Statistica Sinica **16**(2006), 323-351

OBSERVATIONS ON BAGGING

Andreas Buja and Werner Stuetzle

University of Pennsylvania and University of Washington

Abstract: Bagging is a device intended for reducing the prediction error of learning algorithms. In its simplest form, bagging draws bootstrap samples from the training sample, applies the learning algorithm to each bootstrap sample, and then averages the resulting prediction rules. More generally, the resample size M may be different from the original sample size N, and resampling can be done with or without replacement. We investigate bagging in a simplified situation: the prediction rule produced by a learning algorithm is replaced by a simple real-valued U-statistic of i.i.d. data. U-statistics of high order can describe complex dependencies, and yet they admit a rigorous asymptotic analysis. We show that bagging U-statistics often *but not always* decreases variance, whereas it always increases bias. The most striking finding, however, is an equivalence between bagging based on resampling *with* and *without* replacement: the respective resample sizes $M_{with} = \alpha_{with} N$ and $M_{w/o} = \alpha_{w/o} N$ produce very similar bagged statistics if $\alpha_{with} = \alpha_{w/o}/(1 - \alpha_{w/o})$. While our derivation is limited to U-statistics, the equivalence seems to be universal. We illustrate this point in simulations where bagging is applied to CART trees.

Key words and phrases: Bagging, CART, U-statistics.

1. Introduction

Bagging, short for "bootstrap aggregation", was introduced by Breiman (1996) as a device for reducing the prediction error of learning algorithms. Bagging is performed by drawing bootstrap samples from the training sample, applying the learning algorithm to each bootstrap sample, and averaging/aggregating the resulting prediction rules, that is, averaging or otherwise aggregating the predicted values for test observations. Breiman presents empirical evidence that bagging can indeed reduce prediction error. It appears to be most effective for CART trees (Breiman et al. (1984)). Breiman's heuristic explanation is that CART trees are highly unstable functions of the data — a small change in the training sample can result in a very different tree — and that averaging over bootstrap samples reduces the variance component of the prediction error.

The goal of the present article is to contribute to the theoretical understanding of bagging. We investigate bagging in a simplified situation: the prediction rule produced by a learning algorithm is replaced by a simple real-valued statistic of i.i.d. data. While this simplification does not capture some characteristics of

function fitting, it still enables us, for example, to analyze the conditions under which variance reduction occurs. The claim that bagging always reduces variance is in fact not true.

We start by describing bagging in operational terms. Bagging a statistic $\theta(X_1, \ldots, X_N)$ is defined as averaging it over bootstrap samples $X_1^*, \ldots, X_N^*$:

$$\theta^B(X_1, \ldots, X_N) = \text{ave}_{X_1^*, \ldots, X_N^*} \theta(X_1^*, \ldots, X_N^*),$$

where the observations X_i^* are i.i.d. draws from $\{ X_1, \ldots, X_N \}$. The bagged statistic can also be written as

$$\theta^B(X_1, \ldots, X_N) = \frac{1}{N^N} \sum_{i_1, \ldots, i_N} \theta(X_{i_1}, \ldots, X_{i_N})$$

because there are N^N sets of bootstrap samples, each having probability $1/N^N$. For realistic sample sizes N, the N^N sets cannot be enumerated in actual computations, hence one resorts to sampling a smaller number, often as few as 50.

Our analysis covers several variations on bagging. Instead of averaging the values of a statistic over bootstrap samples of the same size N as the original sample, we may choose the resample size M to be smaller, or even larger, than N. Another alternative covered by our analysis is resampling without replacement.

The statistics we consider here are U-statistics. While they do not capture the statistical properties of CART trees, U-statistics can model complex interactions and yet they allow for a rigorous second order analysis. (For an approach tailored to tree-based methods, see Buhlmann and Yu (2002).)

