Model selection, estimation, and bootstrap smoothing

Bradley Efron*
Stanford University

Abstract

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 bootstrap smoothing, also known as bagging, to tame the erratic discontinuities of selection-based estimators. A projection theorem then provides standard errors for the smoothed estimators. Two examples, non-parametric 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.

Keywords: model averaging, C_p, Lasso, bagging, ABC intervals, importance sampling

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 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 y was the *decrease* in blood-level cholesterol measured from the beginning to the end of the trial,

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

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

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

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

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

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

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

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

*Research supported in part by NIH grant 8R01 EB002784 and by NSF grant DMS 1208787.

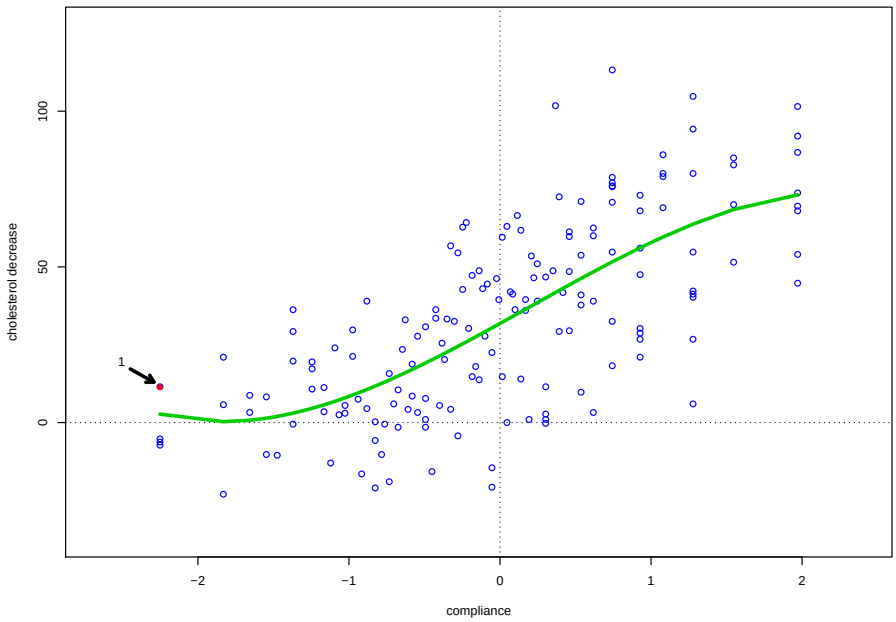

Figure 1: *Cholesterol Data* Cholesterol decrease plotted versus adjusted compliance for 164 men in Treatment arm of the cholostyramine study (Efron and Feldman, 1991). Solid curve is OLS cubic regression, as selected by the C_p criterion. How accurate is the curve, taking account of model selection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

We wish to assign standard errors to the estimates $\hat{\mu}_j$ read from the regression curve in Figure 1. A non-parametric bootstrap estimate $\widetilde{\mathrm{sd}}_j$ of standard deviation for $\hat{\mu}_j$, 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 $\overline{\mathrm{sd}}_j$ obtained from standard OLS calculations, assuming that the cubic model was pre-selected. The ratio $\widetilde{\mathrm{sd}}_j/\overline{\mathrm{sd}}_j$ has median value 1.56; so at least in this case, ignoring model selection can be deceptively optimistic.

The article proceeds as follows: Section 2 describes non-parametric bootstrap smoothing, a model-averaging device that is needed to improve assessments of accuracy in model selection problems. Section 3 then develops simple computational formulas for the standard deviation of a smoothed estimate. A parametric bootstrap version of the smoothing theory is described in Sections 4 and 5. Parametric modeling allows more refined results, permiting 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.

Berk, Brown, Buja, Zhang and Zhao (2012) develop conservative 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 paper assume that the model selection procedure *is* known, yielding smaller standard error estimates and shorter confidence intervals.

Hjort and Claeskens (2003) construct 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 ideal solution to model selection problems, but, as Hjort and Claeskens point out, it requires an intimidating amount of prior knowledge from the statistician. The present article is frequentist in its methodology.

