Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 669–683
journal of statistical planning and inference
www.elsevier.com/locate/jspi

On bagging and nonlinear estimation

Jerome H. Friedman[a,*], Peter Hall[b]

[a] *Statistics Department, Stanford University, Stanford, CA 94305, USA*
[b] *CSIRO Mathematical Sciences and Centre for Mathematics and its Application, Australian National University, Canberra ACT 0200, Australia*

Available online 4 August 2006

Abstract

We propose an elementary model for the way in which stochastic perturbations of a statistical objective function, such as a negative log-likelihood, produce excessive nonlinear variation of the resulting estimator. Theory for the model is transparently simple, and is used to provide new insight into the main factors that affect performance of bagging. In particular, it is shown that if the perturbations are sufficiently symmetric then bagging will not significantly increase bias; and if the perturbations also offer opportunities for cancellation then bagging will reduce variance. For the first property it is sufficient that the third derivative of a perturbation vanish locally, and for the second, that second and fourth derivatives have opposite signs. Functions that satisfy these conditions resemble sinusoids. Therefore, our results imply that bagging will reduce the nonlinear variation, as measured by either variance or mean-squared error, produced in an estimator by sinusoid-like, stochastic perturbations of the objective function. Analysis of our simple model also suggests relationships between the results obtained using different with-replacement and without-replacement bagging schemes. We simulate regression trees in settings that are far more complex than those explicitly addressed by the model, and find that these relationships are generally borne out.

MSC: Primary 62G09; secondary 62E20

Keywords: Bias; Bootstrap; Half-sampling; Regression tree; Variance reduction; With-replacement sampling; Without-replacement sampling

1. Introduction

Bagging was introduced by Breiman (1996) as a means for improving the accuracy of estimators of functions $\theta(\mathbf{x})$ of data $\mathbf{x} = \{x_1, \ldots, x_N\}$,

$$\hat{\theta}(\mathbf{x}) = \arg \min_{\theta(\mathbf{x}) \in \Theta} L(\theta(\mathbf{x})).$$

Here, Θ denotes a function class representable by the estimator, such as neural networks or decision trees. The objective function $L(\theta(\mathbf{x}))$ is a data-based estimate of the expected value of some functional such as negative log-likelihood or other loss function. 'Bagging' involves repeatedly drawing random resamples $\mathbf{x}_b$ of the data, and either optimizing the value of $L(\theta(\mathbf{x}_b))$ averaged over the resamples, or averaging the resample values of $\hat{\theta}$. That is, we define the bagged

* Corresponding author.
E-mail addresses: jhf@stat.stanford.edu (J.H. Friedman), peter.hall@anu.edu.au (P. Hall).

0378-3758/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.jspi.2006.06.002

estimator to be either

$$\hat{\theta}_{\mathrm{bagg}}(\mathbf{x}) = \arg \min_{\theta(\mathbf{x})\in\Theta} \frac{1}{B}\sum_{b=1}^{B} L(\theta(\mathbf{x}_b)) \quad \text{or} \quad \hat{\theta}_{\mathrm{bagg}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}(\mathbf{x}_b), \tag{1.1}$$

where $\hat{\theta}(\mathbf{x}_b)$ is the version of $\hat{\theta}(\mathbf{x})$ computed from the bth resample. Since its introduction, considerable evidence has been accumulated that clearly demonstrates the effectiveness of bagging some estimators, such as decision trees and neural networks; it is now routinely used. The underlying reasons for its success have been less clear.

Bagging the objective function $L(\theta)$ often has the effect of reducing stochastic 'bumpiness' from which the function tends to suffer. As a result the bagged objective function is better approximated locally, in the neighbourhood of the true parameter value, by a paraboloid whose optimum occurs relatively close to the linear component of the statistic. Likewise, the operation of bagging the estimator itself often reduces variability by averaging out stochastic fluctuations caused by bumps; see for example Breiman (1999). These empirical results raise the question of whether the performance of bagging can be understood relatively simply by theoretically modelling, in an elementary and intuitive but nevertheless mathematically rigorous way, the effects of stochastic bumps on an essentially quadratic objective function. In the present paper we suggest such a model, describe the insights that its analysis provides into bagging, and relate that to the results of numerical experiments in the more complex, and less readily accessible, setting of regression trees.

In particular, in Section 2 we give an elementary argument that points to why bagging works in contexts that are substantially more complex than that of the model on which the argument is based. Of course, specific problems can be addressed in greater detail. That is the route taken, to good effect, by Bühlmann and Yu (2000) in their more sophisticated account, based on cube-root asymptotics, of decision trees and related topics. But we argue that the general principles behind the performance of bagging are less elaborate, and may be understood more readily. To make this point we construct a naive model for a quadratic surface with randomly located bumps, and discuss the performance of both types of bagging (see (1.1)) in reducing the effects that bumpiness has on performance of the estimator. From this viewpoint, two main requirements emerge as the key to reducing variance and mean-squared error by bagging.

First, bagging reduces variance if the shapes of the bumps provide opportunities for cancellation. Sinusoidal bumps are a good example—shifting the location of a sinusoid (by computing either the objective function or the estimator for different resamples), and forming the average over the different results, affords considerable scope for reducing variability. On the other hand, if the stochastic 'contamination' that causes the objective function to depart from a quadratic is in the form of a non-oscillatory function, then surprisingly, averaging over it will primarily increase the *variance*, not the bias, of the bagged estimator.

Second, if the contaminating bumps are not symmetric in shape then bagging will introduce a new bias term, and may increase variance as well. Intuitively, this property is to be expected. Bagging works by producing additional noise (conditional on the data), and then averaging over it; asymmetries in the bumps result in the added noise being averaged in a biased way, giving rise to the new bias term.

The remarks just above call to mind the known result that bagging can substantially increase the bias of nonlinear estimators, not least because it adds an extra quadratic term. In particular, if we take the square of the sample mean $\bar{X}$ to be an estimator of the square of the population mean μ, then its conventional bagged form (i.e. the expected value of the squared bootstrap mean $(\bar{X}^*)^2$, conditional on the data) has virtually twice the bias of $\bar{X}^2$, since

$$E\{(\bar{X}^*)^2\} - \mu^2 = (2 - N^{-1})\{E(\bar{X}^2) - \mu^2\},$$

where N denotes sample size. If the noisy perturbations that exacerbate variability of an estimator have a significant quadratic component (deriving from a cubic term in the corresponding contaminations of the objective function) then the perturbations will make themselves felt as inflated bias and possibly increased variance in the bagged estimator, potentially with dire consequences. Remarks made in the previous paragraph, about the need for the high-variability contaminating perturbations to be reasonably symmetric, reflect this result. Indeed, our mathematical arguments in Section 4 show that the requisite condition for symmetry is precisely that local cubic terms in contamination of the objective function vanish, or equivalently, the contaminants add no new quadratic terms to the estimator.

The latter result links our work to recent research of Buja and Steutzle (2000a,b), who studied properties of bagged quadratics and related functions. One of their results, that bagging can increase mean-squared error in part by increasing

bias, is effectively equivalent to our observation that asymmetric departures from linearity at the level of quadratic terms in the estimator, or cubic terms in the objective function, can result in poor performance of the bagged estimator.

One way of presenting these results is in the form of Taylor expansions of estimators, using them to elucidate the properties discussed above. However, the properties follow much more transparently from analysis of the estimating equation in the presence of stochastic contamination of the objective function. There, the variance-reducing effect of bagging when it works, or variance-inflating effect when bagging fails, appear as factors that multiply the term in the equation that adds nonlinear variability to the estimator. Provided the contamination adds no extra quadratic terms to the estimating equation (i.e. adds no cubic terms to the objective function), the factor is less than 1 when bagging works. And it is greater than 1 when bagging fails. We shall, however, also develop theory in more general settings; see Section 4. There we shall use a detailed Taylor-expansion approach to confirm results that are intuitively clear from the estimating function viewpoint.

The simple theoretical model on which our results are based is introduced in Section 2.1, and our main conclusions are drawn there. They foreshadow results about relative performances of different approaches to bagging, based on with-replacement and without-replacement resampling, respectively. These techniques are discussed in Section 2.2, where they are linked to the conclusions in Section 2.1. It is argued, on the basis of on our theory, that certain different bagging approaches (for example, n-out-of-n with-replacement bagging, and $\frac{1}{2}n$-out-of-n without-replacement bagging) can be expected to perform similarly.