The most striking effect we observe, both theoretically and in simulations, is a correspondence between bagging based on resampling with and without replacement. The two modes of resampling produce very similar bagged statistics if the resampling fractions $\alpha_{w/o} = M_{w/o}/N$ for sampling without replacement and $\alpha_{with} = M_{with}/N$ for sampling with replacement satisfy the relation

$$\alpha_{with} = \frac{\alpha_{w/o}}{1 - \alpha_{w/o}}, \quad \text{or} \quad \frac{1}{\alpha_{with}} = \frac{1}{\alpha_{w/o}} - 1.$$

This equivalence holds to order N^{-2} under regularity assumptions. The equivalence is implicit in one form or another in previous work: Friedman and Hall (2000, Sec. 2.6) notice it for a type of polynomial expansion, but they do not make use of it other than noting that half-sampling without replacement ($\alpha_{w/o} = 1/2$) and standard bootstrap sampling ($\alpha_{with} = 1$) yield very similar bagged statistics. Knight and Bassett (2002, Sec. 4) note the equivalence for half-sampling and bootstrap in the case of quantile estimators. In the present article we show the equivalence for U-statistics of fixed but arbitrary order. We also illustrate it

in simulations for bagged trees where it holds with surprising accuracy, hinting at a much greater range of validity.

Other observations about the effects of bagging concern the variance, squared bias, and mean squared error (MSE) of bagged U-statistics. Similar to Chen and Hall (2003) and Knight and Bassett (2002), we obtain effects that are only of order $O(N^{-2})$. This may seem small with the sometimes strong effects of bagging on CART trees in mind, but it should be recalled that the implicit N of a tree-based estimate $\hat{f}(x)$ of a function $f(x)$ at x is often small, namely, in the order of the terminal node size. For small N, however, even effects of order N^{-2} can be sizable. We also find that, with decreasing resample size, squared bias always increases and variance often *but not always* decreases. More precisely, the difference between bagged and unbagged for squared bias is an increasing quadratic function of

$$g := \frac{1}{\alpha_{with}} = \frac{1}{\alpha_{w/o}} - 1,$$

and for the variance it is an often, but not always, decreasing linear function of g. Therefore, the only possible beneficial effect of bagging stems from variance reduction. In those situations where variance is reduced, the combined effect of bagging is to reduce the MSE in an interval of g near zero; equivalently, the MSE is reduced for α_{with} near infinity and correspondingly for $\alpha_{w/o}$ near 1. For the standard bootstrap ($\alpha_{with} = 1$) and half-sampling ($\alpha_{w/o} = 1/2$), improvements in MSE are obtained only if the resample sizes fall in the respective critical intervals. However, there can arise odd situations in which the MSE is improved only for $\alpha_{with} > 1$ and $\alpha_{w/o} > 1/2$.

We finish this article with some illustrative simulations of bagged CART trees. A purpose of these illustrations is to gain some understanding of the peculiarities of trees in light of the fact that bagging often shows dramatic improvements that apparently go beyond the effects described by $O(N^{-2})$ asymptotics. An important point to keep in mind is that there are two notion of bias.

- If we regard $\theta(F_N)$ as a plug-in estimate of $\theta(F)$, then the *plug-in bias* is $\mathbf{E}\,\theta(F_N) - \theta(F)$.
- If we regard $\theta(F_N)$ as an estimate for some parameter μ, then the *estimation bias* is $\mathbf{E}\,\theta(F_N) - \mu$.

The second notion of bias – estimation bias – is the one commonly used in function estimation. Our theory of bagged U-statistics, however, is concerned with plug-in bias, *not* with estimation bias. The same applies to Chen and Hall's (2003) theory of bagging estimating equations, as well as Knight and Bassett's (2002) theory of bagged quantiles. This point even applies to Buhlmann and Yu's (2002) treatment of bagged stumps and trees because their notion of bias refers not

to the true underlying function but to the optimal asymptotic target, that is, the asymptotically best fitting stump or tree. Their theory therefore explains bagging's effect on the variance of stumps and trees (better than any of the other theories, including ours), but it, too, has nothing to say about bias in the usual sense of function fitting.

An interesting observation we make in the simulations is that for smooth underlying $f(x)$, bagging not only decreases variance, but it can reduce estimation bias as well. This should not be too surprising because, according to Buhlmann and Yu's theory, the effect of bagging is essentially to replace fitting a stump with fitting a stump convolved with a narrow-bandwidth kernel. The convolved stump is smooth and has a chance to reduce estimation bias when the underlying $f(x)$ is smooth.