Hurvich and Tsai (1990) provide a nice discussion of what "frequentist" might mean in a model selection framework. (Here I am following their "overall" interpretation.) The non-parametric bootstrap approach

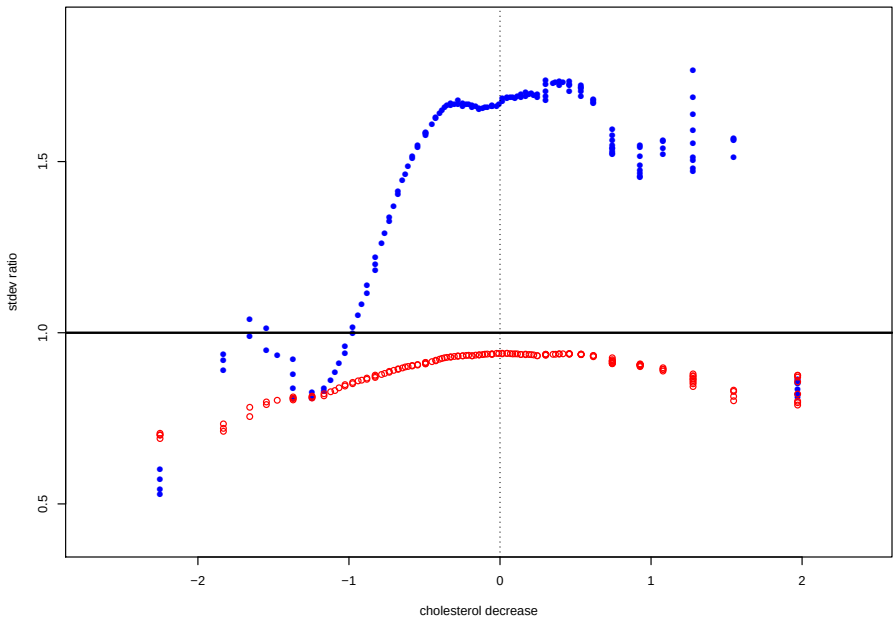

Figure 2: *Solid points*: ratio of standard deviations, taking account of model selection or not, for the 164 values $\hat{\mu}_j$ from the regression curve in Figure 1. Median ratio equals 1.56. Standard deviations including model selection are the smoothed bootstrap estimates $\widetilde{\mathrm{sd}}_B$ of Section 3. *Open points*: ratio of $\widetilde{\mathrm{sd}}_B$ to $\widehat{\mathrm{sd}}_B$, the unsmoothed bootstrap sd estimates as in (2.4), median 0.91.

in Buckland, Burnham and Augustin (1997) has a similar flavor to the computations in Section 2. What I am calling bootstrap smoothing or model averaging is perhaps better known as "bagging" (Breiman, 1996), an excellent recent reference being Buja and Stuetzle (2006).

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 has

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

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

In a non-parametric 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 non-parametric 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 empirical standard deviation of B such draws,

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

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

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

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

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

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

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: 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%)

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.

$B = 4000$ non-parametric bootstrap replications of the C_p-OLS regression curve — several times more than necessary, see Section 3 — were generated: starting with a bootstrap sample $\boldsymbol{y}^*$ (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

