Estimation and Accuracy After Model Selection

Bradley EFRON

Classical statistical theory ignores model selection in assessing estimation accuracy. Here we consider bootstrap methods for computing standard errors and confidence intervals that take model selection into account. The methodology involves bagging, also known as bootstrap smoothing, to tame the erratic discontinuities of selection-based estimators. A useful new formula for the accuracy of bagging then provides standard errors for the smoothed estimators. Two examples, nonparametric and parametric, are carried through in detail: a regression model where the choice of degree (linear, quadratic, cubic, ...) is determined by the C_p criterion and a Lasso-based estimation problem.

KEY WORDS: ABC intervals; Bagging; Bootstrap smoothing; C_p; Importance sampling; Lasso; Model averaging.

1. INTRODUCTION

Accuracy assessments of statistical estimators customarily are made ignoring model selection. A preliminary look at the data might, for example, suggest a cubic regression model, after which the fitted curve's accuracy is computed as if "cubic" were pre-chosen. Here we will discuss bootstrap standard errors and approximate confidence intervals that take into account the model-selection procedure.

Figure 1 concerns the *Cholesterol data*, an example investigated in more detail in Section 2: $n = 164$ men took cholestyramine, a proposed cholesterol-lowering drug, for an average of seven years each; the response variable was the *decrease* in blood-level cholesterol measured from the beginning to the end of the trial,

$$d = \text{cholesterol decrease}; \tag{1.1}$$

also measured (by pill counts) was *compliance*, the proportion of the intended dose taken,

$$c = \text{compliance}, \tag{1.2}$$

ranging from zero to full compliance for the 164 men. A transformation of the observed proportions has been made here so that the 164 c values approximate a standard normal distribution,

$$c \doteq \mathcal{N}(0, 1). \tag{1.3}$$

The solid curve is a regression estimate of decrease d as a cubic function of compliance c, fit by ordinary least squares (OLS) to the 164 points. "Cubic" was selected by the C_p criterion (Mallows 1973), as described in Section 2. The question of interest for us is *how accurate is the fitted curve*, taking account of the C_p model-selection procedure as well as OLS estimation?

More specifically, let μ_j be the expectation of cholesterol decrease for subject j given his compliance c_j,

$$\mu_j = E\{d_j | c_j\}. \tag{1.4}$$

We wish to assign standard errors to estimates of μ_j read from the regression curve in Figure 1. A nonparametric bootstrap estimate $\widetilde{\text{sd}}_j$ of standard deviation, taking account of model selection, is developed in Sections 2 and 3. Figure 2 shows that this is usually, but not always, greater than the naive estimate $\overline{\text{sd}}_j$ obtained from standard OLS calculations, assuming that the cubic model was preselected. The ratio $\widetilde{\text{sd}}_j / \overline{\text{sd}}_j$ has median value 1.52; so at least in this case, ignoring model selection can be deceptively optimistic.

Data-based model selection can produce "jumpy" estimates that change values discontinuously at the boundaries between model regimes. *Bagging* (Breiman 1996), or *bootstrap smoothing*, is a model-averaging device that both reduces variability and eliminates discontinuities. This is described in Section 2, and illustrated on the Cholesterol data.

Our key result is a new formula for the delta-method standard deviation of a bagged estimator. The result, which applies to general bagging situations and not just regression problems, is described in Section 3. Stated in projection terms (see Figure 4), it provides the statistician a direct assessment of the cost in reduced accuracy due to model selection.

A parametric bootstrap version of the smoothing theory is described in Sections 4 and 5. Parametric modeling allows more refined results, permitting second order-accurate confidence calculations of the BCa or ABC type, as in DiCiccio and Efron (1992), Section 6. Section 7 concludes with notes, details, and deferred proofs.

Bagging (Breiman 1996) has become a major technology in the prediction literature, an excellent recent reference being Buja and Stuetzle (2006). The point of view here agrees with that in Bühlmann and Yu (2002), though their emphasis is more theoretical and less data-analytic. They employ bagging to "change hard thresholding estimators to soft thresholding," in the same spirit as our Section 2.

Berk et al. (2012) developed conservative normal-theory confidence intervals that are guaranteed to cover the true parameter value regardless of the preceding model-selection procedure. Very often it may be difficult to say just what selection procedure was used, in which case the conservative intervals are appropriate. The methods of this article assume that the model-selection procedure *is* known, yielding smaller standard error estimates and shorter confidence intervals.

Hjort and Claeskens (2003) constructed an ambitious large-sample theory of frequentist model-selection estimation and model averaging, while making comparisons with Bayesian methods. In theory, the Bayesian approach offers an

Bradley Efron is Professor, Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065 (E-mail: *brad@stat.stanford.edu*). Research supported in part by NIH grant 8R01 EB002784 and by NSF grant DMS 1208787. The author is grateful to Stefan Wager for the bias correction idea in Remark J of this article.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2013.823775

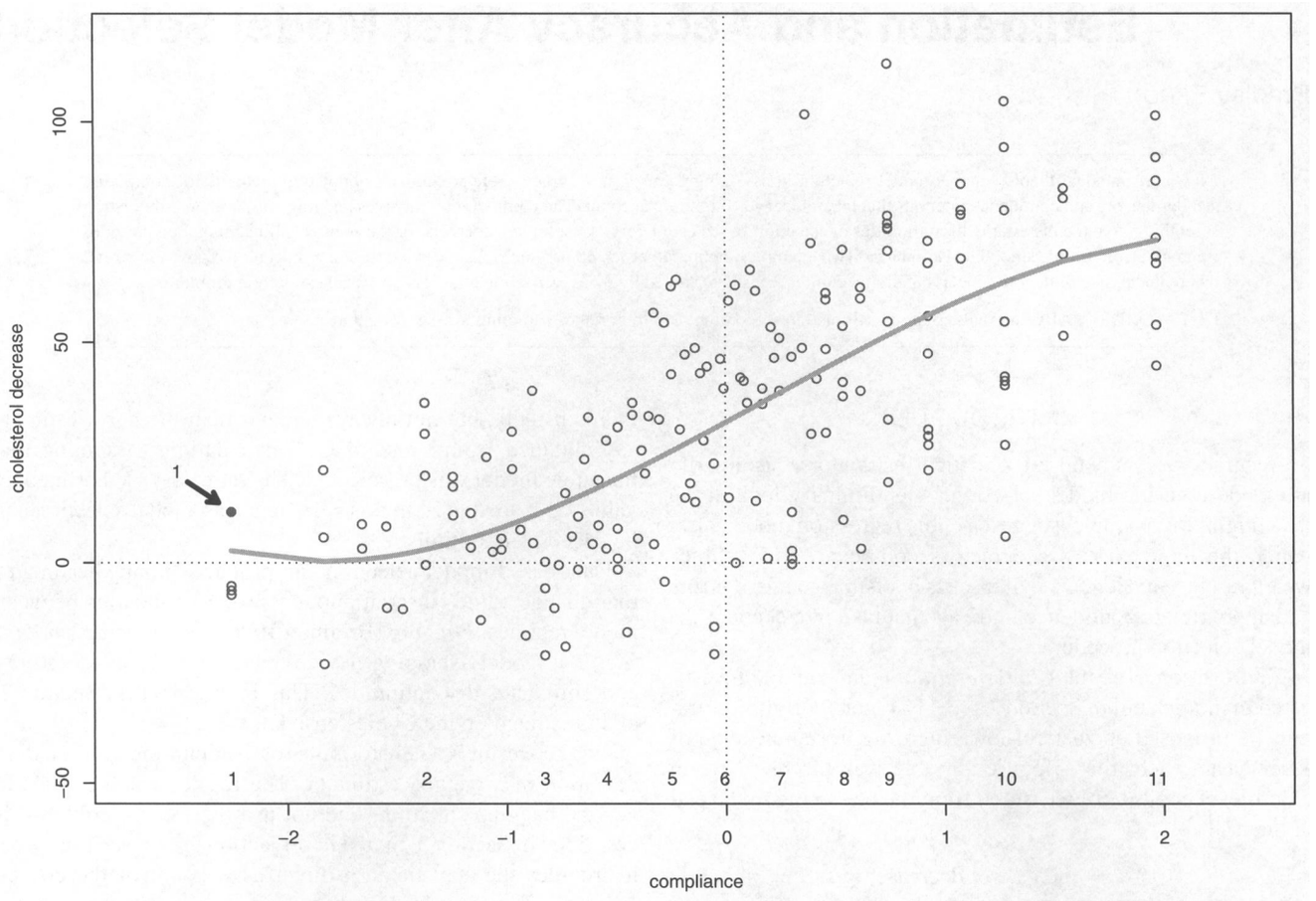

Figure 1. *Cholesterol data* Cholesterol decrease plotted versus adjusted compliance for 164 men in Treatment arm of the cholostyramine study (Efron and Feldman 1991). Solid curve is OLS cubic regression, as selected by the C_p criterion. How accurate is the curve, taking account of model selection as well as least squares fitting? (Solid arrowed point is Subject 1, featured in subsequent calculations. Bottom numbers indicate compliance for the 11 subjects in the simulation trial of 5.)

ideal solution to model-selection problems, but, as Hjort and Claeskens pointed out, it requires an intimidating amount of prior knowledge from the statistician. The present article is frequentist in its methodology.

Hurvich and Tsai (1990) provided a nice discussion of what "frequentist" might mean in a model-selection framework. (Here I am following their "overall" interpretation.) The nonparametric bootstrap approach in Buckland, Burnham, and Augustin (1997) has a similar flavor to the computations in Section 2. A particularly apt reference is Sexton and Laake (2009), who also provide modified bootstrap estimates of accuracy for estimators involving model selection.

Classical estimation theory ignored model selection out of necessity. Armed with modern computational equipment, statisticians can now deal with model-selection problems more realistically. The limited, but useful, goal of this article is to provide a general tool for the assessment of standard errors in such situations. Simple parameters like (1.4) are featured in our examples, but the methods apply just as well to more complicated functionals, for instance the maximum value of a regression surface, or a tree-based estimate.

2. NONPARAMETRIC BOOTSTRAP SMOOTHING

For the sake of simple notation, let $\boldsymbol{y}$ represent all the observed data, and $\hat{\mu} = t(\boldsymbol{y})$ an estimate of a parameter of interest μ. The Cholesterol data have

$$\boldsymbol{y} = \{(c_j, d_j),\ j = 1, 2, \ldots, n = 164\}. \tag{2.1}$$

If $\mu = \mu_j$ (1.4) we might take $\hat{\mu}_j$ to be the height of the C_p-OLS regression curve measured at compliance $c = c_j$.

In a nonparametric setting, we have data

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n), \tag{2.2}$$

where the y_j are independent and identically distributed (iid) observations from an unknown distribution F, a two-dimensional distribution in situation (2.1). The parameter is some functional $\mu = T(F)$, but the plug-in estimator $\hat{\mu} = T(\hat{F})$, where $\hat{F}$ is the empirical distribution of the y_j values, is usually what we hope to improve upon in model-selection situations.

A nonparametric bootstrap sample

$$\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_n^*) \tag{2.3}$$

consists of n draws *with replacement* from the set $\{y_1, y_2, \ldots, y_n\}$, yielding bootstrap replication $\hat{\mu}^* = t(\boldsymbol{y}^*)$. The

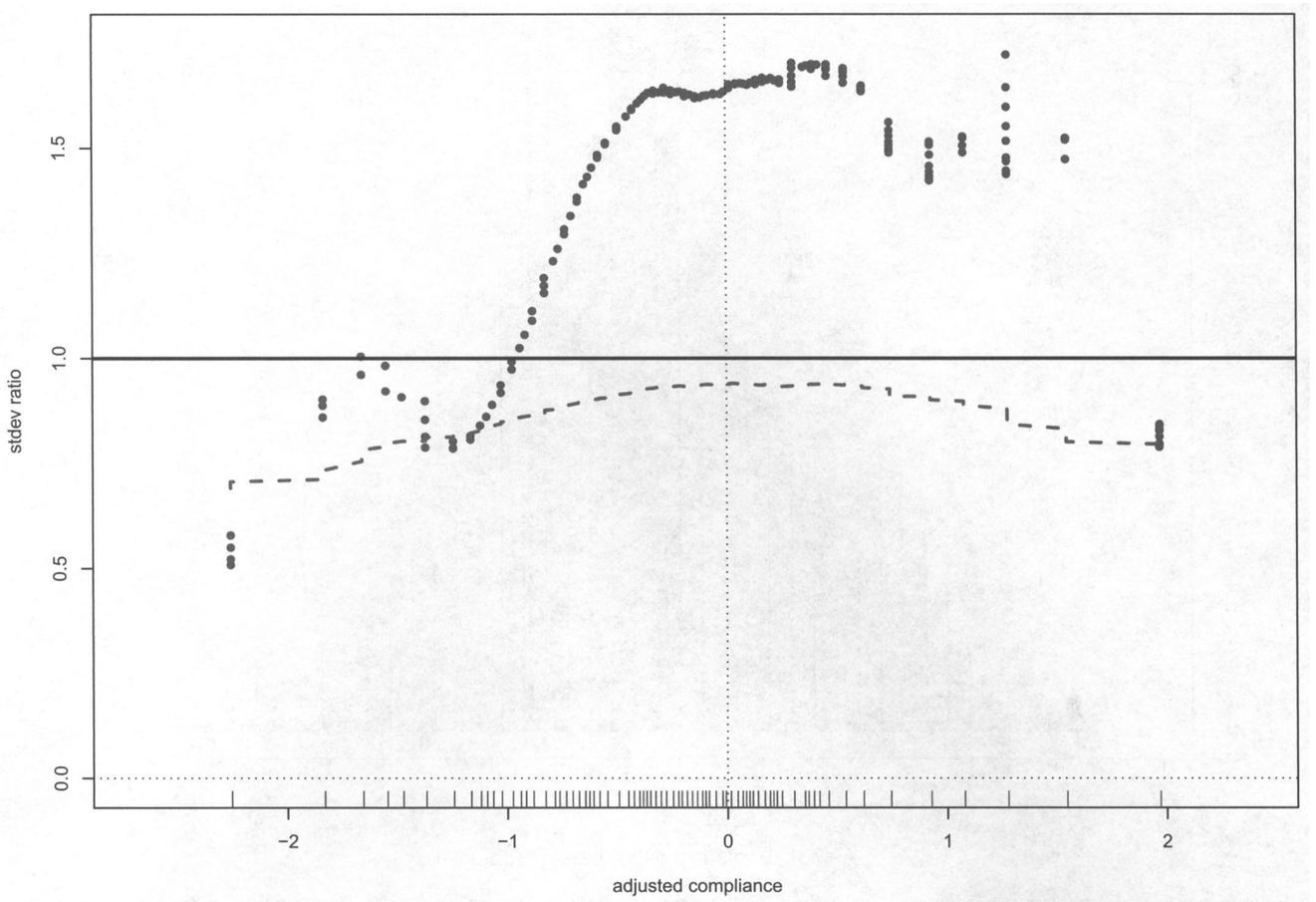

Figure 2. *Solid points*: ratio of standard deviations, taking account of model selection or not, for the 164 values $\hat{\mu}_j$ from the regression curve in Figure 1. Median ratio equals 1.52. Standard deviations including model selection are the smoothed bootstrap estimates $\widetilde{\mathrm{sd}}_B$ of Section 3. *Dashed line*: ratio of $\widetilde{\mathrm{sd}}_B$ to $\widehat{\mathrm{sd}}_B$, the unsmoothed bootstrap sd estimates as in (2.4), median 0.91.

empirical standard deviation of B such draws,

$$\widehat{\mathrm{sd}}_B = \left[\sum_{i=1}^{B}(\hat{\mu}_i^* - \hat{\mu}_\cdot^*)^2 \Big/ (B-1)\right]^{1/2}, \qquad \left(\hat{\mu}_\cdot^* = \sum \hat{\mu}_i^*/B\right), \quad (2.4)$$

is the familiar nonparametric bootstrap estimate of standard error for $\hat{\mu}$ (Efron 1979); $\widehat{\mathrm{sd}}_B$ is a dependable accuracy estimator in most standard situations but, as we will see, it is less dependable for setting approximate confidence limits in model-selection contexts.

