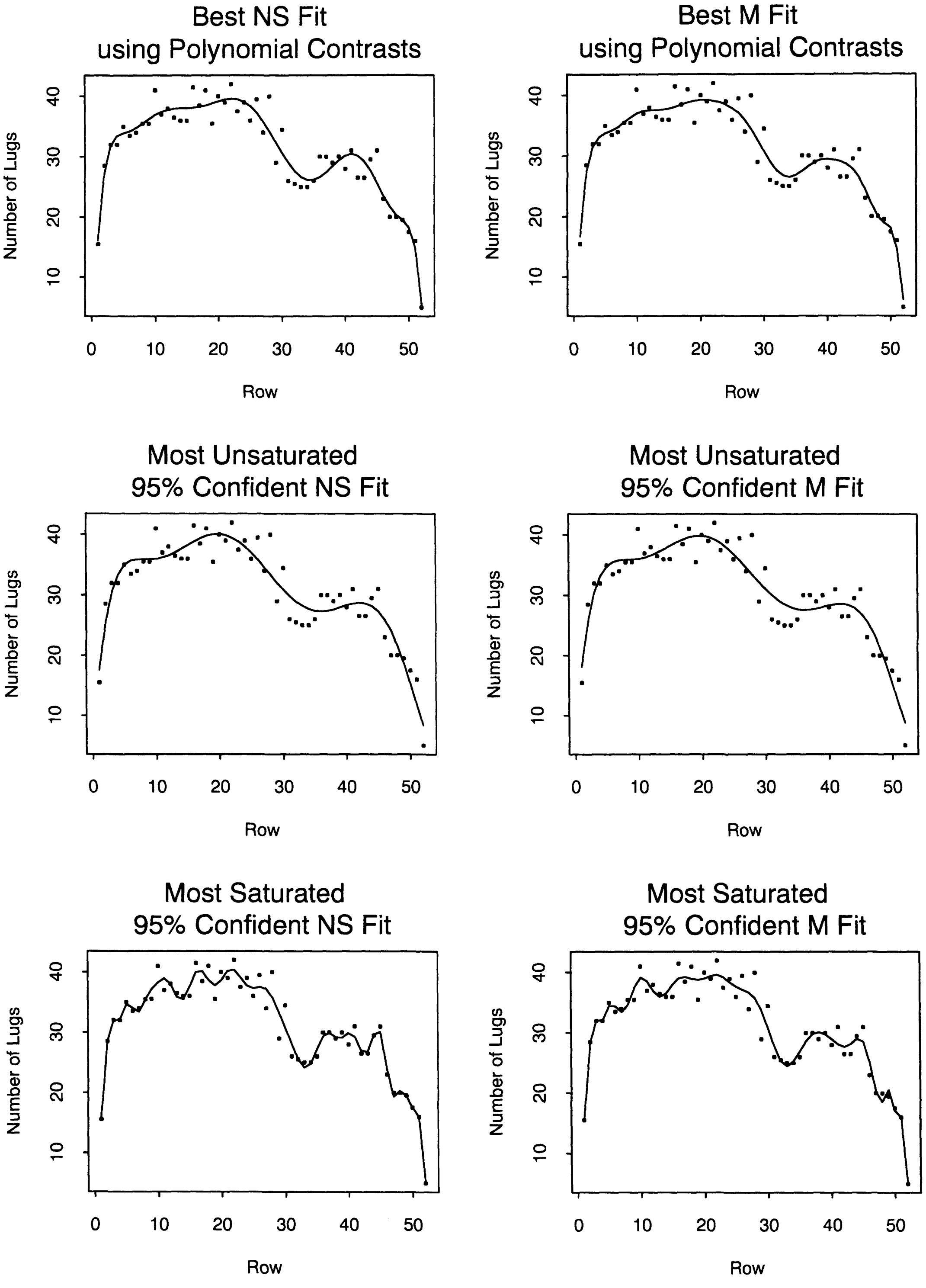

Figure 3. REACT fits and extremal confident fits to the vineyard harvest data.

and $\hat{\eta}_M$. That both extreme unsaturated fits retain the middle-aged dip in income makes it safe to conclude that mean log(income) η has this feature. Possible reasons for the dip include mid-life changes in career, re-entry of women into the work-force after child-rearing, down-sizing of middle-management positions, and so forth.

