Model search and inference by bootstrap "bumping"

Robert Tibshirani and Keith Knight

Department of Statistics

and

Department of Public Health Sciences

University of Toronto

November 18, 1997

Abstract

We propose a bootstrap-based method for searching through a space of models. The technique is well suited to complex, adaptively fitted models: it provides a convenient method for finding better local minima, for resistant fitting, and for optimization under constraints. Applications to regression, classification and density estimation are described. The collection of models can also be used to form a confidence set for the true underlying model, using a generalization of Efron's percentile interval. We also provide results on the asymptotic behaviour of bumping estimates.

Key words and phrases: bootstrap, stochastic search, resistant fitting, confidence sets

1 Introduction

The bootstrap (Efron (1979)) was introduced as a general method for assessing the statistical accuracy of an estimator. See Hall (1992) and Efron & Tibshirani (1993) for discussions of the bootstrap.

Breiman (1996) showed how one can use the bootstrap for the more primary purpose of producing a better estimator. Breiman's *bagging* procedure applies a given estimator $\hat{\theta}$ to each of B bootstrap samples, and then averages the B values to produce a new estimator $\tilde{\theta}$. In a number of experiments involving trees, subset selection and ridge regression, Breiman demonstrated that the bagged estimate $\tilde{\theta}$ often has smaller mean squared error than the original $\hat{\theta}$. The largest gains occurred for unstable estimators $\hat{\theta}$, like subset selection and trees, for which small changes in the data can produce large changes in the estimate. The improvement in mean squared error is mostly due to a reduction in variance.

Unfortunately the averaging process that produces the bagged estimate $\tilde{\theta}$ also destroys any simple structure that is present in the original estimate $\hat{\theta}$. That is, a bagged subset regression is no longer a subset, and a bagged tree is not a tree. And the reason that such estimators are unstable in the first place is that they seek a simple model for the data. Thus the estimators that are most helped by bagging in terms of prediction error are most hurt in terms of the interpretability of the final model. An exception is neural network models, which are neither interpretable nor stable! According to Breiman (personal communication) bagging helps the prediction accuracy of neural networks, and doesn't hurt them since interpreting their structure is rarely possible anyway.

In this paper we propose a different use of the bootstrap: we use bootstrap samples to provide candidate models for our model search. This has the advantage that it preserves the structure of the estimator while still inducing stability. Section 2 defines the bumping procedure. Applications of the idea to finding better local minima and resistant fitting are given in sections 3 and 4. Section 5 discusses constrained optimization problems. Some asymptotic analysis of bumping is given in section 6. Section 7 shows how to form confidence sets from bumping, while some further issues are covered in section 8.

2 The bumping procedure

We begin with a training sample $\mathbf{z} = (z_1, z_2, \ldots z_N)$ independent and identically distributed from a distribution F. We have a model for the data that depends on a set of parameters θ. From the training sample we assume that θ is to be estimated by minimization of a *target criterion*

$$\hat{\theta} = \text{argmin}_\theta \ R(\mathbf{z}, \theta). \tag{1}$$

Suppose also that there is (possibly different) *working criterion* R_0 for which minimization is convenient.

We propose to estimate θ by drawing bootstrap samples $\mathbf{z}^{1*}, \mathbf{z}^{*2}, \ldots \mathbf{z}^{*B}$, estimating $\hat{\theta}$ via R_0 from each sample

$$\hat{\theta}^{*b} = \text{argmin}_\theta \ R_0(\mathbf{z}^{*b}, \theta), \tag{2}$$

and then choosing $\hat{\theta}$ as the value among the $\hat{\theta}^{*b}$ producing the smallest value of $R(\mathbf{z}, \theta)$:

$$\hat{\theta}^B = \hat{\theta}^{*b} \text{ where } \hat{\theta}^{*b} = \text{argmin}_b \ R(\mathbf{z}, \hat{\theta}^{*b}). \tag{3}$$

As a convention, we always include the original sample $\mathbf{z}$ among the B bootstrap samples. We call this procedure "Bumping" for Bootstrap Umbrella of Model Parameters. The value $\hat{\theta}^B$ is the bumping estimate of θ.

There are three distinct scenarios in which bumping may be useful:

1. Problems where R is smooth but possesses many local minima. Then we may choose $R_0 = R$, and hope that the bumping procedure finds a better local minima.

2. Problems in which R is not smooth and hence difficult to minimize numerically. Then minimization of the working criterion R_0 offers a convenient alternative.

3. Constrained optimization problems that make R difficult to minimize numerically. Choosing the working criterion R_0 to be simply the unconstrained version of R offers a convenient numerical approach.

All of these scenarios are illustrated in the next sections.

Our main (but not exclusive) focus is on regression and classification problems, for which each observation z_i has the form $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is a feature vector and y_i a response measurement. We have a function $\eta(\mathbf{x}, \theta)$ that predicts y at $\mathbf{x}$, using parameters θ, and most often R_0 takes the form $R_0(\mathbf{z}, \theta) = \sum_i Q[y_i, \eta(\mathbf{x}_i, \theta)]$. Here Q is a loss function, typically $(y - \eta)^2$ in regression, and $I(y \neq \eta)$ or multinomial log-likelihood in classification. In another example given later, the problem is one of density estimation for a mixture of normal distributions and $-R$ is the log-likelihood function.

3 Bumping for finding better local minima

Suppose we take the working criterion R_0 to be the same as the target criterion R. Then if our procedure for minimizing R_0 always finds its global minimum (such as in linear least squares regression), then bumping will simply give this global minimum. The reason is that the global minimum cannot be improved and the bootstrap samples (by our convention) always include the original sample. However most adaptive procedures only find local minima, so there is potential for bumping to give a better local minimum.

As a specific example, we consider tree-based models for regression or classification as described in the CART (Classification and Regression Trees) work of Breiman, Friedman, Olshen & Stone (1984) and also in Clark & Pregibon (1991). Let T denote a tree and let $R(\mathbf{z}, \theta) = C(\mathbf{z}, T)$ the cost of the tree over the training set $\mathbf{z}$. In regression, $C(\mathbf{z}, T)$ would be taken as residual squared error while in classification, $C(\mathbf{z}, T)$ would be misclassification cost.

Example 1. *CART and local minima.*

Table 1 shows the results of a simulated example. There are 5 independent predictors $x_1, \ldots x_5$, each one uniform on $[-1, 1]$. The binary-valued outcome y equals two if both x_1 and x_2 are less than 0 or both greater than 0, and one otherwise. There are 50 training cases and 500 test cases. This pure interaction is difficult for the CART procedure because there is no information on where to split at the top level.

Table 1 shows that bumping significantly improves the performance of CART's greedy algorithm. Table 1 also shows the results of this procedure applied to some datasets from the University of California-Irvine machine learning database.

Table 1: Test set misclassification error rates for usual and bumped trees. Averages (standard deviation) over 20 simulations.

Dataset	Usual	Bumped
Simulated	0.49 (.03)	0.29 (.05)
Glass	0.35 (.02)	0.30 (.02)
Diabetes	0.24 (.01)	0.25 (.01)
Breast Cancer	0.049 (.003)	0.047 (.003)

For the glass data, Breiman (1996) reports a decrease in error from 32.0% for usual trees to 24.9% for bagged trees; for the diabetes data, the figures are 23.4% and 18.8%; for the breast cancer data, 6.0% and 4.2%. His rates for the usual trees may differ from ours because of differences in software and tuning parameters: however it is clear that bagging provides a larger reduction in error for these datasets.