2. Resampling U-Statistics

Let $X_1, \ldots, X_N$ be i.i.d. random variables. We consider statistics of $X_1, \ldots, X_N$ that are finite sums

$$U = \frac{1}{N}\sum_i A_{X_i} + \frac{1}{N^2}\sum_{i,j} B_{X_i,X_j} + \frac{1}{N^3}\sum_{i,j,k} C_{X_i,X_j,X_k} + \ldots \ ,$$

where the "kernels" B, C, ... are permutation symmetric in their arguments. (We put the arguments in subscripts in order to avoid the clutter caused by frequent parentheses.) The normalizations of the sums are such that, under common assumptions, limits for $N \to \infty$ exist. Strictly speaking, only the off-diagonal part $\sum_{i \neq j} B_{X_i,X_j}$ (e.g.) is a proper U-statistic. Because we include the diagonal $i = j$ in the double sum, this is strictly speaking a V-statistic or von Mises statistic (Serfling (1980, Sec. 5.1.2)), but we use the better known term "U-statistic" anyway. The reason for including the diagonal is that only in this way can U be interpreted as the plug-in estimate $U(F_N)$ of a statistical functional

$$U(F) = \mathbf{E}\, A_X + \mathbf{E}\, B_{X,Y} + \mathbf{E}\, C_{X,Y,Z} + \cdots \ ,$$

where $X, Y, Z, \ldots$ are i.i.d. Knowing what the statistic U estimates is a necessity for bias calculations. A second reason for including the diagonal is that bagging has the effect of creating terms such as B_{X_i,X_i}, so we may as well include such terms from the outset.

It is possible to explicitly calculate the bagged version U^{bag} of a sum of U-statistics U. We can allow bagging based on resampling *with* and *without* replacement as well as arbitrary resample sizes M.

Let $\mathbf{W} = (W_1 \ldots, W_N \geq 0)$ be integer-valued random variables counting the multiplicities of $X_1, \ldots, X_N$ in a *resample*.

- For resampling *with* replacement, that is, *bootstrap*, the distribution of $\mathbf{W}$ is Multinomial$(1/N, \ldots, 1/N; M)$. Conventional bootstrap corresponds to $M = N$, but we allow M to range between 1 and ∞. Although $M > N$ is computationally undesirable, infinity is the conceptually plausible upper bound on M: for $M = \infty$ no averaging takes place because with an "infinite resample" one has $F_M^* = F_N$.
- For resampling *without* replacement, that is, *subsampling*, $\mathbf{W}$ is a hypergeometric random vector where each W_i is Hypergeometric$(N, 1, M)$ with each $i = 1 \ldots N$ being the unique "defective" in turn. Half-sampling, for example, has $M = N/2$, but the resample size M can range between 1 and N. For the upper bound $M = N$ no averaging takes place because the resample is just a permutation of the data, hence $F_M^* = F_N$.

With these facts we can write down the resampled and the bagged version of a U explicitly. We illustrate this for a statistic U with kernels A_{X_i} and B_{X_i,X_j}. For a resample of M with multiplicities $W_1, \ldots, W_N$, the value of U is

$$U^{resample} = \frac{1}{M} \sum W_i A_{X_i} + \frac{1}{M^2} \sum W_i W_j B_{X_i,X_j}.$$

The bagged version of U under either mode of resampling is the expected value with respect to $\mathbf{W}$:

$$\begin{aligned} U^{bag} &= \mathbf{E}_{\mathbf{W}} \Big[\frac{1}{M} \sum W_i A_{X_i} + \frac{1}{M^2} \sum W_i W_j B_{X_i,X_j} \Big] \\ &= \frac{1}{M} \sum \mathbf{E}\,[W_i] A_{X_i} + \frac{1}{M^2} \sum \mathbf{E}\,[W_i W_j] B_{X_i,X_j}. \end{aligned}$$

From the form of U^{bag} it is apparent that the only relevant quantities are moments of $\mathbf{W}$:

$$\begin{aligned} \mathbf{E}\,W_i &= \frac{M}{N} \quad \text{with and w/o} \\ \mathbf{E}\,W_i^2 &= \begin{cases} \text{with: } \frac{M}{N} + \frac{M(M-1)}{N^2} \\ \text{w/o: } \frac{M}{N} \end{cases} \\ \mathbf{E}\,W_i W_j &= \begin{cases} \text{with: } \frac{M(M-1)}{N^2} \\ \text{w/o: } \frac{M(M-1)}{N(N-1)} \end{cases} \quad (i \neq j). \end{aligned}$$

