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,

$$c \mathrel{\dot\sim} \mathcal{N}(0, 1). \tag{1.3}$$

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

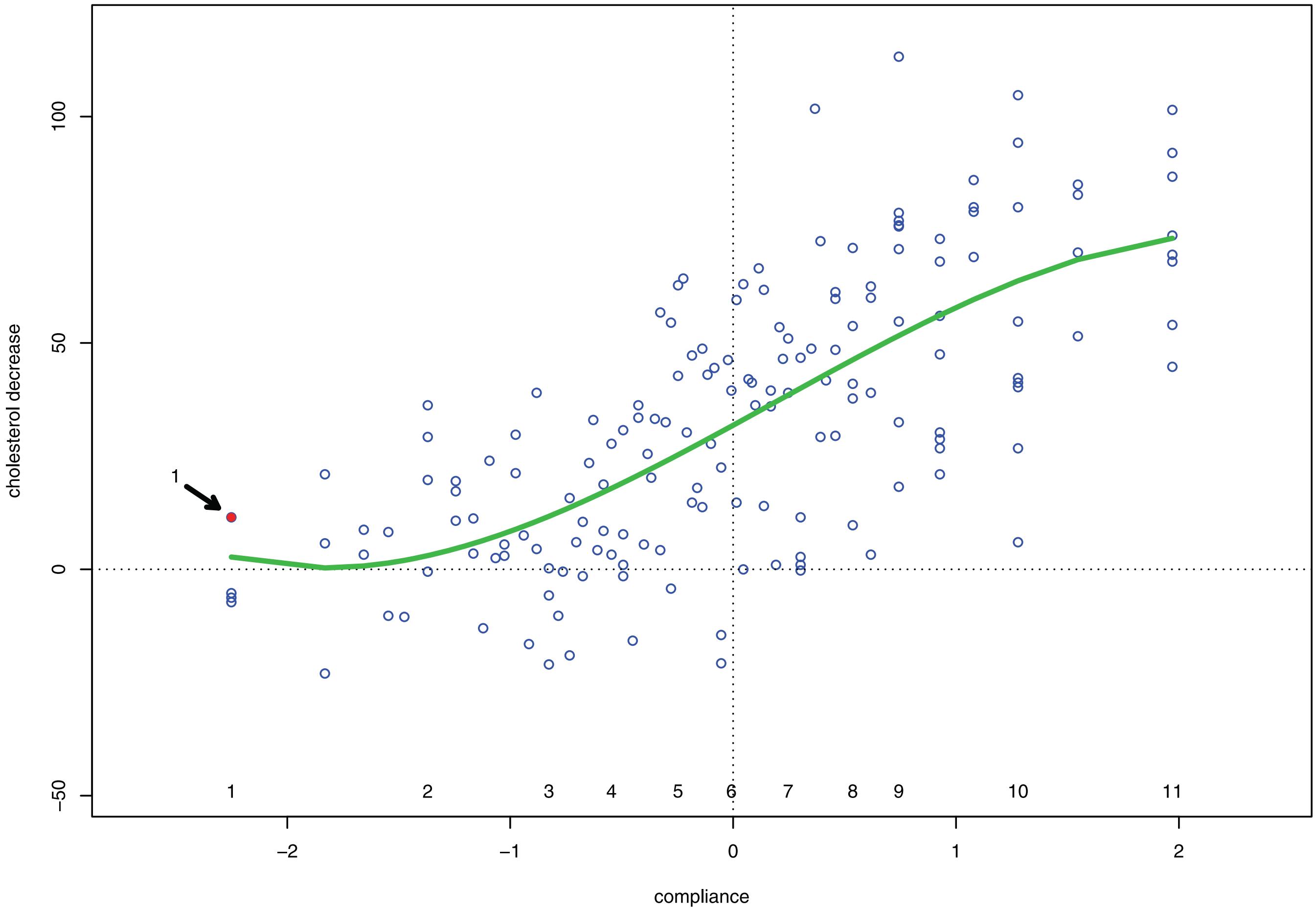

Figure 1. *Cholesterol data* Cholesterol decrease plotted versus adjusted compliance for 164 men in Treatment arm of the cholostyramine study (Efron and Feldman 1991). Solid curve is OLS cubic regression, as selected by the C_p criterion. How accurate is the curve, taking account of model selection as 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 $\hat{\boldsymbol{\mu}} = t(\boldsymbol{y})$ 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

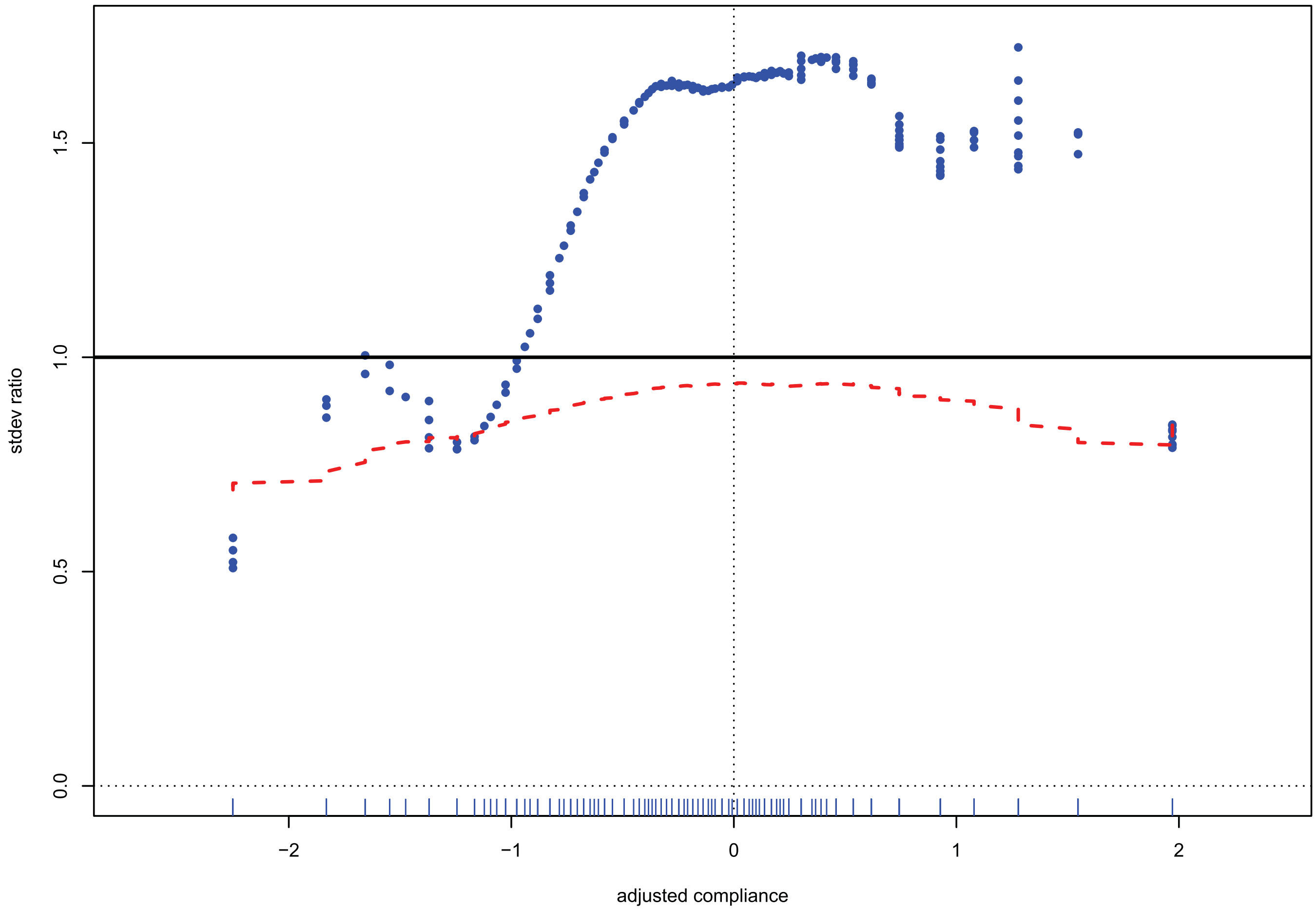

Figure 2. *Solid points*: ratio of standard deviations, taking account of model selection or not, for the 164 values $\hat{\mu}_j$ from the regression curve in Figure 1. Median ratio equals 1.52. Standard deviations including model selection are the smoothed bootstrap estimates $\widetilde{\text{sd}}_B$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

empirical standard deviation of B such draws,

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

is the familiar nonparametric bootstrap estimate of standard error for $\hat{\mu}$ (Efron 1979); $\widehat{\text{sd}}_B$ 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

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

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

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

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%)

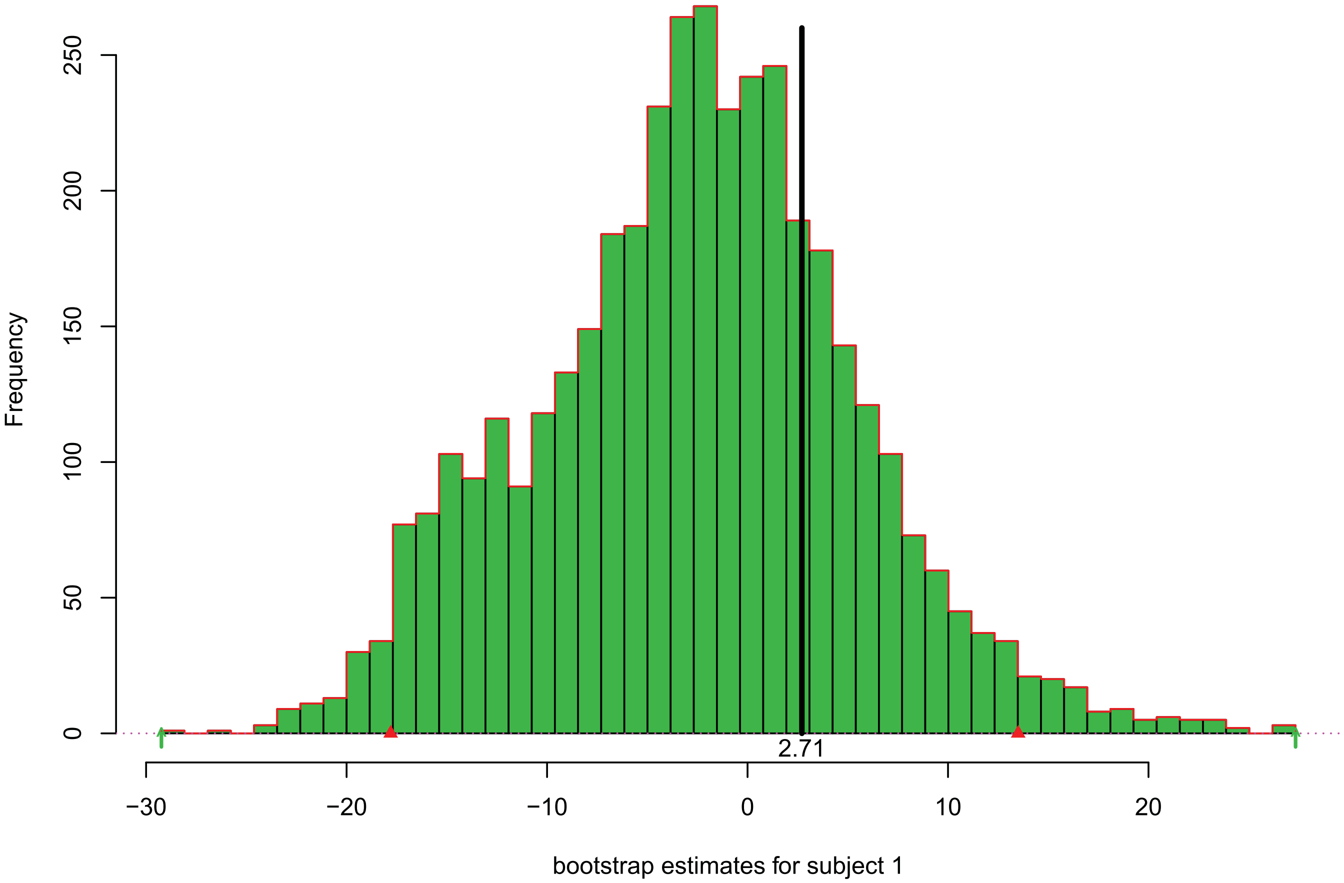

Figure 3. $B = 4000$ bootstrap replications $\hat{\mu}_1^*$ of the C_p-OLS regression estimate for Subject 1. The original estimate $t(\boldsymbol{y}) = \hat{\mu}_1$ is 2.71, exceeding 76% of the replications. Bootstrap standard deviation (2.4) equals 8.02. Triangles indicate 2.5th and 97.5th percentiles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

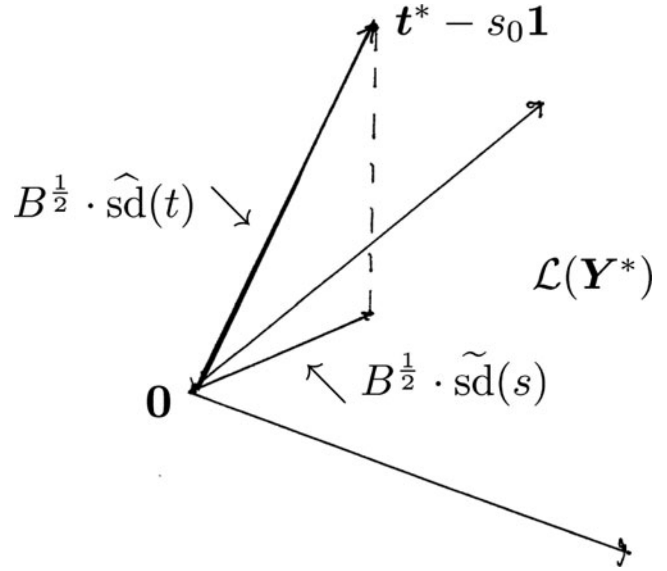
$t^* - s_0\mathbf{1}$
$B^{\frac{1}{2}} \cdot \widehat{\text{sd}}(t)$
$\mathcal{L}(\boldsymbol{Y}^*)$
$\mathbf{0}$
$B^{\frac{1}{2}} \cdot \widetilde{\text{sd}}(s)$