3.1 Controlling model complexity

In the previous examples, we used pruning in CART to control the tree size. The issue of model complexity in the bumping procedure is an important one, so we now give more details. In order for the bumping procedure to make sense in general, the comparison of different $\hat{\theta}$ values with respect to their R value must make sense. Hence either the $\hat{\theta}$ values must be of the same complexity or the criterion R must include a factor for complexity of the model. Specifically for tree-based models, if $|T|$ denotes the number of terminal nodes in a tree T, CART minimizes the cost-complexity criterion

$$C_{\lambda}(\mathbf{z}, T) = C(\mathbf{z}, T) + \lambda|T| \tag{4}$$

where $\lambda > 0$ is a penalty parameter usually estimated by cross-validation. Denote the minimizer by $T_{\lambda}(\mathbf{z})$. Possible approaches include

1. Estimate $\lambda = \hat{\lambda}$ in the usual way from the training set by cross-validation. Compute $T_{\hat{\lambda}}(\mathbf{z}^*)$ from each bootstrap sample and find which tree minimizes the criterion $R(T, \mathbf{z})$.

2. As above, but use $R = C_{\hat{\lambda}}(\mathbf{z}, T) = C(\mathbf{z}, T) + \hat{\lambda}|T|$.

3. As in 1) or 2), but incorporate bumping into the initial cross-validation step for estimating λ.

The difference between (1) and (2) is subtle: if the number of terminal nodes in each tree $T_{\hat{\lambda}}(\mathbf{z}^*)$ was the same, then the two proposals would be the same. However, the number of terminal nodes can vary somewhat, despite the fact that the same $\hat{\lambda}$ is used each time. We found that proposal (1) gave better results in practice: the results reported in the paper use this proposal. Proposal (3) recognizes the fact that bumping is part of the fitting process, and hence its effect should be incorporated into the choice of λ. This proposal is extremely compute-intensive, and would likely not have a large effect on the choice of λ. We have not pursued it here.

In different models, the fixed complexity strategy (1) is implemented differently. For example in stepwise regression, we choose the number of predictors once and for all from the training data, and then fix that number in each bootstrap sample. In neural nets, we can choose the weight decay parameter from the training data, and then fix it in the bumping process.

4 Bumping for resistant fitting

Suppose we have data points $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots N$ where $\mathbf{x}_i$ is a vector of predictors (typically having first component equal to 1), and y_i are the response values. In a linear regression of y_i on $\mathbf{x}_i$, the usual linear least squares estimate is not resistant to the influence of individual data points. One remedy to this problem is to replace the sum of squares criterion by $R(\mathbf{z}, \theta) = \text{median}(y_i - \mathbf{x}_i^T\theta)^2$, whose minimizer the "least median of squares" (LMS) estimator has breakdown roughly 50% (Rousseeuw (1984)). This means that changes to less than 50% of the data cannot unduly influence the estimator. In the bumping framework, we take $R_0 = \sum(y_i - \mathbf{x}_i^T\theta)^2$, and we obtain an approximate version of the LMS estimator. There is no guarantee that the resulting $\hat{\theta}$ will minimize R, even if we take infinite number of bootstrap samples. However since any outlier in the sample will by chance be left out of some bootstrap samples, the least squares estimate for such a bootstrap sample will be close to the true minimizer of R.

Since each observation appears in roughly $1 - (1 - 1/N)^N \approx 1 - e^{-1} \approx 63.2\%$ percent of the observations, the number of bootstrap samples required so that at least one will not contain any of k given points is e^k. For $k = 1, 2, 3, 4, 5, 6$ this equals approximately $3, 7, 20, 55, 148, 403, 1096$. Hence for a reasonable number of bootstrap samples one would only achieve protection against a few masked outliers. Using smaller bootstrap samples can help this a bit. For bootstrap samples of size $\alpha \cdot N$, we need $e^{\alpha k}$ samples to have at least one on average that does not contain a given k points.

In fact, this procedure is not very different from the computational procedure proposed by Rousseeuw (1984) for the LMS procedure. He suggests taking samples without replacement of size the number of regressors p, and computing the interpolating plane for each. Then he chooses among these coefficients the values that minimize the median sum of squares over the entire data set.

This technique can be used to "robustify" nonlinear and/or adaptive procedures such as subset selection, additive models, tree-based models and neural networks.

Example 2. *Subset selection in linear regression.*

We generated 20 datasets of size 50 from the model

$$y = \mathbf{x}^T\theta + \epsilon \tag{5}$$

Here $\mathbf{x}$ is a vector of 6 standard normal variates with $\text{corr}(x_i, x_j) = .5^{|i-j|}$, $\theta = (1.5, 0, 1.5, 0, 0, 0, 0)$. The errors come from a mixture of normals: $\epsilon = (1-\delta)Z_1 + \delta \cdot 10 \cdot Z_2$ where Z_1, Z_2 are $N(0,1)$ and $\text{Prob}(\delta = 0) = .95$, $\text{Prob}(\delta = 1) = .05$. Thus on average there are 2.5 outliers per sample. We applied bumping, our estimator $\hat{\theta}$ being the best subset of size 2. Both the best subset estimator applied to the original sample and the bump estimator found the correct in all 20 simulations.

The mean squared error (MSE) for a test sample with no outliers are shown in Table 2; included is a column for the bagging estimate. The MSE is defined as $\text{E}(Y - \hat{Y})^2 - \sigma^2$ where σ^2 is the error variance in the model, and the expectation is taken over test observations Y.

Table 2: Mean squared errors for Example 2. Average (standard deviation) over 20 simulations,

Method	Usual	Bumped	Bagged
(1) Stepwise regression	.46 (.15)	.30 (.04)	.62 (.13)
(2) Neural net (2 hidden units)	7.0 (1.7)	2.6 (.21)	1.9 (.28)
(3) Neural net (2 hidden units, averaged)	3.3 (.52)	1.8 (.17)	2.1 (.19)

Bumping improves on the usual estimate while bagging does a little worse. Note that we used the efficient leaps and bounds procedure (`leaps` in the S language) in this example, made possible by the choice of sum of squared error for R_0. Direct minimization of the median sum of squares to find the best subset of size 2 would be difficult.

Also shown in table 2 are the mean squared errors for a single hidden layer neural network (line 2), with 2 hidden units and a linear output unit. Tis demonstrates how bumping can robustify a neural network, using a standard learning (fitting) method. We used Brian Ripley's "nnet" Splus function, which utilizes a variable metric optimizer and weight decay. Some of the variability in the errors can be due to the choice of starting weights, so we also tried averaging the fits over 3 random choices of starting weights (line 3). Both bumping and bagging improve the usual network, although the improvement is less when the predictions are averaged over the choice of starting weights.

4.1 Parametric density estimation

Here we observe data $z_1, z_2, \ldots z_N$, assumed to come from a density $f_\theta(z)$. Maximum likelihood estimation of θ proceeds by maximizing the log-likelihood $\ell(\theta, \mathbf{z}) = \sum \log f_\theta(z_i)$. To achieve a more resistant fit, we can apply bumping with $R = -\ell$, $R_0 = -\text{median}[\log f_\theta(z_i)]$; in cases where maximum likelihood finds only a local minimum of R, we can apply bumping with $R_0 = R = -\ell$. Both of these applications of bumping are attractive if maximization of ℓ is convenient computationally. We illustrate the resistant fitting application in the next example.

Example 3. *Fitting of mixtures.*

Consider a density estimation problem where each observation z_i is assumed to come from a mixture of normal distributions:

$$Z = \delta W_1 + (1 - \delta) W_2 \tag{6}$$

Here $W_1 \sim N(\mu_1, \sigma_1^2), W_2 \sim N(\mu_2, \sigma_2^2)$ and $\text{Prob}(\delta = 1) = \gamma, \text{Prob}(\delta = 0) = 1 - \gamma$. Letting $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, \gamma)$, the log-likelihood of the data is

$$\ell(\mathbf{z}, \theta) = \sum_{i=1}^{N} \log[\gamma\phi(z_i; \mu_1, \sigma_1) + (1 - \gamma)\phi(z_i; \mu_2, \sigma_2)] \tag{7}$$

where $\phi(x; \mu, \sigma)$ is the $N(\mu, \sigma)$ density evaluated at x. This is mostly conveniently maximized via the EM algorithm.