The bagged functional can now be written down explicitly. It is necessary to distinguish between the two resampling modes: we denote U^{bag} by U^{with} and

$U^{w/o}$ for resampling with and without replacement, respectively.

$$U^{with} = \frac{1}{N}\sum_i \Big(A_{X_i} + \frac{1}{M_{with}} B_{X_i,X_i}\Big) + \frac{1}{N^2}\sum_{i,j}\Big(1 - \frac{1}{M_{with}}\Big) B_{X_i,X_j},$$

$$U^{w/o} = \frac{1}{N}\sum_i \Big(A_{X_i} + \Big(\frac{1 - \frac{M_{w/o}}{N}}{1 - \frac{1}{N}}\Big)\frac{1}{M_{w/o}} B_{X_i,X_i}\Big) + \frac{1}{N^2}\sum_{i,j}\Big(\frac{1 - \frac{1}{M_{w/o}}}{1 - \frac{1}{N}}\Big) B_{X_i,X_j}.$$

Analogous calculations can be carried out for statistics with U-terms of orders higher than two. We summarize as follows.

Proposition 1. *A bagged sum of U-statistics is also a sum of U-statistics. For a statistic with kernels A_x and $B_{x,y}$ only, the bagged terms A_x^{with}, $B_{x,y}^{with}$ and $A_x^{w/o}$, $B_{x,y}^{w/o}$, respectively, depend on A_x and $B_{x,y}$, with*

$$A_x^{with} = A_x + \frac{1}{M_{with}} B_{x,x}, \qquad B_{x,y}^{with} = \Big(1 - \frac{1}{M_{with}}\Big) B_{x,y},$$

$$A_x^{w/o} = A_x + \Big(\frac{1 - \frac{M_{w/o}}{N}}{1 - \frac{1}{N}}\Big)\frac{1}{M_{w/o}} B_{x,x}, \quad B_{x,y}^{w/o} = \Big(\frac{1 - \frac{1}{M_{w/o}}}{1 - \frac{1}{N}}\Big) B_{x,y}.$$

For U-statistics with terms of first and second order, the proposition is a direct result of the preceding calculations. For general U-statistics of arbitrary order, the proposition is a consequence of the proofs in the appendix of the online version of this article.

We see from the proposition that the effect of bagging is to remove mass from the proper U-part of B ($\sum_{i \neq j}$) and shift it to the diagonal ($\sum_{i=j}$), thus increasing the importance of the linear part. Similar effects take place in higher orders where variability is shifted to lower orders.

3. Equivalence of Resampling With and Without Replacement

Proposition 1 yields a heuristic for an important fact: bagging based on resampling *with* replacement yields results very similar to bagging based on resampling *without* replacement *if* the resample sizes M_{with} and $M_{w/o}$ are suitably matched up. The required correspondence can be derived by equating $A^{with} = A^{w/o}$ and/or $B^{with} = B^{w/o}$ in Proposition 1; both equations yield the identical condition.

Corollary. *Bagging a sum of U-statistics of first and second order yields identical results under the two resampling modes if*

$$\frac{N-1}{M_{with}} = \frac{N}{M_{w/o}} - 1.$$

For a general finite sum of U-statistics of arbitrary order, we do not obtain an identity but an approximate equivalence.

Proposition 2. *Bagging a finite sum of U-statistics of arbitrary order under either resampling mode yields the same results up to order $O(N^{-2})$ if*

$$\frac{N}{M_{with}} = \frac{N}{M_{w/o}} - 1,$$

assuming the kernels are bounded. If the kernels are not bounded but have moments of order q, the approximation is to order $O(N^{-2/p})$, where $1/p + 1/q = 1$.

We will similarly see that variance, squared bias and hence MSE of bagged U-statistics all agree in the N^{-2} term in the two resampling modes under corresponding resample sizes.