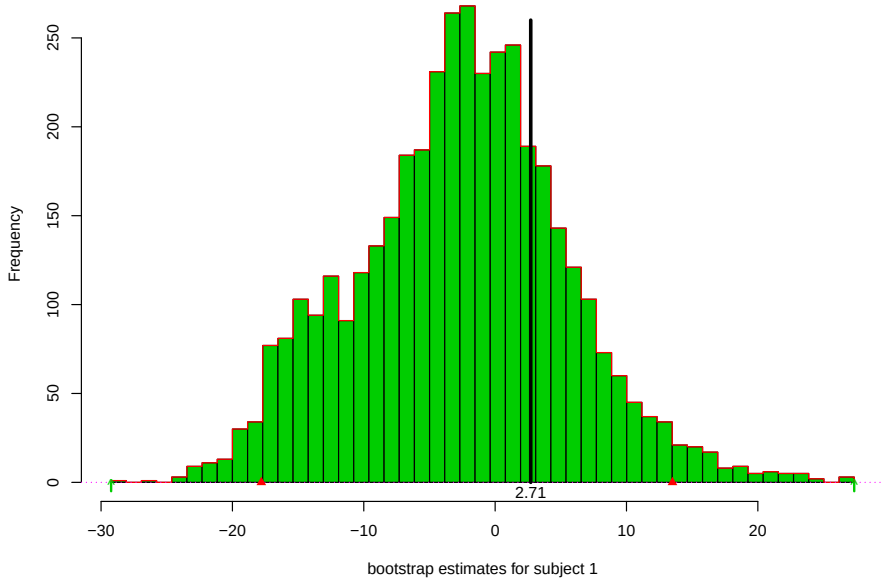

Figure 3: $B = 4000$ bootstrap replications $\hat{\mu}_1^*$ of the C_p-OLS regression estimate for Subject 1. The original estimate $t(\boldsymbol{y}) = \hat{\mu}_1$ is 2.71, exceeding 76% of the replications. Bootstrap standard deviation (2.4) equals 8.02. Triangles indicate 2.5th and 97.5th percentiles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

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.

Figure 4 shows why. The cases when m^* equaled 2 or 6, that is, where C_p selected a linear or quintic model, tended toward smaller $\hat{\mu}_1^*$ values than with cubic selection, the quadratic, quartic, and sextic selections also tending smaller. The actual data point $\boldsymbol{y}$ fell into the cubic selection region, yielding a large value of $\hat{\mu}_1^*$. The bootstrap replications suggest that things might very well have turned out otherwise. Model selection can make the original estimator for μ, $\hat{\mu} = t(\boldsymbol{y})$, "jumpy" and erratic.

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 dangerious jumpiness of $\hat{\mu} = t(\boldsymbol{y})$. The *percentile interval*, Section 13.3 of Efron and Tibshirani (1993),

$$\left[\hat{\mu}^{*(.025)}, \hat{\mu}^{*(.975)}\right], \tag{2.10}$$

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).)

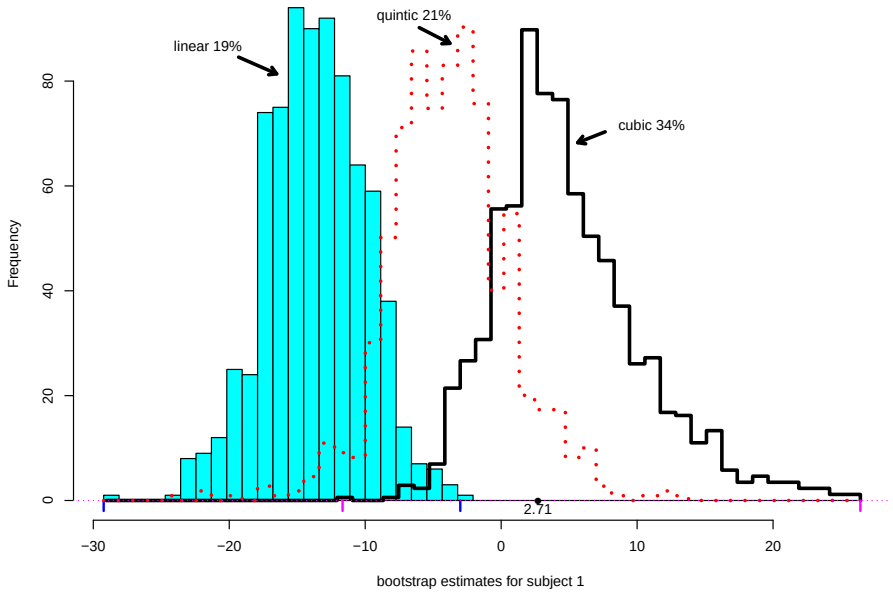

Figure 4: The bootstrap replications $\hat{\mu}_1^*$ separated into those cases where $m^* = 2$ (linear model preferred), $m^* = 6$ (quintic model), and $m^* = 4$ (cubic model). Cubic cases were generally larger, also being larger than the quadratic, quartic, and sextic cases.