We generated 10 samples of size 50 from this distribution, with $\mu_1 = 0, \mu_2 = 4, \sigma_1 = 1, \sigma_2 = 1, \gamma = .5$. To each sample we added an outlier at $z = -8$. To estimate the parameters, we applied both maximum likelihood estimation via EM, and bumping with

$$R_0 = -\ell; \quad R = -\text{median}\{\log[\gamma\phi(z_i; \mu_1, \sigma_1) + (1 - \gamma)\phi(z_i; \mu_2, \sigma_2)]\}. \tag{8}$$

This choice of R should give a more resistant fit, while facilitating use of the convenient EM procedure through the choice of the working criterion R_0. The top panel of Figure 1 shows a typical sample, with the outlier at $y = -8$ excluded. The middle panel shows the true density (solid line), and the median and robust 90% standard error bands from the maximum likelihood fit (broken lines). The bottom panel shows the same for the resistant bumped fit. The overall error is significantly less in the bottom panel.

This example might be viewed as contrived, with an outlier that would be obvious from a plot of the data. However in high dimensional density estimation, such outliers are difficult to detect and can have a strong effect on the maximum likelihood fit.

Local minima also play a role in the mixture problem. In general, the global maximum of the log-likelihood occurs when one of the components is centred at an observation and has zero variance. This maximum is not of interest, and the EM algorithm typically finds the local maximum having both σ_1 and $\sigma_2 > 0$, when appropriate starting values are used. We found that bumping, with $R_0 = R = -\ell$ and no outliers in the sample, did not improve the EM algorithm for this problem.

5 Estimation under constraints

Some straightforward estimation problems are made much more difficult by the inclusion of constraints on the solution. In this situation, bumping can provide a simple and often satisfactory computational solution.

To be specific, consider a problem where our target criterion is to be minimized under constraints $\theta \in K$. The constraints may be of any type- equality constraints, inequality

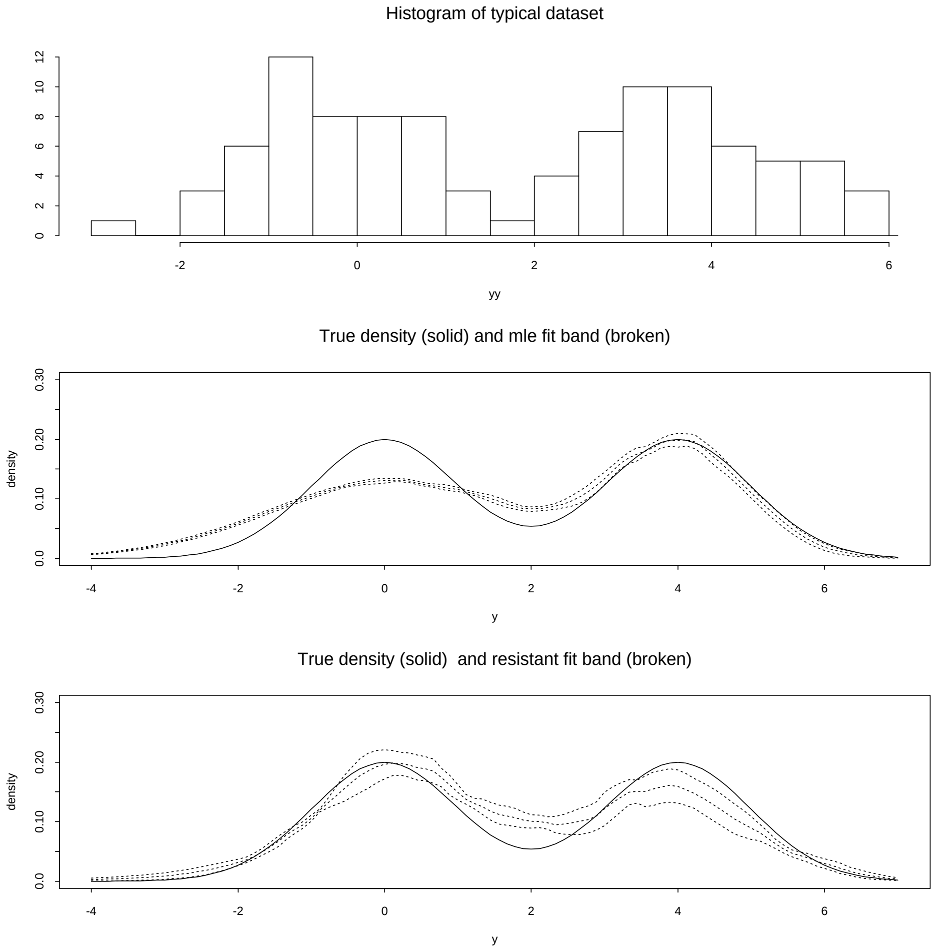

Figure 1: Results for Example 3 (Gaussian mixtures): the top panel shows a typical sample, with the outlier at $y = -8$ excluded. The middle panel shows the true density (solid line) and the median and robust 90% standard error bands from the maximum likelihood fit (broken lines); the bottom panel shows the same for the resistant bumped fit.

constraints, etc. To formally fit this problem into the bumping framework, we can express our target criterion in a Lagrangian penalty form

$$R(\mathbf{z},\theta) = \tilde{R}(\mathbf{z},\theta) + \gamma I(\theta \notin K) \tag{9}$$

with $\gamma > 0$. Here $\tilde{R}(\mathbf{z},\theta)$ is a common criterion such as squared error. To apply bumping, we simply choose the working criterion $R_0(\mathbf{z},\theta)$ equal to $\tilde{R}(\mathbf{z},\theta)$. The bumping procedure therefore reduces to 1) generating unconstrained solutions $\hat{\theta}^*$ from bootstrap samples, and then 2) choosing the one with smallest value of $\tilde{R}(\mathbf{z},\theta)$ satisfying the constraints $\theta \in K$. We illustrate this procedure below.

Example 4. *Monotone spline smoothing.*

Give data pairs $(x_1, y_1), \dots (x_N, Y_N)$, a cubic smoothing splines produces an estimate of $f(x) = \mathrm{E}(Y|x)$ by minimization of the penalized least squared criterion

$$J(f) = \sum_{i=1}^{B}(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt \tag{10}$$

The parameter $\lambda > 0$ trades off goodness of fit of the estimate with its wiggleness as measured by the integrated squared 2nd derivative. For fixed λ, the solution is most conveniently expressed as a linear combination of so-called “B splines”, and efficient algorithms are available for the calculation of the B-spline coefficients. Assuming all x_i values in the training set are unique, the solution involves N B-spline basis functions; surprisingly, the coefficient solutions are a linear function of the responses y_i. Now the smoothing problem becomes much harder if we require that the solution $\hat{f}(x)$ be a monotone increasing (or decreasing) function of x. This imposes N inequality constraints among the B-spline coefficients. Kelly & Rice (1990) suggest simplifying the computation by instead performing a constrained regression on a subset of the B-spline bases. This regression spline approach is reasonable: however it lacks the convenience of cubic smoothing splines where there is no need to pick a subset of the bases and the smoothness is controlled by a single parameter λ.

We apply bumping by fitting standard cubic smoothing splines to the bootstrap samples, and then choosing the monotone function having smallest residual sum of squares over the original training sample. To fix the complexity of each smooth, we fixed the degrees of freedom of each fit (rather than λ): this measure is less dependent on the range of the abscissa values of the data: see Hastie & Tibshirani (1990).