Section 3 takes this matter up in a more complex setting, by treating substantially more sophisticated problems than can be addressed explicitly at the level of Section 2. It is shown there that some of the main conclusions of our theoretical analysis are borne out by numerical work in the case of regression trees. Our theoretical model for the effects of bagging on estimator performance is admittedly a toy one, and does not correspond explicitly to more complex settings were bagging is generally used. But there is nevertheless significant connection between the levels of performance that our model suggests, and those observed numerically.

With-replacement resampling is of course in the spirit of the contemporary bootstrap. Without-replacement methods were employed in early approaches to resampling, for example those of Mahalanobis (1946) and Hartigan (1969, 1971). McCarthy (1966) was an early proponent of without-replacement resampling using half-samples. Efron (1982) discussed related issues, including (Efron, 1982, pp. 62–64) the estimation of variance by half-sampling.

2. Interpretations of bagging

2.1. Noisy perturbations of objective functions

The theoretical arguments given here and in Section 4 require only Taylor expansion, and so there are no more than notational differences between one- and arbitrary-dimensional cases. The numerical study in Section 3 will explore high-dimensional settings. In the present section, to make our technical manipulations more transparent, we shall confine attention to one dimension.

In an ideal, no-noise case the objective function L will be a simple quadratic having its minimum at the true parameter value. Without loss of generality the true value is 0, and $L(\theta) = \frac{1}{2}\theta^2$. The 'estimator' of θ, $\hat{\theta} = \arg\ \min_\theta L(\theta)$, is then of course 0. Suppose, however, that the simple quadratic is contaminated by a noisy, 'bumpy' function Ψ: $L(\theta) = \frac{1}{2}\theta^2 + \Psi(\theta)$. If the bumps in Ψ occur consistently in the same place then the effect of the contamination is often more systematic than stochastic, and bagging cannot be expected to be of much help.

We shall instead propose a simple model for Ψ in which the bump *locations* vary randomly:

$$\Psi(\theta) = \psi(\theta + Z),$$

where Z, a function of the data $\mathcal{X}$, is a random variable. We have included stochastic variability in the contamination term Ψ, and not in the idealized quadratic term $\frac{1}{2}\theta^2$, because we are primarily modelling the case where stochastic variability of the nonlinear component of the estimator $\hat{\theta}$ tends to swamp that of the linear component.

The conventional estimator of θ is thus

$$\hat{\theta} = \arg\ \min_\theta \{\tfrac{1}{2}\theta^2 + \psi(\theta + Z)\},$$

i.e. $\hat{\theta}$ is defined by solving the estimating equation

$$\theta + \psi'(\theta + Z) = 0. \tag{2.1}$$

One form of the bagged estimator $\hat{\theta}_{\text{bagg}}$ is obtained by minimizing, with respect to θ, the function

$$E\{\tfrac{1}{2}\theta^2 + \psi(\theta + Z^*)|\mathcal{X}\} = \tfrac{1}{2}\theta^2 + E\{\psi(\theta + Z^*)|\mathcal{X}\},$$

where Z^* is the version of Z computed from a resample and $\mathcal{X}$ denotes the data. (Here, we have effectively taken the number of resamples B, in our discussion in Section 1, to be infinite.) That is, $\theta = \hat{\theta}_{\text{bagg}}$ solves

$$\theta + E\{\psi'(\theta + Z^*)|\mathcal{X}\} = 0. \tag{2.2}$$

The other approach to bagging takes $\hat{\theta}^*$ to be the solution of $\theta + \psi'(\theta + Z^*) = 0$, and then puts $\hat{\theta}_{\text{bagg}} = E(\hat{\theta}^*|\mathcal{X})$. The approaches have virtually identical properties, but theory for the first is more transparent, so we devote more of our attention to it. Results for the second approach will be outlined in this Section, with mathematical details given in Section 4.2.

Assuming the perturbation Z has relatively small variance, we may Taylor-expand the estimating equation (2.2): $\hat{\theta}_{\text{bagg}}$ solves

$$\theta + \psi'(\theta + Z) + \tfrac{1}{2}\, E\{(\Delta^*)^2|\mathcal{X}\}\, \psi'''(\theta + Z) + \tfrac{1}{6}\, E\{(\Delta^*)^3|\mathcal{X}\}\, \psi^{(4)}(\theta + Z) + \cdots, \tag{2.3}$$

where $\Delta^* = Z^* - Z$ and we have taken $E(\Delta^*|\mathcal{X})$ to be 0. Suppose initially that the function ψ is sinusoidal, so that equally spaced bumps—with randomly varying locations—are added to the quadratic. The tendency of bagging to 'cancel out' bumps can be seen particularly clearly in this case, since now $\psi' = -\psi'''$ and so (2.3) reduces to:

$$\theta + (1 - \tfrac{1}{2}\,\hat{\sigma}^2)\, \psi'(\theta + Z) + \tfrac{1}{6}\, E\{(\Delta^*)^3|\mathcal{X}\}\, \psi^{(4)}(\theta + Z) + \cdots, \tag{2.4}$$

where $\hat{\sigma}^2 = E\{(\Delta^*)^2|\mathcal{X}\}$.

Assuming that the noise variable Z is something like a mean of n variables, perhaps each with relatively large variance, the third conditional moment $E\{(\Delta^*)^3|\mathcal{X}\}$ will be an order of magnitude smaller than $\hat{\sigma}^2$. (It will be of order n^{-2} if Z is the aforesaid mean, compared with order n^{-1} for $\hat{\sigma}^2$.) Therefore, the cubic term in (2.4) will be negligible, and so too will be the quartic, etc. (In Section 4.1 we shall make this mathematically explicit.) Thus, to a first approximation the estimating equation has changed from (2.1) to

$$\theta + (1 - \tfrac{1}{2}\,\hat{\sigma}^2)\, \psi'(\theta + Z) = 0. \tag{2.5}$$

The critical feature of (2.5) is of course the way in which bagging has reduced the main effect of the bumpy function, from $\psi'(\theta + Z)$ in the unbagged estimating equation (2.1) to $(1 - \frac{1}{2}\,\hat{\sigma}^2)\, \psi'(\theta + Z)$ after bagging. This results in improved performance, by reducing both variance and mean-squared error relative to those of the unbagged estimator $\hat{\theta}$. See Section 4.1 for a proof.

Eq. (2.5) also neatly quantifies the way in which the main effects of bagging may be predicted in terms of the size of the empirical variance term, $\hat{\sigma}^2$. In particular, if we bag in two quite different ways, for example, using with-replacement sampling on the one hand and without-replacement sampling on the other, with different resample sizes in the two respective cases; but if the value of $\hat{\sigma}^2$ is similar under both regimes; then the extents to which bagging reduces the effects of noisy bumps in the objective function may be predicted to be similar too, provided the stochastic variability associated with the bumpy function is not too great.

We make the latter caveat only because (2.4) may not be explicitly valid if Z is too highly variable. However, for bump functions such as the sinusoid, the variability of $\hat{\theta}$ will still be reduced, as may be seen from numerical experiments. Moreover, even our more general results (discussed later in this section) do not require the variance of Z to converge to 0 in order to be valid; they hold if the variance of Z is sufficiently small, but fixed. It is therefore incorrect to describe these effects as second-order ones. Recall that we are modelling problems where nonlinear perturbations of estimators are of first order. In particular, we have added the 'contaminant' $\psi(\theta + Z)$, which produces those perturbations, directly to a pure quadratic, which produces a noiseless linear component.

The arguments above are of course founded on the assumption that $\psi' = -\psi'''$. If ψ''' resembles ψ', rather than the negative of the latter, then the reverse conclusion will obtain: bagging will impair rather than enhance performance. Lying at the heart of this analysis is the question of whether the bump function ψ tends to self-correct for large perturbations from the origin, or whether large perturbations are actually reinforced by ψ. Note that bagging adds extra noise to the argument of ψ', since it replaces $\psi'(\theta + Z)$ by $\psi'(\theta + Z + \Delta^*)$. If the structure of ψ is such that, after taking conditional expectation, the effects of added noise tend to cancel those of existing noise, then bagging will tend to improve performance. Otherwise the impact of bagging may be deleterious, and both variance and mean-squared error can be increased.