The third row in Figure 1 exhibits the most saturated NS and M fits that lie within the respective 95% confidence balls for η. These extreme saturated fits both point to two features that $\hat{\eta}_M$ and $\hat{\eta}_{NS}$ might have smoothed out: a small dip in mean log(income) around age 30 and greater variability in mean log(income) at ages 60 to 65. Such informed conjectures rest on the hypothesis of homoscedastic errors. Heteroscedastic errors would offer another possible explanation of the variation in observed log(income) between ages 60 to 65.

The $(2, 2)$ cell of Figure 2 compares the REACT shrinkage vector $\hat{f}_M$ (solid) line with the most unsaturated shrinkage vector $\hat{f}_{M,u}$ (dotted line) and the most saturated shrinkage vector $\hat{f}_{M,s}$ (dashed line). The $(2, 1)$ cell of the same figure makes the analogous comparisons of shrinkage vectors for the various NS fits.

The second and third rows of Figure 3 present the most unsaturated and most saturated M and NS fits to the vineyard data. We conclude from the most unsaturated fits that the "valley" in mean harvest around row number 35 cannot be discounted. The most saturated fits indicate additional possible local patterns in how mean harvest depends on row number. More polynomial contrasts are needed for the analysis of the vineyard data than for the earnings data (see the middle row in Figure 4).

4 Technical Matters

The methodology described in the preceding two sections has a firm foundation in asymptotic theory and in computational algorithms for isotonic weighted least squares. This section outlines the most salient points.

4.1 Asymptotics for confidence sets

Underpinning the two confidence sets described in Section 3 are the following theorems that determine the asymptotic distribution of $\hat{t}_{\mathcal{F},i}$ and establish the asymptotic coverage probability and asymptotic loss of $\hat{C}_{\mathcal{F},i}(\alpha)$. Let d be any metric for weak convergence of probability measures on the real line and let $\mathcal{L}(\hat{t}_{\mathcal{F},i})$ denote the distribution of $\hat{t}_{\mathcal{F},i}$ under the model.

Theorem 4.1 *Suppose that $\mathcal{F}$ is either $\mathcal{F}_{NS}$ or $\mathcal{F}_M$. For $i = 1$, assume that $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$, $m = \min(p, n-p)$, and $\lim_{m\to\infty} p/(n-p) = \gamma^2 < \infty$. For $i = 2$, assume that $\hat{\sigma}^2 = \hat{\sigma}^2_H$, $m = \min(p, p-q)$, $\lim_{m\to\infty} p/(p-q) = \beta^2 < \infty$,*