The correspondence between the two resampling modes is more intuitive if one expresses the resample sizes M_{with} and $M_{w/o}$ as fractions/multiples of the sample size N : $\alpha_{with} = M_{with}/N$ $(> 0, \ < \infty)$ and $\alpha_{w/o} = M_{w/o}/N$ $(> 0, \ < 1)$. The condition of Proposition 2 above is equivalent to

$$\alpha_{with} = \frac{\alpha_{w/o}}{1 - \alpha_{w/o}}.$$

It equates, for example, half-sampling without replacement, $\alpha_{w/o} = 1/2$, with conventional bootstrap, $\alpha_{with} = 1$. Subsampling without replacement with $\alpha_{w/o} > 1/2$ corresponds to bootstrap with $\alpha_{with} > 1$, that is, bootstrap resamples larger than the original sample. The correspondence also maps $\alpha_{w/o} = 1$ to $\alpha_{with} = \infty$, both of which mean that the bagged and the unbagged statistic are identical.

4. The Effect of Bagging on Variance, Bias and MSE

We need some notation: For U-statistics $C_{X,Y,Z\ldots}$ of any order we denote partial conditional expectations by appropriate subscripts, thus $C_X = \mathbf{E}\,[C_{X,Y,Z,W,\ldots}|X]$, $C_{X,Y} = \mathbf{E}\,[C_{X,Y,Z,W,\ldots}|X,Y]$, and so on.

4.1. Variance

Variances of U-statistics can be calculated explicitly. For example, for a statistic that has only terms A_X and $B_{X,Y}$, the variance is

$$\begin{aligned}\mathrm{Var}(U) = {} & N^{-1}\mathrm{Var}(A_X + 2B_X) \\ & + N^{-2}\Big(2\,\mathrm{Cov}\,(A_X, B_{X,X}) + 4\,\mathrm{Cov}\,(B_{X,X}, B_X) - 4\,\mathrm{Cov}\,(A_X, B_X) \\ & + 2\mathrm{Var}(B_{X,Y}) - 12\mathrm{Var}(B_X)\Big) \\ & + N^{-3}\Big(\mathrm{Var}(B_{X,X}) - 2\mathrm{Var}(B_{X,Y}) + 8\mathrm{Var}(B_X) - 4\,\mathrm{Cov}\,(B_{X,X}, B_X)\Big).\end{aligned}$$

We are, however, primarily interested not in variances but differences between variances of bagged and unbagged statistics.

Proposition 3. *Let $g = N/M$ for sampling with replacement and $g = (N/M)-1$ for sampling without replacement. Assume g is fixed and $0 < g < \infty$ as $N \to \infty$. Let U be a finite sum of U-statistics of arbitrary order. Then*

$$\mathrm{Var}(U^{bag}) - \mathrm{Var}(U) = \frac{1}{N^2} \cdot 2S_{\mathrm{Var}} \cdot g + O(\frac{1}{N^3})$$

for both sampling with and without replacement. If U has only terms A_X and $B_{X,Y}$, then

$$S_{\mathrm{Var}} = \mathrm{Cov}\,(A_X + 2B_X,\ B_{X,X} - B_X).$$

The proof is in the appendix of the online version of this article, where we also show how to calculate S_{Var} for statistics with U-terms of any order. The effect of bagging on variance is of order $O(N^{-2})$. For statistics of first order ($B_{X,Y} = 0$) we have $S_{Var} = 0$, hence no effect of bagging.

The assumption about g is essential. If it is not satisfied, the order of the asymptotics will be affected. The jackknife is a case in point: it is obtained for $M = N - 1$ and resampling without replacement. This implies $g \to 0$, which violates the assumption of the proposition. It would be easy to cover this type of asymptotics because the calculations can be performed exactly.

There exist situations in which bagging increases the variance, namely, when $S_{\mathrm{Var}} > 0$. If $S_{\mathrm{Var}} < 0$, variance is reduced, and the beneficial effect becomes more pronounced the smaller the resample size. Therefore, the fact that bagging may reduce variance cannot be the whole story: if variance were the only criterion of interest, one should choose the resample size M as low as operationally feasible for maximal variance reduction. Obviously, one has to take into account bias as well.

4.2. Bias

We show that bagging U-statistics *always* increases squared bias (except for linear statistics, where the bias vanishes). Recall that the statistic $U = U(F_N)$ is the plug-in estimator for the functional $U(F)$, so the bias is $\mathbf{E}\, U(F_N) - U(F)$.