To better appreciate what is going on here, let us assume for simplicity that ψ is an even function, so that it may be represented by

$$\psi(\theta) = \psi(0) + \tfrac{1}{2}\,\alpha\theta^2 + \tfrac{1}{6}\,\beta\theta^4 + \cdots \tag{2.6}$$

in the neighbourhood of the origin, where $\alpha = \psi''(0)$ and $\beta = \psi^{(4)}(0)$. To simplify discussion, suppose $\alpha > -1$ and $\alpha \neq 0$. (Indeed, $\alpha = 0$ implies that even the unbagged estimator is superefficient, while $\alpha < -1$ implies that the noiseless objective function $\frac{1}{2}\theta^2 + \psi(\theta)$ is concave downward, rather than upward, near the origin. In such cases, and often too for $\alpha = -1$, we can no longer regard $\theta = 0$ as the true parameter value.) If ψ is a sinusoid then α and β are of opposite signs, and this means that when $\psi(\theta)$ strays too far from $\psi(0)$ the quartic term in (2.6) tugs it back, tending to reduce the value of $\psi(\theta)$ relative to what it would be if α and β were of the same sign.

For general bump functions ψ, this 'self-correcting' property turns out to be fundamental: if the signs of α and β are different then adding extra noise to the bump function tends to smooth out existing perturbations, and as a result bagging reduces both variance and mean-squared error; but if the signs are the same then, for the analogous reason, bagging makes things worse. Moreover, this is true for bagging either the estimating equation or the estimator itself.

Provided the signs of α and β are different, the extent to which bagging reduces variance is virtually proportional to the (expected) value of $\hat{\sigma}^2$, just as is suggested by (2.5) in the simplest case where ψ is a sinusoid. Again this is true both for bagging the estimating equation and for bagging the estimator. See Section 4 for a detailed argument. Therefore, $\hat{\sigma}^2$ is quite generally the key to the amount by which bagging reduces variance. Numerical work in Section 3 will use this property to illustrate the link between our toy theory and more complex settings to which bagging is generally applied.

Matters are somewhat more complex if the function ψ is asymmetric, in particular, if $\psi'''(0) \neq 0$. There the operation of adding extra noise (in the form of Δ^*) and averaging over it tends to introduce bias, in addition to the impact it has on variability. As noted in Section 1, this result is to be expected. Once again, it holds for bagging the estimating equation and for bagging the estimator. It will be discussed in mathematical detail in Section 4.

More generally, these results apply to more complex models for bumps on the objective function, for example, the model $L(\theta) = \frac{1}{2}\theta^2 + \sum_i \psi_i(\theta + Z_i)$ for potentially correlated random variables Z_i (provided the bootstrap step correctly captures the main effects of the dependence structure). To appreciate why, note that if this were the model then, following the argument leading from (2.1) down to (2.5), we would replace (2.5) by

$$\theta + \sum_i \left(1 - \frac{1}{2}\,\hat{\sigma}_i^2\right) \psi_i'(\theta + Z_i) = 0,$$

where the factor $1 - \frac{1}{2}\,\hat{\sigma}_i^2$ reduces the main effect of the bumpy function from $\psi_i'(\theta + Z_i)$ before bagging to $(1 - \frac{1}{2}\,\hat{\sigma}_i^2)\,\psi_i'(\theta + Z_i)$ after. The keys to obtaining improved performance using bagging are: (a) the bumps should have the self-correcting property, so as to reduce variance, and (b) they should be reasonably symmetric, so as not to increase bias.

2.2. Bagging with or without replacement

The method of m-out-of-n with-replacement bagging involves drawing a resample $\mathcal{X}^*$, of size $m \leqslant n$, by sampling with replacement from a data set $\mathcal{X}$. In without-replacement bagging a resample $\mathcal{X}^\dagger$, in this case a subsample, of size $m \leqslant n - 1$ is drawn by sampling without replacement from $\mathcal{X}$.

To appreciate the effects these different methods have on variability of the bagging step, take $Z = n^{-1}\sum_i X_i$ where $\mathcal{X} = \{X_1, \ldots, X_n\}$ is a collection of independent and identically distributed random variables. Let $\hat{\tau}^2 = n^{-1}\sum_i (X_i - \bar{X})^2$,

where $\bar{X} = n^{-1}\sum_i X_i$, denote the variance of $\mathcal{X}$. Write $\bar{X}^*$ and $\bar{X}^\dagger$ for the means of the resamples $\mathcal{X}^*$ and $\mathcal{X}^\dagger$, respectively. Put $\rho_m = n/m \geqslant 1$. We shall show shortly that if $\rho_m \to \rho$ as $n \to \infty$, where $1 \leqslant \rho < \infty$ and $\rho > 1$ in the case of without-replacement bagging, then

$$E\{(\bar{X}^* - \bar{X})^2 | \mathcal{X}\} = n^{-1}\rho_m\,\hat{\tau}^2 \quad \text{and} \quad E\{(\bar{X}^\dagger - \bar{X})^2 | \mathcal{X}\} = n^{-1}\,(\rho_m - 1)\,\hat{\tau}^2 + \cdots, \tag{2.7}$$

where the 'remainder' denoted by '...' represents higher-order terms.

The relevance of these results is that if we are using m-out-of-n with-replacement bagging then the value of $\hat{\sigma}^2$, introduced in Section 2.1, equals the left-hand side of the first formula at (2.7); and if we are using m-out-of-n without-replacement bagging then it equals the left-hand side of the second formula. It therefore follows from (2.7), and the fact that the extent of variance reduction is virtually proportional to $\hat{\sigma}^2$, that if $\rho_{m_1} \approx \rho_{m_2} - 1$ for two resample sizes m_1 and m_2, then the variability of bagging algorithms using m_1-out-of-n with-replacement sampling, and m_2-out-of-n without-replacement sampling, may be expected to be similar. In particular, taking $m \sim \frac{1}{2}n$ in without-replacement bagging will tend to produce an estimator where the effect of the quadratic term is virtually identical to that in n-out-of-n with-replacement bagging. Numerical illustrations of this relationship will be provided in Section 3, in settings well beyond the simplified one of the model being considered here.

There is another, more intuitive interpretation of why n-out-of-n with-replacement bagging, and $\frac{1}{2}n$-out-of-n without-replacement bagging, perform similarly. Note that the effective size of an n-out-of-n with-replacement bootstrap resample $\mathcal{X}^*$, in terms of the amount of information it contains, is given by the ratio

$$\frac{(\sum_{i=1}^n N_i)^2}{\sum_{i=1}^n N_i^2} \sim \frac{n}{2}, \tag{2.8}$$

where N_i denotes the number of times the ith data value X_i is repeated in $\mathcal{X}^*$. In particular, the variance of the mean of N_i copies of independent and identically distributed random variables Y_i, for $1 \leqslant i \leqslant n$, is very nearly equal to twice the variance of the mean of the n independent random variables themselves, for large n.

Finally, we derive the second part of (2.7); the first part is straightforward. Define $\omega_m = (m-1)/(n-1)$. Provided $2 \leqslant m \leqslant n-1$ we obtain, by expressing $(\bar{X}^\dagger - \bar{X})^2$ as a double series,

$$\begin{aligned} E\{(\bar{X}^\dagger - \bar{X})^2\,|\mathcal{X}\} &= m^{-2}\,[m(m-1)\,E\{(X_1^\dagger - \bar{X})\,(X_2^\dagger - \bar{X})|\mathcal{X}\} + m\,E\{(X_1^\dagger - \bar{X})^2|\mathcal{X}\}] \\ &= m^{-2}\left\{\frac{m(m-1)}{n(n-1)}\sum\sum_{i_1 \neq i_2}(X_{i_1} - \bar{X})\,(X_{i_2} - \bar{X}) + \frac{m}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right\} \\ &= m^{-1}\,(1 - \omega_m)\,\hat{\tau}^2. \end{aligned}$$

The exact form of the second result at (2.7) has $\rho_m - 1$ replaced by $\rho_m - (1 - m^{-1})(1 - n^{-1})^{-1}$, in which the '...' terms in the second formula at (2.7) may be dropped. Since $n \sim \rho m$ then the difference between the exact form and the stated one is of smaller order than the first term on the right-hand side in the second formula.