The cubic regression curve in Figure 1 was selected using the C_p criterion. Suppose that under "Model m" we have

$$y = X_m\beta_m + \epsilon \qquad [\epsilon \sim (0, \sigma^2 I)], \quad (2.5)$$

where X_m is a given n by m structure matrix of rank m, and ϵ has mean $\mathbf{0}$ and covariance σ^2 times the Identity (σ assumed known in what follows). The C_p measure of fit for Model m is

$$C_p(m) = \|y - X_m\hat{\beta}_m\|^2 + 2\sigma^2 m \quad (2.6)$$

with $\hat{\beta}_m$ the OLS estimate of β_m; given a collection of possible choices for the structure matrix, the C_p criterion selects the one minimizing C_p.

Table 1 shows C_p results for the Cholesterol data. Six polynomial regression models were compared, ranging from linear ($m = 2$) to sixth degree ($m = 7$); the value $\sigma = 22.0$ was used, corresponding to the standard estimate $\hat{\sigma}$ obtained from the sixth degree model. The cubic model ($m = 4$) minimized $C_p(m)$, leading to its selection in Figure 1.

$B = 4000$ nonparametric bootstrap replications of the C_p-OLS regression curve—several times more than necessary, see Section 3—were generated: starting with a bootstrap sample

Table 1. C_p model selection for the Cholesterol data; measure of fit $C_p(m)$ (2.6) for polynomial regression models of increasing degree. The cubic model minimizes $C_p(m)$. (Value $\sigma = 22.0$ was used here and in all bootstrap replications.) Last column shows percentage each model was selected as the C_p minimizer, among $B = 4000$ bootstrap replications

Regression model	m	$C_p(m) - 80{,}000$	(Bootstrap %)
Linear	2	1132	(19%)
Quadratic	3	1412	(12%)
Cubic	4	**667**	(34%)
Quartic	5	1591	(8%)
Quintic	6	1811	(21%)
Sextic	7	2758	(6%)

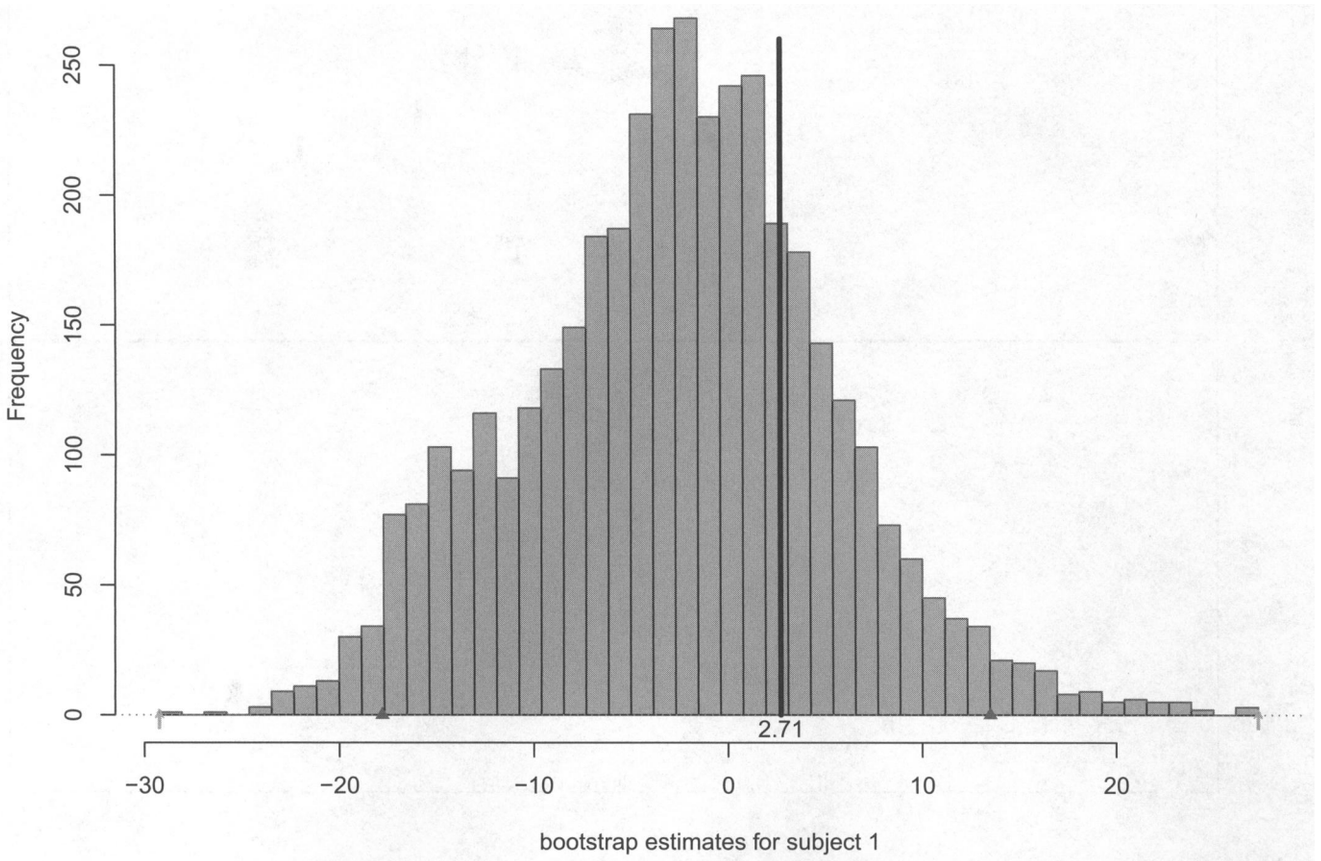

Figure 3. $B = 4000$ bootstrap replications $\hat{\mu}_1^*$ of the C_p-OLS regression estimate for Subject 1. The original estimate $t(\boldsymbol{y}) = \hat{\mu}_1$ is 2.71, exceeding 76% of the replications. Bootstrap standard deviation (2.4) equals 8.02. Triangles indicate 2.5th and 97.5th percentiles of the histogram.

$\boldsymbol{y}^*$ (2.3), the equivalent of Table 1 was calculated (still using $\sigma = 22.0$) and the C_p minimizing degree m^* selected, yielding the bootstrap regression curve

$$\hat{\boldsymbol{\mu}}^* = X_{m^*}\hat{\beta}^*_{m^*}, \tag{2.7}$$

where $\hat{\beta}^*_{m^*}$ was the OLS coefficient vector for the selected model. The last column of Table 1 shows the various bootstrap model-selection percentages: cubic was selected most often, but still only about one-third of the time.

Suppose we focus attention on Subject 1, the arrowed point in Figure 1, so that the parameter of interest μ_1 can be estimated by the C_p-OLS value $t(\boldsymbol{y}) = \hat{\mu}_1$, evaluated to be 2.71. Figure 3 shows the histogram of the 4000 bootstrap replications $t(\boldsymbol{y}^*) = \hat{\mu}_1^*$. The point estimate $\hat{\mu}_1 = 2.71$ is located to the right, exceeding a surprising 76% of the $\hat{\mu}_1^*$ values.

Table 2 shows why. The cases where "Cubic" was selected yielded the largest bootstrap estimates $\hat{\mu}_1^*$. The actual dataset $\boldsymbol{y}$ fell into the cubic region, giving a correspondingly large estimate $\hat{\mu}_1$. Things might very well have turned out otherwise, as the bootstrap replications suggest: model selection can make an estimate "jumpy" and erratic.

Table 2. Mean and standard deviation of $\hat{\mu}_1^*$ as a function of the selected model, 4000 nonparametric bootstrap replications; Cubic, Model 3, gave the largest estimates

Model	1	2	3	4	5	6
Mean	−13.69	−3.69	**4.71**	−1.25	−3.80	−3.56
St. dev.	3.64	3.48	5.43	5.28	4.46	4.95

We can smooth $\hat{\mu} = t(\boldsymbol{y})$ by averaging over the bootstrap replications, defining

$$\tilde{\mu} = s(\boldsymbol{y}) = \frac{1}{B}\sum_{i=1}^{B} t(\boldsymbol{y}^*). \tag{2.8}$$

Bootstrap smoothing (Efron and Tibshirani 1996), a form of model averaging, is better known as "bagging" in the prediction literature; see Breiman (1996) and Buja and Stuetzle (2006). There its variance reduction properties are emphasized. Our example will also show variance reductions, but the main interest here lies in smoothing; $s(\boldsymbol{y})$, unlike $t(\boldsymbol{y})$, does not jump as $\boldsymbol{y}$ crosses region boundaries, making it a more dependable vehicle for setting standard errors and confidence intervals. Suppose, for definiteness, that we are interested in setting approximate 95% bootstrap confidence limits for parameter μ. The usual "standard interval"

$$\hat{\mu} \pm 1.96\,\widehat{\mathrm{sd}}_B \tag{2.9}$$

($= 2.71 \pm 1.96 \cdot 8.02$ in Figure 3) inherits the dangerous jumpiness of $\hat{\mu} = t(\boldsymbol{y})$. The *percentile interval*, section 13.3 of Efron and Tibshirani (1993),

$$\left[\hat{\mu}^{*(0.025)}, \hat{\mu}^{*(0.975)}\right], \tag{2.10}$$

the 2.5th and 97.5th percentiles of the B bootstrap replications, yields more stable results. (Notice that it does not require a central point estimate such as $\hat{\mu}$ in (2.9).)

Table 3. Three approximate 95% bootstrap confidence intervals for μ_1, the response value for Subject 1, Cholesterol data

	Interval	Length	Center point
Standard interval (2.9)	(−13.0, 18.4)	31.4	2.71
Percentile interval (2.10)	(−17.8, 13.5)	31.3	−2.15
Smoothed standard (2.11)	(−13.3, 8.0)	21.3	−2.65

A third choice, of particular interest, here, is the *smoothed interval*

$$\tilde{\mu} \pm 1.96\,\widetilde{\mathrm{sd}}_B, \tag{2.11}$$

where $\tilde{\mu} = s(\boldsymbol{y})$ is the bootstrap smoothed estimate (2.8), while $\widetilde{\mathrm{sd}}_B$ is given by the projection formula discussed in Section 3. Interval (2.11) combines stability with reduced length.

Table 3 compares the three approximate 95% intervals for μ_1. The reduction in length is dramatic here, though less so for the other 163 subjects; see Section 3.

The BCa-ABC system goes beyond (2.9)–(2.11) to produce bootstrap confidence intervals having second-order accuracy, as in DiCiccio and Efron (1992). Section 6 carries out the ABC calculations in a parametric bootstrap context.

3. ACCURACY OF THE SMOOTHED BOOTSTRAP ESTIMATES

The smoothed standard interval $\tilde{\mu} \pm 1.96\,\widetilde{\mathrm{sd}}_B$ requires a standard deviation assessment $\widetilde{\mathrm{sd}}_B$ for the smoothed bootstrap estimate (2.8). A brute force approach employs a second level of bootstrapping: resampling *from* $\boldsymbol{y}_i^*$ (2.3) yields a collection of B second-level replications $\boldsymbol{y}_{ij}^{**}$, from which we calculate $s_i^* = \sum t(\boldsymbol{y}_{ij}^{**})/B$; repeating this whole process for many replications of $\boldsymbol{y}_i^*$ provides bootstrap values s_i^* from which we calculate its bootstrap standard deviation.

The trouble with brute force is that it requires an enormous number of recomputations of the original statistic $t(\cdot)$. This section describes an estimate $\widetilde{\mathrm{sd}}_B$ that uses only the original B bootstrap replications $\{t(\boldsymbol{y}_i^*),\ i = 1, 2, \ldots, B\}$.

The theorem that follows will be stated in terms of the "ideal bootstrap," where B equals all n^n possible choices of $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$ from $\{y_1, y_2, \ldots, y_n\}$, each having probability $1/B$. It will be straightforward then to adapt our results to the nonideal bootstrap, with $B = 4000$ for instance.

Define

$$t_i^* = t(\boldsymbol{y}_i^*) \qquad [\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*, \ldots, y_{ik}^*, \ldots, y_{in}^*)], \tag{3.1}$$

the ith bootstrap replication of the statistic of interest, and let

$$Y_{ij}^* = \#\{y_{ik}^* = y_j\}, \tag{3.2}$$

the number of elements of $\boldsymbol{y}_i^*$ equaling the original data point y_j. The vector $\boldsymbol{Y}_i^* = (Y_{i1}^*, Y_{i2}^*, \ldots, Y_{in}^*)$ follows a multinomial distribution with n draws on n categories each of probability $1/n$, and has mean vector and covariance matrix

$$\boldsymbol{Y}_i^* \sim (\mathbf{1}_n, \boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n'/n), \tag{3.3}$$

$\mathbf{1}_n$ the vector of n 1's and $\boldsymbol{I}$ the $n \times n$ identity matrix.

Theorem 1. The nonparametric delta-method estimate of standard deviation for the ideal smoothed bootstrap statistic $s(\boldsymbol{y}) = \sum_{i=1}^B t(\boldsymbol{y}_i^*)/B$ is

$$\widetilde{\mathrm{sd}} = \left[\sum_{j=1}^n \mathrm{cov}_j^2\right]^{1/2}, \tag{3.4}$$

where

$$\mathrm{cov}_j = \mathrm{cov}_*(Y_{ij}^*, t_i^*), \tag{3.5}$$

the bootstrap covariance between Y_{ij}^* and t_i^*.

(The proof appears later in this section.)

The estimate of standard deviation for $s(\boldsymbol{y})$ in the nonideal case is the analogue of (3.4),

$$\widetilde{\mathrm{sd}}_B = \left[\sum_{j=1}^n \widehat{\mathrm{cov}}_j^2\right]^{1/2}, \tag{3.6}$$

where

$$\widehat{\mathrm{cov}}_j = \sum_{i=1}^n (Y_{ij}^* - Y_{\cdot j}^*)(t_i^* - t_\cdot^*)/B \tag{3.7}$$

with $Y_{\cdot j}^* = \sum_{i=1}^B Y_{ij}^*/B$ and $t_\cdot^* = \sum_{i=1}^B t_i^*/B = s(\boldsymbol{y})$. Remark J concerns a bias correction for (3.6) that can be important in the non-ideal case (it wasn't in the Cholesterol example). All of these results *apply generally to bagging estimators*, and are not restricted to regression situations.