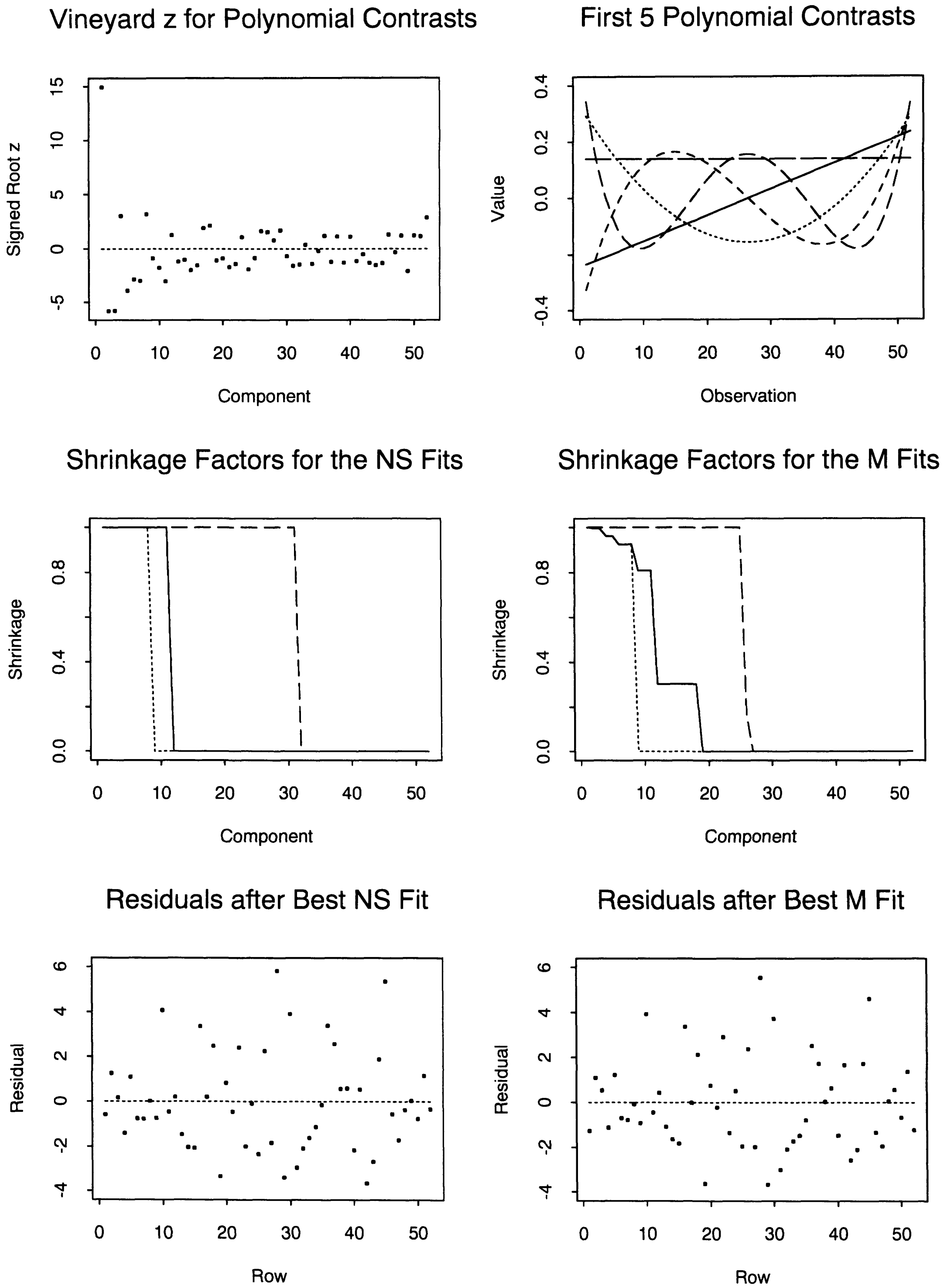

Figure 4. Diagnostic plots for REACT fits to the vineyard harvest data.

and $\lim_{m\to\infty} p^{-1/2}\sum_{i=q+1}^{p}\xi_i^2 = 0$. *Then, for every* $r > 0$ *and every* $\sigma^2 > 0$,

$$\lim_{m\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} d[\mathcal{L}(\hat{t}_{\mathcal{F},i}), N(0,\tau^2_{\mathcal{F},i})] = 0, \tag{4.1}$$

where

$$\begin{aligned}\tau^2_{\mathcal{F},1} = 2\sigma^4\text{ave}[(2\tilde{f}-1)^2] + 2[p/(n-p)]\sigma^4[\text{ave}(2\tilde{f}-1)]^2\\ + 4\sigma^2\text{ave}[\xi^2(1-\tilde{f})^2]\end{aligned} \tag{4.2}$$

and

$$\tau^2_{\mathcal{F},2} = 2\sigma^4\text{ave}(\tilde{h}_1^2) + 4\sigma^2\text{ave}(\tilde{h}_2^2\xi^2) \tag{4.3}$$

with

$$\begin{aligned}\tilde{h}_1 &= 2\tilde{f} - 1 + [p/(p-q)][\text{ave}(1-2\tilde{f})](1-e(q))\\ \tilde{h}_2 &= \tilde{f} - 1 + [p/(p-q)][\text{ave}(1-2\tilde{f})](1-e(q)).\end{aligned} \tag{4.4}$$