3. Numerical experiments

In order to gain further insights we present the results of several simulation experiments. All involve estimating a function of a multivariate argument in noisy settings using regression trees. Regression trees (Breiman et al., 1984) represent a very nonlinear estimation method. Data $\{y_i, \mathbf{x}_i\}_1^n$ were generated according to the model

$$y_i = f(\mathbf{x}_i) + \sigma\varepsilon_i, \tag{3.1}$$

with each $\mathbf{x}_i$ independently generated from a 10-dimensional uniform distribution, $\mathbf{x}_i \backsim \mathrm{U}^{10}[0, 1]$. Each ε_i was randomly drawn from a standard normal distribution. Three 'target' functions $f(\mathbf{x})$ were considered:

- constant: $f(\mathbf{x}) = 0,\ \sigma = 1$,
- piecewise-constant: $f(\mathbf{x}) = \prod_{j=1}^5 1(x_j \geqslant 0.13),\ \sigma = 0.5$,
- linear: $f(\mathbf{x}) = \sum_{j=1}^5 j \cdot x_j,\ \sigma = 3$.

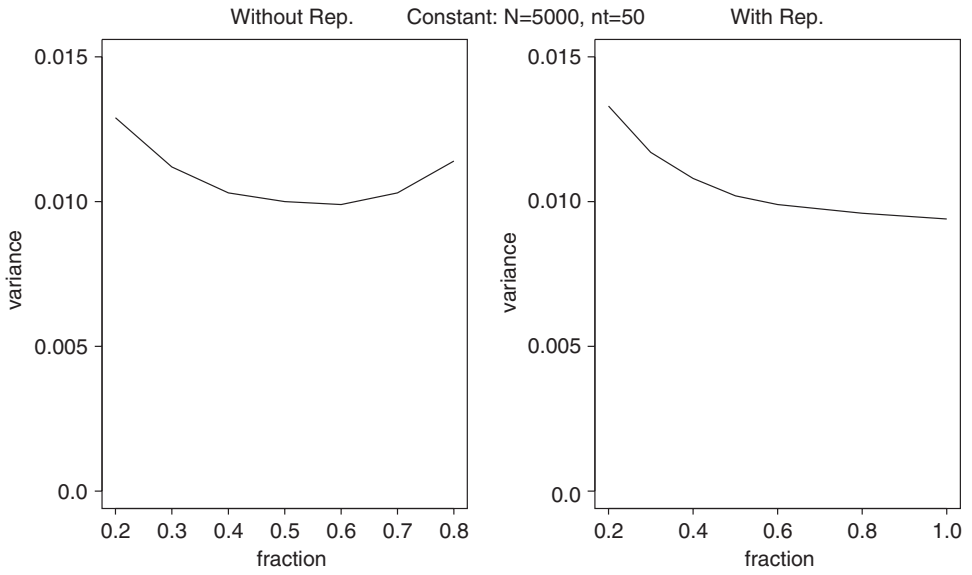

Fig. 1. Average variance as a function of sampling fraction m/n without (left frame) and with (right frame) replacement, for $n = 5000$, with a constant target. The variance of the unbagged estimate is 0.0856.

The truncation point 0.13 for the piecewise-constant function is chosen since $(1 - 0.13)^5 = \frac{1}{2}$, to two significant figures. This ensures that the piecewise-constant target equals 1 for half its volume, and equals zero for the other half.

Note that the last two targets are functions of only five of the 10 predictor variables, so that the others represent irrelevant 'noise' variables. Each generated data set was used to induce a 50-terminal node regression tree, producing a corresponding function estimate $\hat{f}(\mathbf{x})$. Average bias-squared

$$B^2 = E_{\mathbf{x}}\{f(\mathbf{x}) - E_{y\mathbf{x}}\hat{f}(\mathbf{x})\}^2 \tag{3.2}$$

and average variance

$$V = E_{y\mathbf{x}}\{\hat{f}(\mathbf{x}) - E_{y\mathbf{x}}\hat{f}(\mathbf{x})\}^2 \tag{3.3}$$

were computed by averaging over 100 independent trials for each experiment. In all experiments $B = 50$ bagging iterations were employed.

3.1. Constant target

In this setting we study the effect of various forms of bagging on variance alone, since here regression trees are unbiased. Fig. 1 shows the average variance (3.3) of the bagged regression tree estimator $\hat{f}(\mathbf{x})$ as a function of the sampling fraction m/n, without (left frame) and with (right frame) replacement. Training samples of $n = 5000$ were used to induce the trees. The variance of the original 'unbagged' estimate was 0.0856. Thus, for all sampling fractions shown, both types of bagging dramatically reduce variance. The optimal sampling fraction is approximately 0.6 for without replacement sampling and 1.0 for sampling with replacement. However, the curves are fairly flat near their optima, so that a choice is not critical. Note that smaller fractions require less computation.

These results reflect the theoretical conclusions reached in Section 2, in that (a) the variance of without-replacement bagging is approximately a U-shaped function of the sampling fraction, (b) variance in the with-replacement case is a decreasing function of the sampling fraction, and (c) a sampling fraction of $\frac{1}{2}$ in the case of without-replacement bagging produces almost the same variance as a sampling fraction of 1 in the with-replacement case.

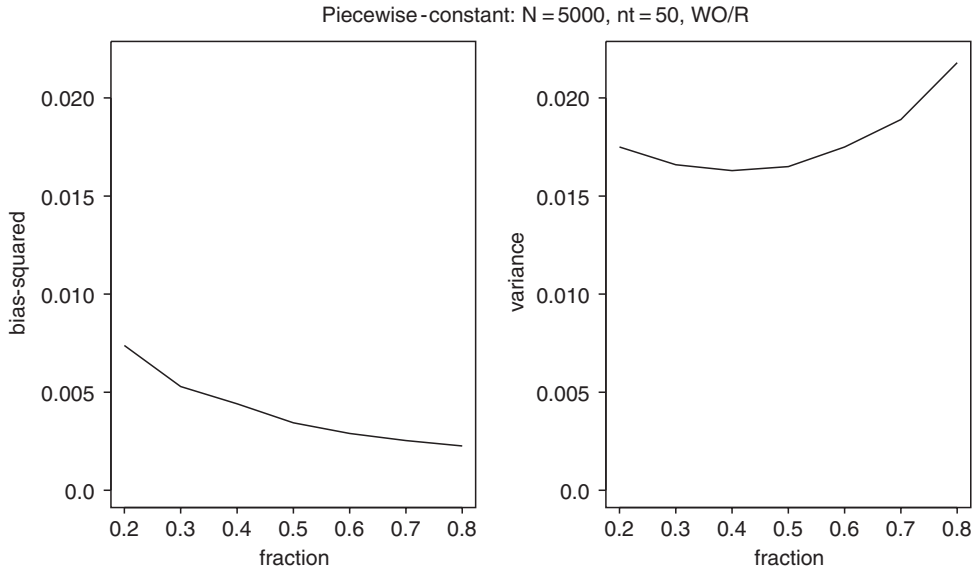

Fig. 2. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for without-replacement sampling and $n = 5000$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863.

3.2. Piecewise-constant target

This represents a situation in which regression trees have the potential to be unbiased because their approximations $\hat{f}(\mathbf{x})$ are piecewise-constant . The target $f(\mathbf{x})$ has the value 1 in the upper corner of a five-dimensional hyper-cube. The responses $\{y_i\}_1^n$ are roughly evenly divided between those that have $Ey_i = 1$, and those with $Ey_i = 0$. A six terminal node regression tree with five optimally placed splits can exactly represent the target.

Fig. 2 shows the bias-squared (3.2) (left frame) and variance (3.3) (right frame) of the without-replacement bagged estimate $\hat{f}(\mathbf{x})$ as a function of the sampling fraction. These values are reported in units of the global target variance $E_{\mathbf{x}}\{f(\mathbf{x}) - E_{\mathbf{x}}f(\mathbf{x})\}^2$. Training samples of $n = 5000$ were used. Although the target lies within the space of the approximating functions, one sees a small bias-squared that decreases with increasing sampling fraction. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863. This bias is due to the fact that the smaller training samples limit the size of the induced trees. Fig. 3 shows the corresponding results for $n = 500$. Here, unbagged trees had a bias-squared of 0.0325 and variance 0.2440. For the bagged trees, the bias-squared is seen to be comparable to the variance, and the latter increases monotonically with the sampling fraction. This monotonic increase of variance with sampling fraction also occurs for the constant target with small training samples (not shown). This effect is also evident with the larger $n = 5000$ sample (Fig. 2) in that the variance is minimized for smaller fractions than with the constant target (Fig. 1, left frame). Thus, with bagging there can be a bias-variance trade-off in choosing the sampling fraction m/n. As with any 'meta'-parameter that controls bias-variance trade-off, an optimal value can be estimated by minimizing an estimate of prediction error through cross-validation or a left out 'test' sample.