Figure 2 shows that $\widetilde{\mathrm{sd}}_B$ is less than $\widehat{\mathrm{sd}}_B$, the bootstrap estimate of standard deviation for the unsmoothed statistic,

$$\widehat{\mathrm{sd}}_B = \left[\sum (t_i^* - t_\cdot^*)^2/B\right]^{1/2}, \tag{3.8}$$

for all 164 estimators $t(\boldsymbol{y}) = \hat{\mu}_j$. This is no accident. Returning to the ideal bootstrap situation, let $\mathcal{L}(\boldsymbol{Y}^*)$ be the $(n-1)$-dimensional subspace of $\mathcal{R}^B$ spanned by the columns of the $B \times n$ matrix having elements $Y_{ij}^* - 1$. [Notice that $\sum_{i=1}^B Y_{ij}^*/B = 1$ according to (3.3).] Also define $s_0 = \sum_{i=1}^B t_i^*/B$, the ideal bootstrap smoothed estimate, so

$$\boldsymbol{U}^* \equiv \boldsymbol{t}^* - s_0\mathbf{1} \tag{3.9}$$

is the B-vector of mean-centered replications $t_i^* - s_0$. *Note*: Formula (3.6) is a close cousin of the "jackknife-after-bootstrap" method of Efron (1992), the difference being the use of jackknife rather than our infinitesimal jackknife calculations.

Corollary 1. The ratio $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ is given by

$$\frac{\widetilde{\mathrm{sd}}_B}{\widehat{\mathrm{sd}}_B} = \frac{\|\hat{\boldsymbol{U}}^*\|}{\|\boldsymbol{U}^*\|} \tag{3.10}$$

where $\hat{\boldsymbol{U}}^*$ is the projection of $\boldsymbol{U}^*$ into $\mathcal{L}(\boldsymbol{Y}^*)$.

(See Remark A in Section 7 for the proof. Remark B concerns the relation of Theorem 1 to the *Hájek projection*.)

The illustration in Figure 4 shows $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ as the cosine of the angle between $\boldsymbol{t}^* - s_0\mathbf{1}$ and $\mathcal{L}(\boldsymbol{Y}^*)$. The ratio is a measure of the nonlinearity of t_i^* as a function of the bootstrap counts Y_{ij}^*. Model selection induces discontinuities in $t(\cdot)$, increasing the nonlinearity and decreasing $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$. The 164 ratios shown as the dashed line in Figure 2 had median 0.91, mean 0.89.

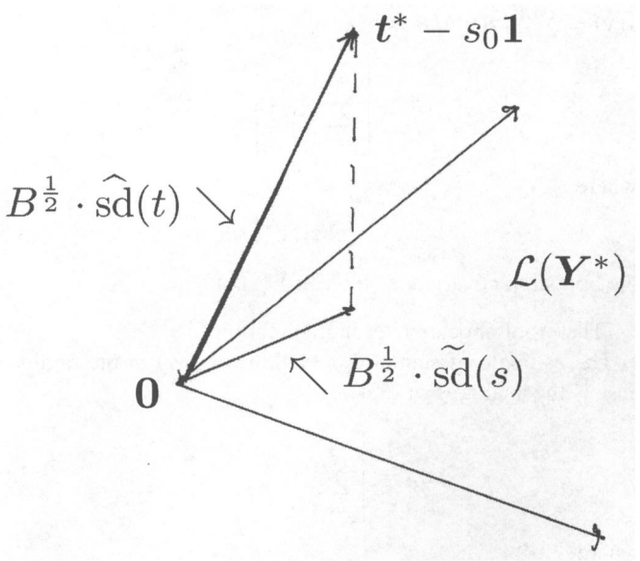

Figure 4. Illustration of Corollary 1. The ratio $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ is the cosine of the angle between $t^* - s_0\mathbf{1}$ (3.9) and the linear space $\mathcal{L}(\boldsymbol{Y}^*)$ spanned by the centered bootstrap counts (3.2). Model-selection estimators tend to be more nonlinear, yielding smaller ratios, that is, greater gains from smoothing.

How many bootstrap replications B are necessary to ensure the accuracy of $\widetilde{\mathrm{sd}}_B$? The jackknife provides a quick answer: divide the B replications into J groups of size B/J each, and let $\widetilde{\mathrm{sd}}_{Bj}$ be the estimate (3.6) computed with the jth group removed. Then

$$\widetilde{\mathrm{cv}}_B = \left[\frac{J}{J-1}\sum_{j=1}^{J}\left(\widetilde{\mathrm{sd}}_{Bj} - \widetilde{\mathrm{sd}}_{B\cdot}\right)^2\right]^{1/2} \Big/ \widetilde{\mathrm{sd}}_B, \quad (3.11)$$

$\widetilde{\mathrm{sd}}_{B\cdot} = \sum \widetilde{\mathrm{sd}}_{Bj}/J$, is the jackknife estimated coefficient of variation for $\widetilde{\mathrm{sd}}_B$. Applying (3.11) with $J = 20$ to the first $B = 1000$ replications (of the 4000 used in Figure 2) yielded $\widetilde{\mathrm{cv}}_B$ values of about 0.05 for each of the 164 subjects. Going on to $B = 4000$ reduced the $\widetilde{\mathrm{cv}}_B$'s to about 0.02. Stopping at $B = 1000$ would have been quite sufficient. *Note*: $\widetilde{\mathrm{cv}}_B$ applies to the *bootstrap* accuracy of $\widetilde{\mathrm{sd}}_B$ as an estimate of the ideal value $\widetilde{\mathrm{sd}}$ (3.4), not to sampling variability due to randomness in the original data $\boldsymbol{y}$, while $\widetilde{\mathrm{sd}}_B$ itself *does* refer to sampling variability.

Proof of Theorem 1. The "nonparametric delta method" is the same as the *influence function* and *infinitesimal jackknife* methods described in Chapter 6 of Efron (1982). It is appropriate here because $s(\boldsymbol{y})$, unlike $t(\boldsymbol{y})$, is a smooth function of $\boldsymbol{y}$. With the original data vector $\boldsymbol{y}$ (2.2) fixed, we can write bootstrap replication $t_i^* = t(\boldsymbol{y}_i^*)$ as a function $T(\boldsymbol{Y}_i^*)$ of the count vector (3.2). The ideal smoothed bootstrap estimate s_0 is the multinomial expectation of $T(\boldsymbol{Y}^*)$,

$$s_0 = E\{T(\boldsymbol{Y}^*)\}, \qquad \boldsymbol{Y}^* \sim \mathrm{Mult}_n(n, \boldsymbol{p}_0), \quad (3.12)$$

$\boldsymbol{p}_0 = (1/n, 1/n, \ldots, 1/n)$, the notation indicating a multinomial distribution with n draws on n equally likely categories.

Now let $S(\boldsymbol{p})$ denote the multinomial expectation of $T(\boldsymbol{Y}^*)$ if the probability vector is changed from $\boldsymbol{p}_0$ to $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$,

$$S(\boldsymbol{p}) = E\{T(\boldsymbol{Y}^*)\}, \qquad \boldsymbol{Y}^* \sim \mathrm{Mult}_n(n, \boldsymbol{p}), \quad (3.13)$$

so $S(\boldsymbol{p}_0) = s_0$. Define the directional derivative

$$\dot{S}_j = \lim_{\epsilon\to 0}\frac{S(\boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)) - S(\boldsymbol{p}_0)}{\epsilon}, \quad (3.14)$$

$\boldsymbol{\delta}_j$ the jth coordinate vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$, with 1 in the jth place. Formula (6.18) of Efron (1982) gives

$$\left(\sum_{j=1}^{n}\dot{S}_j^2\right)^{1/2} \Big/ n \quad (3.15)$$

as the delta method estimate of standard deviation for s_0. It remains to show that (3.15) equals (3.4).

Define $w_i(\boldsymbol{p})$ to be the ratio of the probabilities of $\boldsymbol{Y}_i^*$ under (3.13) compared to (3.12),

$$w_i(\boldsymbol{p}) = \prod_{k=1}^{n}(np_k)^{Y_{ik}^*}, \quad (3.16)$$

so that

$$S(\boldsymbol{p}) = \sum_{i=1}^{B} w_i(\boldsymbol{p})t_i^*/B \quad (3.17)$$

(the factor $1/B$ reflecting that under $\boldsymbol{p}_0$, all the $\boldsymbol{Y}_i^*$'s have probability $1/B = 1/n^n$).

For $\boldsymbol{p}(\epsilon) = \boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)$ as in (3.14), we calculate

$$w_i(\boldsymbol{p}) = (1 + (n-1)\epsilon)^{Y_{ij}^*}(1-\epsilon)^{\sum_{k\neq j} Y_{ik}^*}. \quad (3.18)$$

Letting $\epsilon \to 0$ yields

$$w_i(\boldsymbol{p}) \doteq 1 + n\epsilon(Y_{ij}^* - 1) \quad (3.19)$$

where we have used $\sum_k Y_{ik}^*/n = 1$. Substitution into (3.17) gives

$$\begin{aligned} S(\boldsymbol{p}(\epsilon)) &\doteq \sum_{i=1}^{B}\left[1 + n\epsilon(Y_{ij}^* - 1)\right]t_i^*/B \\ &= s_0 + n\epsilon\,\mathrm{cov}_j \end{aligned} \quad (3.20)$$

as in (3.5). Finally, definition (3.14) yields

$$\dot{S}_j = n\,\mathrm{cov}_j \quad (3.21)$$

and (3.15) verifies Theorem 1 (3.4). □

The validity of an approximate 95% interval $\hat{\theta} \pm 1.96\hat{\sigma}$ is compromised if the standard error σ is itself changing rapidly as a function of θ. *Acceleration* $\hat{a}$ (Efron 1987) is a measure of such change. Roughly speaking,

$$\hat{a} = \left.\frac{d\sigma}{d\theta}\right|_{\hat{\theta}}. \quad (3.22)$$

If $\hat{a} = 0.10$ for instance, then at the upper endpoint $\hat{\theta}_{\mathrm{up}} = \hat{\theta} + 1.96\,\hat{\sigma}$ the standard error will have increased to about $1.196\hat{\sigma}$, leaving $\hat{\theta}_{\mathrm{up}}$ only 1.64, not 1.96, σ-units above $\hat{\theta}$. (The 1987 article divides definition (3.22) by 3, as being appropriate after a normalizing transformation.)

Acceleration has a simple expression in terms of the covariances $\widehat{\mathrm{cov}}_j$ used to calculate $\widetilde{\mathrm{sd}}_B$ in (3.6),

$$\hat{a} = \frac{1}{6}\left[\sum_{j=1}^{n}\widehat{\mathrm{cov}}_j^3 \Big/ \left(\sum\widehat{\mathrm{cov}}_j^2\right)^{3/2}\right], \quad (3.23)$$

Equation (7.3) of Efron (1987). The $\hat{a}$'s were small for the 164 $\widetilde{\text{sd}}_B$ estimates for the Cholesterol data, most of them falling between -0.02 and 0.02, strengthening belief in the smoothed standard intervals $\tilde{\mu}_i \pm 1.96\,\widetilde{\text{sd}}_{Bi}$ (2.11).

Bias is more difficult to estimate than variance, particularly in a nonparametric context. Remark C of Section 7 verifies the following promising-looking result: the nonparametric estimate of bias for the smoothed estimate $\tilde{\mu} = s(\boldsymbol{y})$ (2.8) is

$$\widetilde{\text{bias}} = \frac{1}{2}\,\text{cov}_*(Q_i^*, t_i^*) \qquad \text{where } Q_i^* = \sum_{k=1}^{n}(Y_{nk}^* - 1)^2, \tag{3.24}$$

with cov_* indicating bootstrap covariance as in (3.5). Unfortunately, $\widetilde{\text{bias}}$ proved to be too noisy to use in the Cholesterol example. Section 6 describes a more practical approach to bias estimation in a parametric bootstrap context.

4. PARAMETRIC BOOTSTRAP SMOOTHING

We switch now from nonparametric to parametric estimation problems, but ones still involving data-based model selection. More specifically, we assume that a p-parameter exponential family of densities applies,

$$f_\alpha(\hat{\beta}) = e^{\alpha'\hat{\beta} - \psi(\alpha)} f_0(\hat{\beta}), \tag{4.1}$$

where α is the p-dimensional natural or canonical parameter vector, $\hat{\beta}$ the p-dimensional sufficient statistic vector (playing the role of $\boldsymbol{y}$ in (2.2)), $\psi(\alpha)$ the cumulant generating function, and $f_0(\hat{\beta})$ the "carrying density" defined with respect to some carrying measure (which may include discrete atoms as with the Poisson family). Form (4.1) covers a wide variety of familiar applications, including generalized linear models; $\hat{\beta}$ is usually obtained by sufficiency from the original data, as seen in the next section.

The *expectation parameter* vector $\beta = E_\alpha\{\hat{\beta}\}$ is a one-to-one function of α, say $\beta = \lambda(\alpha)$, having $p \times p$ derivative matrix

$$\frac{d\beta}{d\alpha} = V(\alpha), \tag{4.2}$$

where $V = V(\alpha)$ is the covariance matrix $\text{cov}_\alpha(\hat{\beta})$. The value of α corresponding to the sufficient statistic $\hat{\beta}$, $\hat{\alpha} = \lambda^{-1}(\hat{\beta})$, is the maximum likelihood estimate (MLE) of α.

A *parametric bootstrap sample* is obtained by drawing iid realizations $\hat{\beta}^*$ from the MLE density $f_{\hat{\alpha}}(\cdot)$,

$$f_{\hat{\alpha}}(\cdot) \xrightarrow{\text{iid}} \hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_B^*. \tag{4.3}$$

If $\hat{\mu} = t(\hat{\beta})$ is an estimate of a parameter of interest μ, the bootstrap samples (4.3) provide B parametric bootstrap replications of $\hat{\mu}$,

$$\hat{\mu}_i^* = t(\hat{\beta}_i^*), \qquad i = 1, 2, \ldots, B. \tag{4.4}$$

As in the nonparametric situation, these can be averaged to provide a *smoothed estimate*,

$$\tilde{\mu} = s(\hat{\beta}) = \sum_{i=1}^{B} t(\hat{\beta}_i^*)/B. \tag{4.5}$$

When $t(\cdot)$ involves model selection, $\hat{\mu}$ is liable to an erratic jumpiness, smoothed out by the averaging process.