The top panel Figure 2 shows a plot of some simulated data, along with the usual cubic smoothing spline (broken), and the bumped monotone solution (solid). The solid curve is strictly monotone. The degrees of freedom each fit was fixed at 4, and 500 bootstrap samples were used. The bottom panel shows the bootstrap sample that gave the bumped solution. Hollow points did not appear in the sample; other points were randomly jittered so that duplicates would be visible. In this example, bumping works by changing the weights on the data points, so as to monotonize the cubic spline fit. From the total of 500 bootstrap samples, there were only 6 monotone curves and Figure 2 shows the best fitting one. In general there is no guarantee that a value $\hat{\theta}^*$ satisfying the constraints will emerge, although under parametric bootstrap sampling one can often make such a guarantee. However it may take a very large number of bootstrap samples to obtain an allowable solution, especially if the unconstrained solution does not nearly satisfy the constraints. Table 3 shows the behaviour of bumping, as a function of the number of

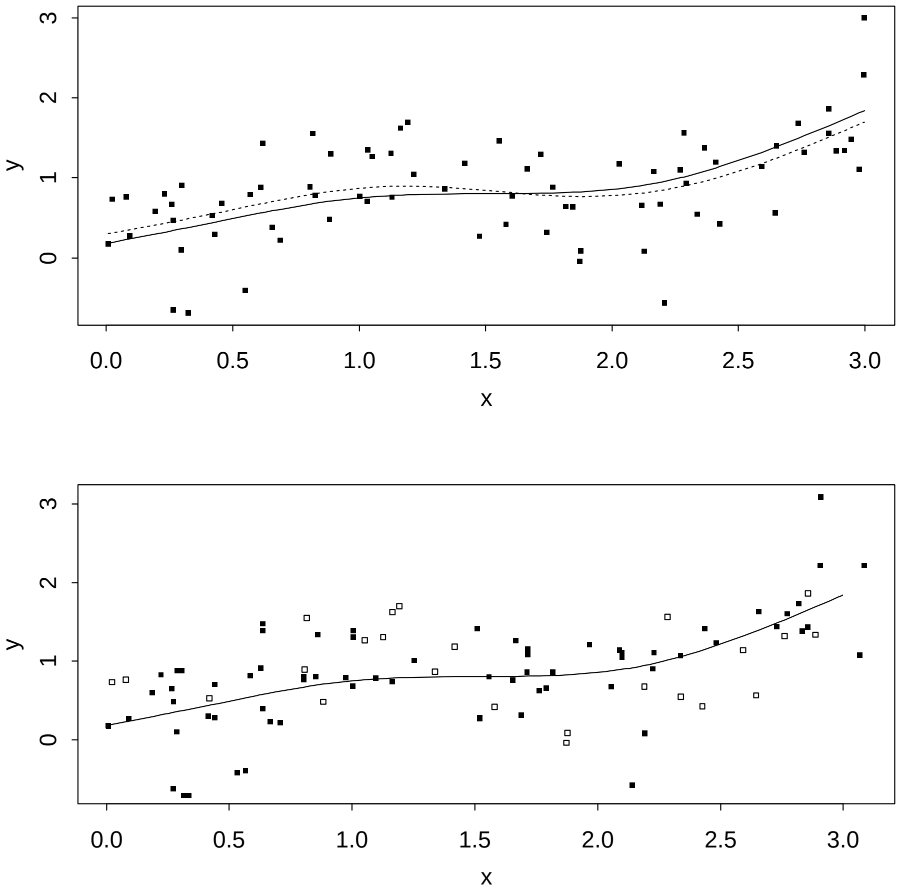

Figure 2: Top panel Figure shows a plot of some simulated data, along with the usual cubic smoothing spline (broken), and the bumped monotone solution (solid). Bottom panel shows the bootstrap sample that gave the bumped solution. Hollow points did not appear in the sample: other points were randomly jittered so that duplicates would be visible.

Table 3: Example 4: Behaviour of bumping as number of bootstrap samples in increased.

B	# of monotone curves	min R
100	2	20.37
500	5	19.90
1000	9	19.75

bootstrap samples B.

6 Some asymptotic analysis of bumping

6.1 Compatibility of R_0 with R

In the bumping procedure, if $\mathbf{z}^*$ is a bootstrap sample, the minima $\hat{\theta}^*$ of the working criterion $R_0(\mathbf{z}^*, \theta)$ are used as candidate values for minimization of the target criterion $R(\mathbf{z}, \theta)$. If we choose R_0 differently from R, we need to ensure that R_0 is in some sense compatible with R so that the $\hat{\theta}^*$ values will serve as reasonable minima for R.

In some situations we might be assured that the minima of R_0 cover the entire parameter space for an infinite number of bootstrap samples: this would typically occur, for example, with parametric bootstrap sampling. However, we would still like to use a small number of bootstrap samples in practice and hence we would like the two criteria to estimate the same parameter in some sense. One reasonable requirement would be

$$\text{argmin}_\theta[\mathrm{E}_F R_0(\mathbf{z}, \theta)] = \text{argmin}_\theta[\mathrm{E}_F R(\mathbf{z}, \theta)] \tag{11}$$

In Example 2, with $R = \sum(y_i - \mathbf{x}_i^T\theta)^2$, $R_0 = \text{median}(y_i - \mathbf{x}_i^T\theta)^2$, (11) is satisfied if $(y - \mathbf{x}^T\theta)^2$ has a symmetric distribution. This will be the case, for example, if y and $\mathbf{x}$ are jointly Gaussian. In the next section, we assume that R_0 and R are compatible in this sense, and examine how the bumping estimates behave asymptotically.

6.2 Asymptotic behaviour of bumping estimates

To gain some insight into the problem, we consider the case of estimating a finite dimensional parameter θ from a training sample with underlying probability measure P. If $\widehat{\theta}$ minimizes our target criterion R, we would like the bumping estimate $\widehat{\theta}^B$ to be "close" to $\widehat{\theta}$ for not-so-large values of B, the number of bootstrap samples. In particular, we would like to be able to carry out inference on θ using $\widehat{\theta}^B$ as if we were using $\widehat{\theta}$.

Define $\widehat{\theta}_n = \widehat{\theta}$ and $\widehat{\theta}_{B,n} = \widehat{\theta}^B$ so that the dependence on n is made explicit. We will assume that there exist constants $\{a_n\}$ such that $U_n = a_n(\widehat{\theta}_n - \theta_0)$ converges in distribution (for some constant θ_0) to some random vector U. (Typically, $a_n = \sqrt{n}$.) If we define $U_{n,B} = a_n(\widehat{\theta}_{n,B} - \theta_0)$ and $U_{n,B} - U_n = a_n(\widehat{\theta}_{B,n} - \widehat{\theta}_n)$ is sufficiently close to 0 (for some B) then it seems reasonable to expect that the asymptotic inference for θ using $\widehat{\theta}_{B,n}$ will be the same as that using $\widehat{\theta}_n$. Of course, $U_{n,B}$ depends not only on the data but on the bootstrap samples used to compute it. Notice that we have centered both U_n and $U_{n,B}$ by the same value θ_0: this assumes that the working function R_0 and the target function R are compatible in the sense of (11).