For completeness Figs. 4 and 5 show the corresponding results for $n = 5000$ and 500, respectively, sampling with replacement. One sees results similar to those for sampling without replacement in the interval $m/n \in [0.2, 0.5]$ of the latter.

The variance plots in Figs. 4 and 5 again lend support to the theoretical conclusions reached in Section 2. In particular, properties (a)–(c) noted in the last paragraph of Section 3.1 apply here, too, except that in the case of with-replacement bagging, variance is not a decreasing function of sampling fraction when the latter is large. Note, however, that in the case $n = 500$, shown in Fig. 5, variance is an increasing function of sampling fraction, but that by increasing sample size to the 'more asymptotic' value $n = 5000$ (Fig. 4) the function has almost, but not quite, turned around.

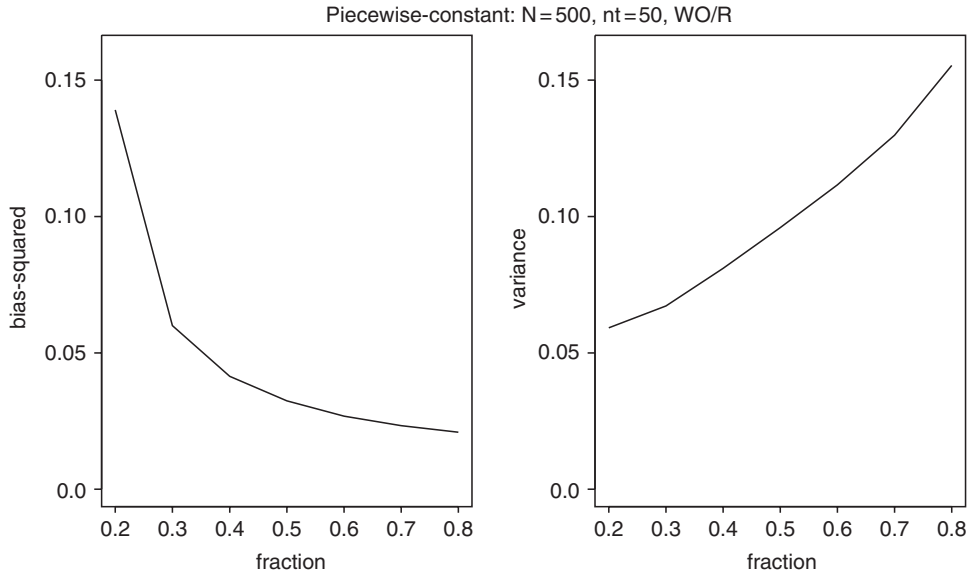

Fig. 3. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for without-replacement sampling and $n = 500$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0325 and variance 0.2440.

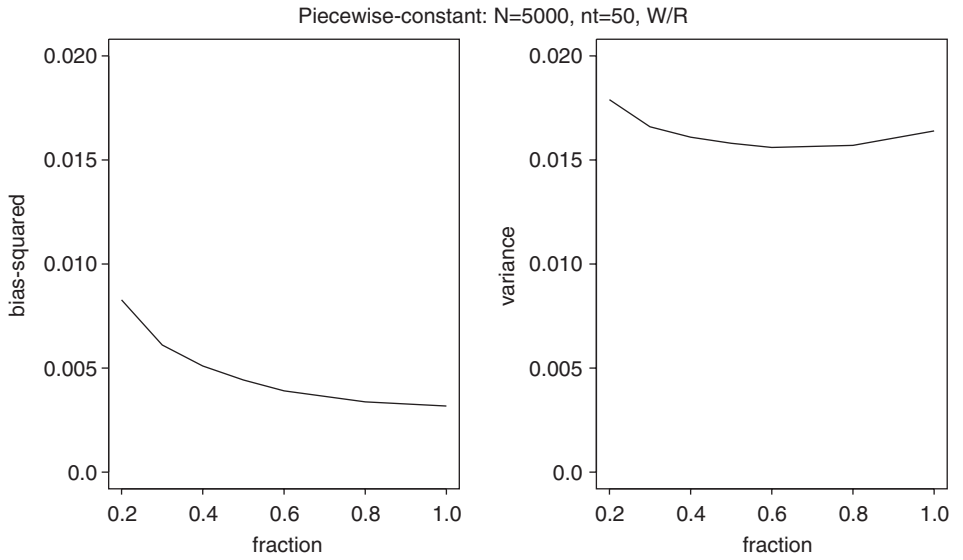

Fig. 4. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for with-replacement sampling and $n = 5000$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0025 and variance 0.0863.

3.3. Linear target

A linear function represents one of the most difficult targets for approximation by regression trees. It does not lie within the space of piecewise-constant functions and its contours are everywhere oblique to the coordinate axes. Fig. 6 shows the bias-squared (3.2) (left frame) and variance (3.3) (right frame) of the without-replacement bagged estimate $\hat{f}(\mathbf{x})$ as a function of the sampling fraction for $n = 5000$, again reported in units of the global target variance. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494. Here one sees the dramatic super-linear increase of variance

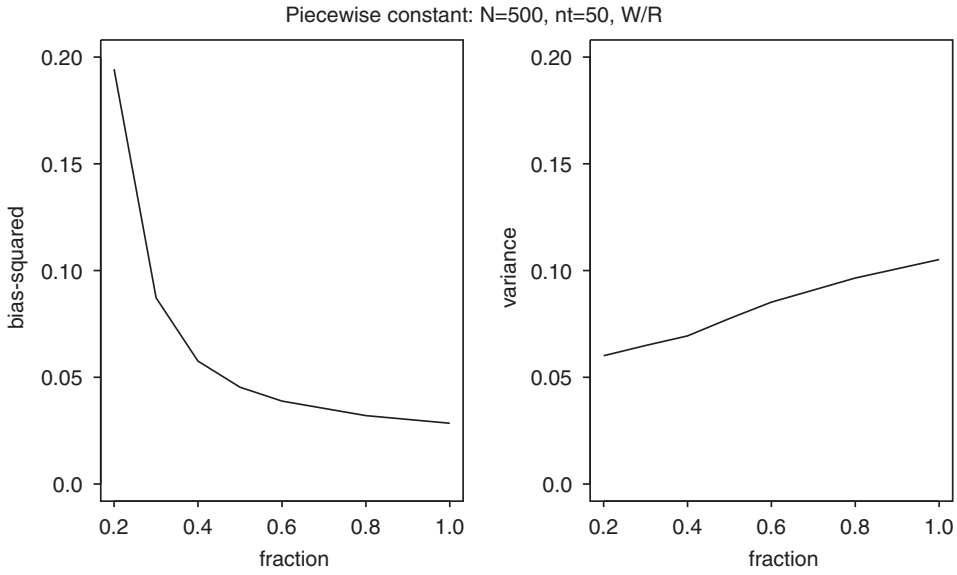

Fig. 5. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for with-replacement sampling and $n = 500$, with a piecewise-constant target. Unbagged trees have a bias-squared of 0.0325 and variance 0.2440.