Proposition 4. *Under the same assumptions as in Proposition* 3*, we have*

$$\mathrm{Bias}^2(U^{bag}) - \mathrm{Bias}^2(U) = \frac{1}{N^2}(g^2 + 2g)S_{\mathrm{Bias}} + O(\frac{1}{N^3})$$

for both sampling with and without replacement. If U has only terms A_X and $B_{X,Y}$, then $S_{\mathrm{Bias}} = (\mathbf{E}\, B_{X,X} - \mathbf{E}\, B_{X,Y})^2$.

Again, the proofs are in the appendix of the online version of this article where we also give a general formula for S_{Bias} for statistics with U-terms of any order. For statistics of first order ($B_{X,Y} = 0$) we have $S_{Bias} = 0$, hence no effect of bagging.

Just as in the comparison of variances, sampling with and without replacement agree in the N^{-2} term modulo differing interpretation of g in the two resampling modes.

4.3. Mean squared error

The mean squared error of $U = U(F_N)$ is

$$MSE(U) = \mathbf{E}\left([U(F_N) - U(F)]^2\right) = \text{Var}(U) + \text{Bias}\,(U)^2.$$

The difference between MSEs of bagged and unbagged functionals is as follows.

Proposition 5. *Under the same assumptions as in Propositions* 3 *and* 4*, we have*

$$MSE(U_M^{bag}(F_N)) - MSE(U(F_N)) = \frac{1}{N^2}(S_{\text{Bias}}\, g^2 + (S_{\text{Var}} + S_{\text{Bias}})2g) + O(\frac{1}{N^3})$$

for both sampling with and without replacement.

4.4. Choice of resample size

In some situations one may obtain a reduction in MSE for some resample sizes M but not for others, while in other situations bagging may never lead to an improvement. The critical factor is the dependence of the MSE difference on g:

$$S_{\text{Bias}}\, g^2 + 2(S_{\text{Var}} + S_{\text{Bias}})g.$$

One immediately reads off the following condition for MSE improvement.

Corollary 5. *There exist resample sizes for which bagging improves the MSE to order N^{-2} iff $S_{\text{Var}} + S_{\text{Bias}} < 0$. Under this condition the range of beneficial resample sizes is characterized by*

$$g < -2\Big(\frac{S_{\text{Var}}}{S_{\text{Bias}}} + 1\Big).$$

The resample size with optimal MSE improvement is

$$g^{opt} = -\Big(\frac{S_{\text{Var}}}{S_{\text{Bias}}} + 1\Big).$$

Conventional bootstrap, $M_{with} = N$, and half-sampling, $M_{w/o} = N/2$, (both characterized by $g = 1$) are beneficial iff $S_{\text{Var}}/S_{\text{Bias}} < -3/2$, and they are optimal iff $S_{\text{Var}}/S_{\text{Bias}} = -2$.

Recall from Proposition 3 that the resample sizes M_{with} and $M_{w/o}$ are expressed in terms of $g_{with} = N/M_{with}$ and $g_{w/o} = N/M_{w/o} - 1$. The corollary therefore prescribes a minimum resample size to achieve MSE reduction. See Figure 4.1 for an illustration.

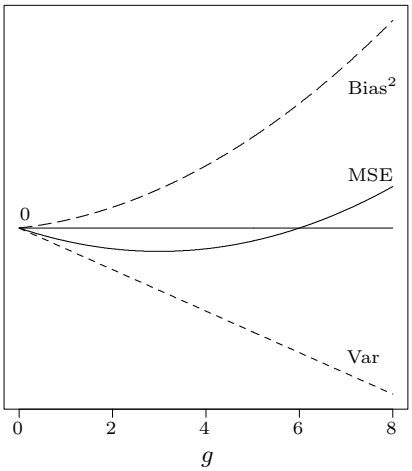

Figure 4.1. Dependence of Variance, Squared Bias and MSE on g. The graph shows the situation for $S_{\text{Var}}/S_{\text{Bias}} = -4$. Bagging is beneficial for $g < 6$, that is, for resample sizes $M_{with} > N/6$ and $M_{w/o} > N/7$. Optimal is $g = 3$, that is, $M_{with} = N/3$ and $M_{w/o} = N/4$.

The intuition that the benefits of bagging arise from variance reduction is thus correct, although it must be qualified: bagging is not always beneficial, but if it is, the reduction in MSE is due to reduction in variance.

Recall that the above statements should be limited to values of g bounded away from zero and infinity. Near either boundary a different type of asymptotics sets in.