Let $\{Z_n\}$ be real-valued continuous random functions (stochastic processes) defined on

R^k, satisfying

$$U_n = a_n(\widehat{\theta}_n - \theta_0) = \text{argmin}_t Z_n(t).$$

(Z_n will depend on the target criterion R.) If $\widehat{\theta}_n^{*1}, \cdots, \widehat{\theta}_n^{*B}$ are (independent) bootstrap estimates of θ, define $T_n^{*b} = a_n(\widehat{\theta}_n^{*b} - \theta_0)$ for $b = 1, \cdots, B$. Then $T_n^{*1}, \cdots, T_n^{*B}$ are independent random variables with (bootstrap) probability distribution Q_n^* ($Q_n^*(A) = P^*(T_n^{*b} \in A)$ where P^* is the bootstrap probability measure) and

$$U_{n,B} = \text{argmin}(Z_n(T_n^{*1}), \cdots, Z_n(T_n^{*B})).$$

We can use either a parametric or non-parametric bootstrap; however, Q_n^* can be any random (i.e. data dependent) or non-random probability distribution. The functions Z_n are chosen so that $\{Z_n(\cdot)\}$ converges in distribution (on the appropriate space of continuous functions) to some other random function $Z(\cdot)$ which has an almost surely unique minimum U. Under regularity conditions U_n will converge in distribution to U (see Kim & Pollard (1990)). Before presenting our main result, we discuss an example that illustrates some of the main issues.

Example. Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. random variables and that we wish to estimate a real-valued parameter θ by minimizing

$$R_n(\theta) = \sum_{i=1}^{n} \rho(\theta; X_i)$$

where $\rho(\theta; x)$ is a smooth function of θ for each x. If $\widehat{\theta}_n = \text{argmin}_\theta R_n(\theta)$ then typically (subject to regularity conditions)

$$\widehat{\theta}_n \rightarrow_p \theta_0 = \text{argmin}_\theta E[\rho(\theta; X_1)].$$

We can then define

$$Z_n(t) = \sum_{i=1}^{n} \left[\rho(\theta_0 + n^{-1/2}t; X_i) - \rho(\theta_0; X_i) \right];$$

note that

$$U_n = \mathrm{argmin}_t Z_n(t) = \sqrt{n}(\widehat{\theta}_n - \theta_0).$$

If $\rho(\theta; x)$ is twice differentiable in θ for each x, we get

$$Z_n(t) = \frac{t}{\sqrt{n}} \sum_{i=1}^{n} \rho'(\theta_0; X_i) + \frac{t^2}{2n} \sum_{i=1}^{n} \rho''(\theta_0; X_i) + \xi_n(t).$$

Under regularity conditions, we get $Z_n(\cdot) \rightarrow_d Z(\cdot)$ where

$$Z(t) = tW + \frac{t^2}{2} E[\rho''(\theta_0; X_1)]$$

and W is normally distributed with mean 0 and variance $\mathrm{Var}[\rho'(\theta_0; X_1)]$. Now $Z(t)$ is minimized at $t = U = -W/E[\rho''(\theta_0; X_1)]$ and so $U_n \rightarrow_d U$ (under appropriate regularity conditions). $\diamond$

The closeness of $U_{n,B}$ to U_n depends on the ability of the distribution Q_n^* to produce T_n^{*b}'s close to U_n. If we can guarantee that at least one T_n^{*b} is close to U_n then $U_{n,B}^*$ should be close to U_n if Z_n is sufficiently smooth. Suppose that $\widetilde{\theta}_n$ is the estimate of θ based on minimizing the working criterion R_0 using the full sample. If the bootstrap is valid then the (bootstrap) distribution of $a_n(\widehat{\theta}_n^{*b} - \widetilde{\theta}_n)$ approaches the asymptotic distribution of $a_n(\widetilde{\theta}_n - \theta_1)$ where θ_1 need not equal θ_0. In other words, we can write

$$T_n^{*b} = a_n(\widehat{\theta}_n^{*b} - \widetilde{\theta}_n) + a_n(\widetilde{\theta}_n - \theta_1) + a_n(\theta_1 - \theta_0).$$

If $\theta_1 = \theta_0$ then the probability that Q_n^* puts in an ϵ-neighbourhood of U_n should be non-negligible. However, if $\theta_1 \neq \theta_0$ then $a_n(\theta_1 - \theta_0)$ is not $O(1)$ and Q_n^* will put negligible probability in an ϵ-neighbourhood of U_n for large n.

The following theorem gives sufficient conditions for bumping to work in the general case.

Theorem. Suppose that

(a) $Z_n(\cdot) \rightarrow_d Z(\cdot)$ on $C(K)$ for any compact set $K \subset R^k$;

(b) $U_n = O_p(1)$ and $U = \text{argmin}_t Z(t)$ is almost surely unique;

(c) There exists a sequence of random functions $\{\phi_n^*(x)\}$ such that

$$\sup_A \left| Q_n^*(A) - \int_A \phi_n^*(x)\, d\lambda(x) \right| \rightarrow_p 0 \quad \text{(where } \lambda = \text{Lebesgue measure)}$$

and for every $\delta > 0$ and compact set K, there exists $\gamma > 0$ such that

$$\liminf_{n\rightarrow\infty} P\left[\inf_{x\in K} \phi_n^*(x) > \gamma\right] \geq 1 - \delta.$$

Then for each $\epsilon > 0$ and $\delta > 0$, there exists $B_{\epsilon,\delta}$ such that

$$P^*\left[|U_n - U_{B,n}| > \epsilon\right] \leq \Delta_n^\epsilon$$

for all $B \geq B_{\epsilon,\delta}$ where $P(|\Delta_n^\epsilon| > \delta) \rightarrow 0$ as $n \rightarrow \infty$.

Proof:

A sketch of the proof goes as follows. Choose a set K such that $P(U_n \in K) \geq 1 - \delta$ and such that $P\left[\inf_{x\in K^\epsilon} \phi_n^*(x) > \gamma\right] \geq 1 - \delta$ where K^ϵ contains all points either in K or within a distance ϵ of K. Then for any subset A of K^ϵ,

$$P\left[Q_n^*(A) \geq \gamma'\lambda(A)\right] \geq 1 - 2\delta$$

for some $\gamma' > 0$. If U_n lies within the set K and $Q_n^*(A) \geq \gamma'\lambda(A)$, we can choose B sufficiently large so that at least one T_n^{*b} lies within ϵ of U_n with P^*-probability arbitrarily close to 1. Thus $U_{n,B}$ will lie in an ϵ-neighbourhood of U_n unless there exists t outside this neighbourhood with $Z_n(t)$ smaller than $Z_n(T_n^{*b})$ for all T_n^{*b} within the neighbourhood. However, if ϵ is sufficiently small then the probability of this latter event can be made arbitrarily small as $n \to \infty$; see, for example, the proof of Theorem 2.7 of Kim & Pollard (1990). $\diamondsuit$

Condition (c) guarantees that any set A which is a subset of a compact set K will have $Q_n^*(A) \geq \gamma\lambda(A)$ with high probability for some $\gamma > 0$. This condition will put at least one T_n^{*b} in an ϵ-neighbourhood of U_n with high bootstrap (P^*) probability for B sufficiently large.

When is condition (c) satisfied? Consider the situation where R and R_0 are compatible (in the sense of (11)). Suppose that θ is real-valued and that $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d N(0, \sigma^2)$ where $\widetilde{\theta}_n$ minimizes the working criterion R_0 for the full sample. If the bootstrap is asymptotically valid (to first order) then the (random) bootstrap distribution of $\sqrt{n}(\widehat{\theta}_n^{*b} - \widetilde{\theta}_n)$ will converge in probability or almost surely to a (fixed) $N(0, \sigma^2)$ distribution. For example, we might have

$$P^*(\sqrt{n}(\widehat{\theta}_n^{*b} - \widetilde{\theta}_n) \in A) \to_p \int_A \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\, dx = \int_A \phi(x)\, dx$$

for any set A. Since $\sqrt{n}(\widetilde{\theta}_n - \theta_0)$ is constant from the point of view of bootstrap sampling, it follows that for large n,

$$P^*(\sqrt{n}(\widehat{\theta}_n^{*b} - \theta_0) \in A) = Q_n^*(A) \approx \int_A \phi\left(x - \sqrt{n}(\widetilde{\theta}_n - \theta_0)\right)\, dx = \int_A \phi_n^*(x)\, dx$$

where $\phi_n^*(x)$ is now a Gaussian density function whose mean is a random variable. Since $\sqrt{n}(\widetilde{\theta}_n - \theta_0) = O_p(1)$, it is easy to verify that condition (c) will hold provided that the convergence of $Q_n^*(A)$ to its limit is uniform over all sets A.