Figure 4. Illustration of Corollary 1. The ratio $\widetilde{\text{sd}}_B/\widehat{\text{sd}}_B$ is the cosine of the angle between $\boldsymbol{t}^* - s_0\mathbf{1}$ (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 $\widetilde{\text{sd}}_B$? The jackknife provides a quick answer: divide the B replications into J groups of size B/J each, and let $\widetilde{\text{sd}}_{Bj}$ be the estimate (3.6) computed with the jth group removed. Then

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

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

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

$\boldsymbol{p}_0 = (1/n, 1/n, \ldots, 1/n)$, 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 $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$,

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

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

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

$\boldsymbol{\delta}_j$ 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

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

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),

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

so that

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

(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

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

Letting $\epsilon \to 0$ yields

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

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

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

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

$$\dot{S}_j = n\,\text{cov}_j \qquad (3.21)$$

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,

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

If $\hat{a} = 0.10$ for instance, then at the upper endpoint $\hat{\theta}_{\text{up}} = \hat{\theta} + 1.96\,\hat{\sigma}$ the standard error will have increased to about $1.196\hat{\sigma}$, leaving $\hat{\theta}_{\text{up}}$ 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 $\widehat{\text{cov}}_j$ used to calculate $\widetilde{\text{sd}}_B$ in (3.6),

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

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):

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

$\boldsymbol{1}_B$ 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 $t^* = t(\boldsymbol{y}^*)$,

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

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,

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

the ratio being

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

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 $\boldsymbol{t}^* - s_0\boldsymbol{1}_B$ 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

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

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

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

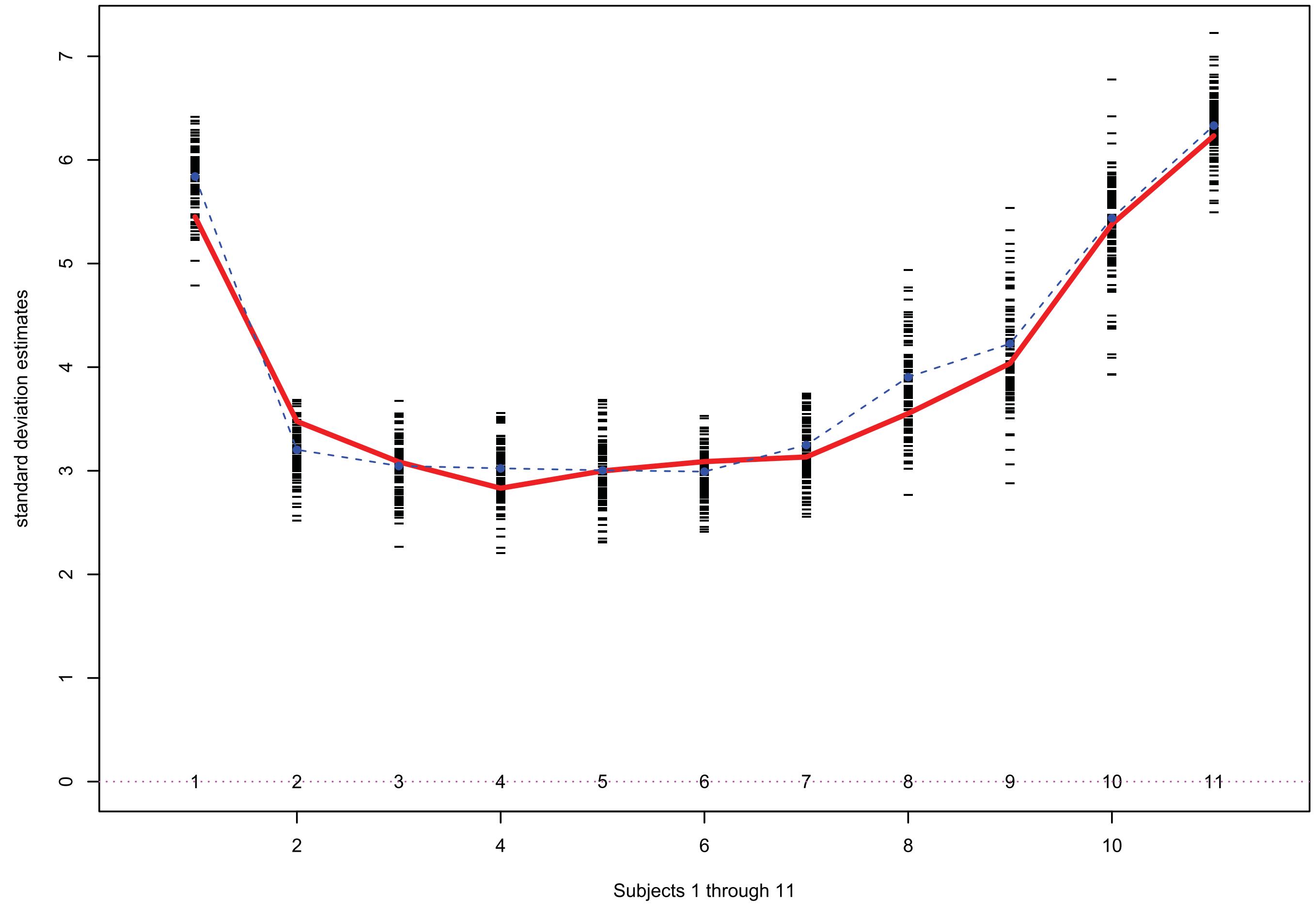

Figure 5. Simulation test of Theorem 2, parametric model (4.16)–(4.18), Cholesterol data; 100 simulations, 1000 parametric bootstraps each, for the 11 subjects indicated at the bottom of Figure 1. Heavy line connects observed empirical standard deviations (4.22); dashes show the 100 estimates $\widetilde{\text{sd}}$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

As $B \to \infty$, $\widehat{\text{cov}} \to \text{cov}_*$, and $\bar{V} \to \hat{V}$, so $\widetilde{\text{sd}}_B \to \widetilde{\text{sd}}$ (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,

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

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

$$\sigma_i = 23.7 + 5.49c - 2.25c^2 - 1.03c^3, \qquad (4.17)$$

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

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

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

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

and covariance matrix (4.2)

$$V = G^{-1}, \qquad (4.20)$$

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 $\hat{\beta}_i^*$,

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

from which smoothed estimate $\tilde{\mu}_i$ (4.5) and estimated standard deviation $\widetilde{\text{sd}}_i$ 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 $\widetilde{\text{sd}}_i$ values for each of the 11 subjects. This is compared with the observed empirical standard deviations of the smoothed estimates,

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

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

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

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

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),

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

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,

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

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

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

so from (4.2),

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

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 $\underset{\sim}{\mathrm{sd}}$ 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

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

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

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

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

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

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

$$\hat{\boldsymbol{\mu}}_{\mathrm{OLS}} = X\hat\alpha_{\mathrm{OLS}}. \tag{5.4}$$

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

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

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

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

$\hat\alpha_\lambda$ 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

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

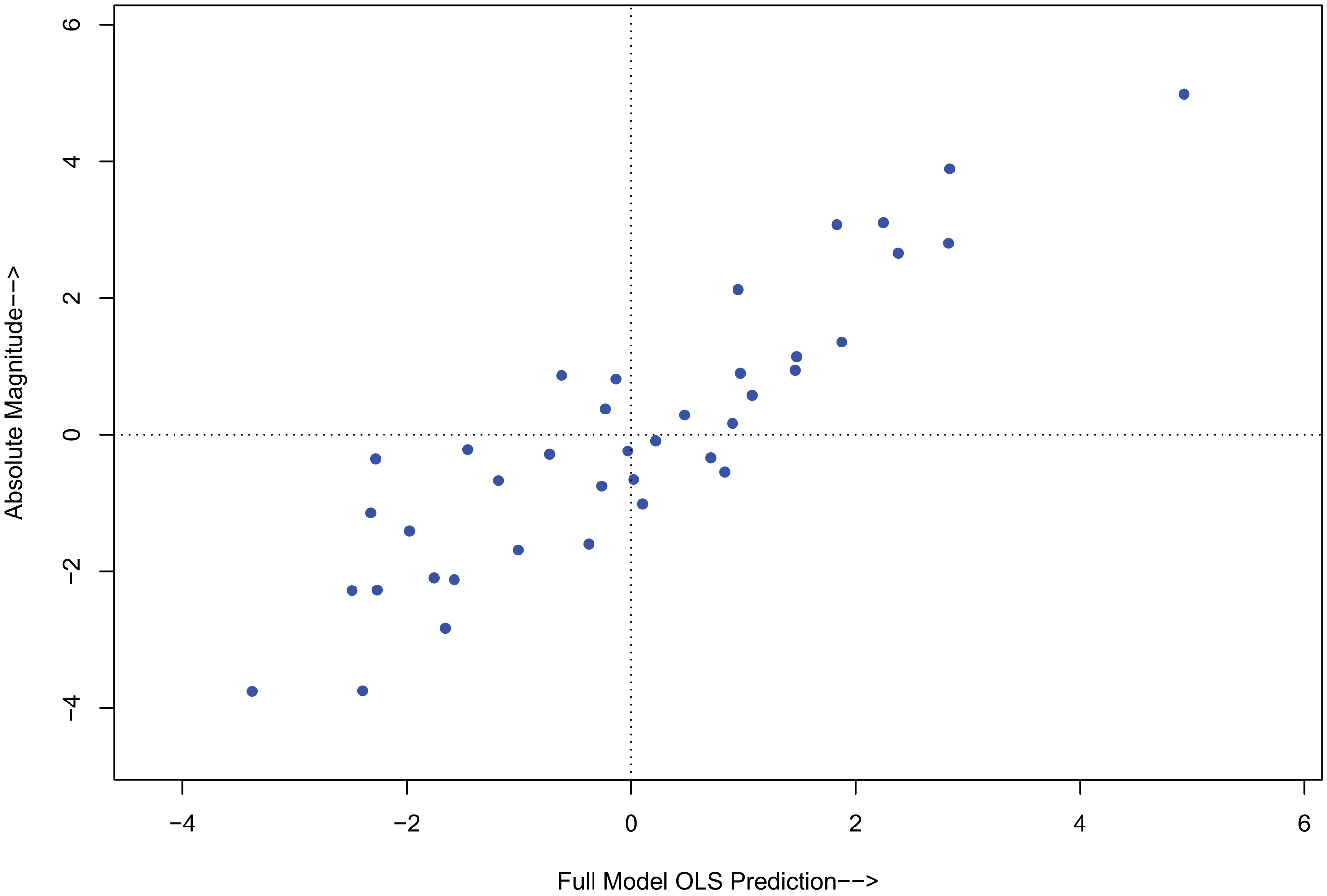

Figure 6. *The Supernova data* Absolute magnitudes of $n = 39$ Type Ia supernovas plotted versus their OLS estimates from the full linear model (5.3); adjusted R^2 (5.5) equals 0.69.

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

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

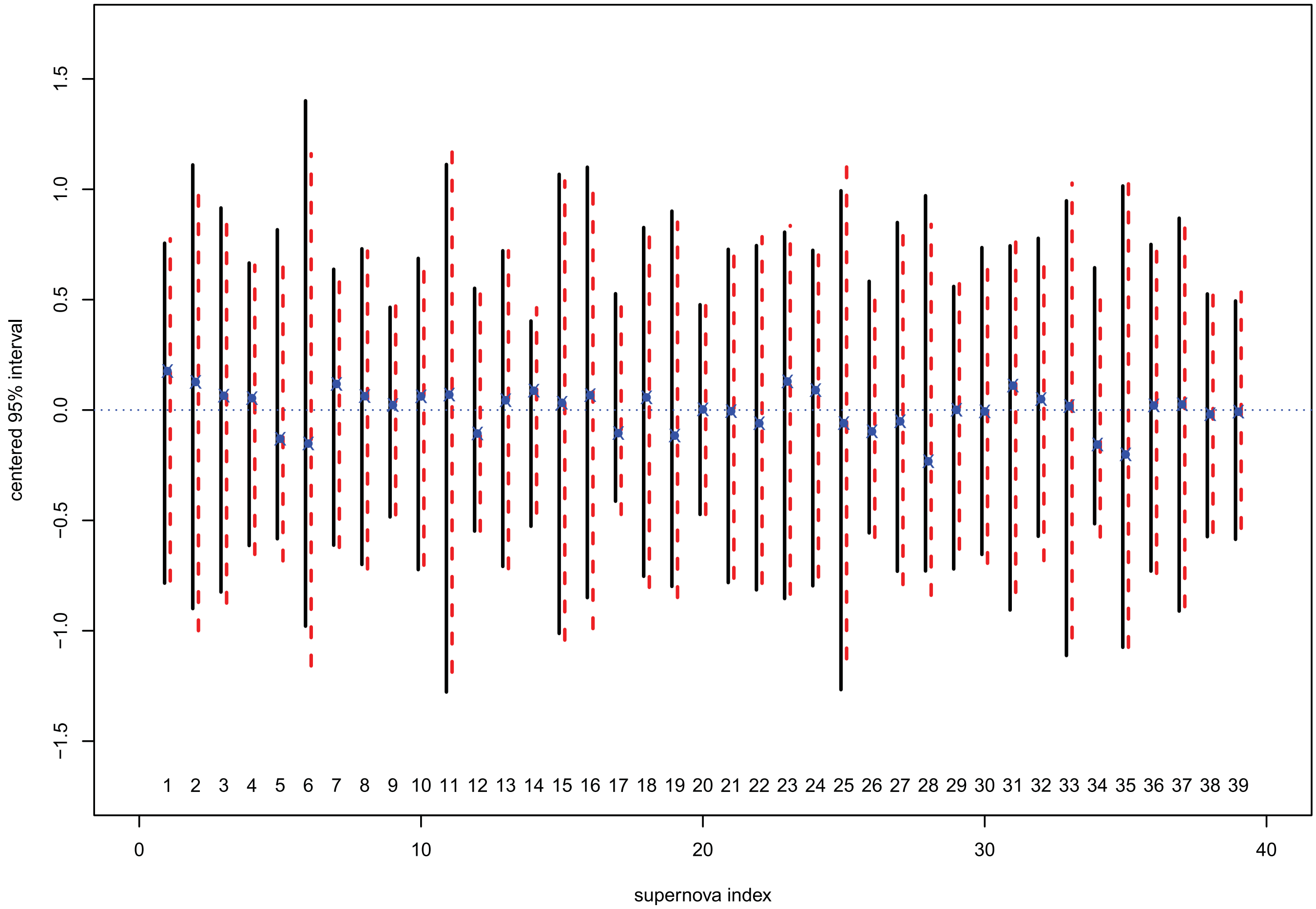

Figure 7. Approximate 95% confidence limits for the 39 supernova magnitudes μ_k (after subtraction of smoothed estimates $\tilde{\mu}_k$ (5.12)); ABC intervals (solid) compared with smoothed standard intervals $\tilde{\mu}_k \pm 1.96\widetilde{\mathrm{sd}}_k$ (dashed). Crosses indicate differences between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