The bootstrap replications $\hat{\beta}^* \sim f_{\hat{\alpha}}(\cdot)$ have mean vector and covariance matrix

$$\hat{\beta}^* \sim (\hat{\beta}, \hat{V}) \qquad [\hat{V} = V(\hat{\alpha})]. \tag{4.6}$$

Let $\boldsymbol{B}$ be the $B \times p$ matrix with ith row $\hat{\beta}_i^* - \hat{\beta}$. As before, we will assume an ideal bootstrap resampling situation where $B \to \infty$, making the empirical mean and variance of the $\hat{\beta}^*$ values exactly match (4.6):

$$\boldsymbol{B}'\mathbf{1}_B/B = \boldsymbol{O} \quad \text{and} \quad \boldsymbol{B}'\boldsymbol{B}/B = \hat{V}, \tag{4.7}$$

$\mathbf{1}_B$ the vector of B 1's.

Parametric versions of Theorem 1 and Corollary 1 depend on the p-dimensional bootstrap covariance vector between $\hat{\beta}^*$ and $t^* = t(\boldsymbol{y}^*)$,

$$\text{cov}_* = \boldsymbol{B}'(\boldsymbol{t}^* - s_0\mathbf{1}_B)/B, \tag{4.8}$$

where $\boldsymbol{t}^*$ is the B-vector of bootstrap replications $t_i^* = t(\boldsymbol{y}^*)$, and s_0 the ideal smoothed estimate (4.5).

Theorem 2. The parametric delta-method estimate of standard deviation for the ideal smoothed estimate (4.5) is

$$\widetilde{\text{sd}} = [\text{cov}_*' \hat{V}^{-1} \text{cov}_*]^{1/2}. \tag{4.9}$$

(Proof given at the end of this section.)

Corollary 2. $\widetilde{\text{sd}}$ is always less than or equal to $\widehat{\text{sd}}$, the bootstrap estimate of standard deviation for the unsmoothed estimate,

$$\widehat{\text{sd}} = [\|\boldsymbol{t}^* - s_0\mathbf{1}_B\|^2/B]^{1/2}, \tag{4.10}$$

the ratio being

$$\widetilde{\text{sd}}/\widehat{\text{sd}} = B^{1/2}[(\boldsymbol{t}^* - s_0\mathbf{1}_B)'\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'(\boldsymbol{t}^* - s_0\mathbf{1}_B)]^{1/2}/\widehat{\text{sd}}. \tag{4.11}$$

In the ideal bootstrap case, (4.7) and (4.9) show that $\widetilde{\text{sd}}$ equals $B^{-1/2}$ times the numerator on the right-hand side of (4.11). This is recognizable as the length of projection of $\boldsymbol{t}^* - s_0\mathbf{1}_B$ into the p-dimensional linear subspace of $\mathcal{R}^B$ spanned by the columns of $\boldsymbol{B}$. Figure 4 still applies, with $\mathcal{L}(\boldsymbol{B})$ replacing $\mathcal{L}(\boldsymbol{Y}^*)$.

If $t(\boldsymbol{y}) = \hat{\mu}$ is multivariate, say of dimension K, then cov_* as defined in (4.8) is a $p \times K$ matrix. In this case

$$\text{cov}_*' \hat{V}^{-1} \text{cov}_* \tag{4.12}$$

(or $\widehat{\text{cov}}'\bar{V}^{-1}\widehat{\text{cov}}$ in what follows) is the delta-method assessment of *covariance* for the smoothed vector estimate $s(\boldsymbol{y}) = \sum t(\boldsymbol{y}_i^*)/B$, also called $t_\cdot^*$ below.

Only minor changes are necessary for realistic bootstrap computations, that is, for $B < \infty$. Now we define $\boldsymbol{B}$ as the $B \times p$ matrix having ith row $\hat{\beta}_i^* - \hat{\beta}_\cdot^*$, with $\hat{\beta}_\cdot^* = \sum \hat{\beta}_i^*/B$, and compute the empirical covariance vector

$$\widehat{\text{cov}} = \boldsymbol{B}'(\boldsymbol{t}^* - t_\cdot^*\mathbf{1}_B)/B \tag{4.13}$$

and the empirical bootstrap variance matrix

$$\bar{V} = \boldsymbol{B}'\boldsymbol{B}/B. \tag{4.14}$$

Then the estimate of standard deviation for the smoothed estimate $\tilde{\mu} = s(\hat{\beta})$ (4.5) is

$$\widetilde{\text{sd}}_B = [\widehat{\text{cov}}'\bar{V}^{-1}\widehat{\text{cov}}]^{1/2}. \tag{4.15}$$

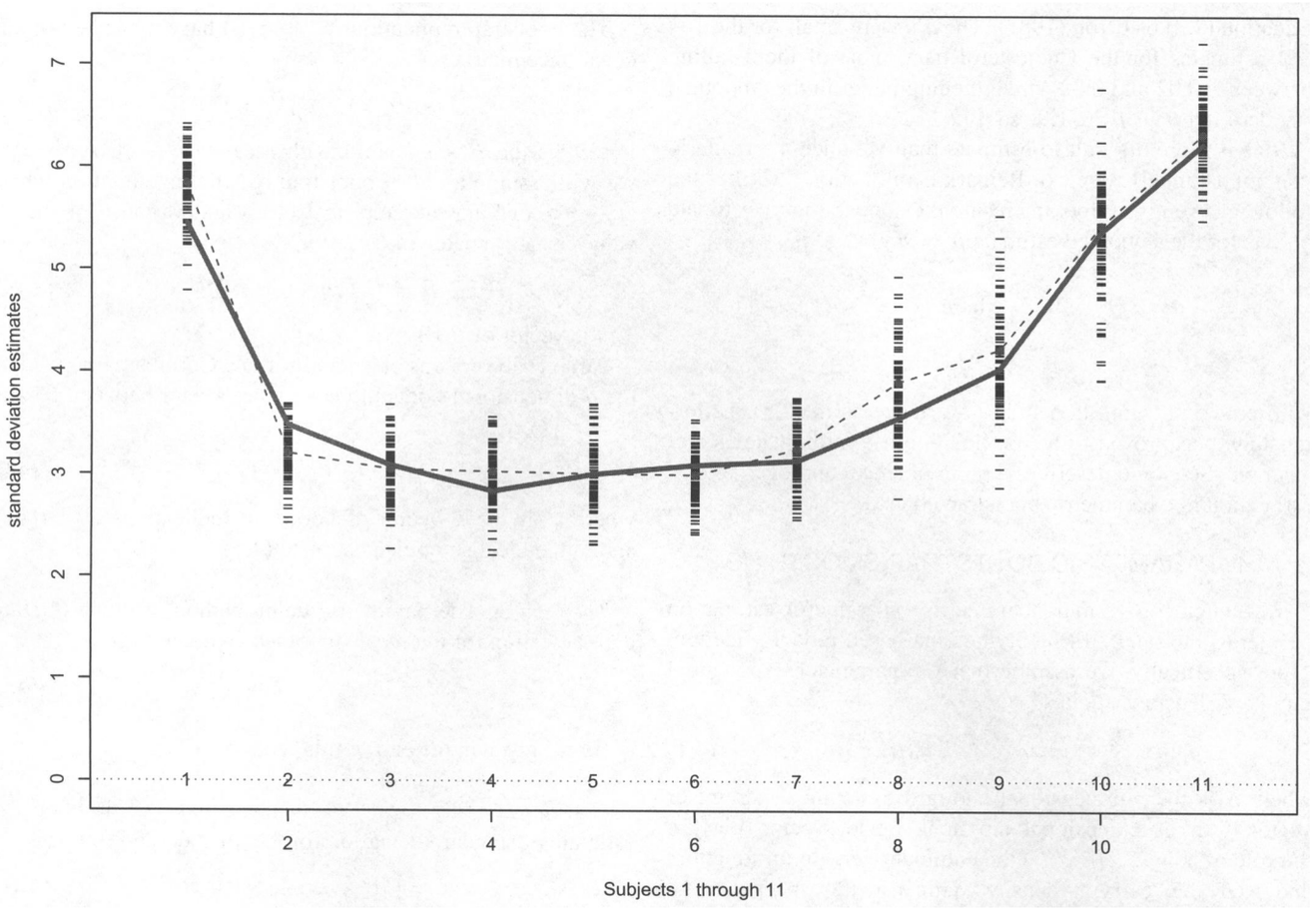

Figure 5. Simulation test of Theorem 2, parametric model (4.16)–(4.18), Cholesterol data; 100 simulations, 1000 parametric bootstraps each, for the 11 subjects indicated at the bottom of Figure 1. Heavy line connects observed empirical standard deviations (4.22); dashes show the 100 estimates $\widetilde{\mathrm{sd}}$ from Theorem 2 (4.15). Light dashed line connects averages of the $\widetilde{\mathrm{sd}}$ values, as discussed in Remark K.

As $B \to \infty$, $\widehat{\mathrm{cov}} \to \mathrm{cov}_*$, and $\bar{V} \to \hat{V}$, so $\widetilde{\mathrm{sd}}_B \to \widetilde{\mathrm{sd}}$ (4.9). Corollary 2, with s_0 replaced by $\tilde{\mu}$ (4.5), remains valid.

Figure 5 reports on a simulation test of Theorem 2. This was based on a parametric model for the Cholesterol data of Figure 1,

$$\boldsymbol{y} \sim \mathcal{N}_{164}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \tag{4.16}$$

where $\boldsymbol{\sigma}^2$ was diagonal, with diagonal elements a cubic function of compliance c (obtained from a regression precentile fit),

$$\sigma_i = 23.7 + 5.49c - 2.25c^2 - 1.03c^3, \tag{4.17}$$

making σ_i about twice as large to the right as to the left. The expectation vector $\boldsymbol{\mu}$ was taken to be

$$\boldsymbol{\mu} = X\hat{\beta}(6) = \hat{\boldsymbol{\mu}}(6), \tag{4.18}$$

the sixth degree OLS fit for cholesterol decrease as a function of compliance in (4.16), with X the corresponding 164×7 structure matrix.

Model (4.16)–(4.18) is a 7-parameter exponential family (4.1), with sufficient statistic

$$\hat{\beta} = G^{-1}X'(\boldsymbol{\sigma}^2)^{-1}\boldsymbol{y} \qquad [G = X'(\boldsymbol{\sigma}^2)^{-1}X] \tag{4.19}$$

and covariance matrix (4.2)

$$V = G^{-1}, \tag{4.20}$$

which is all that is necessary to apply Theorem 2.

The simulation began with 100 draws y_i^*, $i = 1, 2, \ldots, 100$, from (4.16), each of which gave OLS estimate $\hat{\boldsymbol{\mu}}_i^* = X\hat{\beta}_i^*(6)$. Then $B = 1000$ parametric bootstrap draws were generated from $\hat{\beta}_i^*$,

$$\boldsymbol{y}_{ij}^{**} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_i^*, \boldsymbol{\sigma}^2\right), \quad j = 1, 2, \ldots, 1000, \tag{4.21}$$

from which smoothed estimate $\tilde{\mu}_i$ (4.5) and estimated standard deviation $\widetilde{\mathrm{sd}}_i$ were calculated according to (4.15). All of this was done for 11 of the 164 subjects, as indicated in Figure 1.

The dashes in Figure 5 indicate the 100 $\widetilde{\mathrm{sd}}_i$ values for each of the 11 subjects. This is compared with the observed empirical standard deviations of the smoothed estimates,

$$\widetilde{\mathrm{Sd}} = \left[\sum_1^{100} (\tilde{\mu}_i - \tilde{\mu}_\cdot)^2/99\right]^{1/2} \quad \left[\tilde{\mu}_\cdot = \sum_1^{100} \tilde{\mu}_i/100\right], \tag{4.22}$$

connected by the heavy solid curve. The $\widetilde{\mathrm{sd}}$ values from Theorem 2 are seen to provide reasonable estimates of $\widetilde{\mathrm{Sd}}$, though with some bias and variability.

There is more to the story. The empirical standard deviations $\widetilde{\mathrm{Sd}}$ are themselves affected by model-selection problems. Averaging the 100 $\widetilde{\mathrm{sd}}_i$ values (connected by the dashed line

in Figure 5) gives more dependable results, as discussed in Remark K.

Proof of Theorem 2. Suppose that instead of $f_{\hat\alpha}(\cdot)$ in (4.3) we wished to consider parametric bootstrap samples drawn from some other member of family (4.1), $f_\alpha(\cdot)$ (α not necessarily the "true value"). The ratio $w_i = f_\alpha(\hat\beta_i^*)/f_{\hat\alpha}(\hat\beta_i^*)$ equals

$$w_i = c_{\alpha,\hat\alpha} e^{Q_i} \qquad \text{where } Q_i = (\alpha - \hat\alpha)'\left(\hat\beta_i^* - \hat\beta\right), \quad (4.23)$$

with the factor $c_{\alpha,\hat\alpha}$ not depending on $\hat\beta_i^*$. Importance sampling can now be employed to estimate $E_\alpha\{t(\hat\beta)\}$, the expectation under f_α of statistic $t(\hat\beta)$, using only the original bootstrap replications $(\hat\beta_i^*, t_i^*)$ from (4.3),

$$\hat E_\alpha = \sum_{i=1}^B w_i t_i^* \Big/ \sum_{i=1}^B w_i = \sum_{i=1}^B e^{Q_i} t_i^* \Big/ \sum_{i=1}^B e^{Q_*}. \quad (4.24)$$

Notice that $\hat E_\alpha$ is the value of the smoothed estimate (4.5) at parameter α, say s_α. The delta-method standard deviation for our estimate $s_{\hat\alpha}$ depends on the derivative vector $ds_\alpha/d\alpha$ evaluated at $\alpha = \hat\alpha$. Letting $\alpha \to \hat\alpha$ in (4.23)–(4.24) gives,

$$s_\alpha \doteq \frac{\sum(1+Q_i)t_i^*/B}{\sum(1+Q_i)/B} = s_{\hat\alpha} + (\alpha - \hat\alpha)'\,\mathrm{cov}_*, \quad (4.25)$$

where the denominator term $\sum Q_i/B$ equals 0 for the ideal bootstrap according to (4.7). (For the nonideal bootstrap, $\sum Q_i/B$ approaches 0 at rate $O_p(1/\sqrt{B})$.)

We see that

$$\left.\frac{ds_\alpha}{d\alpha}\right|_{\hat\alpha} = \mathrm{cov}_*, \quad (4.26)$$

so from (4.2),

$$\left.\frac{ds_\alpha}{d\beta}\right|_{\hat\alpha} = \hat V^{-1}\,\mathrm{cov}_*\,. \quad (4.27)$$

Since $\hat V$ is the covariance matrix of $\hat\beta^*$, that is, of $\hat\beta$ under distribution $f_{\alpha=\hat\alpha}$, (4.6) and (4.27) verify $\widetilde{\mathrm{sd}}$ in (4.9) as the usual delta-method estimate of standard deviation for $s(\hat\beta)$. □

Theorem 1 and Corollary 1 can be thought of as special cases of the exponential family theory in this section. The multinomial distribution of $\boldsymbol{Y}^*$ (3.12) plays the role of $f_{\hat\alpha}(\hat\beta^*)$; $\hat V$ in (4.9) becomes $I - \mathbf{1}_n\mathbf{1}_n'/n$ (3.3), so that (4.9) becomes (3.4). A technical difference is that the $\mathrm{Mult}_n(n, \boldsymbol{p})$ family (3.13) is singular (that is, concentrated on a $n-1$-dimensional subspace of $\mathcal{R}^n$), making the influence-function argument a little more involved than the parametric delta-function calculations. More seriously, the dimension of the nonparametric multinomial distribution increases with n, while for example, the parametric "Supernova" example of the next section has dimension 10 no matter how many supernovas might be observed. The more elaborate parametric confidence interval calculations of Section 6 failed when adapted for the nonparametric Cholesterol analysis, perhaps because of the comparatively high dimension, 164 versus 10.