We note that this theorem does not cover all of the examples in the paper. For instance the least median of squares regression is not $\sqrt{n}$ consistent, and hence we cannot apply this theorem to establish asymptotic validity of bumping.

7 Confidence sets for models

The bumping idea can be used to form confidence sets for models. Here we choose the working criterion R_0 equal to the target criterion R. Let $\hat{\theta}^{*(1)}, \ldots \hat{\theta}^{*(B)}$ be the values of $\hat{\theta}^{*b}$ ordered from smallest to largest value of $R(\mathbf{z}, \hat{\theta}^{*b})$. Then given a level α and letting $V = [\alpha \cdot B]$, we define

$$\hat{C}_{1-\alpha} = \{\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots \hat{\theta}^{*(V)}\} \tag{12}$$

$\hat{C}_{1-\alpha}$ is an approximate $1-\alpha$ confidence set for the true model parameter θ. The operative idea here is twofold. First we use $R(\mathbf{z}, \theta)$ to order the models $\hat{\theta}^{*b}$; secondly we use bumping to give us a distribution of parameter values from which we form the confidence set.

Example 5. *Prostate cancer data.*

This data comes from a study by Stamey *et. al.* (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy. The factors were:

1. log cancer volume (lcavol)
2. log prostate weight (lweight)
3. age
4. log of benign prostatic hyperplasia amount (lbph)
5. seminal vesicle invasion (svi)
6. log of capsular penetration (lcp)
7. Gleason score (gleason)
8. percent Gleason scores 4 or 5 (pgg45)

We fit a linear model to the log of prostate specific antigen (lpsa) after first standardizing the predictors. Using cross-validation we determined the best subset size to be three; for the original training data the best subset of size three was (1, 2, 5).

Figure 3 shows the results of 100 bumped models, with the subset size fixed at three. The model (1,2,5) is easily favored, followed by (1,4,5) and (1,2,8). The remaining models are much less likely.

Note that Figure 3 shows only which predictors have non-zero coefficients, and doesn't indicate the size of the coefficients. Figure 4 shows boxplots of the 8 coefficients over the 90 models in the confidence set. Predictors 1 and 5 show the strongest consistent effects.

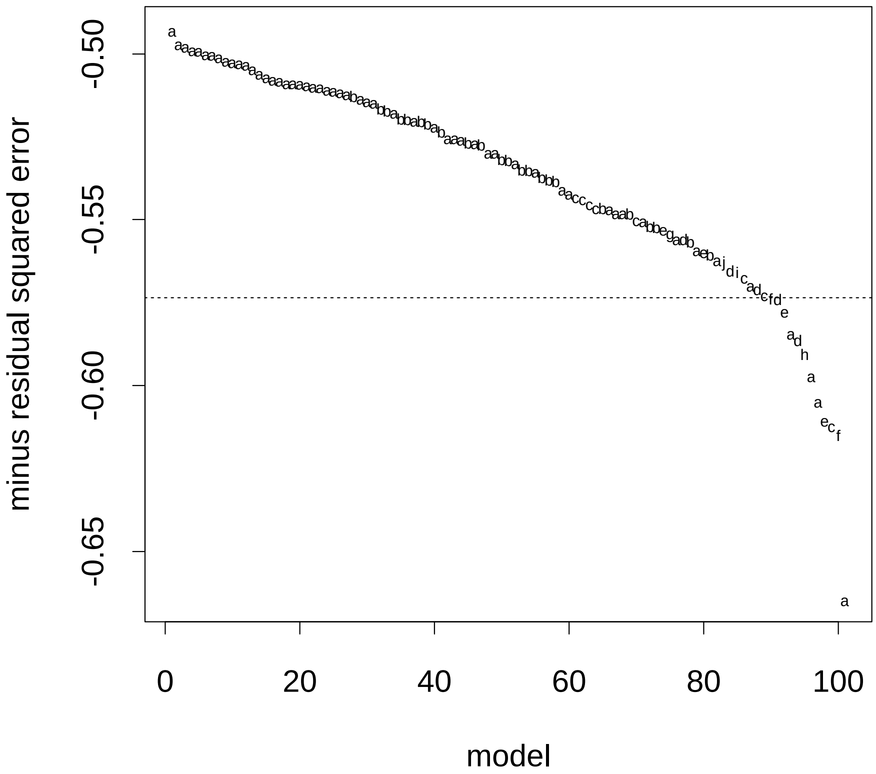

Figure 3: Minus residual squared error for 100 bumped models; legend a=(1,2,5), b=(1,4,5), c=(1,2,8), d=(1,2,3), e=(1,2,6), f=(1,2,7), g=(1,2,4), h=(1,3,4), i=(1,4,8), j=(1,5,8); 90% confidence set lies above the broken line. The frequency of models (a, b, c, d, e, f, g, h, i, j) in the top 90 were (51, 23, 7, 3, 2, 1, 1, 0, 1, 1)

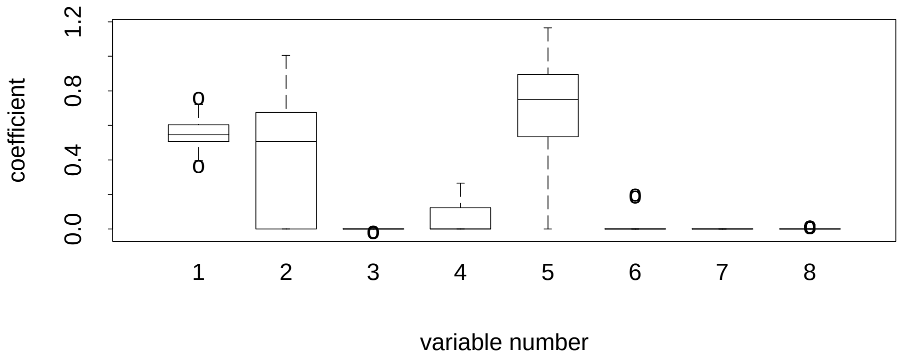

Figure 4: Boxplots of coefficients for the top 90 bumped models, prostate cancer example.

Display and interpretation of the proposed confidence set would be a challenge for many models. For example, with a tree-based model one might display the primary splitting variables among the top trees, or the number of occurrences of each variable anywhere in the trees. Note also that we have conditioned on the model complexity in forming the confidence set: in the above example we fixed the subset size at 3. One could allow this number to vary, to account for the additional uncertainty. However the number of different models in the confidence set would grow considerably making interpretation even more difficult.

In a special case, the set $\hat{C}_{1-\alpha}$ corresponds to a symmetrized version of the bootstrap percentile confidence interval (see Efron & Tibshirani (1993), chapter 13). For real-valued parameters θ, the $1-\alpha$ percentile interval is defined by $(P^{\alpha/2}, P^{1-\alpha/2})$, where P^{γ} is the γ-percentile of the $\hat{\theta}^{*b}$ values. Now if θ is real-valued, and $R(\mathbf{z}, \theta)$ is a convex function, symmetric around its minimum $\hat{\theta}$, then it is easy to see that $\hat{C}_{1-\alpha}$ equals the percentile interval if the bootstrap distribution of $\hat{\theta}^*$ is symmetric around $\hat{\theta}$. In other words, $\hat{C}_{1-\alpha}$ ignores the direction of departure of $\hat{\theta}^*$ from $\hat{\theta}$. When θ is not real-valued, as is the case in this section, the percentile interval is not defined; but the confidence set $\hat{C}_{1-\alpha}$ is well-defined, using R to induce an ordering on the θ values. In the real-valued case, the percentile interval captures shape information that may result in assymmetry of the confidence interval about $\hat{\theta}$; when θ is not real-valued, the notion of assymmetry is not meaningful.