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 2: 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

Table 2 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 $\boldsymbol{y}_j^{**}$, from which we calculate $s_i^* = \sum t(\boldsymbol{y}_j^{**})/B$; 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 non-ideal 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

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

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

Theorem 1. *The non-parametric 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

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

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 non-ideal 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

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

with $Y_{\cdot j}^* = \sum_{i=1}^B Y_{ij}^*/B$ and $t_\cdot^* = \sum_{i=1}^B t_i^*/B = s(\boldsymbol{y})$.

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 $t^*_i - s_0$.

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*.)

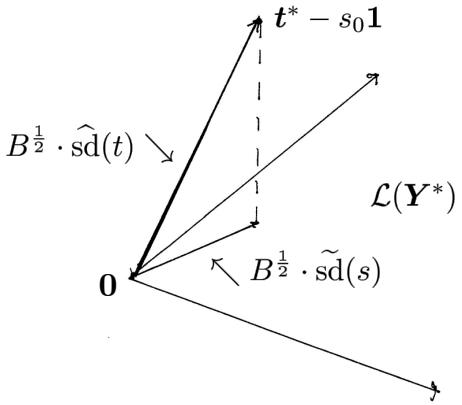

Figure 5: Illustration of Corollary 1. The ratio $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ is the cosine of the angle between $\boldsymbol{t}^* - s_0\mathbf{1}$ (3.9) and the linear space $\mathcal{L}(\boldsymbol{Y}^*)$ spanned by the centered bootstrap counts (3.2). Model selection estimators tend to be more non-linear, yielding smaller ratios.

The illustration in Figure 5 shows that $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ is the cosine of the angle between $\boldsymbol{t}^* - s_0\mathbf{1}$ and $\mathcal{L}(\boldsymbol{Y}^*)$. The ratio is a measure of the non-linearity of t^*_i as a function of the bootstrap counts Y^*_{ij}. Model selection induces discontinuities in $t(\cdot)$, increasing the non-linearity and decreasing $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$. The 164 ratios shown as open circles in Figure 2 had median 0.91, mean 0.89.

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

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

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

Proof of Theorem 1. The "non-parametric delta method" is the same as the *influence function* and *infinitesimal jackknife* methods described in Chapter 6 of Efron (1982). It applies 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 $T(\boldsymbol{Y}^*)$,

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

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

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

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

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

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

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

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),

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

so that

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

(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

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

Letting $\epsilon \to 0$ yields

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

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

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

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

$$\dot{S}_j = n \,\text{cov}_j \tag{3.21}$$

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,

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

If $\hat{a} = 0.10$ for instance, then at the upper endpoint $\hat{\theta}_{\rm up} = \hat{\theta} + 1.96\,\hat{\sigma}$ the standard error will have increased to about $1.196\hat{\sigma}$, leaving $\hat{\theta}_{\rm up}$ only 1.64, not 1.96, σ-units above $\hat{\theta}$.

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

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

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

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

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

with ${\rm cov}_*$ indicating bootstrap covariance as in (3.5). Unfortunately, $\widetilde{\rm bias}$ 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,

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

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 ${\rm cov}_\alpha(\hat{\beta})$. 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 i.i.d. realizations $\hat{\beta}^*$ from the MLE density $f_{\hat{\alpha}}(\cdot)$,

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

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 $\hat{\mu}$,

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

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

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

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

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

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

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

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

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

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

$$\mathrm{cov}_* = \boldsymbol{B}'\boldsymbol{t}^*/B \tag{4.8}$$

where $\boldsymbol{t}^*$ is the B-vector of bootstrap replications $t_i^* = t(\boldsymbol{y}^*)$.

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

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

(Proof given at the end of this section.)

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

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

s_0 *the ideal smoothed estimate* (4.5)*, the ratio being*

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

In the ideal bootstrap case, (4.7) and (4.9) show that $\widetilde{\mathrm{sd}}$ equals $B^{-1/2}$ times the numerator on the right-hand side of (4.11). This is recognizable as the length of projection of $\boldsymbol{t}^* - s_0\mathbf{1}_B$ into the p-dimensional linear subspace of $\mathcal{R}^B$ spanned by the columns of $\boldsymbol{B}$. Figure 5 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

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

(or $\widehat{\mathrm{cov}}'\hat{V}^{-1}\widehat{\mathrm{cov}}$ in what follows) is the delta-method assessment of *covariance* for the smoothed vector estimate $s(\boldsymbol{y}) = \sum t(\boldsymbol{y}_i^*)/B$.