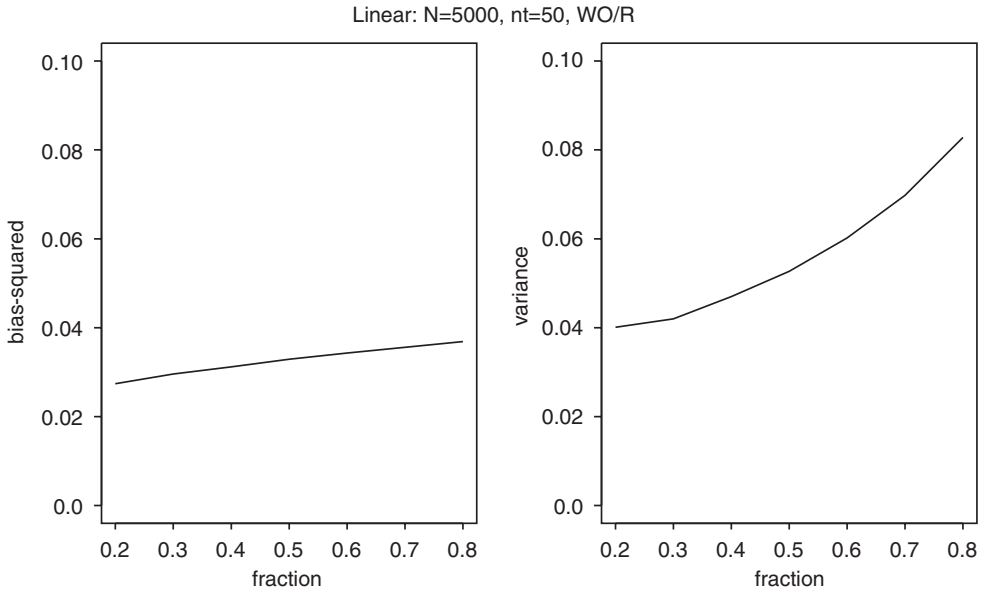

Fig. 6. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for without-replacement sampling and $n = 5000$, with a linear target. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494.

with sampling fraction characteristic of the *smaller* ($n = 500$) training samples above. Also, the bias-squared *increases* with sampling fraction. Here, bagging is reducing *bias-squared* as well as variance, with smaller sampling fractions producing the most improvement in both. Fig. 7 shows the corresponding plot for sampling with replacement. Again the results are similar to the left half (fraction $m/n \leqslant 0.5$) of the without replacement results.

Fig. 8 shows without-replacement results for much larger training samples $n = 50\,000$. Here, unbagged trees have an average bias-squared of 0.0775 and average variance of 0.0821. Comparing to the corresponding (unbagged) $n = 5000$

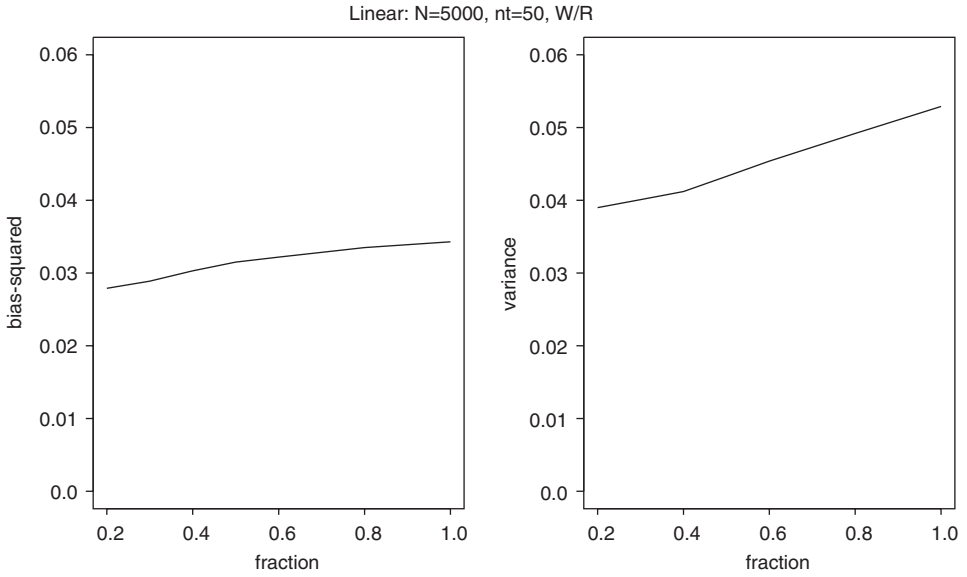

Fig. 7. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for with-replacement sampling and $n = 5000$, with a linear target. Unbagged trees have a bias-squared of 0.0402 and variance 0.2494.

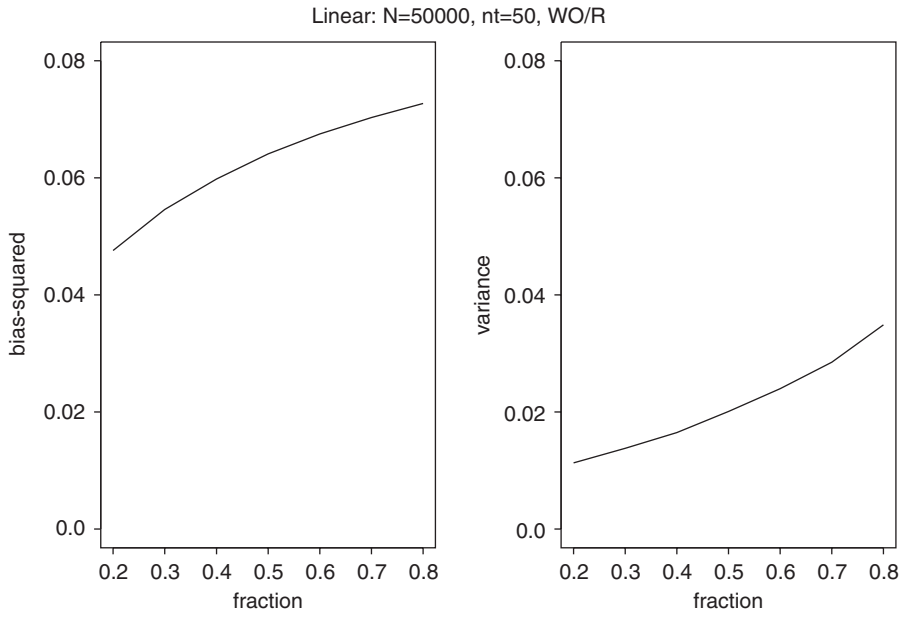

Fig. 8. Average bias-squared (left frame) and variance (right frame) as a function of sampling fraction m/n, for without-replacement sampling and $n = 50\,000$, with a linear target. Unbagged trees have an average bias-squared of 0.0775 and average variance of 0.0821.

results above, one sees that using the larger training sample reduces variance, but by a factor of about $1/\sqrt{10}$. However, the bias-squared has *increased* by almost a factor of two. The dependence of bias-squared and variance on sampling fraction is similar to that for $n = 5000$ shown in Fig. 6; they both decrease with decreasing sampling fraction m/n. However, here the bias-squared dominates mean-squared error of the bagged trees.

Although perhaps counter intuitive, the increase in bias-squared with larger samples for fixed-sized regression trees is easy to understand. Fig. 9 illustrates the concept in an idealized setting. Shown is a hypothetical asymptotic ($n = \infty$)

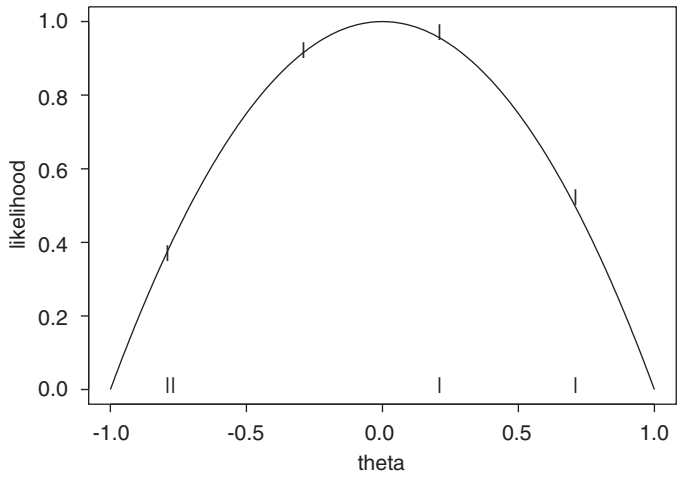

Fig. 9. Hypothetical asymptotic likelihood and an estimator that can only realize a discrete set of values.

likelihood as a function of a parameter θ with maximum at $\theta_0 = 0$. Suppose an estimator $\hat{\theta}$ that can realize a discrete set of values $\hat{\theta} \in \{\hat{\theta}_i\} = \{-0.8, -0.3, 0.2, 0.7\}$ (hash marks), none of which is equal to the population parameter value $\theta_0 = 0$. In the asymptotic limit the value of the estimate is $\hat{\theta}_\infty = 0.2$, with a bias-squared of 0.04. However, with finite samples of size n, the expected value of $\hat{\theta}$ is

$$E_n\hat{\theta} = \sum_{i=1}^{4} p_i(n)\,\hat{\theta}_i \tag{3.4}$$

with $p_i(n)$ being the probability that a likelihood based on a sample of size n is maximized at $\hat{\theta}_i$. The dispersion of the $p_i(n)$-values will tend to increase with decreasing n. This enlarges the set of values that can be realized by (3.4), enabling it to achieve values closer to θ_0, thereby potentially decreasing the bias.