5. THE SUPERNOVA DATA

Figure 6 concerns a second example we will use to illustrate the parametric bootstrap theory of the previous section, the *Supernova data*: the absolute magnitude y_i has been determined for $n = 39$ Type Ia supernovas, yielding the data

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'. \quad (5.1)$$

Each supernova has also had observed a vector of spectral energies $\boldsymbol{x}_i$ measured at $p = 10$ frequencies,

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i10}) \quad (5.2)$$

for supernova i. The 39×10 covariate matrix X, having $\boldsymbol{x}_i$ as its ith row, will be regarded as fixed.

We assume a standard normal linear regression model

$$\boldsymbol{y} = X\alpha + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_{39}(\boldsymbol{0}, I), \quad (5.3)$$

referred to as the *full model* in what follows. [For convenient discussion, the y_i have been rescaled to make (5.3) appropriate.] It has exponential family form (4.1), $p = 10$, with natural parameter α, $\hat\beta = X'\boldsymbol{y}$, and $\psi = \alpha' X' X\alpha/2$.

Then $(X'X)^{-1}\hat\beta = \hat\alpha$, the MLE of α, which also equals $\hat\alpha_{\mathrm{OLS}}$, the ordinary least squares estimate of α in (5.3), yielding the full-model vector of supernova brightness estimates

$$\hat{\boldsymbol{\mu}}_{\mathrm{OLS}} = X\hat\alpha_{\mathrm{OLS}}. \quad (5.4)$$

Figure 6 plots y_i versus its estimate $\hat\mu_{\mathrm{OLS},i}$. The fit looks good, having an unadjusted R^2 of 0.82. Adjusting for the fact that we have used $m = 10$ parameters to fit $n = 39$ data points yields the more realistic value

$$R^2_{\mathrm{adj}} = R^2 - 2\cdot(1-R^2)\frac{m}{n-m} = 0.69; \quad (5.5)$$

see Remark D.

Type Ia supernovas were used as "standard candles" in the discovery of dark energy and the cosmological expansion of the universe (Riess et al. 1998; Perlmutter et al. 1999). Their standardness assumes a constant absolute magnitude. This is not exactly true, and in practice regression adjustments are made. Our 39 supernovas were close enough to Earth to have their absolute magnitudes ascertained independently. The spectral measurements $\boldsymbol{x}$, however, can be made for *distant* Type Ia supernovas, where independent methods fail, the scientific goal being a more accurate estimation function $\hat\mu(\boldsymbol{x})$ for their absolute magnitudes, and improved calibration of cosmic expansion.

We will use the Lasso (Tibshirani 1996) to select $\hat\mu(\boldsymbol{x})$. For a given choice of the nonnegative "tuning parameter" λ, we estimate α by the Lasso criterion

$$\hat\alpha_\lambda = \arg\min_\alpha \left\{\|\boldsymbol{y} - X\alpha\|^2 + \lambda\sum_{k=1}^p |\alpha_k|\right\}; \quad (5.6)$$

$\hat\alpha_\lambda$ shrinks the components of $\hat\alpha_{\mathrm{OLS}}$ toward zero, some of them all the way. As λ decreases from infinity to 0, the number m of nonzero components of $\hat\alpha_\lambda$ increases from 0 to p. Conveniently enough, it turns out that m also nearly equals the effective degrees of freedom for the selection of $\hat\alpha_\lambda$ (Efron et al. 2004). In what follows we will write $\hat\alpha_m$ rather than $\hat\alpha_\lambda$.

Table 4 shows a portion of the Lasso calculations for the Supernova data. Its last column gives R^2_{adj} (5.5) with R^2 having the usual form

$$R^2 = 1 - \frac{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_m\|^2}{\|\boldsymbol{y} - \bar y\mathbf{1}\|^2} \quad \left(\hat{\boldsymbol{\mu}}_m = X\hat\alpha_m,\ \bar y = \sum y_i/n\right). \quad (5.7)$$

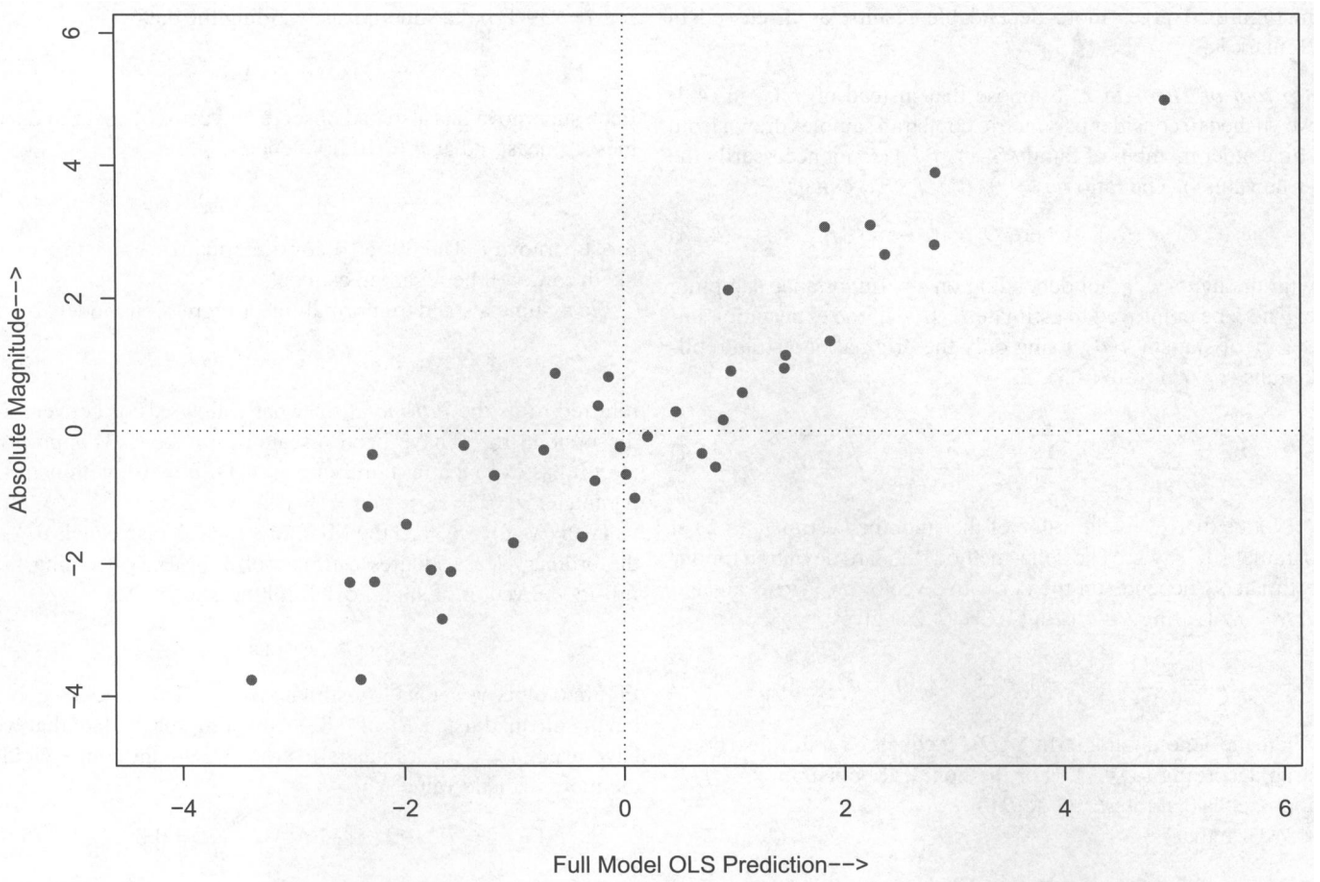

Figure 6. *The Supernova data* Absolute magnitudes of $n = 39$ Type Ia supernovas plotted versus their OLS estimates from the full linear model (5.3); adjusted R^2 (5.5) equals 0.69.

The choice $\hat{m} = 7$ maximizes R^2_{adj},

$$\hat{m} = \arg\max_m \{R^2_{\text{adj}}\}, \tag{5.8}$$

yielding our selected coefficient vector $\hat{\alpha}_{\hat{m}}$ and the corresponding vector of supernova estimates

$$\hat{\boldsymbol{\mu}} = X\hat{\alpha}_{\hat{m}}; \tag{5.9}$$

note that $\hat{\alpha}_{\hat{m}}$ is *not* an OLS estimate.

$B = 4000$ bootstrap replications $\hat{\boldsymbol{\mu}}^*$ were computed (again many more than were actually needed): bootstrap samples $\boldsymbol{y}^*$ were drawn using the full OLS model,

$$\boldsymbol{y}^* \sim \mathcal{N}_{39}(\hat{\boldsymbol{\mu}}_{\text{OLS}}, \boldsymbol{I}); \tag{5.10}$$

see Remark E. The equivalent of Table 4, now based on data $\boldsymbol{y}^*$, was calculated, the R^2_{adj} maximizer $\hat{m}^*$ and $\hat{\alpha}^*_{\hat{m}^*}$ selected, giving

$$\hat{\boldsymbol{\mu}}^* = X\hat{\alpha}^*_{\hat{m}^*}. \tag{5.11}$$

Averaging the 4000 $\hat{\boldsymbol{\mu}}^*$ vectors yielded the smoothed vector estimates

$$\tilde{\boldsymbol{\mu}} = \sum_{i=1}^{B} \hat{\boldsymbol{\mu}}^*_i / B. \tag{5.12}$$

Standard deviations $\widetilde{\text{sd}}_{Bj}$ for supernova j's smoothed estimate $\tilde{\mu}_j$ were then calculated according to (4.15), $j = 1, 2, \ldots, 39$. The ratio of standard deviations $\widetilde{\text{sd}}_B/\widehat{\text{sd}}_B$ for the 39 supernovas ranged from 0.87 to 0.98, with an average of 0.93. Jackknife calculations (3.11) showed that $B = 800$ would have been enough for good accuracy.

At this point it pays to remember that $\widetilde{\text{sd}}_B$ is a delta-method shortcut version of a full bootstrap standard deviation for the smoothed estimator $s(\boldsymbol{y})$. We would prefer the latter if not for the computational burden of a second level of bootstrapping. As a check, a full second-level simulation was run, beginning with simulated data vectors $\boldsymbol{y}^* \sim \mathcal{N}_{39}(\hat{\boldsymbol{\mu}}_{\text{OLS}}, I)$ (5.10), and for each $\boldsymbol{y}^*$ carrying through calculations of s^* and $\widetilde{\text{sd}}^*_B$ based on $B = 1000$ second-level bootstraps. This was done 500 times,

Table 4. Lasso model selection for the Supernova data. As the regularization parameter λ in (5.6) decreases from infinity to zero, the number m of nonzero coordinates of $\hat{\alpha}_m$ increases from 0 to 10. The choice $m = 7$ maximizes the adjusted R^2 value (5.7), making it the selected model

λ	m	R^2	R^2_{adj}	
∞	0	0	0	
63	1	0.17	0.12	
19.3	3	0.74	0.70	
8.2	5	0.79	0.73	
0.496	7	0.82	**0.735**	(Selected)
0.039	9	0.82	0.71	
0	10	0.82	0.69	(OLS)

Table 5. Percentage of the 4000 bootstrap replications selecting m nonzero coefficients for $\hat{\alpha}^*$ in (5.11), $m = 1, 2, \ldots, 10$. The original choice $m = 7$ is not quite modal

m	1	2	3	4	5	6	**7**	8	9	10
%	0	1	8	13	16	18	**18**	14	9	2

yielding 500 values s_k^* for each of the 39 supernovas, which provided direct bootstrap estimates say $\widetilde{\mathrm{Sd}}_k$ for s_k. The $\widetilde{\mathrm{Sd}}_k$ values averaged about 7.5% larger than the delta-method approximations $\widetilde{\mathrm{sd}}_{Bk}$. Taking this into account, the reductions in standard deviation due to smoothing were actually quite small, the ratios averaging about 98%; see the end of Remark H.

Returning to the original calculations, model selection was highly variable among the 4000 bootstrap replications. Table 5 shows the percentage of the 4000 replications that selected m nonzero coefficients for $\hat{\alpha}^*$ in (5.11), $m = 1, 2, \ldots, 10$, with the original choice $m = 7$ not quite being modal. Several of the supernovas showed effects like that in Figure 3.

Model averaging, that is bootstrap smoothing, still has important confidence interval effects even though here it does not substantially reduce standard deviations. This is shown in Figure 7 of the next section, which displays approximate 95% confidence intervals for the 39 supernova magnitudes.

Other approaches to bootstrapping Lasso estimates are possible. Chatterjee and Lahiri (2011), referring back to work by Knight and Fu (2000), resample regression residuals rather than using the full parametric bootstrap (5.10). The "m out of n" bootstrap is featured in Hall, Lee, and Park (2009). Asymptotic performance, mostly absent here, is a central concern of these papers; also, they focus on estimation of the regression coefficients, α in (5.3), a more difficult task than estimating $\boldsymbol{\mu} = X\alpha$.

6. BETTER BOOTSTRAP CONFIDENCE INTERVALS

The central tactic of this article is the use of bootstrap smoothing to convert an erratically behaved model selection-based estimator $t(\cdot)$ into a smoothly varying version $s(\cdot)$. Smoothing makes the good asymptotic properties of the bootstrap, as extensively developed in Hall (1992), more credible for actual applications. This section carries the smoothing theme further, showing how $s(\cdot)$ can be used to form *second-order accurate* intervals.