7.1 Some theory of confidence sets obtained from bumping

To motivate the interval $\hat{C}_{1-\alpha}$, consider the case of linear least squares regression with p predictors having true value θ_0, and assume that $y_i \sim N(\mathbf{x}_i^T\theta, \sigma^2)$ with σ^2 known. Letting $R(\mathbf{z}, \theta) = \sum(y_i - \mathbf{x}_i^T\theta)^2$, we have

$$R(\mathbf{z}, \hat{\theta}) - R(\mathbf{z}, \theta_0) \sim \sigma^2\chi_p^2 \qquad (13)$$

where sampling is under $y_i \sim N(\mathbf{x}_i^T\theta_0, \sigma^2)$. An exact confidence set for θ can be obtained by inverting (13), but let's consider another (roughly equivalent) approach. We can find a monotone increasing transformation $g[\cdot]$ so that $g[R(\mathbf{z}, \hat{\theta})] - g[R(\mathbf{z}, \theta_0)]$ is approximately normal with mean zero and constant variance (the transformation is roughly a cube-root: see Efron (1987)). Using this we can construct a standard normal confidence set on the g scale, and transform it back to obtain a set for θ. Now if the bootstrap analogue of (13) also holds

$$R(\mathbf{z}, \hat{\theta}^*) - R(\mathbf{z}, \hat{\theta}) \sim \sigma^2\chi_p^2, \qquad (14)$$

(at least as $B \to \infty$), then the bumping set (12) will coincide with the normal set on the g scale. Since percentiles map to percentiles under monotone increasing transformations, the bumping set will also coincide with the exact set on the original θ scale.

We now formalize the preceding argument for the general case. It is convenient to talk in terms of maximizing rather than minimizing a cost function, so we work with $-R(\mathbf{z}, \theta)$. Denote by $C_{1-\alpha}$ the "true" bootstrap set obtained by letting $B \to \infty$ in the definition of $\hat{C}_{1-\alpha}$. Specifically, suppose α is such that there exists a value $r^{(1-\alpha)}$ satisfying [1]

$$\text{Prob}_*(-R(\mathbf{z}, \hat{\theta}^*) \geq r^{(1-\alpha)}) = 1 - \alpha. \tag{15}$$

Here and henceforth, Prob_* refers to bootstrap sampling with the training set $\mathbf{z}$ fixed. Then we define

$$C_{1-\alpha} = \{\hat{\theta}^*; -R(\mathbf{z}, \hat{\theta}^*) \geq r^{(1-\alpha)}\} \tag{16}$$

Let θ_0 be the true value of θ and suppose there exists a monotone increasing transformation $g[\cdot]$ such that

$$\begin{aligned} h(\hat{\theta}, \theta_0) &= g[-R(\mathbf{z}, \hat{\theta})] - g[-R(\mathbf{z}, \theta_0)] \sim H \\ h(\hat{\theta}^*, \hat{\theta}) &= g[-R(\mathbf{z}, \hat{\theta}^*)] - g[-R(\mathbf{z}, \hat{\theta})] \sim H \end{aligned} \tag{17}$$

where H is a fixed distribution, symmetric around zero. In the first line of (17), sampling is under the model θ_0; in the second line sampling is under model $\hat{\theta}$, with $\hat{\theta}$ is fixed. Letting $H^{(1-\alpha)}$ be the $1 - \alpha$ percentile of the H, a confidence set of the desired coverage $1 - \alpha$ is

[1] This assumption is made to avoid a technical complication: in the nonparametric bootstrap case, due to discreteness (15) will hold exactly only for some α.

given by

$$D_{1-\alpha} = \{\theta; h(\hat{\theta}, \theta) \leq H^{(1-\alpha)}\} \tag{18}$$

Finally, if we let Θ^* be all of the possible bootstrap values of $\hat{\theta}^*$, we have the result

$$C_{1-\alpha} = D_{1-\alpha} \cap \Theta^* \tag{19}$$

In other words, the confidence set $C_{1-\alpha}$ equals the correct set $D_{1-\alpha}$ except for the fact that under (nonparametric) bootstrap sampling, the bootstrap values might not obtain all of the θ values in $D_{1-\alpha}$. A proof appears in the Appendix. This result is a generalization of the percentile lemma that appears in Efron & Tibshirani (1993), page 173. Note that under parametric bootstrap sampling, Θ^* will often equal the entire parameter space, and hence $C_{1-\alpha} = D_{1-\alpha}$.

7.2 Relationship with likelihood-based confidence regions

Consider parametric sampling under a model governed by a p-dimensional parameter θ, with log-likelihood $\ell(\theta; \mathbf{z})$. Suppose we choose $R_0 = R = -2\ell(\theta; \mathbf{z})$. Then with $g[\cdot]$ in (17) equal to the identity, H is (asymptotically) equal to the χ^2_p distribution, and the set $D_{1-\alpha}$ is a standard likelihood ratio confidence region. The above arguments show that under parametric bootstrap sampling, the bumping region $C_{1-\alpha}$ agrees asymptotically with the likelihood ratio region. In finite samples, we would expect the bumping region to exhibit more accurate coverage since it does not rely on the asymptotic Chi-squared approximation.

In a similar fashion we can view $\hat{C}_{1-\alpha}$ as a Bayesian highest posterior density region, given a flat prior for θ.

8 Adaptive bumping

Rather than sampling with replacement from the data, it might be helpful to vary the sampling probabilities as a function of the current fitted model. By doing so in the right way, we might reduce the number of samples needed to find a good minimum of R. We investigate this briefly here.

In general, if we have generated b bumping sets, the sampling probabilities for the $b+1$st sample could be a function of $R(\mathbf{z}, \hat{\theta}^{*k})$, $R_0(\mathbf{z}, \hat{\theta}^{*k})$ for $k = 1, 2, \ldots b$, or any of the quanities that make up R and R_0. Here is a specific proposal that seems sensible. We restrict attention to scenarios where the working function $R_0(\mathbf{z}, \theta)$ is a sum over the observations, that is $R_0(\mathbf{z}, \theta) = \sum r_0(z_i, \theta)$ for some function r. Then we model the sampling probabilities for the ith data point in the $b+1$st sample as

$$p_{ib} = f[r_0(z_i, \hat{\theta}^{*b}), R(\mathbf{z}, \hat{\theta}^{*b})] \tag{20}$$

where f is a function to be chosen.

Here is an example. We generated data from the model

$$Y_i = X_i + .1Z_i + 3 \cdot I(i > N - 5); \quad i = 1, 2, \ldots N \tag{21}$$

where X is uniform on $[0, 1]$ and Z is standard normal. A plot of a typical sample is shown in Figure 5. We consider least median of squares fitting as in section 4, where $R(\mathbf{z}, \theta) = \text{median}(y_i - \mathbf{x}_i^T\theta)^2$, $R_0(\mathbf{z}, \theta) = \sum(y_i - \mathbf{x}_i^T\theta)^2$, $r_0(z_i, \theta) = (y_i - \mathbf{x}_i^T\theta)^2$.