Only minor changes are necessary for realistic bootstrap computations, i.e., 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

$$\widehat{\mathrm{cov}} = \boldsymbol{B}'\boldsymbol{t}^*/B \tag{4.13}$$

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

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

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

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

$$w_i = c_{\alpha,\hat{\alpha}} e^{Q_i} \qquad \text{where } Q_i = (\alpha - \hat{\alpha})'\,\hat{\beta}_i^*, \tag{4.16}$$

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),

$$\hat{E}_\alpha = \sum_{i=1}^B w_i t_i^* \Big/ \sum_{i=1}^B w_i = \sum_{i=1}^B e^{Q_i} t_i^* \Big/ \sum_{i=1}^B e^{Q_*}. \tag{4.17}$$

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 (4.16)–(4.17) gives,

$$s_\alpha \doteq \frac{\sum(1+Q_i)t_i^*/B}{\sum(1+Q_i)/B} = s_{\hat{\alpha}} + (\alpha - \hat{\alpha})'\,\mathrm{cov}_* \tag{4.18}$$

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

We see that

$$\left.\frac{ds_\alpha}{d\alpha}\right|_{\hat{\alpha}} = \mathrm{cov}_*, \tag{4.19}$$

so from (4.2),

$$\left.\frac{ds_\alpha}{d\beta}\right|_{\hat{\alpha}} = \hat{V}^{-1}\,\mathrm{cov}_*\,. \tag{4.20}$$

Since $\hat{V}$ is the covariance matrix of $\hat{\beta}^*$, that is, of $\hat{\beta}$ under distribution $f_{\alpha=\hat{\alpha}}$, (4.6) and (4.20) verify $\widetilde{\mathrm{sd}}$ in (4.9) as the usual delta-method estimate of standard deviation for $s(\hat{\beta})$. ■

Theorem 1 and Corollary 1 can be thought of as special cases of the exponential family theory in this section. The multinomial distribution of $\boldsymbol{Y}^*$ (3.12) plays the role of $f_{\hat{\alpha}}(\hat{\beta}^*)$; $\hat{V}$ in (4.9) becomes $I - \mathbf{1}_n\mathbf{1}_n'/n$ (3.3), so that (4.9) becomes (3.4). A technical difference is that the $\mathrm{Mult}_n(n, \boldsymbol{p})$ family (3.13) is singular (that is, concentrated on a $n-1$-dimensional subspace of $\mathcal{R}^n$), making the influence function argument a little more involved than 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 Cholesterol Data, perhaps because of the comparatively high dimension, 164 versus 10.

5 The Supernova Data

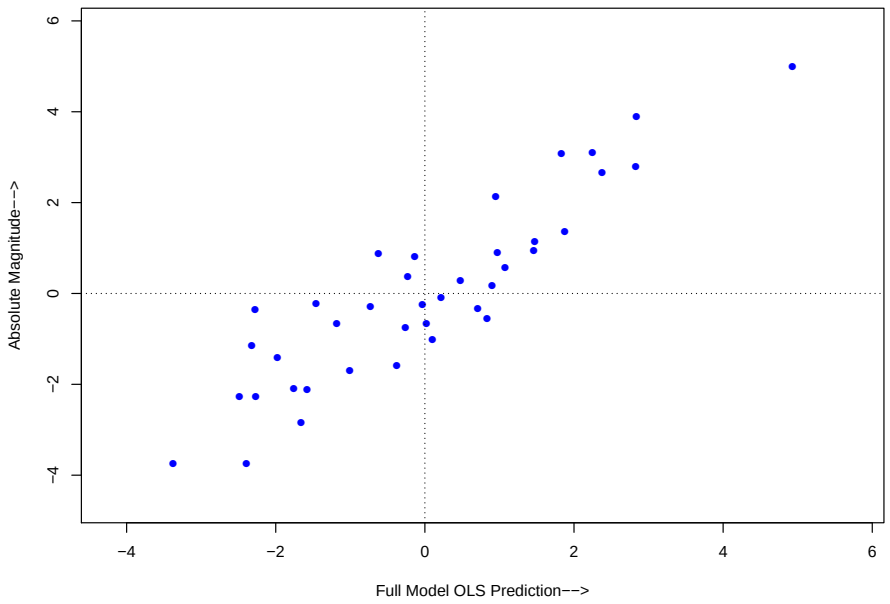

Figure 6: *The Supernova Data* Absolute magnitudes of $n = 39$ Type Ia supernovas plotted versus their OLS estimates from the full linear model (5.3); adjusted R^2 (5.5) equals 0.69.