An L-terminal node regression tree $\hat{f}_L(\mathbf{x})$ cannot get arbitrarily close to a continuous function $f(\mathbf{x})$ such as the linear target used here. Thus, there is an asymptotic average bias-squared characteristic of the closest possible L-terminal node tree. As the sample size is reduced, the distance to the target of the expected approximation $\bar{f}_L(\mathbf{x}) = E_n\hat{f}_L(\mathbf{x})$ becomes smaller, reducing bias-squared. Of course, this expected approximation $\bar{f}_L(\mathbf{x})$ is itself not realizable as a (finite sized) regression tree. Bagging reduces bias-squared in such situations simply by reducing the sample size; each bagged tree is trained on a subset of the complete training set (2.8). The averaging aspect of bagging has no effect on bias-squared, but sharply reduces the nonlinear component of variance, thereby producing the win–win situation observed here. However, this argument does not explain why variance increases monotonically with increasing sampling fraction. It may be that in this situation the linear component of the estimator is so small that it does not play a significant role. In such cases the theory of Section 2 cannot provide much insight.

3.4. Bootstrap versus half-sampling

One of the results of the theory is that half-sampling ($m = n/2$) without replacement should produce similar results to full ($m = n$) bootstrap sampling with replacement. In the case of variance, confirmation of this property can be deduced directly from the figures. Table 1 shows that the relationship also extends to bias-squared, and thus to root-mean-squared error. The first column identifies each example by the figure(s) in which it was presented. The second and third columns give the corresponding root-mean-squared estimation error ($\sqrt{\text{bias-squared} + \text{variance}}$) for half-sampling without replacement and full sampling with replacement, respectively. The results are seen to verify the theory in this respect.

Table 1
Root-mean-squared estimation errors in the case of half-sampling without replacement (see the column headed $n/2$: W/O) or full sampling with replacement (column headed n: W)

Fig.	$n/2$: W/O	n: W
1	0.100	0.097
2,4	0.140	0.139
3,5	0.356	0.364
6,7	0.292	0.295

Half-sampling of course has a computational advantage, especially if the implementation of the estimator does not support observation weights. More generally, one can see by comparing the relevant figures that m-out-of-n with, and $\frac{1}{2}m$-out-of-n without, replacement sampling give fairly similar results.

4. Technical properties

4.1. Theory for solutions of bagged estimating equations

Here, we demonstrate theoretically the properties of bagging discussed heuristically in Section 2.1. The following notation is used: $\alpha = \psi''(0)$, $\alpha_j = \psi^{(j+2)}(0)/(j+1)!$, $\beta = \psi^{(4)}(0)$, $\beta_j = \psi^{(j+4)}(0)/(j+1)!$ and $\gamma = \psi'''(0)$.

We take Z to be the mean of n independent and identically distributed random variables with zero mean, and consider theory as n increases. Further we assume ψ is smooth in a neighbourhood of the origin, so that Taylor expansion may be conducted there, and that $\psi'(0) = 0$ and $-1 < \psi''(0) \neq 0$. The constraint on $\psi'(0)$ serves only to ensure that bumpiness of the objective function L does not add a systematic error to the estimator $\hat{\theta}$. That assumption is not essential to our analysis, but it simplifies the algebra.

The conditions on $\psi''(0)$ are more substantive. If $\psi''(0) = 0$ then the unbagged estimator becomes superefficient, in the sense that its convergence rate is faster than $O_p(n^{-1/2})$. (For example, the rate is $O_p(n^{-1})$ if $\psi''(0) = 0$ and $\psi'''(0) \neq 0$.) And if $\psi''(0) = -1$ then the objective function is no longer quadratic near the origin, since the component $\frac{1}{2}\theta^2$ cancels completely with the quadratic term in $\psi(\theta)$. As a result the convergence rate is again different; depending on high-order derivatives of ψ, the estimator $\hat{\theta}$ may now be inconsistent. If $\psi''(0) < -1$ then $\theta = 0$ produces a local maximum, rather than a local minimum, of the noiseless objective function $\frac{1}{2}\theta^2 + \psi(\theta)$. In this case the cup shape of the objective function has been completely overwhelmed, and reversed, by the added bumpy function.

The first step in our analysis is to derive properties of the bagged estimator. Put $\hat{\sigma}^2 = E\{(\Delta^*)^2|\mathcal{X}\}$ and $\eta = \frac{1}{2}\hat{\sigma}^2$, and note that since $E(\Delta^*|\mathcal{X}) = 0$,

$$\begin{aligned} E\{\psi'(\theta + Z^*)|\mathcal{X}\} &= \psi'(\theta + Z) + \tfrac{1}{2} E\{(\Delta^*)^2|\mathcal{X}\}\,\psi'''(\theta + Z) + \tfrac{1}{6} E\{(\Delta^*)^3|\mathcal{X}\}\,\psi^{(4)}(\theta + Z) + \cdots \\ &= \psi'(\theta + Z) + \eta\psi'''(\theta + Z) + O_p(n^{-2}). \end{aligned}$$

Therefore, the bagged estimating equation (2.2) may be expanded as

$$\theta + \gamma\eta + (\alpha + \beta\eta)\,(\theta + Z) + (\alpha_1 + \beta_1\eta)\,(\theta + Z)^2 + (\alpha_2 + \beta_2\eta)\,(\theta + Z)^3 + \cdots + O_p(n^{-2}) = 0. \tag{4.1}$$

Taking $\theta = (1 + \alpha + \beta\eta)^{-1}\{\xi - (\alpha + \beta\eta)\,Z - \gamma\eta\}$ in (4.1) we deduce that

$$\theta + Z = \frac{\xi - (\alpha + \beta\eta)\,Z - \gamma\eta + (1 + \alpha + \beta\eta)\,Z}{1 + \alpha + \beta\eta} = \frac{\xi + Z - \gamma\eta}{1 + \alpha + \beta\eta}.$$

Substituting this formula for $\theta + Z$ in (4.1) we find that (4.2), we deduce that ξ satisfies

$$\xi + (\alpha_1 + \beta_1\eta)\left(\frac{\xi + Z - \gamma\eta}{1+\alpha+\beta\eta}\right)^2 + (\alpha_2 + \beta_2\eta)\left(\frac{\xi + Z - \gamma\eta}{1+\alpha+\beta\eta}\right)^3 + \cdots + \mathrm{O_p}(n^{-2}) = 0. \tag{4.2}$$

Since $\eta = \mathrm{O_p}(n^{-1})$ and $Z = \mathrm{O_p}(n^{-1/2})$, then (4.2) implies that $\xi = \mathrm{O_p}(n^{-1})$. Substituting the results $\eta = \mathrm{O_p}(n^{-1})$, $Z = \mathrm{O_p}(n^{-1/2})$ and $\xi = \mathrm{O_p}(n^{-1})$ into all but the first term (i.e. the ξ) on the left-hand side of (4.2), we deduce that ξ satisfies

$$\xi = -\alpha_1\left(\frac{Z}{1+\alpha}\right)^2 + \mathrm{O_p}(n^{-3/2}) = \xi_0 + \frac{2\alpha_1\gamma}{(1+\alpha)^2}\,\eta Z + \mathrm{O_p}(n^{-2}),$$

where ξ_0 is the value assumed by ξ in the unbagged case, i.e. where $\theta = (1+\alpha)^{-1}(\xi_0 - \alpha Z)$ solves (2.1). Therefore, the solution $\hat\theta_{\mathrm{bagg}}$ of (4.1) satisfies

$$\begin{aligned}\hat\theta_{\mathrm{bagg}} &= \frac{1}{1+\alpha+\beta\eta}\left\{\xi_0 - (\alpha+\beta\eta)\,Z + \frac{2\alpha_1\gamma}{(1+\alpha)^2}\,\eta Z - \gamma\eta\right\} + \mathrm{O_p}(n^{-2})\\ &= \left(1 + \frac{\beta\eta}{\alpha(1+\alpha)}\right)\hat\theta + \frac{\gamma}{1+\alpha}\left(\frac{2\alpha_1}{(1+\alpha)^2}\,Z - 1\right)\eta + \mathrm{O_p}(n^{-2}),\end{aligned} \tag{4.3}$$

where $\theta = \hat\theta$, the unbagged estimator, solves (2.1).