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

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

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

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

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

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

$j = 1, 2, \dots, n$ (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,

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

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

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

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

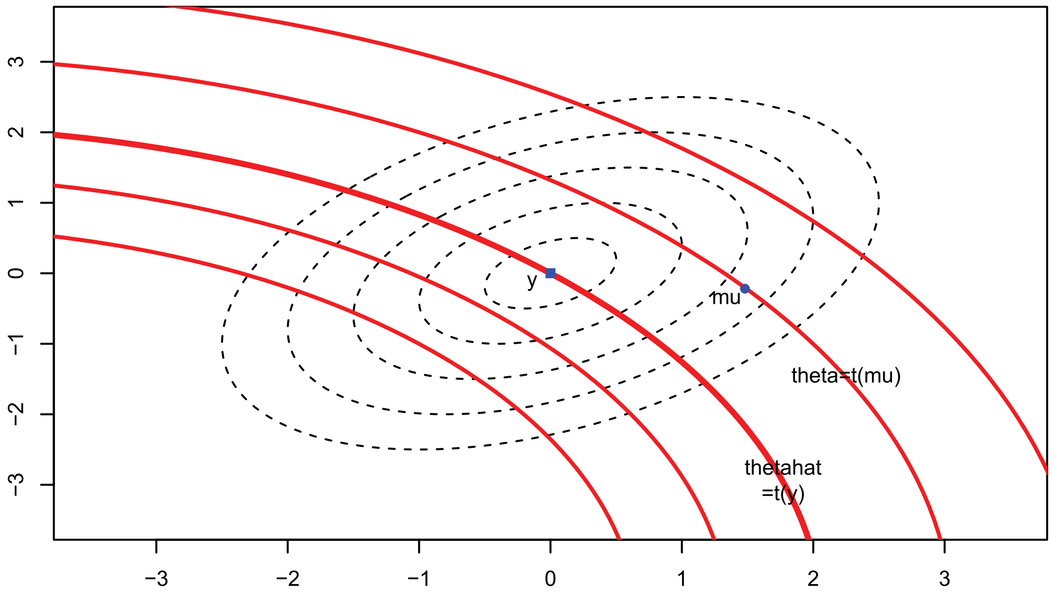

Figure 8. Schematic diagram of large-sample bootstrap estimation in situations without model selection. Observed vector $\boldsymbol{y}$ has expectation $\boldsymbol{\mu}$. Ellipses indicate bootstrap distribution of $\boldsymbol{y}^*$ given $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. Parameter of interest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

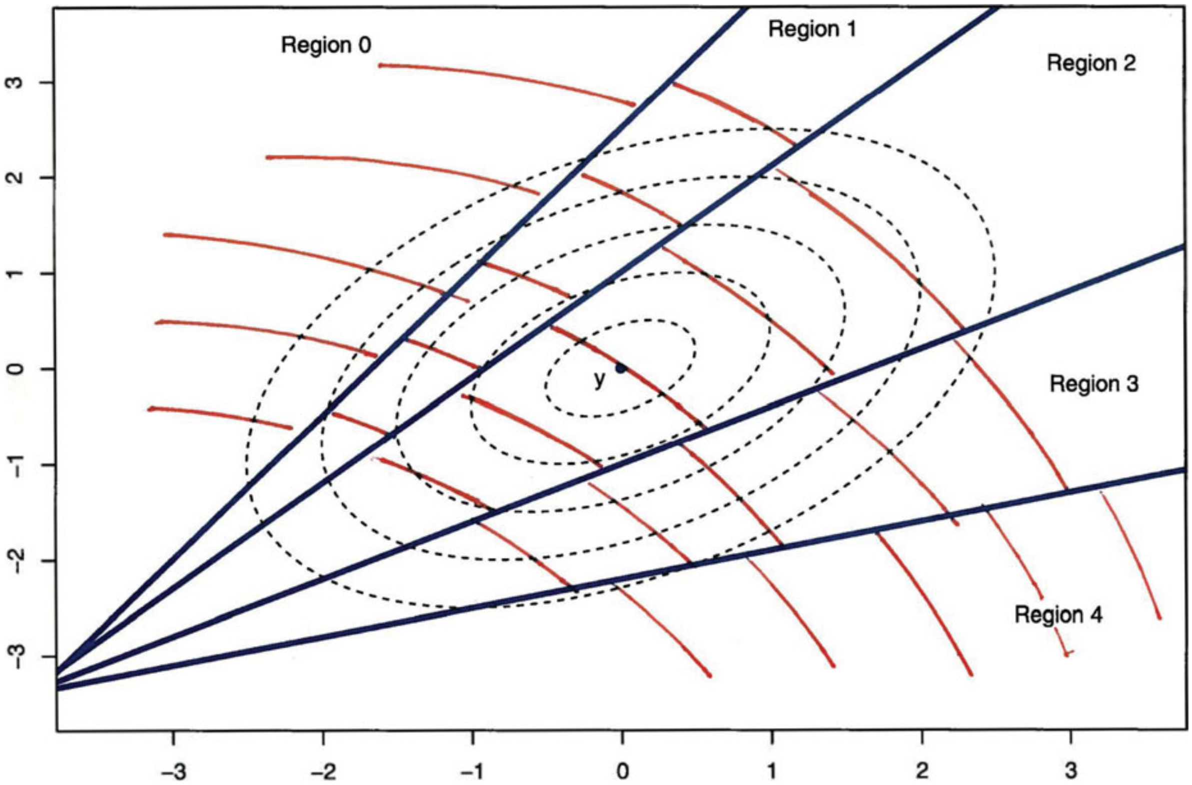

Figure 9. Estimation after model selection. The regions indicate different model choices. Now the curves of constant estimation jump discontinuously as $\boldsymbol{y}$ crosses regional boundaries.

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

Of course $\widetilde{\text{sd}}_{\cdot}$ 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, $\widetilde{\text{sd}}_{\cdot}$ 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.

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

Table 1. Linear regression

Method	Interval type	Center	Length	Coverage
C_p	New	−1.58	3.25	0.98
	Percentile	−1.58	3.46	0.98
	Standard	−1.54	3.46	1.00
AIC	New	−1.58	3.26	0.97
	Percentile	−1.58	3.46	0.98
	Standard	−1.54	3.46	1.00
BIC	New	−1.56	2.93	0.98
	Percentile	−1.56	3.23	0.98
	Standard	−1.52	3.22	0.99
LASSO	New	−1.47	3.33	0.96
	Percentile	−1.49	3.39	0.97
	Standard	−1.40	3.38	0.95
SCAD	New	−1.55	2.97	0.98
	Percentile	−1.55	3.25	0.98
	Standard	−1.51	3.24	0.99

We consider the 95% confidence interval for estimating the conditional mean at $X = (-2.5, -2.5, -2.5, -2.5, -2.5, -2.5)'$. The results are summarized in Table 1 based on 4000 bootstrap samples. We assess the performance by the length of the confidence interval and its coverage probability (reported in the last two columns of the table). The third column reports the center of the confidence interval.

Example 2. (Poisson regression). The response variable is generated from the model $Y \mid X \sim \text{Poisson}(e^{1+X-X^2})$, where X has a standard normal distribution. The sample size is $n = 400$. We use AIC and BIC for model selection. For candidate models, we consider different polynomial degrees of X, from linear to sextic. The results for the confidence interval for estimating $E[Y \mid X = -2]$ are reported in Table 2 based on 6000 bootstrap runs.

Example 3. (Quantile regression). The response variable is generated from the heteroscedastic regression model $Y = 1 + 3X_1 - 1.5X_3 + 2X_6 + (1 + X_2)\epsilon$, where the X_i's, $i = 1, \ldots, 6$, are independent and uniformly distributed on (0, 1). The random error ϵ has a standard normal distribution and is independent of the X_i's. The sample size is $n = 200$.

We considered AIC and BIC for model selection, which are based on the quantile loss function and programmed in the *quantreg* package in R. Penalized quantile regression with LASSO or SCAD penalty is also considered. The results for the confidence interval for estimating the 0.7 conditional quantile at $X = (0.9, \ldots, 0.9)'$ are reported in Table 3 based on 4000 bootstrap runs.

Table 2. Poisson regression

Method	Interval type	Center	Length	Coverage
AIC	New	20.13	3.09	0.97
	Percentile	20.14	3.75	0.99
	Standard	20.18	3.71	0.99
BIC	New	20.12	2.07	0.97
	Percentile	20.13	2.56	0.97
	Standard	20.11	2.54	0.98

Table 3. Quantile regression

Method	Interval type	Center	Length	Coverage
AIC	New	5.10	1.85	0.95
	Percentile	5.11	2.12	0.98
	Standard	5.08	2.12	1.00
BIC	New	5.07	1.77	0.94
	Percentile	5.09	2.12	0.97
	Standard	5.05	2.13	0.97
LASSO	New	5.00	1.73	0.94
	Percentile	5.02	2.00	0.95
	Standard	5.03	2.02	0.96
SCAD	New	5.06	1.77	0.94
	Percentile	5.08	2.11	0.98
	Standard	5.05	2.12	0.97

Example 4. (Nonparametric regression). The response variable is generated from the regression model $Y = 1 + X^2\exp(X) + \epsilon$, where X is uniformly distributed on (0, 1). The random error ϵ is normally distributed with mean zero and standard deviation 2, and is independent of X. The sample size is $n = 100$.

We estimate the nonparametric regression function via B-spline regression. We select the number of knots (ranging from 1 to 5) by a BIC criterion. More specifically, let ν represent the number of degrees of freedom of a candidate model and let $\hat{\sigma}_\nu^2$ be the estimate of σ^2 for the corresponding model. We select the model that minimizes $\text{BIC}(\nu) = n\log(\hat{\sigma}_\nu^2) + \nu\log(n)$, see, for example, He and Shi (1996). The results for the confidence interval for estimating the conditional mean at $X = 0.9$ are reported in Table 4 based on 4000 bootstrap runs.

1.2 Observations From the Numerical Examples

In the above examples, we observe that the new confidence interval proposed in Efron's article provides a more accurate confidence interval for all cases and keeps better coverage rates for most cases than the standard interval and the percentile interval when the estimator is obtained after variable selection.

From our limited simulation experience, we note that the choice of the number of bootstrap samples is important to the performance of the new method. A suitable choice of B can vary depending on the underlying model and the amount of noise in the data. We find that $B = 4000$ works reasonably well for most of the situations we have considered.

An interesting observation from our simulations is that the new method can also be useful for regularized procedures, in particular SCAD, when the tuning parameter is chosen in a data-driven fashion. It is known that the "*bootstrap smoothed*" (or "*bagged*") estimators are most valuable when hard decision rules (such as best subset selection, decision trees) are involved, which result in instability in prediction. In practice, when a

Table 4. Nonparametric regression

Interval type	Center	Length	Coverage
New	2.99	1.83	0.93
Percentile	2.99	2.18	0.97
Standard	2.99	2.16	0.97

regularization procedure such as LASSO or SCAD is applied, the tuning parameter is often selected by cross-validation or a modified BIC, which introduces extra variability in the final estimator. Although the improvement over LASSO is sometimes marginal as Efron has pointed out, it may still be worthwhile (in the quantile regression example, we observe a 15% reduction of interval length for LASSO). For SCAD, with the tuning parameter being selected by BIC, the improvement is more significant. Our simulation experience, including that not reported here due to space limitation, indicates that the gain of the new method is more pronounced when the sample size is smaller and the data are noisier.

2. FURTHER INSIGHT FROM A SIMPLE EXAMPLE

Next, we will consider Efron's main example in the orthogonal regression case, which sheds some light on its performance. Let Y be the $n \times 1$ vector of responses and $X = (X_1, \ldots, X_p)^T$ be the design matrix. It is assumed that $X^T X = nI_n$, where I_n is the $n \times n$ identity matrix. The least-square estimator for β_j is $\widehat{\beta}_j = n^{-1}X_j^T Y$.

For a given model M, where M denotes an index set for the covariates in the model, Mallow's C_p is defined as $C_p(M) = (Y - X_M\widehat{\beta}_M)^T(Y - X_M\widehat{\beta}_M) + 2\sigma^2|M|$, where X_M denotes the submatrix of X corresponding to M, and β_M denotes the least-square estimator for model M. In the orthogonal regression case, it is easy to see

$$C_p(M) = Y^T Y + \sum_{j\in M}\big(-n\widehat{\beta}_j^2 + 2\sigma^2\big).$$

As a result, C_p selects all X_j such that $-n\widehat{\beta}_j^2 + 2\sigma^2 < 0$. Hence, given a vector of covariates $x = (x^{(1)}, \ldots, x^{(p)})$, the estimator of $E(Y|X = x)$ obtained after applying Mallow's C_p criterion can be written as

$$\sum_{j=1}^{p} x^{(j)}\widehat{\beta}_j I(|\widehat{\beta}_j| > \sigma\sqrt{2/n}).$$

Since the effect of each covariate is additive, we consider the univariate case in the following discussion. The post-selection estimator of the conditional mean at x is

$$t_n(Y|x) = x\widehat{\beta}I(|\widehat{\beta}| > \sigma\sqrt{2/n}).$$

The bootstrap smoothed estimator given by Efron is

$$s_n(Y|x) = B^{-1}\sum_{i=1}^{B} t_n(Y^*|x),$$

where Y^* is the bootstrap sample.