Figure 6 concerns an 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. (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 $\hat{\alpha}_{\rm OLS}$, the ordinary least squares estimate of α in (5.3), yielding the full-model vector of supernova brightness estimates

$$\hat{\boldsymbol{\mu}}_{\rm OLS} = X\hat{\alpha}_{\rm OLS}. \tag{5.4}$$

Figure 6 plots y_i versus its estimate $\hat{\mu}_{{\rm OLS},i}$. 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

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

see Remark D.

Type Ia supernovas were used as "standard candles" in the discovery of dark energy and the cosmological expansion of the universe (Perlmutter et al., 1999; Riess et al., 1998). 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 non-negative "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 $\hat{\alpha}_{\rm OLS}$ toward zero, some of them all the way. As λ decreases from infinity to 0, the number m of non-zero 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, Hastie, Johnstone and Tibshirani, 2004). In what follows we will write $\hat{\alpha}_m$ rather than $\hat{\alpha}_\lambda$.

Table 3: Lasso model selection for the Supernova Data. As the regularization parameter λ in (5.6) decreases from infinity to zero, the number m of non-zero 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	$R^2_{\rm adj}$	
∞	0	0	0	
63	1	.17	.12	
19.3	3	.74	.70	
8.2	5	.79	.73	
.496	7	.82	**.735**	(selected)
.039	9	.82	.71	
0	10	.82	.69	(OLS)

Table 3 shows a portion of the Lasso calculations for the Supernova Data. Its last column gives $R^2_{\rm adj}$ (5.5) with R^2 having the usual form

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

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

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

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,

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

see Remark E. The equivalent of Table 3, 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, yielding 500 values s^*_k for each of the 39 supernovas, which provided direct bootstrap estimates say $\widetilde{\text{Sd}}_k$ for s_k. The $\widetilde{\text{Sd}}_k$ values averaged about 7.5% larger than the delta-method approximations $\widetilde{\text{sd}}_{Bk}$. 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.

Table 4: Percentage of the 4000 bootstrap replications selecting m non-zero 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

Returning to the original calculations, model selection was highly variable among the 4000 bootstrap replications. Table 4 shows the percentage of the 4000 replications that selected m non-zero 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 4.

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 method of this paper 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

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

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

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

Here $\hat{\alpha}$ is 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

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

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 (4.16)–(4.17), 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),

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

without requiring any further evaluations of $t(\cdot)$. (Note that $\hat{f}_\alpha(\hat{\beta}_i^*)$ is proportional to w_i in (4.17).) 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) in order 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 *non-normality* 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).