What form should we use for the probablities f? Depending on the current fitted line, the five outliers could have large or small values of the squared residual $r_0(z_i, \theta)$. We chose

$$p_{ib} = \frac{k}{1 + \exp(5q_{ib})} \tag{22}$$

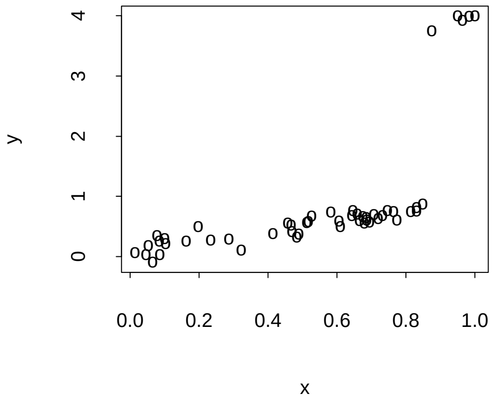

Figure 5: Typical sample for adaptive bumping example.

where k is chosen so that $\sum_i p_{ib}$=1 and three choices were tried for q_{ib}, namely

$$q_{ib} = C[r_0(z_i, \hat{\theta}^{*b}) - R(\mathbf{z}, \hat{\theta}^{*b})]/R(\mathbf{z}, \hat{\theta}^{*b}) \tag{23}$$

where $C = +, -$, or absolute value. These three correspond to downweighting data points with large, small and (large or small) squared residuals. The value 5 was picked empirically so that each bumping sample had a reasonable number (.6N) of unique data points.

We generated 50 realizations from model (21) for dataset sizes $N = 20$ and 50, and tried the various bumping procedures on each. Figure 6 shows the number of bumping samples needed to obtain a sample containing none of the outliers, using standard bumping and the three adaptive bumping procedures. Downweighting the large residuals worked very well for N=50, but got stuck in local minima for N=20. Downweighting small residuals worked well, and downweighting both large and small worked best, requiring only about 2

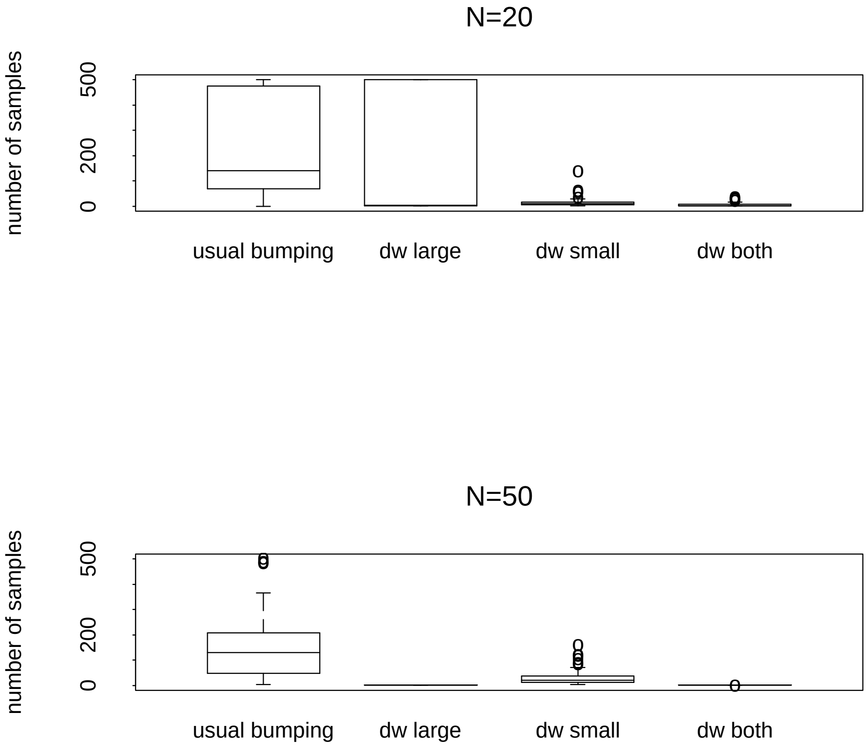

Figure 6: Adaptive bumping for resistant fitting: number of samples B required to obtain a sample containing no outliers. The boxes correspond to standard bumping, and downweighting data points with large, small or (large and small) squared residuals. B was truncated at 500 for the graphical display.

or 3 bumping samples on the average.

9 Some further issues

9.1 Relationship to empirical likelihood

Functions produced by the bumping procedure (like that in Figure 3) are a type of non parametric likelihood for the model parameters θ. In this respect, bumping is similar to the empirical likelihood method of Owen (1988). A key difference in bumping is the use of the real-valued function $R(\mathbf{z}, \theta)$ to induce a one-dimensional ordering of the models. The empirical likelihood for a p-dimensional model parameter is a p-dimensional surface: for the models considered here, this would be very difficult to compute and interpret.

9.2 Number of bootstrap samples

In our experiments, we found that 20 or 30 bootstrap samples were adequate for the local minima search and resistant. For other application of bumping (constrained optimization, confidence sets) a larger number might be needed, at least in the 100-200 range.

Acknowledgments

We would like to thank David Andrews, Jerry Friedman, Trevor Hastie, Radford Neal, and John Rice for enjoyable and helpful discussions. The authors were supported by the Natural Sciences and Engineering Research Council of Canada and an IRIS center of Excellence grant.

Appendix: proof of result (19).

We assume (17), namely

$$\begin{aligned} h(\hat\theta, \theta_0) &= g[-R(\mathbf{z}, \hat\theta)] - g[-R(\mathbf{z}, \theta_0)] \sim H \\ h(\hat\theta^*, \hat\theta) &= g[-R(\mathbf{z}, \hat\theta^*)] - g[-R(\mathbf{z}, \hat\theta)] \sim H \end{aligned}$$

and (15)

$$\mathrm{Prob}_*(-R(\mathbf{z}, \hat\theta^*) \geq r^{(1-\alpha)}) = 1 - \alpha.$$

Then

$$\begin{aligned} \mathrm{Prob}[R(\mathbf{z}, \theta) \geq r^{1-\alpha)}] &= \mathrm{Prob}[g(-R(\mathbf{z}, \theta)) \geq g(r^{(1-\alpha)})] \\ &= \mathrm{Prob}[g(-R(\mathbf{z}, \hat\theta)) - g(-R(\mathbf{z}, \theta))] < g(-R(\mathbf{z}, \hat\theta)) - g(r^{(1-\alpha)})) \\ &= \mathrm{Prob}_*[g(-R(\mathbf{z}, \hat\theta^*)) - g(-R(\mathbf{z}, \hat\theta))] < g(-R(\mathbf{z}, \hat\theta^*)) - g(r^{(1-\alpha)})) \\ &= \mathrm{Prob}_*[g(-R(\mathbf{z}, \hat\theta^*)) - g(-R(\mathbf{z}, \hat\theta))] \geq -g(-R(\mathbf{z}, \hat\theta^*)) + g(r^{(1-\alpha)})) \\ &= \mathrm{Prob}_*[g(-R(\mathbf{z}, \hat\theta^*)) \geq g(r^{(1-\alpha)})] = 1 - \alpha \end{aligned}$$

The next to last line follows because of the assumed symmetry of H around zero. This shows that the R endpoints of $C_{1-\alpha}$ and $D_{1-\alpha}$ agree, and hence (19) follows.

References

Breiman, L. (1996), 'Bagging predictors', *Machine Learning*.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Clark, L. & Pregibon, D. (1991), Tree-based models, *in* J. Chambers & T. Hastie, eds, 'Statistical models in S', Wadsworth.

Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *Annals of Statistics* **7**, 1–26.

Efron, B. (1987), 'Better bootstrap confidence intervals (with discussion)', *J. Amer. Statist. Assoc.* **82**, 171–200.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall.

Kelly, C. & Rice, J. (1990), 'Monotone smoothing with application to dose-response curves and the assessment of synergism', *Biometrics* **46**, 1071–1085.

Kim, J. & Pollard, D. (1990), 'Cube root asymptotics', *Annals of Statistics* **11**(1), 191–219.

Owen, A. (1988), 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika* **75**, 237–249.

Rousseeuw, P. (1984), 'Least median of squares regression', *J. Amer. Statist. Assoc.* **79**, 871–880.