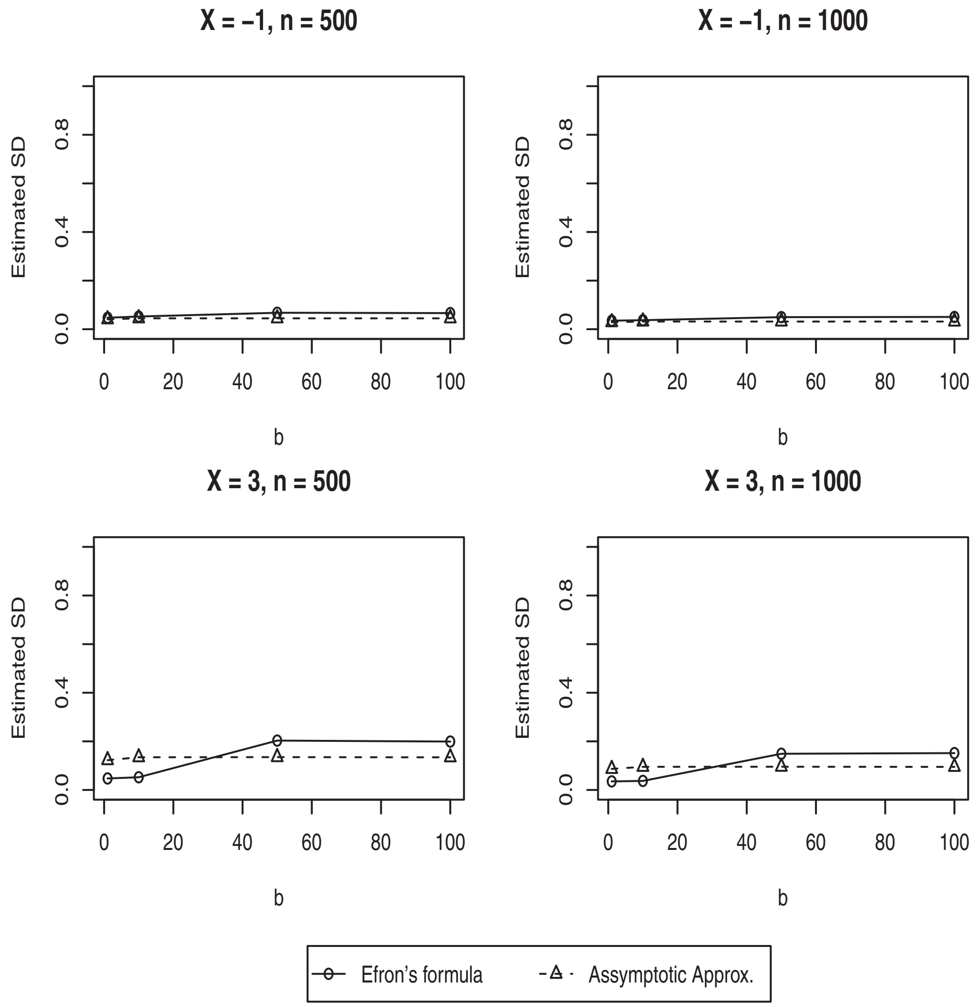

Figure 1. Comparing Efron's estimator with the theoretical value.

The asymptotic distribution of $s_n(Y|x)$ is known under a local asymptotic framework. Assume that $Y_i = \beta X_i + \epsilon_i$, where $\beta = \beta_n(b) = b\sigma n^{-1/2}$ for some constant b, $X_1, \dots, X_n$ are iid random variables with $\mathrm{E}(X_i^2) = 1$, $\epsilon_1, \dots, \epsilon_n$ are iid and independent from the X_i's, $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2 < \infty$. It follows Proposition 2.2 of Bühlmann and Yu (2002) that,

$$n^{1/2}\sigma^{-1}s_n(Y|x) \to g_B(Z_b|x)$$

in distribution, where $Z_b = b + Z$, $Z \sim N(0, 1)$, and $g_B(z|x) = (z - \{z\Phi(\sqrt{2} - z) - \phi(\sqrt{2} - z) - z\Phi(-\sqrt{2} - z) + \phi(-\sqrt{2} - z)\})x$, with Φ and ϕ denoting the distribution function and density function of the standard normal distribution, respectively. The theory thus suggests that the bootstrap smoothed estimator has approximate standard deviation $n^{-1/2}\sigma \times \mathrm{sd}(g_B(z|x))$, where $\mathrm{sd}(g_B(z|x))$ denotes the standard deviation of the distribution given by $g_B(z|x)$.

In Figure 1, we compare the estimated standard deviation of $s_n(Y|x)$ using Efron's formula with that obtained from the above asymptotic distribution (based on simulating the distribution of $g_B(z|x)$) for different values of b at $x = -1$ and 3, for sample sizes $n = 500$ and 1000. The two curves are quite close to each other, suggesting that Efron's estimator performs well in this setting. It is noted that AIC and BIC can be analyzed similarly in the orthogonal design case.

3. CONCLUSIONS

Two intriguing questions about Efron's new procedure are: (1) Is it possible to derive the asymptotic property, such as consistency? (2) Can the nonparametric delta method used for deriving the standard deviation formula be extended to the case the number of covariates p_n grows with n? Positive answers to these questions will greatly extend the scope of the application of the new method.

As the bootstrap-smoothed estimator combines estimators from different candidate models, it may be applicable to situations where we would like to seek inference for a particular parameter of one selected model, unless such a parameter is common to all models. However, we demonstrated that Efron's estimator is useful in a variety of settings when prediction is the goal. Even for a "soft" procedure such as LASSO or SCAD, it can sometimes have notable improvement over existing procedures, when the tuning parameter of such a procedure is selected by a data-driven method.

We greatly appreciate the opportunity of discussing this stimulating work and congratulate the author for his important contributions to this challenging problem.

REFERENCES

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [1010]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1007]

He, X., and Shi, P. (1996), "Bivariate Tensor-Product b-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181. [1008]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1007]

Wang, H., Li, R., and Tsai, C. L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [1007]

Comment

Dimitris N. Politis

1. WHICH BOOTSTRAP?

Professor Brad Efron, a pioneer in the recasting of modern statistics in its current computer-intensive framework, has given us another important and thought-provoking piece of work. To discuss it, consider the standard additive regression model

$$Y_j = \mu_p(\underline{x}_j) + \varepsilon_j \quad \text{for } j = 1, \dots, n, \tag{1}$$

where $Y_1, \dots, Y_n$ are the data, ε_j are the errors assumed iid $(0, \sigma^2)$, and $\underline{x}_j$ is a length p vector of explanatory (predictor) variables associated with the observation Y_j. The function $\mu_p(\cdot)$ is unknown but assumed to belong to a certain class of functions, which is either finite-dimensional or not. For simplicity, let us focus on the simple case where $\mu_p(\cdot)$ is affine in its arguments, that is, $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j'\underline{\beta}_p$ with $\underline{\beta}_p = (\beta_1, \dots, \beta_p)'$. Also for simplicity assume that the p coordinates of $\underline{x}_j$ are ranked in terms of their importance so that model selection is tantamount to choosing the order p; this is the case with the polynomial regression example of the cholesterol data.

In the above, the regressor $\underline{x}_j$ is most often thought of as deterministic, and $\mu_p(\underline{x}_j)$ has the interpretation of expected value of the response Y_j associated with regressor $\underline{x}_j$. But if the regressors are random, then Efron's setup where the pairs

$$(Y_j, \underline{x}_j) \text{ for } j = 1, \dots, n \text{ are iid} \tag{2}$$

is appropriate. In that case, Equation (1) still applies by defining $\mu_p(\underline{x}_j)$ to be the (theoretical) orthogonal projection of Y_j onto the linear span of the elements of $\underline{x}_j$ (plus a constant), and letting Equation (1) serve as the definition of ε_j, which would then be uncorrelated with $\underline{x}_j$. Of course, under joint normality of $(Y_j, \underline{x}_j)$, the projection $\mu_p(\underline{x}_j)$ would equal the conditional

Dimitris N. Politis is Professor of Mathematics and Adjunct Professor of Economics, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (E-mail: *dpolitis@ucsd.edu*).

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.905788

expectation $E(Y_j|\underline{x}_j)$ since the latter would be affine as a function of $\underline{x}_j$; in this case the ε_j would be normal as well, and independent of $\underline{x}_j$.

In the cholesterol example, it is obvious that even though the (transformed) compliance variable may be normally distributed, when raised to the power of two or higher it will not be. However, the assumption of normality is innocuous if it is just used as a trick to derive the orthogonal projection via the simple formula for Gaussian conditional expectations. Note that it is possible to avoid the assumption of normality but still retain the linearity of $E(Y_j|\underline{x}_j)$. A simple way of doing that is to assume that Equation (1) is true with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j'\underline{\beta}_p$ and ε_j being iid $(0, \sigma^2)$ as before, coupled with the "exogeneity" assumption that $\underline{x}_1, \ldots, \underline{x}_n$ are iid and independent of $\{\varepsilon_1, \ldots, \varepsilon_n\}$.

The two aforementioned viewpoints on the same scatterplot motivate the two most popular bootstrap methods for homoscedastic regression, namely, the *residual bootstrap* and the *pairs bootstrap*. The latter is a straightforward implication of Equation (2). By contrast, the residual bootstrap keeps the $\underline{x}_j$ fixed, and creates pseudo-data using Equation (1), that is, letting

$$Y_j^* = \hat{\beta}_0 + \underline{x}_j'\hat{\underline{\beta}}_p + \varepsilon_j^* \text{ for } j = 1, \ldots, n, \qquad (3)$$

where $\hat{\beta}_0, \hat{\underline{\beta}}_p$ are the least-square (LS) estimators of $\beta_0, \underline{\beta}_p$, and ε_j^* is a random draw from the set of fitted residuals $\{e_1, \ldots, e_n\}$ with $e_j = Y_j - \hat{\beta}_0 - \underline{x}_j'\hat{\underline{\beta}}_p$. Bose and Chatterjee (2002) reviewed and compared several different resampling methods for linear regression including the pairs and residual bootstraps.

Efron uses the pairs bootstrap in the article, and I do not think this is simply a matter of taste. Doing the residual bootstrap presupposes a choice of the order p, that is, model selection. Suppose $\hat{p}$ is a data-based selector of the order p, then the residual bootstrap would generate data from a model with dimension $\hat{p}$, and model selection procedures when applied in the bootstrap world would disproportionally often select the same $\hat{p}$ again. To elaborate, suppose that with this type of data your favorite model selection procedure, say Mallows C_p, would select $\hat{p} = k$ with sampling probability p_k, that is, $100p_k\%$ of such scatterplots would result into $\hat{p} = k$. The sampling probability p_k would not be well captured/replicated when pseudo-scatterplots are generated by residual bootstrap that always uses the order $\hat{p}$ (say $\hat{p} = 3$) that was chosen based on the original data.

The question arises: can we still employ a residual bootstrap in such a case where model selection is also involved? The answer appears to be yes but it may be quite more cumbersome. To start with, one can use the pairs bootstrap (or other considerations) to estimate the aforementioned sampling probability p_k. The important thing here is not to underestimate model order; so one can probably afford to be slightly less parsimonious at the stage of estimating p_k. Then, use a two-step residual bootstrap: first generate the order, say p^*, using the discrete distribution that puts mass p_k on the number k, and then generate a pseudo-scatterplot via a residual bootstrap based on order p^*. The collection of many bootstrap scatterplots generated this way should reflect well the variability associated with model selection. If the regressors are not ranked, that is, the models are not nested, then one may associate sampling probability p_k with candidate model k, and modify the above two-step procedure accordingly.

2. ESTIMATION AND PREDICTION

Efron focuses on the linear combination $\mu_p(\underline{x}) = \beta_0 + \underline{x}'\underline{\beta}_p$ as the parameter of interest. As previously mentioned, $\mu_p(\underline{x})$ has the interpretation of the mean response when the regressor vector takes the value $\underline{x}$. As such, it is a quantity that has precise meaning for *all* models considered; indeed, all models should be able to capture such a quantity regardless of whether individual β-parameters are zeroed out or not. Interestingly, $\mu_p(\underline{x})$ has an additional interpretation: it is the L_2-optimal (linear) *predictor* of the future response Y_{n+1} that is associated with a regression vector $\underline{x}_{n+1}$ that is equal to $\underline{x}$.

Estimation and prediction often go hand-by-hand. It is not a coincidence that popular model selection methods, such as Mallows C_p or cross-validation, rank models in terms of their predictive ability. On the other hand, prediction is typically conducted using an estimated model, which implies a preliminary step of model fitting. Since fitting a model gives the practitioner the ability to predict future responses one can ask if the converse is also true. The answer is yes: if one is able to predict the future response that is associated with *any* regressor value $\underline{x}$, then an implied model fitting is taking place as the curve explaining/predicting Y on the basis of $\underline{x}$ is being constructed.

But how can one predict without a model? The *model-free (MF) prediction principle* of Politis (2013) substitutes the notion of *transformation* in place of a model, and places the emphasis on observable quantities, that is, current and future data, as opposed to unobservable model parameters and estimates thereof. To briefly state it, consider the vector of responses $\underline{Y}_m = (Y_1, \ldots, Y_m)'$, where Y_j is associated with regressor $\underline{x}_j$; the latter can be assumed deterministic for the time being. Thus, $\underline{Y}_n$ contains the already observed responses while $\underline{Y}_{n+1}$ contains $\underline{Y}_n$ plus the future (yet unobserved) response Y_{n+1} associated with regressor value $\underline{x}_{n+1}$.

If the Y_i's were iid, then prediction would be trivial: the L_2-optimal predictor of Y_{n+1} would simply be given by the common mean of the Y_i's, totally disregarding the regressor value $\underline{x}_{n+1}$. Since the Y_i's are not iid, the MF prediction principle amounts to using the structure of the problem—that also uses the regressors—to find an *invertible transformation* H_m that can map the non-iid vector $\underline{Y}_m$ to a vector $\underline{\epsilon}_m = (\epsilon_1, \ldots, \epsilon_m)'$ that has iid components; here m could be taken equal to either n or $n + 1$ as needed. Letting H_m^{-1} denote the inverse transformation, we would then have $\underline{\epsilon}_m = H_m(\underline{Y}_m)$ and $\underline{Y}_m = H_m^{-1}(\underline{\epsilon}_m)$, that is,

$$\underline{Y}_m \overset{H_m}{\longmapsto} \underline{\epsilon}_m \text{ and } \underline{\epsilon}_m \overset{H_m^{-1}}{\longmapsto} \underline{Y}_m. \qquad (4)$$

If the practitioner is successful in implementing the MF procedure, that is, in identifying the transformation H_m to be used, then the prediction problem is reduced to the trivial one of predicting iid variables. To see why, note that Equation (4) with $m = n + 1$ yields $\underline{Y}_{n+1} = H_{n+1}^{-1}(\underline{\epsilon}_{n+1}) = H_{n+1}^{-1}(\underline{\epsilon}_n, \epsilon_{n+1})$. But $\underline{\epsilon}_n$ can be treated as known given the data $\underline{Y}_n$; just use Equation (4) with $m = n$. Since the unobserved Y_{n+1} is just the $(n + 1)$th coordinate of vector $\underline{Y}_{n+1}$, the former can also be expressed as a function of the unobserved ϵ_{n+1}. Finally, note that predicting a function, say $g(\cdot)$, of an iid sequence $\epsilon_1, \ldots, \epsilon_n, \epsilon_{n+1}$ is straightforward because $g(\epsilon_1), \ldots, g(\epsilon_n), g(\epsilon_{n+1})$ is simply another sequence of iid random variables.