The improved confidence intervals depend on the properties of bootstrap samples from exponential families (4.1). We define an "empirical exponential family" $\hat{f}_\alpha(\cdot)$ that puts probability

$$\hat{f}_\alpha(\hat{\beta}_i^*) = e^{(\alpha-\hat{\alpha})'\hat{\beta}_i^* - \hat{\psi}(\alpha)}\frac{1}{B} \tag{6.1}$$

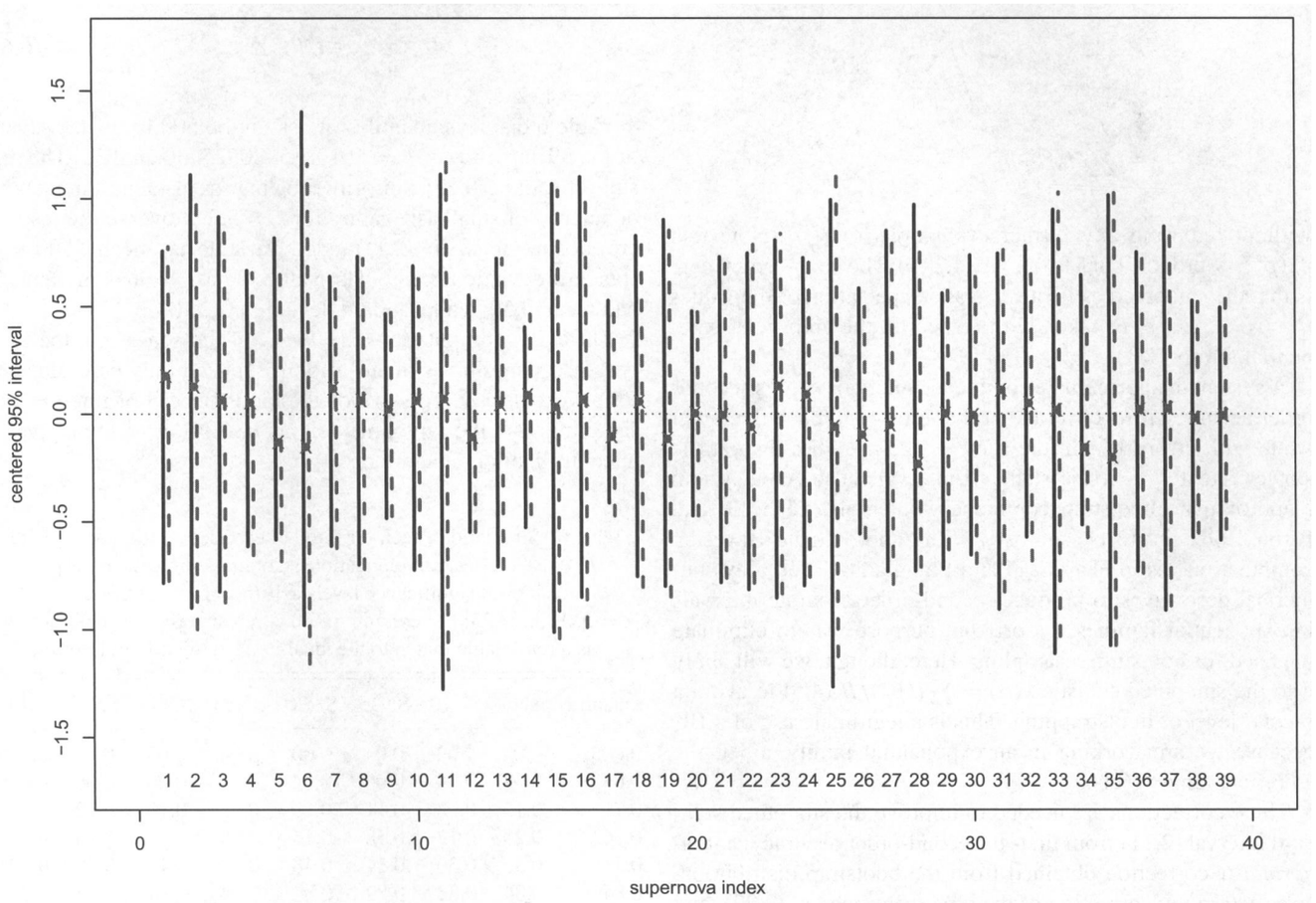

Figure 7. Approximate 95% confidence limits for the 39 supernova magnitudes μ_k (after subtraction of smoothed estimates $\tilde{\mu}_k$ (5.12)); ABC intervals (solid) compared with smoothed standard intervals $\tilde{\mu}_k \pm 1.96\widetilde{\mathrm{sd}}_k$ (dashed). Crosses indicate differences between unsmoothed and smoothed estimates, (5.9) minus (5.12).

on bootstrap replication $\hat{\beta}_i^*$ (4.3) for $i = 1, 2, \ldots, B$, where

$$\hat{\psi}(\alpha) = \log\left(\sum_{i=1}^{B} e^{(\alpha-\hat{\alpha})'\hat{\beta}_i^*}/B\right). \tag{6.2}$$

Here $\hat{\alpha}$ is fixed as the MLE of α in the original family (4.1), $\hat{\alpha} = \lambda^{-1}(\hat{\beta})$ in the notation following (4.2).

The choice of $\alpha = \hat{\alpha}$ makes $\hat{f}_{\hat{\alpha}}(\hat{\beta}_i^*) = 1/B$ for $i = 1, 2, \ldots, B$; in other words, it yields the empirical probability distribution of the bootstrap sample (4.3) in $\mathcal{R}^p$. Other choices of α "tilt" the empirical distribution in direction $\alpha - \hat{\alpha}$; (6.1) is a direct analogue of the original exponential family (4.1), which can be re-expressed as

$$f_\alpha(\hat{\beta}^*) = e^{(\alpha-\hat{\alpha})'\hat{\beta}^* - (\psi(\alpha)-\psi(\hat{\alpha}))} f_{\hat{\alpha}}(\hat{\beta}^*), \tag{6.3}$$

now with $\hat{\alpha}$ fixed and $\hat{\beta}^*$ the random variable. Notice that $\hat{\psi}(\hat{\alpha}) = 0$ in (6.2). Taking this into account, the only difference between the original family (6.3) and the empirical family (6.1) is the change in support, from $f_{\hat{\alpha}}(\cdot)$ to the empirical probability distribution. Under mild regularity conditions, family $\hat{f}_\alpha(\cdot)$ approaches $f_\alpha(\cdot)$ as the bootstrap sample size B goes to infinity.

As in (4.23)–(4.24), let s_α be the value of the smoothed statistic we would get if bootstrap samples were obtained from f_α rather than $f_{\hat{\alpha}}$. We can estimate s_α from the original bootstrap samples (4.3) by importance sampling in family (4.1),

$$\begin{aligned} s_\alpha &= \sum_{i=1}^{B} e^{(\alpha-\hat{\alpha})'\hat{\beta}_i^*} t_i^* \Big/ \sum_{i=1}^{B} e^{(\alpha-\hat{\alpha})'\hat{\beta}_i^*} \\ &= \sum_{i=1}^{B} \hat{f}_\alpha(\hat{\beta}_i^*) t_i^* \end{aligned} \tag{6.4}$$

without requiring any further evaluations of $t(\cdot)$. (Note that $\hat{f}_\alpha(\hat{\beta}_i^*)$ is proportional to w_i in (4.24).) The main point here is that the smoothed estimate s_α is the expectation of the values t_i^*, $i = 1, 2, \ldots, B$, taken with respect to the empirical exponential family (6.1).

A system of approximate confidence intervals enjoys second-order accuracy if its coverage probabilities approach the target value with errors $1/n$ in the sample size n, rather than at the slower rate $1/\sqrt{n}$ of the standard intervals. The ABC system ("approximate bootstrap confidence" intervals, DiCiccio and Efron 1992, not to be confused with "approximate Bayesian computation" as in Fearnhead and Prangle 2012) employs numerical derivatives to produce second-order accurate intervals in exponential families. Its original purpose was to eliminate the need for bootstrap resampling. Here, though, we will apply it to the smoothed statistic $s(\hat{\beta}) = \sum t(\hat{\beta}_i^*)/B$ (4.5) to avoid a *second* level of bootstrapping. This is a legitimate use of ABC because we are working in an exponential family, albeit the empirical family (6.1).

Three corrections are needed to improve the smoothed standard interval (2.11) from first- to second-order accuracy: a *nonnormality* correction obtained from the bootstrap distribution, an *acceleration* correction of the type mentioned at (3.22), and a *bias-correction*. ABC carries these out via $p + 2$ numerical second derivatives of $\hat{s}_\alpha$ in (6.4), taken at $\alpha = \hat{\alpha}$, as detailed in Section 2 of DiCiccio and Efron (1992). The computational burden is effectively nil compared with the original bootstrap calculations (4.3).

Figure 7 compares the ABC 95% limits for the supernova brightnesses μ_k, $k = 1, 2, \ldots, 39$, solid lines, with parametric smoothed standard intervals (2.11), dashed lines. [The smoothed estimates $\tilde{\mu}_k$ (5.12) have been subtracted from the endpoints to put all the intervals on the same display.] There are a few noticeable discrepancies, for supernovas 2, 6, 25, and 27 in particular, but overall the smoothed standard intervals hold up reasonably well.

Smoothing has a moderate effect on the Supernova estimates, as indicated by the values of $\hat{\mu}_k - \tilde{\mu}_k$, (5.11) minus (5.12), the crosses in Figure 7. A few of the intervals would be much different if based on the unsmoothed estimates $\hat{\mu}_k$, for example, supernovas 1, 12, 17, and 28. Remark I says more about the ABC calculations.

As a check on the ABC intervals, the "full simulation" near the end of Section 4, with $B = 1000$ bootstrap replications for each of 500 trials, was repeated. For each trial, the 1000 bootstraps provided new ABC calculations, from which the "achieved significance level" asl_k^* of the original smoothed estimate $\tilde{\mu}_k$ (5.12) was computed: that is,

$$\text{asl}_k^* = \text{ bootstrap ABC confidence level for } (-\infty, \tilde{\mu}_k). \tag{6.5}$$

If the ABC construction were working perfectly, asl_k^* would have a uniform distribution,

$$\text{asl}_k^* \sim U(0, 1) \tag{6.6}$$

for $k = 1, 2, \ldots, 39$.

Table 6 displays quantiles of asl_k^* in the 500 trials, for seven of the 39 supernovas, $k = 5$, 10, 15, 20, 25, 30, and 35. The results are not perfectly uniform, showing for instance a moderate deficiency of small asl_k^* values for $k = 5$, but overall the results are encouraging. A $U(0, 1)$ random variable has mean 0.500 and standard deviation 0.289, while all 3500 asl_k^* values in Table 6 had mean 0.504 and standard deviation 0.284.

The ABC computations are *local*, in the sense that the importance sampling estimates s_α in (6.4) need only be evaluated for α very near $\hat{\alpha}$. This avoids the familiar peril of importance sampling, that the sampling weights in (6.4) or (4.1) may vary uncontrollably in magnitude.

Table 6. Simulation check for ABC intervals; 500 trials, each with $B = 1000$ bootstrap replications. Columns show quantiles of achieved significance levels asl_k^* (6.5) for supernovas $k = 5, 10, \ldots, 35$; last column for all seven supernovas combined. It is a reasonable match to the ideal uniform distribution (6.6)

Quantile	SN5	SN10	SN15	SN20	SN25	SN30	SN35	ALL
0.025	0.04	0.02	0.04	0.00	0.04	0.03	0.02	0.025
0.05	0.08	0.04	0.06	0.04	0.08	0.06	0.06	0.055
0.1	0.13	0.08	0.11	0.10	0.12	0.10	0.12	0.105
0.16	0.20	0.17	0.18	0.16	0.18	0.18	0.18	0.175
0.5	0.55	0.50	0.54	0.48	0.50	0.48	0.50	0.505
0.84	0.84	0.82	0.82	0.84	0.84	0.84	0.84	0.835
0.9	0.90	0.88	0.90	0.88	0.90	0.90	0.90	0.895
0.95	0.96	0.94	0.96	0.94	0.94	0.94	0.94	0.945
0.975	0.98	0.97	0.98	0.98	0.96	0.98	0.97	0.975

If one is willing to ignore the peril, full bootstrap standard errors for the smoothed estimates $\tilde{\mu}$ (4.5), rather than the delta-method estimates of Theorem 2, become feasible: in addition to the original parametric bootstrap samples (4.3), we draw J more times, say

$$f_{\hat{\alpha}}(\cdot) \longrightarrow \tilde{\beta}_1^*, \tilde{\beta}_2^*, \ldots, \tilde{\beta}_J^*, \tag{6.7}$$

and compute the corresponding natural parameter estimates $\tilde{\alpha}_j^* = \lambda^{-1}(\tilde{\beta}_j^*)$, as following (4.2). Each $\tilde{\alpha}_j^*$ gives a bootstrap version of the smoothed statistic $s_{\tilde{\alpha}_j^*}$, using (6.4), from which we calculate the usual bootstrap standard error estimate,

$$\widetilde{\mathrm{sd}}_{\mathrm{boot}} = \left[\sum_{j=1}^{J}(s_{\tilde{\alpha}_j^*} - s_\cdot)^2/(J-1)\right]^{1/2}, \tag{6.8}$$

where $s_\cdot = \sum s_{\tilde{\alpha}_j^*}/J$. Once again, no further evaluations of $t(\cdot)$ beyond the original ones in (4.5) are required.

Carrying this out for the Supernova data gave standard errors $\widetilde{\mathrm{sd}}_{\mathrm{boot}}$ a little smaller than those from Theorem 2, as opposed to the somewhat larger ones found by the full simulation near the end of Section 5. Occasional very large importance sampling weights in (6.4) did seem to be a problem here.

Compromises between the delta method and full bootstrapping are possible. For the normal model (5.3) we have $\tilde{\beta}_j^* \sim \mathcal{N}(\hat{\beta}, X'X)$ in (6.7). Instead we might take

$$\hat{\beta}_j^* \sim \mathcal{N}\left(\hat{\beta}, cX'X\right) \tag{6.9}$$

with c less than 1, placing $\tilde{\alpha}_j^*$ nearer $\hat{\alpha}$. Then (6.8) must be multiplied by $1/\sqrt{c}$. Doing this with $c = 1/9$ gave standard error estimates almost the same as those from Theorem 2.

7. REMARKS, DETAILS, AND PROOFS

This section expands on points raised in the previous discussion.

A. Proof of Corollary 1 With $\boldsymbol{Y}^* = (Y_{ij}^*)$ as in (3.2), let $\boldsymbol{X} = \boldsymbol{Y}^* - \mathbf{1}_B\mathbf{1}_n' = (Y_{ij}^* - 1)$. For the ideal bootstrap, $B = n^n$,

$$\boldsymbol{X}'\boldsymbol{X}/B = \boldsymbol{I} - \mathbf{1}_n'\mathbf{1}_n, \tag{7.1}$$

the multinomial covariance matrix in (3.3). This has $(n-1)$ nonzero eigenvalues all equaling 1, implying that the singular value decomposition of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \sqrt{B}\boldsymbol{L}\boldsymbol{R}', \tag{7.2}$$

$\boldsymbol{L}$ and $\boldsymbol{R}$ orthonormal matrices of dimensions $B \times (n-1)$ and $n \times (n-1)$. Then the B-vector $\boldsymbol{U}^* = (t_i^* - s_0)$ has projected squared length into $\mathcal{L}(\boldsymbol{X})$

$$\begin{aligned}\boldsymbol{U}^{*\prime}\boldsymbol{L}\boldsymbol{L}'\boldsymbol{U}^* &= B\boldsymbol{U}^{*\prime}\frac{\boldsymbol{L}\sqrt{B}\boldsymbol{R}'\boldsymbol{R}\sqrt{B}\boldsymbol{L}'}{B^2}\boldsymbol{U}^* \\ &= B(\boldsymbol{U}^{*\prime}\boldsymbol{X}/B)(\boldsymbol{X}'\boldsymbol{U}^*/B) = B\widetilde{\mathrm{sd}}^2,\end{aligned} \tag{7.3}$$

verifying (3.10).