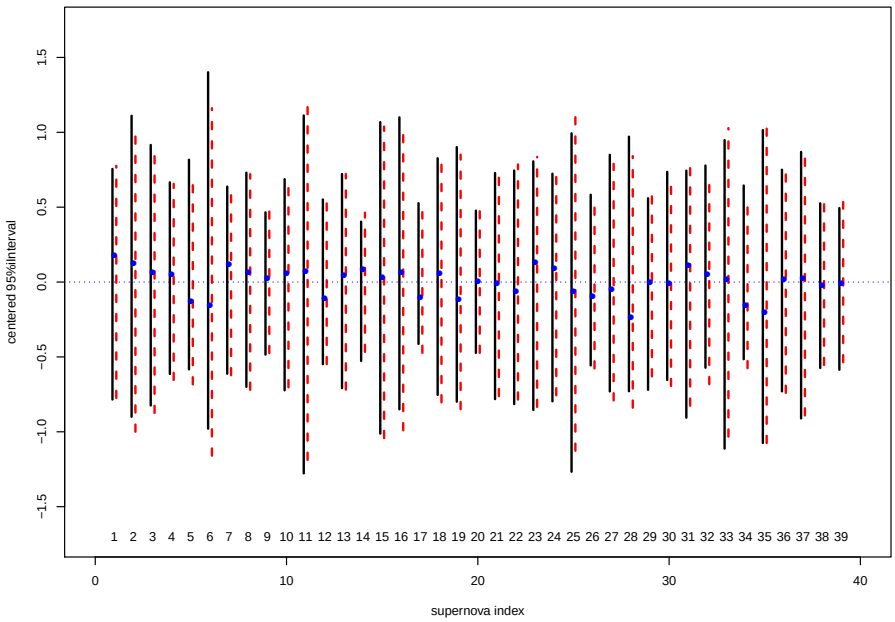

Figure 7: Approximate 95% confidence limits for the 39 supernova magnitudes μ_k (after subtraction of smoothed estimates $\tilde{\mu}_k$ (5.12)); ABC intervals (solid) compared with smoothed standard intervals $\tilde{\mu}_k \pm 1.96\widetilde{\text{sd}}_k$ (dashed). Dots indicate differences between unsmoothed and smoothed estimates, (5.11) minus (5.12).

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 in order 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 solid dots in Figure 7. A few of the intervals would be much different if based on the unsmoothed estimates $\hat{\mu}_k$, e.g., 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 5 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 5 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 size.

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

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

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

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

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

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

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 $\widetilde{\text{sd}}_{\text{boot}}$ 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 $(n-1)$ non-zero 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 $\mathcal{L}(\boldsymbol{X})$

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

verifying (3.4).

***B**. Hájek projection and ANOVA decomposition* For the ideal non-parametric bootstrap of Section 3, define the conditional bootstrap expectations

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

$j = 1, 2, \ldots, n$ (not depending on k). The bootstrap ANOVA decomposition of Efron (1983, Sect. 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

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

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,

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

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

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

***C**. Nonparametric bias estimate* There is a non-parametric 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 non-parametric 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

$$\ddot{S}_j = \lim_{\epsilon\to 0} \frac{S\left(\boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)\right) - 2S(\boldsymbol{p}_0) + S\left(\boldsymbol{p}_0 - \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)\right)}{\epsilon^2}. \tag{7.12}$$

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:

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

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} is motivated by OLS estimation in a homoskedastic 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

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

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

$$R^2 = 1 - \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 \Big/ \sum (y_j - \bar{y})^2 \tag{7.16}$$

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

$$R^2_{\text{adj}} = 1 - \left\{\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2m\hat{\sigma}^2\right\} \Big/ \sum (y_j - \bar{y})^2, \tag{7.17}$$

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 non-parametric 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

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

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* 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.

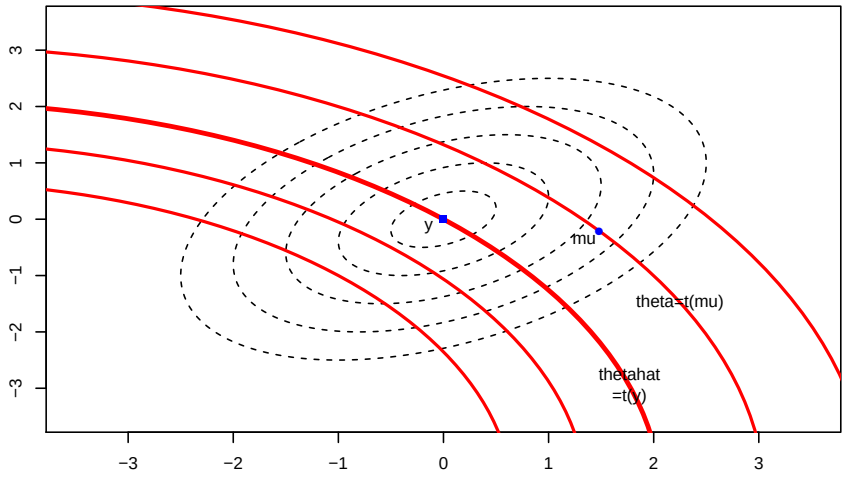

Figure 8: Schematic diagram of large-sample bootstrap estimation. Observed vector $\boldsymbol{y}$ has expectation $\boldsymbol{\mu}$. Ellipses indicate bootstrap distribution of $\boldsymbol{y}^*$ given $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. Parameter of interest $\theta = t(\boldsymbol{\mu})$ is estimated by $\hat{\theta} = t(\boldsymbol{y})$. Solid curves indicate surfaces of constant value of $t(\cdot)$.