Under regularity conditions, such a transformation H_m always exists although it is not unique. The challenge to the skills and expertise of the statistician is to be able to devise and estimate a workable such transformation for the problem at hand; see Politis (2013) for a complete treatment of the regression paradigm. Note, however, that having mapped our data onto the iid variables $\epsilon_1, \ldots, \epsilon_n$, an *MF bootstrap* scheme readily presents itself, namely: (a) generate bootstrap variables $\epsilon_1^*, \ldots, \epsilon_n^*$ by random drawing (without replacement) from the set $\{\epsilon_1, \ldots, \epsilon_n\}$, and (b) generate a pseudo-response vector $\underline{Y}_n^* = \hat{H}_n^{-1}(\underline{\epsilon}_n^*)$, where $\underline{\epsilon}_n^* = (\epsilon_1^*, \ldots, \epsilon_n^*)'$ and $\hat{H}_n^{-1}$ is the estimated (inverse) transformation.

The MF bootstrap can be viewed as an extension of the residual bootstrap to settings where a model is not available. To see why, note that if the additive model (1) is actually available, then the transformation H_n can be readily estimated by first estimating $\mu_p(\cdot)$. For example, constructing the fitted residuals $e_j = Y_j - \hat{\beta}_0 - \underline{x}_j'\hat{\underline{\beta}}_p$ can be viewed as a transformation of the $\underline{Y}_n$ data toward (approximate) iid-ness; recall that the residuals are approximately iid being proxies for the true errors.

However, this is not the only possible transformation; for instance, one can define $\epsilon_j = Y_j - \hat{\beta}_0^{(j)} - \underline{x}_j'\hat{\underline{\beta}}_p^{(j)}$, where $\hat{\beta}_0^{(j)}, \hat{\underline{\beta}}_p^{(j)}$ are the LS estimates obtained from the *delete-one* dataset $\{(Y_t, \underline{x}_t)$ for $t = 1, \ldots, n$ but with $t \neq j\}$. In the above, the ϵ_j are nothing more than the *predictive* residuals that are typically used in cross-validation; see, for example, Geisser (1993) and the references therein. Politis (2013) gave an argument based on the MF prediction principle that favors using the predictive (as opposed to the fitted) residuals for resampling; doing so appears to partially correct the under-coverage of bootstrap prediction intervals noticed early on by Efron (1983) and Stine (1985).

In any case, when model selection is also involved, that is, when the number p of regressors to be used in the transformation H_n is up for debate, the analogy between the residual bootstrap and the MF bootstrap suggests that a similar trick as the one suggested at the end of last section may be helpful. To elaborate, one can use a two-step resampling procedure: (a) generate the model order, say p^*, using some estimated distribution (say p_k), and then generate a pseudo-scatterplot via the MF bootstrap based on an estimated transformation $\hat{H}_n$ that uses p^* regressors.

Nevertheless, there is nothing to stop the MF practitioner from using the pairs bootstrap in this setting; this could be done just to obtain an estimate of the sampling distribution p_k needed above, or to carry out the complete task of capturing the variability of an estimator that includes the model selection step. But using the pairs bootstrap is associated with an assumption that the regressors $\underline{x}_j$ are random, and furthermore that the pairs $(Y_1, \underline{x}_1), (Y_2, \underline{x}_2), \ldots$ are iid as in Equation (2). In the case of random regressors, the MF prediction principle can be simply restated by conditioning on the regressor values. In other words, the transformation H_m of Equation (4) would be constructed conditionally on the values $\{\underline{x}_1, \ldots, \underline{x}_m\}$, and the goal of the MF practitioner is to render the transformed variables $\epsilon_1, \ldots, \epsilon_m$ as close to iid as possible conditionally on $\{\underline{x}_1, \ldots, \underline{x}_m\}$.

3. MODELS VERSUS TRANSFORMATIONS: A RECONCILIATION

The MF approach can form the basis for a complete statistical inference that includes point estimators and predictors in addition to confidence and prediction intervals without assuming an additive model such as (1); see Politis (2013, 2014) for details. Interestingly, however, when an additive model is known to hold true, there is no discrepancy if one adheres to the MF approach, that is, tries to find a transformation toward "iid-ness."

To see why, let us assume Equation (1) with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}_j'\underline{\beta}_p$. The essence of this model—as far as MF prediction is concerned—is that the variables $\epsilon_j \equiv Y_j - \underline{x}_j'\underline{\beta}_p$ are iid albeit with (possibly) nonzero mean β_0. Thus, a candidate transformation to "iid-ness" may be constructed by letting $r_j = Y_j - \underline{x}_j'\hat{\underline{\beta}}_p$, where $\hat{\underline{\beta}}_p$ is a candidate vector. The MF principle now mandates choosing $\hat{\underline{\beta}}_p$ with the objective of having the r_j's become as close to iid as possible. However, under the stated regression model, the r_j's would be iid if only their first moment was properly adjusted.

To elaborate, a homoscedastic regression model such as (1) implies that all central moments of order two or higher are constant; the only non-iid feature is in the first moment. So, in this case, the MF principle suggests choosing $\hat{\underline{\beta}}_p$ in such a way as to make $r_1, \ldots, r_n$ have (approximately) the *same* first moment. Noting that the first moment—if it is common—would be naturally approximated by the empirical value $\hat{r} = n^{-1}\sum_{i=1}^n r_i$, we can use a *subsampling* construction to make this happen.

To fix ideas, assume for simplicity that $p = 1$, and that the univariate design points $x_1, \ldots, x_n$ are sorted in ascending order. Then compute the overlapping block means

$$\bar{r}_{k,b} = b^{-1}\sum_{j=k}^{k+b-1} r_j \quad \text{for} \quad k = 1, \ldots, q, \tag{5}$$

where b is the block size, and $q = n - b + 1$ is the number of available blocks.

Note that $\bar{r}_{k,b}$ is an estimate of the first moment of the r_i's found in the kth block. To achieve the target requirement that all $r_1, \ldots, r_n$ have first moment that is the same (and thus approximately equal to $\hat{r}$), the MF practitioner may choose $\hat{\beta}_1$ that minimizes

$$\mathrm{LS}(b) = \sum_{k=1}^{q}(\bar{r}_{k,b} - \hat{r})^2 \quad \text{or} \quad L1(b) = \sum_{k=1}^{q}|\bar{r}_{k,b} - \hat{r}| \tag{6}$$

according to whether an L_2 or L_1 loss criterion is preferred.

Instead of $\hat{r}$, we could equally use the mean of means, that is, $\bar{\bar{r}} = q^{-1}\sum_{k=1}^{q}\bar{r}_{k,b}$ as the centering value in Equation (6). If $b = 1$, then $\hat{r} = \bar{\bar{r}}$; if $b > 1$, then $\hat{r} = \bar{\bar{r}} + O_P(b/n)$ so the difference is negligible provided b is small as compared to n. Recall that in the typical application of subsampling for variance or distribution estimation, it is suggested to take the block size b to be large (but still of smaller order than n); this is for the purpose of making the subsample statistics $\bar{r}_{k,b}$ have asymptotically the same distribution as the statistic $\hat{r}$ computed from the full sample, see, for example, Politis, Romano and Wolf (1999).

Nevertheless, it is not crucial in our current setting that each of the $\bar{r}_{k,b}$ has asymptotically the same distribution as $\hat{r}$. What is important is that all the $\bar{r}_{k,b}$ (for $k = 1, \ldots, q$) have approximately the same distribution whatever that may be. Therefore, it is not necessary in Equation (6) to use a large value for b. Even the value $b = 1$ is acceptable, in which case we have

$$\frac{d}{d\hat{\beta}_1}\mathrm{LS}(1) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

In other words, the MF fitting procedure (6) with L_2 loss and $b = 1$ is reassuringly $identical$ to the usual LS estimator! Note that the r_i's serve as proxies for the unobservable ε_i's, which have expected value β_0 under model (1). Hence, β_0 is naturally estimated by the sample mean of the r_i's, that is,

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 x_i) = \bar{Y} - \hat{\beta}_1\bar{x},$$

which is again the LS estimator.

Minimizing LS(b) with $b > 1$ gives a more robust way of doing LS in which the effect of potential outliers is diminished by the local averaging of b neighboring values; details are omitted due to lack of space. Similarly to the above, minimizing $L1(1)$ is equivalent to L_1 regression, whereas minimizing $L1(b)$ with $b > 1$ gives additional robustness.

Finally, let us revisit the general case of model (1) with $\mu_p(\underline{x}_j) = \beta_0 + \underline{x}'_j\underline{\beta}_p$. When $p > 1$, the regressors $\underline{x}_j$ cannot be sorted in ascending order. One could instead use a local-averaging or nearest-neighbor technique to compute the sample means. But no such trick is needed in the most interesting case of $b = 1$ since the quantities LS(1) and $L1(1)$ are unequivocally defined as

$$\mathrm{LS}(1) = \sum_{k=1}^{n}(r_k - \hat{r})^2 \text{ and } L1(1) = \sum_{k=1}^{n}|r_k - \hat{r}|. \tag{7}$$

It is now easy to see that the MF practitioner that chooses the β's to minimize LS(1) or $L1(1)$ is effectively doing LS or L_1 regression, respectively. Hence, when an additive model is available, there is no discrepancy between the MF approach and traditional model fitting. Nevertheless, the MF approach can still lend some insights such as the aforementioned use of predictive residuals in connection with the model-based residual bootstrap.

REFERENCES

Bose, A., and Chatterjee, S. (2002), "Comparison of Bootstrap and Jackknife Variance Estimators in Linear Regression: Second Order Results," *Statistica Sinica*, 12, 575–598. [1011]

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331. [1012]

Geisser, S. (1993), *Predictive Inference: An Introduction*, New York: Chapman and Hall. [1012]

Politis, D. N. (2013), "Model-Free Model-Fitting and Predictive Distributions" (with discussion), *Test*, 22, 183–250. [1011,1012]

——— (2014), "Bootstrap Confidence Intervals in Nonparametric Regression Without an Additive Model," in *Topics in Nonparametric Statistics: Proceedings of the First Conference of the International Society for NonParametric Statistics*, eds. M. G. Akritas, S. N. Lahiri, and D. N. Politis, New York: Springer. [1012]

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, New York: Springer. [1012]

Stine, R. A. (1985), "Bootstrap Prediction Intervals for Regression," *Journal of the American Statistical Association*, 80, 1026–1031. [1012]

Comment

Shuva GUPTA and S. N. LAHIRI

1. INTRODUCTION

This is an interesting and stimulating article by Professor Bradley Efron on the important topic of accuracy of bootstrap estimation after model selection. While the bootstrap is routinely used in many problems where model selection forms a part of the exploratory data analysis, the effect of model selection on subsequent inference is often conveniently ignored. The current article clearly points out the perils of this naive approach with both parametric and nonparametric bootstrap. The naive (or standard) bootstrap confidence intervals (CIs) are unstable as the centers may fluctuate erratically based on the ordinary least-square (OLS) estimators under the selected models. Stability can be ensured by bootstrap model averaging (or the bagging) as suggested in the article, as was also noted by Bühlmann and Yu (2002). One of the major contributions of the article is an elegant derivation of the delta-method estimate of the standard error of the bagging estimator, which will be useful in other applications as well.

Although the main article considers both the parametric and the nonparametric bootstrap methods, our discussion here will be restricted to the nonparametric bootstrap only. Specifically, we shall suppose that $\{y_i, c_i)\}$ are independent and identically distributed (iid) random vectors satisfying

$$y_i = \beta_0 + \beta_1 c_i + \cdots + \beta_p c_i^p + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

Shuva Gupta (E-mail: *sgupta22@ncsu.edu*) is Post-Doctoral Fellow, and S. N. Lahiri (E-mail: *snlahiri@ncsu.edu*) is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. Research partially supported by grants NSF DMS 1310068 and NSA H98230-11-1-0130.

© 2014 American Statistical Association
Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.905789

where $\beta_i \in \mathbb{R}$ with $\beta_{p_0} \neq 0$ for some $1 \leq p_0 \leq p$ and $\beta_i = 0$ for all $i = p_0 + 1, \ldots, p$, and where ϵ_i are zero mean random variables with $E\epsilon_1^2 = \sigma^2 \in (0, \infty)$. The resampling is done with replacement from the observed pairs of variables $\{(y_i, c_i) : i = 1, \ldots, n\}$ to generate the nonparametric bootstrap resamples. As highlighted in the main article, a naive application of the bootstrap with the model selection step repeated on each resample often does not produce a desirable result, particularly when the model selection criterion itself is not very accurate. In this case, the model selection step selects different models, which in turn add to the bias and variability of the resulting bootstrap estimates of the target parameter. In particular, this naive approach fails to reproduce the sampling distribution under the true model. In the following, we describe two alternative methods of constructing CIs for the parameters when true model is not prespecified and the user must determine the true model as well as carry out inference on functions of the true parameters using the data $\{(y_i, c_i) : i = 1, \ldots, n\}$.

The first approach is based on the adaptive Lasso (ALASSO) method of Zou (2006) that is known to enjoy the oracle property (see Fan and Li 2001). Here, the ALASSO method performs variable selection and parameter estimation simultaneously and therefore, also identifies the true model with high probability. As a result, the standard bootstrap method can be applied with the ALASSO to produce CIs for parameters like $\mu_i = E(y_i|c_i)$, a linear function of $(\beta_0, \beta_1, \ldots, \beta_{p_0})$. Here, the bootstrap does not suffer from the erratic discontinuities of selection-based estimators. The variable selection consistency essentially guarantees the stability over different resamples. See Section 2.1 for more details.

The second, which we will call the maximum frequency bootstrap (MFB) approach, is qualitatively different and it is designed to deal with the variability associated with the model selection criterion itself. The key idea here is that although the model selection criterion may not identify the true model with high probability, it may still be able to provide important clues to the correct model when applied to the resamples. Thus, applying the model selection step to the resamples, first we attempt to identify the true model and then only use the subset of the resamples corresponding to the selected model. This way the sampling variability of the model selection procedure and the erratic discontinuity of the selection-based estimators are significantly reduced. We describe the details of the construction of MFB CIs in Section 2.2. Results from the simulation study show that the performance of the MFB method is very good.

The rest of the discussion is organized as follows. In Section 2, we describe in details the two approaches to constructing CIs for parameters of interest when model selection is involved. In Section 3, we present the results from a small simulation study illustrating finite sample performance of the MFB method.