The variance $\tau^2_{\mathcal{F},i}$ depends on p, ξ^2, σ^2, and on $n-p$ or $p-q$ according to i. The estimators $\hat{\tau}^2_{\mathcal{F},i}$ defined in (3.2) and (3.5) substitute $z^2 - \hat{\sigma}^2$ for ξ^2, $\hat{f}$ for $\tilde{f}$, and $\hat{\sigma}^2$ for σ^2, constraining the estimator of the last term on the right of (4.2) and (4.3) to be non-negative and using the appropriate definition of $\hat{\sigma}^2$. The next theorem establishes that the α-th quantile of the $N(0,\tau^2_{\mathcal{F},i})$ distribution is consistently estimated by $\hat{\tau}_{\mathcal{F},i}\Phi^{-1}(\alpha)$. This leads to the definitions of the confidence sets $\hat{C}_{\mathcal{F},i}(\alpha)$ in (3.3) and (3.6).

Let $\hat{r}^2_{\mathcal{F},i} = \hat{\rho}(\hat{f}) + p^{-1/2}\hat{\tau}_{\mathcal{F},i}\Phi^{-1}(\alpha)$. Of interest are two properties of $\hat{C}_{\mathcal{F},i}(\alpha)$: the coverage probability $\mathrm{P}(\eta \in \hat{C}_{\mathcal{F},i}(\alpha))$ and the geometrical quadratic loss

$$L(\hat{C}_{\mathcal{F},i}(\alpha),\eta) = \sup_{\theta\in\hat{C}_{\mathcal{F},i}(\alpha)} p^{-1}|\theta-\eta|^2 = [p^{-1/2}|\hat{\eta}_{\mathcal{F}} - \eta| + \hat{r}_{\mathcal{F},i}]^2. \tag{4.5}$$

Treating $\hat{C}_{\mathcal{F},i}(\alpha)$ as a set-valued estimator of η, this geometrical loss measures how poorly elements of the confidence set can estimate η.

Theorem 4.2 *Under the hypotheses of Theorem 4.1, for every* $r > 0$ *and every* $\sigma^2 > 0$,

$$\begin{aligned}\lim_{m\to\infty,\,K\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} \mathrm{P}[|L(\hat{C}_{\mathcal{F},i}(\alpha),\eta) - 4\rho(\tilde{f},\xi^2,\sigma^2)| \ge Kp^{-1/2}] = 0\\ \lim_{m\to\infty,\,K\to\infty} \sup_{\text{ave}(\xi^2)\le\sigma^2 r} \mathrm{P}[|\hat{r}^2_{\mathcal{F},i} - \rho(\tilde{f},\xi^2,\sigma^2)| \ge Kp^{-1/2}] = 0.\end{aligned} \tag{4.6}$$

For every $\epsilon > 0$,

$$\lim_{m\to\infty} \mathrm{P}[|\hat{\tau}^2_{\mathcal{F},i} - \tau^2_{\mathcal{F},i}| > \epsilon] = 0. \tag{4.7}$$

Moreover,

$$\liminf_{m\to\infty} \inf_{\mathrm{ave}(\xi^2)\le\sigma^2 r} \tau^2_{\mathcal{F},i} > 0 \tag{4.8}$$

and

$$\lim_{m\to\infty} \sup_{\mathrm{ave}(\xi^2)\le\sigma^2 r} |\mathrm{P}(\eta \in \hat{C}_{\mathcal{F},i}(\alpha)) - \alpha| = 0. \tag{4.9}$$

Theorems 3.1 and 3.2 in Beran and Dümbgen (1998) imply the two theorems above for the case $i = 1$. The results for $i = 2$ are proved by straightforward modification of the argument, the salient difference being that

$$\hat{\sigma}^2_H - \sigma^2 = [p/(p-q)]\{\mathrm{ave}[\bar{e}(q)W_1] + 2\mathrm{ave}[\bar{e}(q)W_2] + \mathrm{ave}[\bar{e}(q)\xi^2]\} \tag{4.10}$$

with $\bar{e}(q) = 1 - e(q)$, $W_1 = (z-\xi)^2 - \sigma^2$ and $W_2 = \xi(z-\xi)$.