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

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

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

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

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 "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 $\tilde{\mu} \pm 1.96\widetilde{\mathrm{se}}_B$ 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.

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, e.g., 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

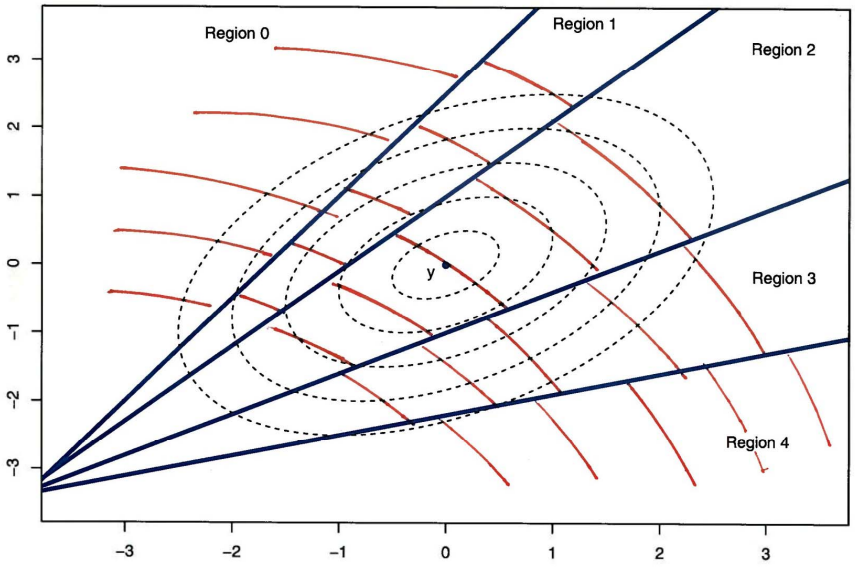

Figure 9: Estimation after model selection. The regions indicate different model choices. Now the curves of constant estimation jump discontinuously as $\boldsymbol{y}$ crosses regional boundaries.

(7.20)–(7.21) are less convincing when the local geometry near the observed $\boldsymbol{y}$ can change wildly 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 paper, 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 non-differentiable, across region boundaries, giving a picture somewhere between Figure 8 and Figure 9. This might help explain the smooth estimators' relatively modest reductions in standard error for the Supernova analysis.)

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 non-linearity 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 in order 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 6 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.

References

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2012). Valid post-selection inference. *Ann. Statist.* Submitted. `http://stat.wharton.upenn.edu/~buja/PoSI.pdf`.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24: 123–140.

Table 6: 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	.00	.00	.00	.00
stdev	.01	.13	.04	.06
lowest	−.01	−.21	−.07	−.14
highest	.01	.27	.09	.12

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* 53: 603–618.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statist. Sinica* 16: 323–351.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* 106: 608–625.

DiCiccio, T. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* 79: 231–245.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* 7: 1–26.

Efron, B. (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).

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* 78: 316–331.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* 82: 171–200, with comments and a rejoinder by the author.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.* 86: 9–17.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* 32: 407–499, with discussion, and a rejoinder by the authors.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability *57*. New York: Chapman and Hall.

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* 24: 2431–2461.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Roy. Statist. Soc. Ser. B* 74: 419–474.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist* 39: 325–346.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. New York: Springer-Verlag.

Hall, P., Lee, E. R. and Park, B. U. (2009). Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statist. Sinica* 19: 449–471.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* 98: 879–899.

Hurvich, C. M. and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician* 44: 214–217.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* 28: 1356–1378.

Mallows, C. L. (1973). Some Comments on C_p. *Technometrics* 15: 661–675.

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. *Astrophys. J.* 517: 565–586.

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. *Astron. J.* 116: 1009–1038.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58: 267–288.