B. Hájek projection and ANOVA decomposition For the ideal nonparametric bootstrap of Section 3, define the conditional bootstrap expectations

$$e_j = E_*\{t(\boldsymbol{y}_i^*)|y_{ik}^* = y_j\}, \tag{7.4}$$

$j = 1, 2, \ldots, n$ (not depending on k). The bootstrap ANOVA decomposition of Efron (1983, sec. 7) can be used to derive an orthogonal decomposition of $t(\boldsymbol{y}^*)$,

$$t(\boldsymbol{y}_i^*) = s_0 + L_i^* + R_i^*, \tag{7.5}$$

where $s_0 = E_*\{t(\boldsymbol{y}^*)\}$ is the ideal smoothed bootstrap estimate, and

$$L_i^* = \sum_{j=1}^{n} Y_{ij}^*(e_j - s_0), \tag{7.6}$$

while R_i^* involves higher-order ANOVA terms such as $e_{jl} - e_j - e_l + s_0$ with

$$e_{jl} = E_*\{t(\boldsymbol{y}_i^*)|y_{ik}^* = y_j \text{ and } y_{im}^* = y_k\}. \tag{7.7}$$

The terms in (7.5) satisfy $E_*\{L^*\} = E_*\{R^*\} = 0$ and are orthogonal, $E_*\{L^*R^*\} = 0$. The bootstrap *Hájek projection* of $t(\boldsymbol{y}^*)$ (Hájek 1968) is then the first two terms of (7.5), say

$$H_i^* = s_0 + L_i^*. \tag{7.8}$$

Moreover, it can be shown that

$$L_i^* = \sum_{j=1}^{n} Y_{ij}^* \operatorname{cov}_j \tag{7.9}$$

from (3.5) and the ratio of smoothed-to-unsmoothed standard deviation (3.10) equals

$$\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B = [\mathrm{var}_*\{L_i^*\}/(\mathrm{var}_*\{L_i^*\} + \mathrm{var}_*\{R_i^*\})]^{1/2}. \tag{7.10}$$

C. Nonparametric bias estimate There is a nonparametric bias estimate $\widetilde{\mathrm{bias}}_B$ for the smoothed statistic $s(\boldsymbol{y})$ (2.8) corresponding to the variability estimate $\widetilde{\mathrm{sd}}_B$. In terms of $T(\boldsymbol{Y}^*)$ and $S(\boldsymbol{p})$ (3.13)–(3.14), the nonparametric delta method gives

$$\widetilde{\mathrm{bias}}_B = \frac{1}{2}\sum_{j=1}^{n}\frac{\ddot{S}_j}{n^2}, \tag{7.11}$$

where $\ddot{S}_j$ is the second-order influence value

$$\begin{aligned}\ddot{S}_j = &\lim_{\epsilon\to 0} \\ &\times \frac{S(\boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)) - 2S(\boldsymbol{p}_0) + S(\boldsymbol{p}_0 - \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0))}{\epsilon^2}.\end{aligned} \tag{7.12}$$

See section 6.6 of Efron (1982).

Without going into details, the Taylor series calculation (3.18)–(3.19) can be carried out one step further, leading to the following result:

$$\widetilde{\mathrm{bias}}_B = \mathrm{cov}_*(D_i^*, t_i^*), \tag{7.13}$$

where $D_i^* = \sum_1^n (Y_{ij}^* - 1)^2$.

This looks like a promising extension of Theorem 1 (3.4)–(3.5). Unfortunately, (7.13) proved unstable when applied to the Cholesterol data, as revealed by jackknife calculations like (3.11). Things are better in parametric settings; see Remark I. There is also some question of what "bias" means with model selection-based estimators; see Remark G.

D. Adjusted R^2 Formula (5.5) for R^2_{adj}, not the usual definition, is motivated by OLS estimation and prediction in a

homoscedastic model. We observe

$$\boldsymbol{y} \sim (\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \tag{7.14}$$

and estimate $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y}$, where the $n \times n$ symmetric matrix $\boldsymbol{M}$ is idempotent, $\boldsymbol{M}^2 = \boldsymbol{M}$. Then $\hat{\sigma}^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2/(n-m)$, m the rank of $\boldsymbol{M}$, is the usual unbiased estimate of σ^2. Letting $\boldsymbol{y}^\circ$ indicate an independent new copy of $\boldsymbol{y}$, the expected prediction error of $\hat{\boldsymbol{\mu}}$ is

$$E\{\|\boldsymbol{y}^\circ - \hat{\boldsymbol{\mu}}\|^2\} = E\{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2m\hat{\sigma}^2\} \tag{7.15}$$

as in (2.6). Finally, the usual definition of R^2,

$$R^2 = 1 - \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 \,/\, \|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2 \tag{7.16}$$

is adjusted by adding the amount suggested in (7.15),

$$R^2_{\text{adj}} = 1 - \{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2m\hat{\sigma}^2\}/\,\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2\,, \tag{7.17}$$

and this reduces to (5.5).

E. Full-model bootstrapping The bootstrap replications (5.10) are drawn from the full model, $\boldsymbol{y}^* \sim \mathcal{N}_{39}(\hat{\boldsymbol{\mu}}_{\text{OLS}}, \boldsymbol{I})$, rather than say the smoothed Lasso choice (5.12), $\boldsymbol{y}^* \sim \mathcal{N}_{39}(\tilde{\boldsymbol{\mu}}, \boldsymbol{I})$. This follows the general development in Section 4 (4.3) and, less obviously, the theory of Sections 2 and 3, where the "full model" is the usual nonparametric one (2.3).

An elementary example, based on section 10.6 of Hjort and Claeskens (2003), illustrates the dangers of bootstrapping from other than the full model. We observe $y \sim \mathcal{N}(\mu, 1)$, with MLE $\hat{\mu} = t(y) = y$, and consider estimating μ with the shrunken estimator $\tilde{\mu} = s(y) = cy$, where c is a fixed constant $0 < c < 1$, so

$$\tilde{\mu} \sim \mathcal{N}(c\mu, c^2). \tag{7.18}$$

Full-model bootstrapping corresponds to $y^* \sim \mathcal{N}(\hat{\mu}, 1)$, and yields $\tilde{\mu}^* = cy^* \sim \mathcal{N}(c\hat{\mu}, c^2)$ as the bootstrap distribution. However the "model-selected bootstrap" $y^* \sim \mathcal{N}(\tilde{\mu}, 1)$ yields

$$\tilde{\mu}^* \sim \mathcal{N}(c^2\hat{\mu}, c^2), \tag{7.19}$$

squaring the amount of shrinkage in (7.18).

Returning to the Supernova example, the Lasso is itself a shrinkage technique. Bootstrapping from the Lasso choice $\tilde{\boldsymbol{\mu}}$ would shrink twice, perhaps setting many more of the coordinate estimates to zero.

F. Bias of the smoothed estimate In situations without model selection there is a simple asymptotic expression for the bias of the bootstrap smoothed estimator in exponential families, following DiCiccio and Efron (1992). The schematic diagram of Figure 8 shows the main elements: the observed vector $\boldsymbol{y}$, expectation $\boldsymbol{\mu}$, generates the bootstrap distribution of $\boldsymbol{y}^*$, indicated by the dashed ellipses. A parameter of interest $\theta = t(\boldsymbol{\mu})$ has MLE $\hat{\theta} = t(\boldsymbol{y})$. Isoplaths of constant value for $t(\cdot)$ are indicated by the solid curves in Figure 8.

The asymptotic mean and variance of the MLE $\hat{\theta} = t(\boldsymbol{y})$ as sample size n grows large are of the form

$$\hat{\theta} \sim \left(\theta + \frac{b(\boldsymbol{\mu})}{n}, \frac{c^2(\boldsymbol{\mu})}{n}\right) + O_p(n^{-3/2}). \tag{7.20}$$

Here the bias $b(\boldsymbol{\mu})/n$ is determined by the curvature of the level surfaces near $\boldsymbol{\mu}$. Then it is not difficult to show that the ideal smoothed bootstrap estimate $\tilde{\theta} = \sum t(\boldsymbol{y}_i^*)/B$, $B \to \infty$, has mean and variance

$$\tilde{\theta} \sim \left(\theta + 2\frac{b(\boldsymbol{\mu})}{n}, \frac{c^2(\boldsymbol{\mu})}{n}\right) + O_p(n^{-3/2}). \tag{7.21}$$

So smoothing *doubles the bias* without changing variance. This just says that smoothing cannot improve on the MLE $\hat{\theta}$ in the already smooth asymptotic estimation context of Figure 8.

G. Two types of bias The term $b(\boldsymbol{\mu})/n$ in (7.20) represents "statistical bias," the difference between the expected value of $t(\hat{\boldsymbol{\mu}})$ and $t(\boldsymbol{\mu})$. Model-selection estimators also involve

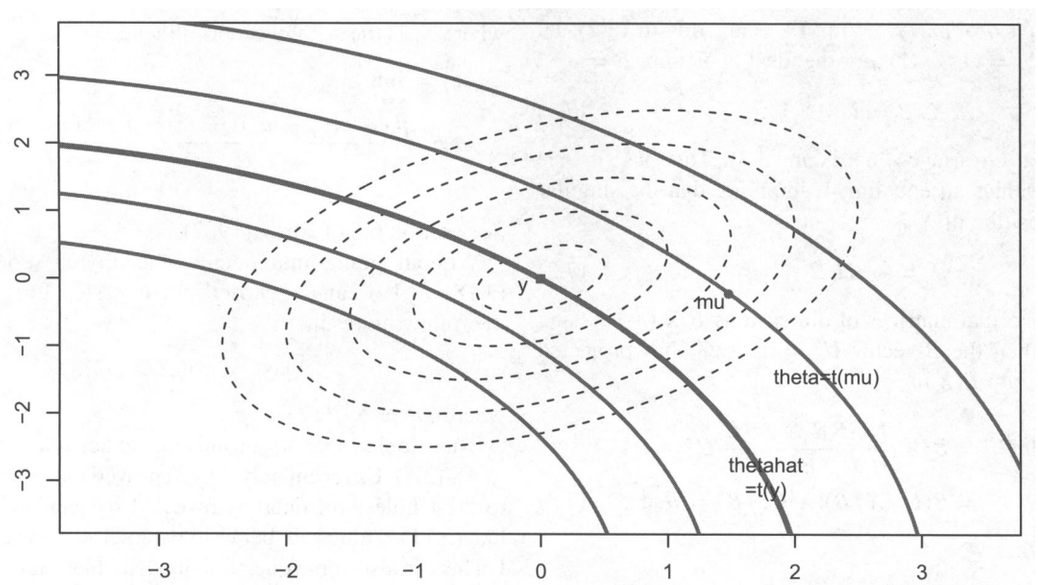

Figure 8. Schematic diagram of large-sample bootstrap estimation in situations without model selection. Observed vector $\boldsymbol{y}$ has expectation $\boldsymbol{\mu}$. Ellipses indicate bootstrap distribution of $\boldsymbol{y}^*$ given $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. Parameter of interest $\theta = t(\boldsymbol{\mu})$ is estimated by $\hat{\theta} = t(\boldsymbol{y})$. Solid curves indicate surfaces of constant value of $t(\cdot)$.

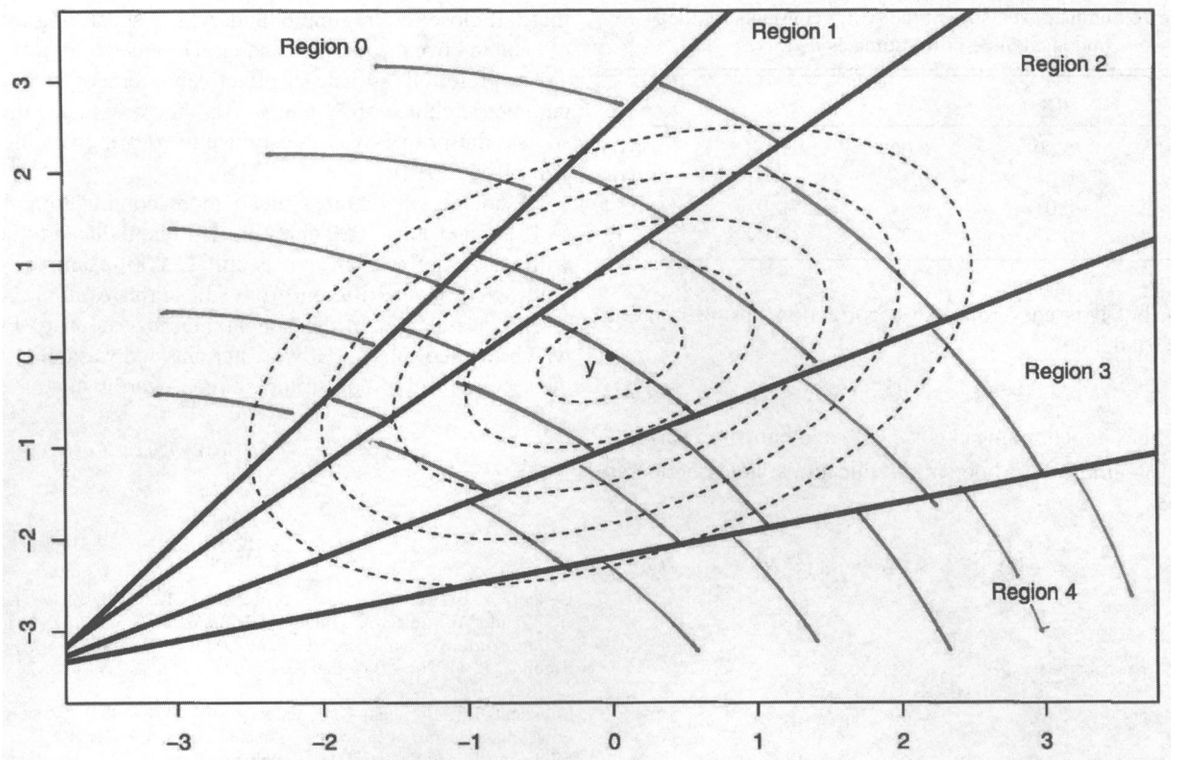

Figure 9. Estimation after model selection. The regions indicate different model choices. Now the curves of constant estimation jump discontinuously as $\boldsymbol{y}$ crosses regional boundaries.

"definitional bias": we wish to estimate $\theta = T(\boldsymbol{\mu})$, but for reasons of robustness or efficiency we employ a different functional $\hat{\theta} = t(\boldsymbol{y})$, a homely example being the use of a trimmed mean to estimate an expectation. The ABC bias correction mentioned in Section 6 is correcting the smoothed standard interval $\tilde{\mu} \pm 1.96\widetilde{\text{se}}_B$ for statistical bias. Definitional bias can be estimated by $t(\boldsymbol{y}) - T(\boldsymbol{y})$, but this is usually too noisy to be of help. Section 2 of Berk et al. (2012) makes this point nicely (see their discussion of "target estimation") and I have followed their lead in not trying to account for definitional bias. See also Bühlmann and Yu (2002), Definition 1.2, for an asymptotic statement of what is being estimated by a model-selection procedure.