2. TWO ALTERNATIVE CIs

2.1 CIs Based on the ALASSO

The ALASSO method of Zou (2006) estimates the regression parameters $\beta = (\beta_0, \ldots, \beta_p)$ in model (1) using a preliminary estimator $\tilde{\beta}_n$, such as the OLS estimator of β, and a weighted ℓ_1-penalty. Specifically, the ALASSO estimator of β is defined as the minimizer of the penalized least-square criterion function,

$$\hat{\beta}_n = \text{argmin}_{u \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i' u)^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\tilde{\beta}_{j,n}|^{\gamma}},$$

where $\lambda_n > 0$ is a regularization parameter, $\gamma > 0$, $x_i = (1, c_i, \ldots, c_i^p)'$ and where $\tilde{\beta}_{j,n}$ and u_j, respectively, denote the jth component of $\tilde{\beta}_n$ and u, respectively. Zou (2006) showed that under mild regularity conditions, the ALASSO selects the exact set of relevant variables (corresponding to $\beta_j \neq 0$ in (1)) with probability tending to one. The residuals from the ALASSO fit can then be used for bootstrap resampling and the common bootstrap CIs can be constructed from these resamples. Chatterjee and Lahiri (2013) showed that under some suitable regularity conditions, the bootstrap-t CIs for linear combinations of the regression parameter β are second-order accurate. They also present numerical results illustrating good finite sample properties of different bootstrap CIs. As a result, the bootstrap based on the ALASSO (as well as other penalize regression methods having the oracle property) are viable methods for constructing reasonably accurate CIs in regression models where the true model is not prespecified.

2.2 Maximum Frequency Bootstrap-*t* CIs

We now describe a different approach to bootstrapping in post-model selection inference that may be applied with many standard model selection methods. Given the data $\mathcal{D}_n \equiv \{(y_i, c_i) : i = 1, \ldots, n\}$, generate B bootstrap replicates $\mathcal{D}_n^{*b}$, $b = 1, \ldots, B$. For each of the bootstrap resamples, apply the given model selection criterion to select one of the p models. For $j = 1, \ldots, p$, let $\hat{f}_j$ denote the proportion of replicates (out of B) where model j was selected. Next, let

$$j_0 = \text{argmax}_j \hat{f}_j$$

be the model that has the maximum frequency of getting selected among the B replicates. Also, let $\mathcal{B}_0$ denote the collection

Table 1. Selection frequencies and standard deviations of the 10 models based on the CP, the AIC, and the BIC based on 500 simulation runs and 1000 bootstrap replicates

	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6	Mod 7	Mod 8	Mod 9	Mod 10
AIC (freq)	0.000	0.004	0.329	0.103	0.092	0.086	0.080	0.082	0.102	0.125
C_p (freq)	0.000	0.000	0.373	0.111	0.095	0.085	0.077	0.077	0.087	0.097
BIC (freq)	0.000	0.000	0.849	0.077	0.034	0.020	0.009	0.005	0.004	0.002
AIC (sd)	0.000	0.000	0.164	0.083	0.083	0.077	0.059	0.058	0.089	0.085
C_p (sd)	0.000	0.000	0.180	0.088	0.085	0.081	0.061	0.060	0.081	0.0740
BIC (sd)	0.000	0.000	0.140	0.085	0.055	0.047	0.014	0.008	0.011	0.003

Table 2. Coverage accuracy and average lengths of MFB CIs for μ_1 for the CP, the AIC, and the BIC methods based on 500 simulation runs and 1000 bootstrap replicates. The nominal confidence level is 95

	Length of CIs	Coverage probability
AIC	3.140	0.927
C_p	3.161	0.960
BIC	3.182	0.980

of bootstrap replicates (among B many) that resulted in selection of the model j_0. Then, the MFB method makes use of the subcollection

$$\{\mathcal{D}_n^{*b} : b \in \mathcal{B}_0\}$$

of resamples to carry out bootstrap-based inference. For example, bootstrap CIs for linear combinations of the regression parameter vector can be obtained by using the bootstrap-*t* method applied only to the resamples $\{\mathcal{D}_n^{*b} : b \in \mathcal{B}_0\}$. Since all replicates in this collection correspond to a single model, the extra variability that results from the model selection step in different resamples is eliminated. In fact, this MFB approach was used for constructing percentile-*t* CIs for the parameter μ_1 in Section 3. Although the respective model selection methods have considerable variability in selecting the true model among B resamples, the empirical coverage accuracy of the MFB approach reported therein appears reasonable for each of the three model selection methods. Theoretical properties of the MFB method is currently under investigation.

3. NUMERICAL RESULTS

Here, we report results from a small simulation study on the MFB method. We consider model (1) with $p = 10$ and $p_0 = 3$ (a cubic model), where $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.4$, $\beta_3 = 5.0$, and $\beta_i = 0$ for all $i = 4, \ldots, 10$. We generated the variables (c_i, ϵ_i) as iid bivariate normal vectors with zero mean vector and identity covariance matrix. The sample size considered was $n = 200$. The MFB method was used to construct bootstrap CIs for the parameter $\mu_1 = E(y_1|c_1)$ where the model selection was performed with the CP, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) methods. The results from the model selection step applied to the bootstrap resamples are summarized in Table 1. The first three rows of the table give the frequencies of the different models, which were selected by each of the three methods over 600 simulation runs. The last three rows give the associated standard deviations. It is evident from the table that except for the BIC, which is known to be consistent for model selection, the other two methods selected the true model with low empirical probability. As a result, the use of either of these model methods in the naive approach would produce very distorted results. However, by using the MFB approach, even in such situations, we are able to identify the true model. The empirical coverage accuracy and the average lengths of a nominal 95% CI for μ_1 are reported in Table 2. The coverage is evidently very good irrespective of the model selection performance of the three model selection methods.

REFERENCES

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [1013]

Chatterjee, A., and Lahiri, S. N. (2013), "Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *The Annals of Statistics*, 41, 1055–1692. [1014]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1014]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1014]

Comment

Andrew Gelman and Aki Vehtari

1. ACCOUNTING FOR MODEL SELECTION IN STATISTICAL INFERENCE

How can one proceed with predictive inference and assessment of model accuracy if we have selected a single model from some collection of models? Selecting a single model instead of model averaging can be useful as it makes the model easier to explain, and in some cases that single model gives similar predictions as the model averaging.

The selection process, however, causes overfitting and biased estimates of prediction error; thus much work has gone into estimating predictive accuracy given available data (e.g., Gelman, Hwang, and Vehtari 2013). In Efron's article, bagging is used to average over different models, and the main contribution is providing a useful new formula estimating the accuracy of bagging in this situation.

It makes sense that bagging should work for the smooth unstable ("jumpy") estimates in the examples shown. Full Bayesian inference should also be able to handle these problems, but it can be useful to have different approaches based on different principles.

Andrew Gelman is Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *gelman@stat.columbia.edu*). Aki Vehtari is Adjunct Professor, Department of Biomedical Engineering and Computational Science, Aalto University, Espoo, Finland (E-mail: *aki.vehtari@aalto.fi*).

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.906153

One of the appeals of the bootstrap is its generality (as, in a completely different way, with Bayes; see Gelman 2011). Any estimate can be bootstrapped; all that is needed are an estimate and a sampling distribution. The very generality of the boostrap creates both opportunity and peril, allowing researchers to solve otherwise intractable problems but also sometimes leading to an answer with an inappropriately high level of certainty.

We demonstrate with two examples from our own research: one problem where bootstrap smoothing was effective and led us to an improved method, and another case where bootstrap smoothing would not solve the underlying problem. Our point in these examples is not to disparage bootstrapping but rather to gain insight into where it will be more or less effective as a smoothing tool.

2. AN EXAMPLE WHERE BOOTSTRAP SMOOTHING WORKS WELL

Bayesian posterior distributions are commonly summarized using Monte Carlo simulations, and inferences for scalar parameters or quantities of interest can be summarized using 50% or 95% intervals. A $1-\alpha$ interval for a continuous quantity is typically constructed either as a central probability interval (with probability $\alpha/2$ in each direction) or a highest posterior density interval (which, if the marginal distribution is unimodal, is the shortest interval containing $1-\alpha$ probability). These intervals can in turn be computed using posterior simulations, either using order statistics (e.g., the lower and upper bounds of a 95% central interval can be set to the 25th and 976th order statistics from 1000 simulations) or the empirical shortest interval (e.g., the shortest interval containing 950 of the 1000 posterior draws).

For large models or large datasets, posterior simulation can be costly, the number of effective simulation draws can be small, and the empirical central or shortest posterior intervals can have a high Monte Carlo error, especially for wide intervals such as 95% that go into the tails and thus sparse regions of the simulations. We have had success using the bootstrap, in combination with analytical methods, to smooth the procedure and produce posterior intervals that have much lower mean squared error compared with the direct empirical approaches (Liu, Gelman, and Zheng 2013).

3. AN EXAMPLE WHERE BOOTSTRAP SMOOTHING IS UNHELPFUL

When there is separation in logistic regression, the maximum likelihood estimate of the coefficients diverges to infinity. Gelman et al. (2008) illustrated with an example of a poll from the 1964 U.S. presidential election campaign, in which none of the black respondents in the sample supported the Republican candidate, Barry Goldwater. As a result, when presidential preference was modeled using a logistic regression including several demographic predictors, the maximum likelihood for the coefficient of "black" was $-\infty$. The posterior distribution for this coefficient, assuming the usual default uniform prior density, had all its mass at $-\infty$ as well. In our article, we recommended a posterior mode (equivalently, penalized likelihood) solution based on a weakly informative Cauchy (0, 2.5) prior distribution that pulls the coefficient toward zero. Other, similar, approaches to regularization have appeared over the years. We justified our particular solution based on an argument about the reasonableness of the prior distribution and through a cross-validation experiment. In other settings, regularized estimates have been given frequentist justifications based on coverage of posterior intervals (see, e.g., the arguments given by Agresti and Coull 1998, in support of the binomial interval based on the estimate $\hat{p} = \frac{y+2}{n+4}$).

Bootstrap smoothing does not solve problems of separation. If zero black respondents in the sample supported Barry Goldwater, then zero black respondents in any bootstrap sample will support Goldwater as well. Indeed, bootstrapping can exacerbate separation by turning near-separation into complete separation for some samples. For example, consider a survey in which only one or two of the black respondents support the Republican candidate. The resulting logistic regression estimate will be noisy but it will be finite. But, in bootstrapping, some of the resampled data will happen to contain zero black Republicans, hence complete separation, hence infinite parameter estimates. If the bootstrapped estimates are regularized, however, there is no problem.

The message from this example is that, perhaps paradoxically, bootstrap smoothing can be more effective when applied to estimates that have already been smoothed or regularized.

REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate is Better Than Exact for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126. [1016]

Gelman, A. (2011), "The Pervasive Twoishness of Statistics; in Particular, the Sampling Distribution and the Likelihood are Two Different Models, and That is a Good Thing," Statistical Modeling, Causal Inference, and Social Science Blog, 20 June. Available at *http://andrewgelman.com/2011/06/20/the_sampling_di_1/*. [1016]

Gelman, A., Hwang, J., and Vehtari, A. (2013), "Understanding Predictive Information Criteria for Bayesian Models," *Statistics and Computing*, doi: 10.1007/sl 1222-013-9416-2. [1015]

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *Annals of Applied Statistics*, 2, 1360–1383. [1016]

Liu, Y., Gelman, A., and Zheng, T. (2013), "Simulation-Efficient Shortest Probability Intervals," Technical report, Department of Statistics, Columbia University. [1016]

Comment

Nils Lid Hjort

Reaching accurate inference statements after model selection continues to be challenging, from both practical and theoretical perspectives. That the issue needs the attention of statisticians is clear, in a world with papers titled "I just ran two million regressions" and similar (Sala-i-Martin 1997). Also, the widespread use of selecting one model out of a great many candidates and then somehow ignoring the initial model selection step, when reporting one's findings, is troublesome; see what Breiman (1992) called "the quiet scandal of statistics." Efron's article is a welcome and significant contribution to this area, and I appreciate being given the opportunity to comment. There are many aspects and details I would wish to comment on and pursue further, but due to considerations of space I shall focus on (1) distribution theory for postselection and model average estimators, (2) relevant applications of this theory, related to both confidence limits and to specially designed model selection schemes, and (3) similar distribution theory for bagging, leading to clear performance comparisons for bootstrap smoothing and classes of similar methods.

1. DISTRIBUTION THEORY FOR POSTSELECTION AND MODEL AVERAGE ESTIMATORS

Suppose μ is a parameter one wishes to estimate, via a list of candidate models for the data at hand, say models $M_1, \ldots, M_q$. Thus there is an estimate $\widehat{\mu}_j$ based on model M_j, say reached via maximum likelihood, and the final estimate is $\widehat{\mu} = \sum_{j=1}^q \widehat{\mu}_j I\{M_j \text{ wins}\}$. Importantly, the parameter μ ought to have clear statistical interpretation across all candidate models, perhaps with a separate formula for each model, in terms of that model's parameters. I shall first briefly summarize some precise large-sample theory regarding the behavior of $\widehat{\mu}$. It will be convenient to do so inside the regression framework of Efron's Section 2, though results presented below generalize to almost arbitrary parametric models.

Suppose, then, that $y_i = m(x_i) + \varepsilon_i$, where different nested models are under consideration for the mean function $m(x)$, from a "narrow model" with $m(x_i) = x_i^{\rm t}\beta$ to a "wide model" with $m(x_i) = x_i^{\rm t}\beta + z_i^{\rm t}\gamma$, where x_i and z_i are covariate vectors of dimensions say p and q. Though one would often choose to work with all 2^q subsets formed by inclusion and exclusion of γ_j parameters, I restrict attention here to the nested situation with $q+1$ candidate models $M_0, M_1, \ldots, M_q$, corresponding to including, respectively, none, the first, the two first, etc., of $\gamma_1, \ldots, \gamma_q$. This fits Efron's setup of sec. 2, where $x_i = (1, c_i)^{\rm t}$ is of dimension 2 and $z_i = ((c_i - \bar{c})^2, \ldots, (c_i - \bar{c})^6)^{\rm t}$ is of dimension 5. Also, the ε_i above are modelled as iid $\mathrm{N}(0, \sigma^2)$. Incidentally, the σ is not merely a Greek letter but a parameter with a slightly different interpretation for the different models, becoming smaller with inclusion of more terms in the regression function, so it is slightly misleading to use the same σ across models. From a pedantic viewpoint we might thus operate with and estimate σ_j for candidate model M_j, though this is seen to change the analysis in only a minor fashion for the cholesterol dataset.