Note too that

$$\hat\theta = -\frac{\alpha Z}{1+\alpha} + \mathrm{O_p}(n^{-1}) \quad \text{and} \quad \eta = \frac{1}{2}n^{-1}(\tau^2 + Z_1) + \mathrm{O_p}(n^{-2}), \tag{4.4}$$

where τ^2 is the variance of each of the data of which Z is the mean, and Z_1 is the average of the sum of the squares of those data, centered at its expected value. Assume for the time being that $\gamma = 0$. Then (4.3) simplifies significantly, and in company with (4.4) it implies that

$$E(\hat\theta^2_{\mathrm{bagg}}) = E\left\{\left(1 + \frac{2\beta\eta}{\alpha(1+\alpha)}\right)\hat\theta^2\right\} + \mathrm{O}(n^{-3}) = \left(1 + \frac{\beta\tau^2}{\alpha(1+\alpha)\,n}\right)E(\hat\theta^2) + \mathrm{O}(n^{-3}). \tag{4.5}$$

It may also be proved that when $\gamma = 0$,

$$E(\hat\theta) = \mathrm{O}(n^{-2}) \quad \text{and} \quad E(\hat\theta_{\mathrm{bagg}}) = \mathrm{O}(n^{-2}). \tag{4.6}$$

From (4.5) we deduce that, provided β and α have different signs, $E(\hat\theta^2_{\mathrm{bagg}}) < E(\hat\theta^2)$ for all sufficiently large n. Moreover, to first order the amount by which $E(\hat\theta^2)$ exceeds $E(\hat\theta^2_{\mathrm{bagg}})$ increases in direct proportion to $\tau^2 = \frac{1}{2}\lim_{n\to\infty} n\,E(\hat\sigma^2)$. The first of these results confirms the potential reductions in mean-squared error offered by bagging, and the second argues that if the variance of the bumpy perturbations is not too large then the improvement tends to be in proportion to the extent to which bagging reduces the variability of locations of bumps.

On the other hand, if β and α have the same sign then $E(\hat\theta^2_{\mathrm{bagg}}) > E(\hat\theta^2)$, and so bagging increases, rather than reduces, mean-squared error. Again this effect is in proportion to the variance of the location of the bumps, to first order. Eqs. (4.5) and (4.6) also imply that the results in this paragraph and the previous one remain true if we replace 'mean-squared error' by 'variance' throughout.

The case where $\psi'''(0) \neq 0$ is more complex. There we can see from (4.3) that the bagged estimator has acquired an extra bias term, equal to $-\gamma(1+\alpha)^{-1}E(\eta) + \mathrm{O}(n^{-2})$ and of size n^{-1}. This can increase mean-squared error of the bagged estimator, at the same level (i.e. n^{-2}) as any improvements offered by bagging. Also, contributions to variance from the term in Z at (4.3) will affect performance at the level n^{-2}. The net influence of these contributions is that overall mean-squared error performance of the bagged estimator, relative to its unbagged counterpart, can no longer be explained solely in terms of the sign of α/β. Bagging can either improve or degrade performance, depending on the relationship among α, β and γ.

The difficulty when $\psi'''(0) \neq 0$ is caused by the fact that ψ is significantly asymmetric, and of course bagging the estimating equation compounds asymmetry. That is the source of the bias term, and indirectly of the component involving Z.

4.2. Theory for bagged solutions of estimating equations

For brevity we treat only the case $\gamma = 0$, noting that the main effect of the level of asymmetry implied by $\gamma \neq 0$ is exactly that reported in Section 4.1: a bias component of size n^{-1} is added to the bagged estimator, and variance of the estimator alters a little, at the level n^{-2}. Now, the unbagged estimator is defined by $\hat{\theta} = (1+\alpha)^{-1}(\xi_0 - \alpha Z)$, where ξ_0 is obtained by solving (2.1) for $\theta = \hat{\theta}$. This leads to the formula

$$\hat{\theta} = -\frac{1}{1+\alpha}\left(\alpha Z + \frac{\beta}{6(1+\alpha)^3} Z^3\right) + \mathrm{O_p}(n^{-2}).$$

Replacing Z by $Z^* = Z + \Delta^*$ in this expression, taking expectation conditional on $\mathcal{X}$, putting $\eta = \frac{1}{2} E\{(\Delta^*)^2 | \mathcal{X}\} = \mathrm{O_p}(n^{-1})$, and noting that $E\{(\Delta^*)^j | \mathcal{X}\} = \mathrm{O_p}(n^{-2})$ for $j \geqslant 3$, we deduce that the bagged solution of Eq. (2.1) satisfies

$$\begin{aligned}\hat{\theta}_{\mathrm{bagg}} &= -\frac{1}{1+\alpha}\left(\alpha Z + \frac{\beta}{6(1+\alpha)^3} Z^3 + \frac{\beta\eta Z}{(1+\alpha)^3}\right) + \mathrm{O_p}(n^{-2}) \\ &= \left(1 + \frac{\beta\eta}{\alpha(1+\alpha)^3}\right)\hat{\theta} + \mathrm{O_p}(n^{-2}). \end{aligned} \tag{4.7}$$

Moreover, the analogue of (4.6) holds: when $\gamma = 0$, $E(\hat{\theta}) = \mathrm{O}(n^{-2})$ and $E(\hat{\theta}_{\mathrm{bagg}}) = \mathrm{O}(n^{-2})$.

Eq. (4.7) is a direct analogue of (4.3), provided we take $\gamma = 0$ in the latter. It produces the corresponding direct analogue of (4.5):

$$E(\hat{\theta}^2_{\mathrm{bagg}}) = \left(1 + \frac{\beta\tau^2}{2\alpha(1+\alpha)^3 n}\right) E(\hat{\theta}^2) + \mathrm{O}(n^{-3}), \tag{4.8}$$

where τ^2 is the variance of each of the data of which Z is the mean. This implies the main properties noted in the case of bagged estimating equations. In particular, bagging improves performance (by reducing both variance and mean-squared error), provided α and β have different signs (and $-1 < \alpha \neq 0$). The extent of the improvement is once again proportional to τ^2, to first order. And bagging degrades performance if α and β have different signs.

If $\alpha > 0$, meaning that the perturbation $\psi'(\theta + Z)$ of the estimating equation enhances the equation's local convexity, then it can be seen by comparing (4.5) and (4.8) that bagging the estimating equation has a greater effect on performance than bagging the estimator itself.

Acknowledgements

The work of Jerome H. Friedman was partially supported by CSIRO Mathematical and Information Sciences, Australia, the Department of Energy under contract DE-AC03-76SF00515, and by Grant DMS9764431 of the National Science Foundation. We are grateful to Rob Tibshirani for the helpful comments.

References

Breiman, L., 1996. Bagging predictors. Mach. Learning 24, 123–140.
Breiman, L., 1999. Using adaptive bagging to debias regressions. Technical Report No. 547, Department of Statistics, University of California, Berkeley.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.
Bühlmann, P., Yu, B., 2000. Analyzing bagging. Ann. Statist. 30, 927–961.
Buja, A., Steutzle, W., 2000a. Bagging does not always decrease mean squared error. Manuscript.
Buja, A., Steutzle, W., 2000b. Smoothing effects of bagging. Manuscript.
Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia.
Hartigan, J.A., 1969. Using subsample values as typical values. J. Amer. Statist. Assoc. 64, 1303–1317.
Hartigan, J.A., 1971. Error analysis by replaced samples. J. Roy. Statist. Soc. Ser. B 33, 98–110.
Mahalanobis, P.C., 1946. Report on the Bihar crop survey: rabi season 1943–1944. Sankhyā 7, 269–280.
McCarthy, P.J., 1966. Replication (an approach to the analysis of data from complex surveys). National Center for Health Statistics.