H. Selection-based estimation The introduction of model selection into the estimation process disrupts the smooth properties seen in Figure 8. The wedge-shaped regions of Figure 9 indicate different model choices, for example, linear, quadratic, cubic, etc., regressions for the Cholesterol data. Now the surfaces of constant estimation jump discontinuously as $\boldsymbol{y}$ crosses regional boundaries. Asymptotic properties, such as (7.20)–(7.21), are less convincing when the local geometry near the observed $\boldsymbol{y}$ can change abruptly a short distance away.

The bootstrap ellipses in Figure 9 are at least qualitatively correct for the Cholesterol and Supernova examples, since in both cases a wide bootstrap variety of regions were selected. In this article, the main purpose of bootstrap smoothing is to put us back into Figure 8, where for example the standard intervals (2.11) are more believable. (*Note*: Lasso estimates are continuous, though nondifferentiable, across region boundaries, giving a picture somewhere between Figures 8 and 9. This might help explain the smooth estimators' relatively modest reductions in standard error for the Supernova analysis.)

Bagging amounts to replacing the discontinuous isoplaths of $\theta = t(\boldsymbol{\mu})$ with smooth ones, say for $\theta_{\text{bag}} = s(\boldsymbol{\mu})$. The standard deviations and approximate confidence intervals of this article apply to θ_{bag}, ignoring the possible definitional bias.

I. The ABC intervals The approximate bootstrap confidence limits in Figure 7 were obtained using the ABC_q algorithm, as explained in detail in section 2 of DiCiccio and Efron (1992). In addition to the acceleration a and bias-correction constant z_0, ABC_q also calculates c_q: in a one-parameter exponential family (4.1), c_q measures the nonlinearity of the parameter of interest $\theta = t(\beta)$ as a function of β, with a similar definition applying in p dimensions. The algorithm involves the calculation of $p + 2$ numerical second derivatives of s_α (6.4) carried out at $\alpha = \hat{\alpha}$. Besides a, z_0, and c_q, ABC_q provides an estimate of statistical bias for s_α.

If $(\alpha, z_0, c_q) = (0, 0, 0)$, then the ABC_q intervals match the smoothed standard intervals (2.11). Otherwise, corrections are made to achieve second-order accuracy. For instance $(a, z_0, c_q) = (0, -0.1, 0)$ shifts the standard intervals leftwards by $0.1 - \hat{\sigma}$. For all three constants, values outside of ± 0.1 can produce noticeable changes to the intervals.

Table 7 presents summary statistics of a, z_0, c_q, and bias for the 39 smoothed Supernova estimates $\tilde{\mu}_k$. The differences between the ABC_q and smoothed standard intervals seen in Figure 7 were primarily due to z_0.

J. Bias correction for $\widetilde{\text{sd}}_B$ The nonparametric standard deviation estimate $\widetilde{\text{sd}}_B$ (3.7) is biased upward for the ideal value

Table 7. Summary statistics of the ABC_q constants for the 39 smoothed Supernova estimates $\tilde{\mu}_k$ (5.12)

	a	z_0	c_q	Bias
Mean	0.00	0.00	0.00	0.00
St. dev.	0.01	0.13	0.04	0.06
Lowest	−0.01	−0.21	−0.07	−0.14
Highest	0.01	0.27	0.09	0.12

$\widetilde{\text{sd}}$ (3.4), but it is easy to make a correction. Using notation (3.3)–(3.9), define

$$Z_{ij}^* = (Y_{ij}^* - 1)(t_i^* - s_0). \tag{7.22}$$

Then Z_{ij}^* has bootstrap mean cov_j (3.5) and bootstrap variance say Δ_j^2. A sample of B bootstrap replications yields bootstrap moments

$$\widehat{\text{cov}}_j = \frac{1}{B}\sum_{i=1}^{B} Z_{ij}^* \sim_* \left(\text{cov}_j, \Delta_j^2/B\right), \tag{7.23}$$

so

$$E_*\widetilde{\text{sd}}_B^2 = \widetilde{\text{sd}}^2 + \frac{1}{B}\sum_{j=1}^{n}\Delta_j^2. \tag{7.24}$$

Therefore, the bias-corrected version of $\widetilde{\text{sd}}_B^2$ is

$$\widetilde{\text{sd}}_B^2 - \frac{1}{B^2}\sum_{j=1}^{n}\sum_{i=1}^{B}(Z_{ij}^* - \widehat{\text{cov}}_j)^2. \tag{7.25}$$

K. Improved estimates of the bagged standard errors The simulation experiment of Figure 5 can also be regarded as a two-level parametric bootstrap procedure, with the goal of better estimating $\text{sd}(\tilde{\mu}_k)$, the bagged standard deviations for subjects $k = 1, 2, \ldots, 11$ in the Cholesterol study. Two possible estimates are shown: (1) the empirical standard deviation $\widetilde{\text{Sd}}$ (4.22), solid curve, and (2) the average $\widetilde{\text{sd}}$. of the 100 second-level $\widetilde{\text{sd}}_i$ values (4.15), dashed curve. There are two reasons to prefer the latter.

The first has to do with the sampling error of the standard deviation estimates themselves. This was about 10 times larger for $\widetilde{\text{Sd}}$ than $\widetilde{\text{sd}}$., for example, 5.45 ± 0.35 compared to 5.84 ± 0.03 for subject 1. (*Note*: The two curves in Figure 5 do not differ significantly at any point.)

The second and more important reason has to do with the volitility of model-selection estimates and their standard errors. Let $\sigma(\beta)$ denote the standard deviation of a bagged estimator $\tilde{\mu}$ in a parametric model such as (4.16)–(4.17). The unknown true parameter β_0 has yielded the observed value $\hat{\beta}$, and then bootstrap values $\hat{\beta}_i^*$, $i = 1, 2, \ldots, 100$, and second-level bootstraps $\hat{\beta}_{ij}^*$, $j = 1, 2, \ldots, 1000$. The estimate $\widetilde{\text{sd}}_{100}$ obtained from the $\hat{\beta}_i^*$'s (4.15) is a good approximation to $\sigma(\hat{\beta})$. The trouble is that the functional $\sigma(\beta)$ is itself volatile, so that $\sigma(\hat{\beta})$ may differ considerably from the "truth" $\sigma(\beta_0)$.

This can be seen at the second level in Figure 5, where the dashes indicating $\widetilde{\text{sd}}_i$ values, $i = 1, 2, \ldots, 100$, vary considerably. (This is not due to the limitations of using $B = 1000$ replications; the bootstrap "internal variance" component accounts for only about 30% of the spread.) Broadly speaking, $\hat{\beta}_i^*$ values that fall close to a regime boundary, say separating the choice of "Cubic" from "Quartic," had larger values of $\sigma(\hat{\beta}_i^*) \doteq \widetilde{\text{sd}}_i$.

The preferred estimate $\widetilde{\text{sd}}$. effectively averages $\sigma(\hat{\beta}_i^*)$ over the parametric choice of $\hat{\beta}_i^*$ and $\hat{\beta}$. Another way to say this is that $\widetilde{\text{sd}}$. is a flat-prior Bayesian estimate of $\sigma(\beta_0)$, given the data $\hat{\beta}$. See Efron (2012).

Of course $\widetilde{\text{sd}}$. requires much more computation than $\widetilde{\text{sd}}_B$ (4.15). Our 100×1000 analysis could be reduced to 50×500 without bad effect, but that is still 25,000 resamples. In fact, $\widetilde{\text{sd}}$. was not much different from $\widetilde{\text{sd}}_B$ in this example. The difference was larger in the nonparametric version of Figure 5, which showed substantially greater bias and variability, making the second level of bootstrapping more worthwhile.

[Received August 2012. Revised April 2013.]

REFERENCES

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2012), "Valid Post-Selection Inference," *Annals of Statistics*, available at *http://stat.wharton.upenn.edu/ buja/PoSI.pdf* [991,1005]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [991,994]

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [992]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [991,1005]

Buja, A., and Stuetzle, W. (2006), "Observations on Bagging," *Statistica Sinica*, 16, 323–351. [991,994]

Chatterjee, A., and Lahiri, S. N. (2011), "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 106, 608–625. [1001]

DiCiccio, T., and Efron, B. (1992), "More Accurate Confidence Intervals in Exponential Families," *Biometrika*, 79, 231–245. [991,995,1002,1004,1005]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [993]

——— (1982), "The Jackknife, the Bootstrap and Other Resampling Plans," CBMS-NSF Regional Conference Series in Applied Mathematics *38*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). [996,1003]

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-validation," *Journal of the American Statistical Association*, 78, 316–331. [1003]

——— (1987), "Better Bootstrap Confidence Intervals," (with comments and a rejoinder by the author), *Journal of the American Statistical Association*, 82, 171–200. [996]

——— (1992), "Jackknife-after-Bootstrap Standard Errors and Influence Functions," *Journal of the Royal Statistical Society,* Series B, 54, 83–127. [995]

——— (2012), "Bayesian Inference and the Parametric Bootstrap," *Annals of Applied Statistics*, 6, 1971–1997. [1006]

Efron, B., and Feldman, D. (1991), "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association*, 86, 9–17. [992]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with comments and a rejoinder by the authors), *The Annals of Statistics*, 32, 407–499. [999]

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall. [994]

——— (1996), "Using Specially Designed Exponential Families for Density Estimation," *The Annals of Statistics*, 24, 2431–2461. [994]

Fearnhead, P., and Prangle, D. (2012), "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation," *Journal of the Royal Statistical Society,* Series B, 74, 419–474. [1002]

Hájek, J. (1968), "Asymptotic Normality of Simple Linear Rank Statistics Under Alternatives," *The Annals of Mathematical Statistics*, 39, 325–346. [1003]

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag. [1001]

Hall, P., Lee, E. R., and Park, B. U. (2009), "Bootstrap-Based Penalty Choice for the Lasso, Achieving Oracle Performance," *Statistica Sinica*, 19, 449–471. [1001]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [991,1004]
Hurvich, C. M., and Tsai, C.-L. (1990), "Model Selection for Least Absolute Deviations Regression in Small Samples," *Statistics & Probability Letters*, 9, 259–265. [992]
Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378. [1001]
Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675. [991]
Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R., Nugent, P., Castro, P., Deustua, S., Fabbro, S., Goobar, A., Groom, D., Hook, I., Kim, A., Kim, M., Lee, J., Nunes, N., Pain, R., Pennypacker, C., Quimby, R., Lidman, C., Ellis, R., Irwin, M., McMahon, R., Ruiz-Lapuente, P., Walton, N., Schaefer, B., Boyle, B., Filippenko, A., Matheson, T., Fruchter, A., Panagia, N., Newberg, H., and Couch, W. (1999), "Measurements of Omega and Lambda From 42 High-redshift Supernovae," *The Astrophysical Journal*, 517, 565–586. [999]
Riess, A., Filippenko, A., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P., Gilliland, R., Hogan, C., Jha, S., Kirshner, R., Leibundgut, B., Phillips, M., Reiss, D., Schmidt, B., Schommer, R., Smith, R., Spyromilio, J., Stubbs, C., Suntzeff, N., and Tonry, J. (1998), "Observational Evidence From Supernovae for an Accelerating Universe and a Cosmological Constant," *The Astrophysical Journal*, 116, 1009–1038. [999]
Sexton, J., and Laake, P. (2009), "Standard Errors for Bagged and Random Forest Estimators," *Computational Statistics and Data Analysis*, 53, 801–811. [992]
Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [999]

Comment

Lan WANG, Ben SHERWOOD, and Runze LI

We congratulate Efron for his stimulating and timely work that addresses an important issue on estimation after model selection. In practice, it is typical to ignore the variability of the variable selection step, which could result in inaccurate post-selection inference. Although the flaw of this practice is widely recognized, finding a general solution is extremely challenging. The model selection step is often a complex decision process and can involve collecting expert opinions, preprocessing, applying a variable selection rule, data-driven choice of one or more tuning parameters, among others. Except in simple cases, explicitly characterizing the form of the post-selection estimator is itself difficult. The key result of this article is a closed-form formula for obtaining the standard deviation of a "*bootstrap smoothed*" (or "*bagged*") estimator. This elegant formula is not only simple to implement but also versatile. It indeed provides a general approach for obtaining a confidence interval for a class of parameters of interest while incorporating the variability of variable selection.

Our discussions will focus on two aspects: (1) the generality of the method, and (2) further insight into the performance of the proposed method in a simple but, we hope, informative example.

1. GENERALITY OF THE METHOD

In principle, the standard deviation formula in Efron's Theorem 1 can be applied to general "*bootstrap smoothed*" (or "*bagged*") estimators. As the central example of the article is traditional linear regression, we empirically investigate the performance of the proposed estimator in a variety of regression settings where the proposed method is expected to be useful through Monte Carlo simulations. In particular, we will consider: (1) LASSO (least absolute shrinkage and selection operator; Tibshirani 1996) and SCAD (smoothly clipped absolute deviation; Fan and Li 2001) for linear regression, (2) Poisson regression as a representative example of generalized linear models, (3) quantile regression for predicting a conditional quantile, and (4) nonparametric regression where we apply a data-driven method to select the number of spline basis functions (this last example was motivated by a discussion with Professor Xuming He).

For each of the four cases, we construct confidence intervals for the conditional mean (or quantile) using the new method proposed in Efron's article (denoted by "new"). We compare the new method with the standard bootstrap confidence interval (denoted by "standard") and the percentile interval (denoted by "percentile"), as described in Efron's article.

1.1 Several Numerical Examples

Example 1. (Regularized estimators for linear regression). The response variable is generated from the model $Y = 1 + X_1 - X_3 + X_6 + \epsilon_i$, where the candidate covariates $X_1, \ldots, X_6$ are independent standard normal random variables. The random error ϵ is normally distributed with mean zero and standard deviation 2, and is independent of the covariates. The sample size is $n = 200$. The main goal is to study the proposed method when regularized methods such as LASSO and SCAD are used to obtain the selected model. We implement LASSO using the R package *glmnet* and implement SCAD using the coordinate descent algorithm in the R package *ncvreg*. For LASSO, we use five-fold cross-validation to select the tuning parameter; while for SCAD we apply BIC (Bayesian information criterion; Wang, Li, and Tsai 2007) for selecting the tuning parameter. For completeness, we also include best subset selection procedures based on C_p, Akaike information criterion (AIC), and BIC.

Lan Wang (E-mail: *wangx346@umn.edu*) is Associate Professor and Ben Sherwood (E-mail: *sher0422@umn.edu*) is graduate student, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Runze Li is Distinguished Professor, Department of Statistics and the Methodology Center, the Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rzli@psu.edu*). Wang and Sherwood's research is supported by an NSF grant DMS1308960. Li's research is supported by NIDA, NIH grants P50 DA10075 and P50 DA036107. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.905399