If now $\mu = \mu(\beta, \sigma, \gamma)$ is a focus parameter to be estimated, there are $q+1$ separate estimates $\widehat{\mu}_0, \widehat{\mu}_1, \ldots, \widehat{\mu}_q$, where $\widehat{\mu}_j$ stems from using maximum likelihood with model M_j, which has $\gamma_1, \ldots, \gamma_j$ on board but sets the remaining γ_k to zero. Efron's illustration is of this form, with $\mu = x_0^{\rm t}\beta + z_0^{\rm t}\gamma$, for some given c_0 in $x_0 = (1, c_0)^{\rm t}$ and $z_0 = ((c_0 - \bar{c})^2, \ldots, (c_0 - \bar{c})^6)^{\rm t}$. The final estimate is

$$\widehat{\mu} = \sum_{j=0}^{q} \widehat{\mu}_j I\{M_j \text{ is selected}\} \qquad (1)$$

using Mallows's C_p to determine the winner. Efron does not directly touch on the distribution of this final estimator, but a theorem may be put up, as follows. The framework is of the local large-sample type, where $\gamma_j = \delta_j/\sqrt{n}$, and involves $D_n = \sqrt{n}\widehat{\gamma}$, with $\widehat{\gamma}$ the maximum likelihood estimator in the wide model. The point of this framework is that it makes squared biases and variances exchangeable currencies, both of order $O(1/n)$, leading, as we shall see, to clear large-sample theorems for a wide class of postselection estimators and also model averaging estimators, bootstrap smoothed estimators, etc.

Let $J_n = -n^{-1}\partial^2 \ell_n(\widehat{\alpha})/\partial\alpha\partial\alpha^{\rm t}$ be the normalized Hessian matrix computed at the maximum likelihood position $\widehat{\alpha}$, with $\alpha = (\beta, \sigma, \gamma)$ the full parameter in the wide model. We need the blocks of J_n and its inverse,

$$J_n = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix} \quad \text{and} \quad J_n^{-1} = \begin{pmatrix} J^{n,00} & J^{n,01} \\ J^{n,10} & J^{n,11} \end{pmatrix},$$

and have special use for $Q_n = J^{n,11}$, of size $q \times q$. Under mild conditions, J_n converges in probability to a well-defined positive definite matrix J, hence, also with Q_n converging to the consequent $Q = J^{11}$, as sample size n increases. To complete the description of the limit distributions of estimators of type (1) we shall also need certain matrices $G_0, G_1, \ldots, G_q$, each of size $q \times q$, defined as follows. First, let π_j be the $j \times q$ projection matrix of zeros and ones that maps $v = (v_1, \ldots, v_q)^{\rm t}$ to $\pi_j v = (v_1, \ldots, v_j)^{\rm t}$, and let

$$G_j = \pi_j^{\rm t} Q_j \pi_j Q^{-1}, \quad \text{where } Q_j = (\pi_j Q^{-1} \pi_j^{\rm t})^{-1}, \qquad (2)$$

Nils Lid Hjort, Department of Mathematics University of Oslo, Oslo, Norway (E-mail: *nils@math.uio.no.*). The author gratefully acknowledges support from the Norwegian Research Council in connection with the project "FocuStat: Focus Driven Statistical Inference With Complex Data."

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.923315

with $G_0 = 0$. Finally define

$$\widehat{\tau}_0^2 = \frac{\partial\mu}{\partial\theta}(\widehat{\alpha})^{\rm t} J_{n,00}^{-1} \frac{\partial\mu}{\partial\theta}(\widehat{\alpha}) \quad \text{and} \quad \widehat{\omega} = J_{n,10} J_{n,00}^{-1} \frac{\partial\mu}{\partial\theta}(\widehat{\alpha}) - \frac{\partial\mu}{\partial\gamma}(\widehat{\gamma}), \tag{3}$$

with partial derivatives at the maximum likelihood position, and with $\theta = (\beta, \sigma)$ the "protected parameters" that are included in all candidate models. Under standard conditions, $\widehat{\tau}_0 \to_{\rm pr} \tau_0$ and $\widehat{\omega} \to_{\rm pr} \omega$, say, depending on the focus parameter μ under attention.

First, with $\mu_{\rm true}$ the value of μ computed under $(\beta, \delta/\sqrt{n}, \sigma)$,

$$\sqrt{n}(\widehat{\mu}_j - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - G_j D), \tag{4}$$

in which $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}_q(\delta, Q)$, with these two being independent. This describes the limit distributions for each separate $\widehat{\mu}_j$ in a joint framework. Second, for the wide class of model averaging estimators of the form

$$\widehat{\mu} = \sum_{j=0}^{q} w(j \mid D_n)\widehat{\mu}_j, \tag{5}$$

where $w(j \mid d)$ are weight functions summing to one, we have

$$\sqrt{n}(\widehat{\mu} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\}, \tag{6}$$

in which $\widehat{\delta}(D) = \sum_{j=0}^{q} w(j \mid D) G_j D$. For a proof and wider discussion, see Hjort and Claeskens (2003) and Claeskens and Hjort (2008, chaps. 6 and 7). This result in particular encompasses the case of postselection estimators, as long as the model selection scheme can be given in terms of D_n, or in a form large-sample equivalent to such functions of D_n. The Mallows scheme of C_p used by Efron is of this type, as it can be seen to be first-order equivalent to the AIC method, and with $w(j \mid D_n)$ equal to 1 for the maximizer of

$$\mathrm{AIC}(j, D_n) = D_n^{\rm t} Q^{-1} \pi_j^{\rm t} Q_j \pi_j Q^{-1} D_n - 2j$$

and 0 for the others.

2. APPLYING THE THEORY

Result (6) gives a precise description of the distribution of postselection estimators and also of more general model average estimators, as when $w(j \mid D_n)$ is taken as a function of the C_p or AIC score, for example, giving more weight to the best models and less weight to those doing badly, without being restricted to trusting only the winner. The limits in question are nonlinear mixtures of normals (unless the weights $w(j \mid D_n)$ are fixed, i.e., do not depend on data), and often quite nonnormal. They are also not centered around zero, reflecting modeling bias. Result (6) may also be used as a starting point for constructing confidence intervals and more generally full confidence distributions; see Schweder and Hjort (2014). The limiting distribution of (6) may be simulated for each position δ in the parameter space for $\gamma = \delta/\sqrt{n}$, for example, at the estimated position $\widehat{\delta} = D_n = \sqrt{n}\widehat{\gamma}$. Thus we have machinery for calculating confidence limits for each such position $\delta/\sqrt{n}$, but the issue is more complicated as δ cannot be consistently estimated inside this local large-sample framework. A conservative two-stage method of the Bonferroni kind is outlined in Claeskens and Hjort (2008, chap. 7), but other approximations may be developed, such as those flowing from Efron's work, and then evaluated for accuracy using (6).

Another use of the machinery above is to use specially designed model selectors for different focus parameters. The limiting mean squared error when using model M_j is

$$\mathrm{MSE}_j = \tau_0^2 + \omega^{\rm t} G_j Q G_j^{\rm t} \omega + \{\omega^{\rm t}(I - G_j)\delta\}^2,$$

and estimating this quantity unbiasedly in the limit experiment leads to formulas containing quantities which then may be estimated from data; this essentially yields the focused information criterion (FIC) of Claeskens and Hjort (2003, 2008). The point is then that different optimal models are selected for different purposes, unlike for the C_p, the AIC, the BIC, etc. When running the FIC through the 164 tasks of estimating each person's $\mu = {\rm E}\, y_i = m(x_i)$, with appropriate calculations and computations of the required matrices $G_0, G_1, \ldots, G_5$, along with say $\tau_{0,i}$ and ω_i for each $m(x_i)$, one finds that the simplest model M_0 (linear trend) wins in 50% of the cases, whereas models M_1 (quadratic), M_2 (cubic), M_3 (quartic) win in 8%, 37%, 5% of the cases, respectively, with the two biggest models of polynomial orders 5 and 6 never being selected. In contrast, the C_p and AIC methods deliver one and only one model, to be used for all purposes.

We may also illustrate the nonnormality of the actual distributions involved, by simulating from the precise limit distribution (6), for a given dataset, a focus parameter of interest, and at, for example, the estimated position in the parameter space. Figure 1 displays such densities for $\widehat{\mu}_{\rm AIC} - \mu$ and $\widehat{\mu}_{\rm FIC} - \mu$, for the cholesterol dataset, with focus parameter $\mu = m(x_1)$, following Efron, the mean for the individual with $c = c_1$, the smallest value. Here $\widehat{\mu}_{\rm AIC}$ and $\widehat{\mu}_{\rm FIC}$ are the postselection estimators using, respectively, the AIC and the FIC methods. The densities displayed are those of the exact limit distributions given by (6), with τ_0, δ, Q, and the G_j and $w(j \mid D)$ estimated from the data, and then divided by $\sqrt{n}$ to mimic the random errors $\widehat{\mu}_{\rm AIC} - \mu$ and $\widehat{\mu}_{\rm FIC} - \mu$. The nonnormality is striking for both, hence also pointing to the difficulty of setting good confidence limits. The mean squared errors following AIC and FIC are actually very similar, for this case of the left-most position $c = c_1$, though the error distributions are markedly different. For the clear majority of other positions $c = c_i$, however, the post-FIC error distribution is more tightly concentrated around zero than the post-AIC error distribution.

3. THEORY FOR BAGGING

Efron has aptly argued that passing from a postmodel estimator $\widehat{\mu}$ to its bagged cousin, the bootstrap smoothed $\widetilde{\mu}$, is indeed a smoother function of data and, hence, lending itself more easily to further analysis, including variances and approximate confidence limits. Developing these methods, and demonstrating that they work well, does not touch the perhaps deeper issue, though, which is whether and then in which ways bagging improves on the behavior and precision of the original unbagged estimator. I shall now exhibit a clear large-sample result for such bootstrap smoothed estimators, exploiting the same local large-sample framework as above. The point is partly to make it possible to compare performances, for example, via risk functions, but also to have a framework where accuracy of approximations to confidence distributions may be assessed.

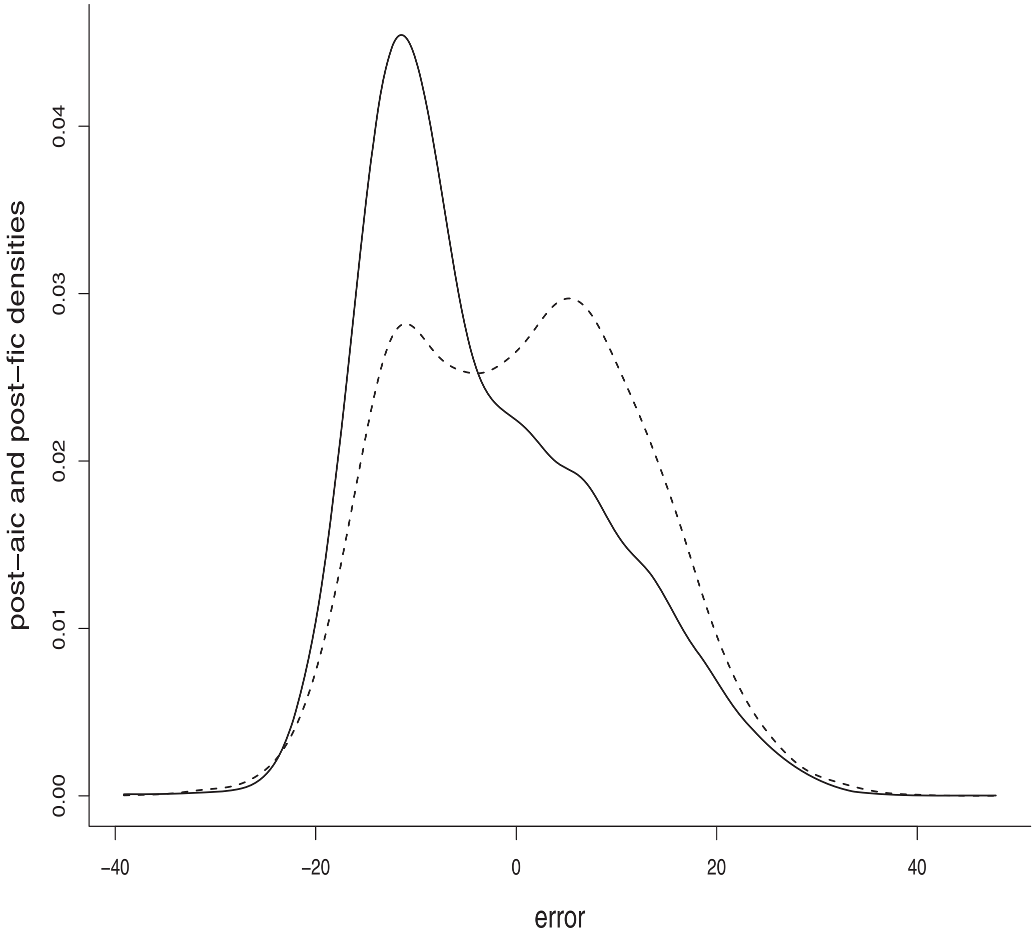

Figure 1. Densities of the limit distributions for $\widehat{\mu}_{\rm AIC} - \mu$ (dotted line) and $\widehat{\mu}_{\rm FIC} - \mu$ (full line), computed at the maximum likelihood estimated position in the parameter space, for the cholesterol dataset, with $\mu = m(x_1)$, the mean at position $c = c_1$.