According to Theorem 4.2, the asymptotic geometrical loss of each confidence set is four times the asymptotic risk of the estimator at the center of the confidence. This is a compelling reason for using confidence sets centered at superefficient REACT estimators in place of the classical confidence set centered at the least squares estimator.

4.2 Computing saturated and unsaturated fits

Computation of $\hat{f}_{NS}$ and of the unsaturated and saturated nested selection shrinkage vectors $\hat{f}_{NS,u}(k)$ and $\hat{f}_{NS,s}(k)$ is accomplished by finite search. To compute $\hat{f}_M$, let $\hat{g} = (z^2 - \hat{\sigma}^2)/z^2$ and observe that

$$\hat{f}_M = \operatorname*{argmin}_{f\in\mathcal{F}_M} \hat{\rho}(f) = \operatorname*{argmin}_{f\in\mathcal{F}_M} \mathrm{ave}[(f-\hat{g})^2 z^2]. \tag{4.11}$$

Let $\mathcal{B} = \{b \in R^p\colon b_1 \ge b_2 \ge \ldots \ge b_p\}$. A further argument given in Beran and Dümbgen (1998) shows that

$$\hat{f}_M = \check{f}_+ \quad \text{with} \quad \check{f} = \operatorname*{argmin}_{b\in\mathcal{B}} \mathrm{ave}[(b-\hat{g})^2 z^2]. \tag{4.12}$$

Algorithms for isotonic weighted least squares yield $\check{f}$.

When f is unsaturated down to order k,

$$\hat{\rho}(f) = p^{-1}\sum_{i=1}^{k-1}[\hat{\sigma}^2 f_i^2 + (z_i^2 - \hat{\sigma}^2)(1-f_i)^2] + p^{-1}\sum_{i=k}^{p}(z_i^2 - \hat{\sigma}^2). \tag{4.13}$$

On the other hand, when f is saturated up to order k,

$$\hat{\rho}(f) = p^{-1}k\hat{\sigma}^2 + p^{-1}\sum_{i=k+1}^{p}[\hat{\sigma}^2 f_i^2 + (z_i^2 - \hat{\sigma}^2)(1-f_i)^2]. \tag{4.14}$$

Thus, as in the preceding paragraph, algorithms for isotonic weighted least squares suffice to compute $\hat{f}_{M,u}(k)$ and $\hat{f}_{M,s}(k)$.

Acknowledgements Professor J. S. Marron introduced the author to the Canadian earnings data. The stimulus provided by participants in a Weekend Seminar at Oberflockenbach in May 1998 is gratefully noted.

REFERENCES

Beran, R., (1996). Confidence sets centered at C_p estimators. *Annals of the Institute of Statistical Mathematics* **48**, 1–15.

Beran, R., (2000). REACT scatterplot smoothers: superefficiency through basis economy. *Journal of the American Statistical Society* **95**, in press.

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

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

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

Donoho, D.L., and Johnstone, I.M., (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

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

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

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

Li, K.-C., (1987). Asymptotic optimality for C_p, C_L and generalized cross-validation: discrete index set. *Annals of Statistics* **15**, 958–976.

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

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

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

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

Stein, C., (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. *Research Papers in Statistics: Festschrift for Jerzy Neyman* (F.N. David, ed.), 351–366. Wiley, London.

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

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

Department of Statistics
University of California, Berkeley
Berkeley, CA 94720–3860
USA
beran@stat.berkeley.edu