Though the theory briefly summarized below generalizes suitably to collections of general parametric models, I shall again be content to use the simpler framework of nested regression models, as above. The narrow model has parameter θ of length p, whereas the wider model has an additional parameter γ of length q. Submodel M_j uses parameters $(\theta, \gamma_1, \ldots, \gamma_j)$, for $j = 0, 1, \ldots, q$, with the narrow model corresponding to these extra parameters γ_j being zero. Consider then a general postselection or model averaging estimator of the form (5), again with $D_n = \sqrt{n}\widehat{\gamma}$ and the weights $w(j \mid D_n)$ summing to one, and for which the limit distribution is described by (6). The parametric bootstrap smoothed version is $\widetilde{\mu} = B^{-1}\sum_{b=1}^{B}\widehat{\mu}(y_b^*)$, with the y_b^* being bootstrap datasets sampled from the estimated wide model, that is, at $(\widehat{\theta}, \widehat{\gamma})$. This is the finite-bootstrap version for our computer, so to speak, whereas the theoretical version may be expressed as $\widetilde{\mu} = {\rm E}_B\widehat{\mu}(y^*)$, with expectation being with respect to the distribution of y^* given the dataset. One may then prove the master theorem

$$\sqrt{n}(\widetilde{\mu} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widetilde{\mu}(D)\}, \qquad (7)$$

as a clear analogue of result (6) for the unsmoothed estimator. Here $\widetilde{\mu}(D)$ is the bootstrap expectation of $\widehat{\mu}(D^*)$, in the limit experiment, where D^* is from the properly estimated version of ${\rm N}_q(\delta, Q)$, that is, $D^* \mid D \sim {\rm N}_q(D, Q)$. In other words,

$$\begin{aligned}\widetilde{\mu}(D) = {\rm E}_B\widehat{\mu}(D^*) &= \int \widehat{\mu}(u)\phi(u - D, Q)\,{\rm d}u\\ &= \int \sum_{j=0}^{q} w(j \mid u)G_j u\phi(u - D, Q)\,{\rm d}u,\end{aligned}$$

writing $\phi(v, Q)$ for the density of ${\rm N}_q(0, Q)$.

Armed with limit distributions (6) and (7) we may now compare performance aspects, for example, via risk functions, which with squared error loss become, respectively,

$$\begin{aligned}r_0(\delta) &= \tau_0^2 + {\rm E}\,[\omega^{\rm t}\{\delta - \widehat{\delta}(D)\}]^2,\\ r(\delta) &= \tau_0^2 + {\rm E}\,[\omega^{\rm t}\{\delta - \widetilde{\delta}(D)\}]^2.\end{aligned}$$

We learn that differences in performance rests with how well the two estimators $\widehat{\psi} = \omega^{\rm t}\widehat{\delta}(D)$ and $\widetilde{\psi} = \omega^{\rm t}\widetilde{\delta}(D)$ perform as estimators of the linear combination parameter $\psi = \omega^{\rm t}\delta$, with known coefficients ω, in the very clean limit experiment where a single $D \sim {\rm N}_q(\delta, Q)$ is observed, with mean δ unknown and variance matrix Q known. Note that these risk functions pan out differently for different focus parameters, via ω of (3), so bootstrap bagging may be more successful for some parameters than for others, even with the same data and the same list of candidate models. We also see that these constructions do not matter so much in cases where τ_0 is large compared to $(\omega^{\rm t}Q\omega)^{1/2}$, also since the individual estimators $\widehat{\mu}_j$ then will be highly correlated,

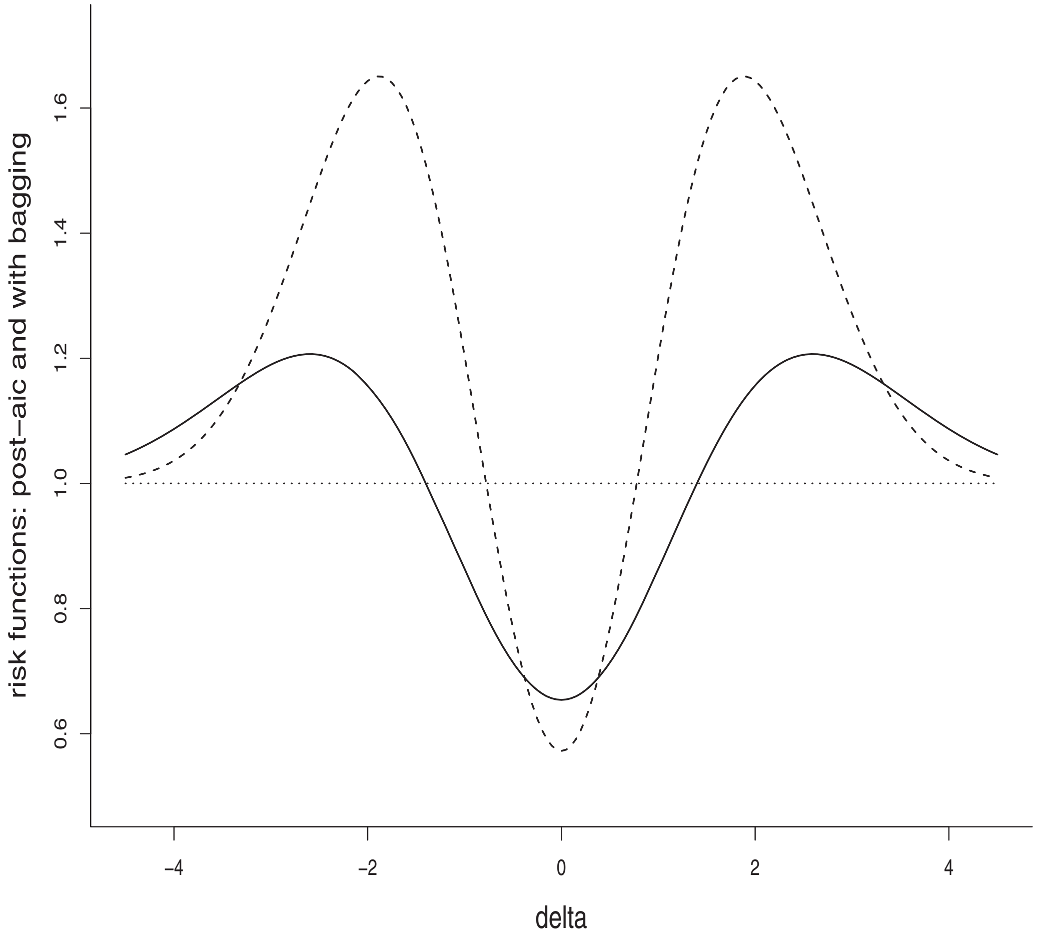

Figure 2. Risk functions for the AIC related postselection estimator $\widehat{\delta} = DI\{|D| \geq \sqrt{2}\}$ (dotted line) and the bagged or bootstrap smoothed version $\widetilde{\delta}$ (full line). Also indicated is the flat risk function for the minimax estimator $\bar{\delta} = D$.

and hence nearly equal with high probability. In cases where τ_0 is relatively smaller, however, differences in performance show up, and the potential gain by careful model averaging is more significant.

As a simple illustration, consider the case where the narrow model has p parameters and the wider model has just one more extra parameter, that is, $q = 1$. In that case the various formulas above simply, for example, to $G_0 = 0$ and $G_1 = 1$, and risk functions differences hinge on the simpler $\mathrm{E}\,(\widehat{\delta} - \delta)^2$ and $\mathrm{E}\,(\widetilde{\delta} - \delta)^2$. A start estimator of the form $\widehat{\delta}(D) = w(D)D$ then needs to be compared to its bootstrap smoothed version $\widetilde{\delta}(D) = \int w(u)u\phi(u - D, Q)\,\mathrm{d}u$; here $1 - w(D_n)$ and $w(D_n)$ are the weights given to the narrow and the wide model, respectively. Figure 2 displays these two risk functions, for the C_p or AIC case, which can be seen to correspond to $\widehat{\delta}(D) = I\{|D|/Q^{1/2} \geq \sqrt{2}\}$, and where I choose $Q = 1$ when displaying the plots. We note that bagging helps significantly, in parts of the parameter space, but not uniformly. Risk function mountains and hills caused by the bumpiness of C_p and AIC shall be made low. The same applies to other choices of $w(D)$. The figure also displays the flat minimax risk function for the estimator $\bar{\delta} = D$, associated with sticking to the widest model; here bagging simply reproduces the same estimator.

Note that results (6) and (7) are generally valid for any model averaging methods, not merely the special cases just pointed to, those associated with postselection inference. We may, for example, compare smoothed AIC with smoothed FIC, along with their bagged versions.

The theory developed here may also be used to assess the accuracy of the confidence limit methods of Efron, for example, the $\widetilde{\mu} \pm 1.96\,\widetilde{\mathrm{sd}}_B$ of Section 2, and to working out alternatives. One would expect there to be certain improvements on that particular ± 1.96 method, as the distribution of $\widetilde{\mu} - \mu$ is typically highly nonnormal, asymmetric, etc.

REFERENCES

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754. [1017]

Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion" (with discussion and rejoinder), *Journal of the American Statistical Association*, 98, 900–916. [1018]

——— (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press. [1018]

Hjort, N. L., and Claeskens, N. L. (2003), "Frequentist Model Averaging Estimators" (with discussion and rejoinder), *Journal of the American Statistical Association*, 98, 879–899. [1018]

Sala-i-Martin, X. X. (1997), "I Just Ran Two Million Regressions," *The American Economic Review*, 87, 178–183. [1017]

Schweder, T., and Hjort, N. L. (2014), *Confidence, Likelihood, Probability*, Cambridge: Cambridge University Press. [1018]

Rejoinder

Bradley EFRON

Model selection was an underdeveloped country on the map of classical statistics. The blame lay with intractable mathematics, of the sort connected with discontinuous functional forms. That excuse has worn thin in an era of virtually infinite computer power. The current article shows some progress being made through a combination of a little mathematics with a lot of computation, while the discussants offer other promising paths forward. It seems a safe bet that model selection, how to do it and what are its effects on inference, will continue to be a major topic for statisticians. Here I have touched on both aspects, bootstrap smoothing for the how-to-do-it part, and two bagging accuracy theorems on its effects.

Professor Hjort, following through on his ambitious series of articles with Claeskens, puts the problem into an asymptotic framework. This necessarily involves making the signal weaker ("$\gamma_j = \delta_j/\sqrt{n}$") as sample size n increases. Otherwise the model-selection aspect disappears: in terms of my schematic Figure 9, the distributional ellipses shrink to lie within a single wedge. Here, we have to worry that changing the signal strength may reduce asymptotics' relevance to the situation at hand. No such change is necessary in the classical picture of Figure 8, where the asymptotics are inherently simpler.

My article avoids asymptotics, or at least the mention of asymptotics. Bootstrap methods are by their nature nonasymptotic, though their formal justification in the literature usually involves large-sample calculations. Professor Politis' *Model-Free Prediction Principle* for regression is justified in terms of transformations that induce heteroscedastic residuals. This is similar in intention to Efron (1987), where bootstrap confidence intervals are justified by hypothetical transformations to normality, avoiding at least a direct appeal to asymptotics.

Politis' discussion is a reminder that neither of my current theorems applies to bootstrapping residuals. On the other hand, they do apply to model-selection situations other than regression—K-means clustering for example, or data-based choice of window width in kernel density estimation.

I was reassured by the new method's good performance in the simulation studies of Professors Wang, Sherwood, and Li, notably more ambitious than the few in my article. The questions raised of the number of bootstrap replications B is an important one. The tactic in my examples, choosing B much bigger than necessary, might be impractical in more complicated problems. Formula (3.11) provides a data-based guide to the choice, supplemented with the bias correction calculations of Remark J.

The bootstrap is fundamentally a plug-in methodology, along the lines of maximum likelihood estimation (MLE). As such I would not expect $\widetilde{\text{se}}_B$ to perform well in the "large p" context of Wang–Sherwood–Li's second question, but I would be happy to be proven wrong.

Bootstrap smoothing, or bagging, can be thought of as a form of nonparametric MLE (Efron and Tibshirani 1997): suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is an iid sample from distribution F, and $\hat{\theta} = t(\boldsymbol{x})$ is an unbiased estimator of a parameter of interest $\theta = \theta(F)$,

$$\theta(F) = E_F\{\hat{\theta}\}.$$

Then the nonparametric MLE of θ is

$$\tilde{\theta} = \theta(\hat{F}) = E_{\hat{F}}\{\hat{\theta}^*\},$$

where $\hat{F}$ is the empirical distribution corresponding to $\boldsymbol{x}$; in other words $\tilde{\theta}$ is the bagged version of $\hat{\theta} = t(\boldsymbol{x})$.

Professors Gelman and Veharti's first example is one in which bagging works well. Figure 4 implies that their statistic must have been quite nonlinear as a function of the bootstrap counts. Remark F emphasizes the point that bagging is *not* helpful in the case where $\hat{\theta}$ is already linear.

Blind application of bootstrap smoothing is not going to work if $\hat{\theta}^* = t(\boldsymbol{x}^*)$ can take on infinite values. This is a sign that the statistic $t(\cdot)$ is unstable in the relevant neighborhood of the sample space, and that some regularization, of the kind Gelman and Veharti use in the "bad" example, is called for. In other words, one should not throw out the bootstrap with the bad statistic's bathwater.

Professors Gupta and Lahiri suggest two more approaches to the post model-selection accuracy problem. Zou's interesting ALASSO method asymptotically selects the right model. In terms of Figure 9, the sampling ellipses around the ALASSO estimate must shrink to lie entirely within a single wedge, and in fact the *correct* one. In the examples I have looked at, admittedly not all that many, model selection was realistically far more random than that.

A quite different approach is suggested as the *maximum frequency* (MF) method: only employ those bootstrap replications that fall into the same wedge as the original data, thereby avoiding model-selection jumpiness. For the Cholesterol example of Table 3, MF gives approximate 95% interval

$$4.71 \pm 1.96 \cdot 5.43 = [-5.93, 15.35],$$

far to the right of the smoothed interval $[-13.3, 8.0]$.

This raises the question of *conditionality*: perhaps the statistician should condition on the observed selected model (though this raises the peril of ignoring model selection effects, our original objection to classical practice).

Bayesian estimates of accuracy are automatically conditional. For model selection problems, however, they require

Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305-4065 (E-mail: *brad@stat.stanford.edu*).

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.932172

a fearsome amount of prior specification: prior probabilities for the different models, and then informative prior distributions within each model. In two current articles (Efron 2012, 2014), I have drawn connections between "objective" Bayes' analysis and bootstrap estimates of variability. That line of thinking supports the unconditional kinds of bootstrap smoothing suggested in the current article, but so far it is only a suggestion.

My thanks go to the discussants and editors for an informative exchange on an important topic.

REFERENCES

Efron, B. (1987), "Better Bootstrap Confidence Intervals" (with comments and a rejoinder by the author), *Journal of the American Statistical Association*, 82, 171–200. [1021]

——— (2012), "Bayesian Inference and the Parametric Bootstrap," *The Annals of Applied Statistics*, 6, 1971–1997. [1022]

——— (2014), "Frequentist Accuracy of Bayesian Estimates," *Journal of the Royal Statistical Society* , Series B. Available at *http://statistics.stanford.edu/˜brad/papers/*. [1022]

Efron, B., and Tibshirani, R. (1997), "Improvements on Cross-Validation: The 0.632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548–560. [1021]