4.5. An example: quadratic functionals

Consider as a concrete example of U-statistics the case of quadratic functions: $A_X = a \cdot X^2$ and $B_{X,Y} = b \cdot XY$, that is,

$$U = a \cdot \frac{1}{N} \sum X_i^2 + b \cdot (\frac{1}{N} \sum X_i)^2.$$

In order to determine the terms S_{Var} and S_{Bias}, we need the first four moments of X. Let $\mu = \mathbf{E}\, X$, $\sigma^2 = \mathbf{E}\,[(X-\mu)^2]$, $\gamma = \mathbf{E}\,[(X-\mu)^3)]/\sigma^3$ and $\kappa = \mathbf{E}\,[(X-\mu)^4]/\sigma^4$ be expectation, variance, skewness and kurtosis, respectively. Then $S_{\text{Var}} = (2\mu\gamma\sigma^3 + (\kappa-1)\sigma^4)ab + 2\mu\gamma\sigma^3 b^2$ and $S_{\text{Bias}} = b^2\sigma^4$. It is

convenient to write the criterion for the existence of resample sizes with beneficial effect on the MSE as $S_{\text{Var}}/S_{\text{Bias}} + 1 < 0$:

$$\Big(2\frac{\mu}{\sigma}\gamma + (\kappa - 1)\Big)\frac{a}{b} + \Big(2\frac{\mu}{\sigma}\gamma + 1\Big) < 0.$$

If $\mu = 0$ or $\gamma = 0$, this simplifies to $(\kappa - 1)a/b + 1 < 0$. Since $\kappa > 1$ for all distributions except a balanced 2-point mass, the condition becomes $a/b < -1/\kappa - 1$. For $a = 1,\ b = -1$, that is, the empirical variance $U = \text{mean}(X^2) - \text{mean}(X)^2$, beneficial effects of bagging exist iff $\kappa > 2$. For $a = 0$, that is, the squared mean $U = \text{mean}(X)^2$, no beneficial effects exist.

5. Simulation Experiment

The prinicipal purpose of the experiments presented here is is to demonstrate the correspondence between resampling with and without replacement in the non-trivial setting of bagging CART trees.

Scenarios. We consider four scenarios, differing in the size N of the training sample, the dimension p of the predictor space, the noise variance σ^2, the number K of leaves of the CART tree, and the true regression function f. The scenarios are adapted from Friedman and Hall (2000).

Scenario	N	p	X	σ^2	K	$f(x)$
1	800	1	$U[0,1]$	1	2	$I(x > 0.5)$
2	800	1	$U[0,1]$	1	2	x
3	8000	10	$U[0,1]^{10}$	0.25	50	$\prod_{i=1}^{5} I(x_i > 0.13)$
4	8000	10	$U[0,1]^{10}$	0.25	50	$\sum_{i=1}^{5} i\, x_i$

We grew all trees in Scenarios 3 and 4 in a stagewise forward manner without pruning; at each stage we split the node that resulted in the largest reduction of the residual sum of squares, this until the desired number of leaves was reached.

Performance Measures. Let $T_\alpha^{w/o}(\cdot;\mathcal{L})$ be the bagged tree obtained by averaging CART trees grown on resamples of size $\alpha\, N$ drawn without replacement from a training sample $\mathcal{L} = (\vec{x}_1, y_1), \ldots, (\vec{x}_N, y_N)$, and let $T_\alpha^{wi}(\cdot;\mathcal{L})$ be the bagged tree obtained by averaging over resamples of size $\alpha\, N/(1-\alpha)$ drawn with replacement. The mean squared error (MSE) of $T_\alpha^{w/o}$ is

$$\begin{aligned}
\text{MSE}\,(T_\alpha^{w/o}) &= \mathbf{E}_X[\,\mathbf{E}_{\mathcal{L}}((T_\alpha^{w/o}(X;\mathcal{L}) - f(X))^2)] \\
&= \mathbf{E}_X[\,\mathbf{E}_{\mathcal{L}}((T_\alpha^{w/o}(X;\mathcal{L}) - \mathbf{E}_{\mathcal{L}}(T_\alpha^{w/o}(X;\mathcal{L})))^2)] \\
&\quad + \mathbf{E}_X[\,(\mathbf{E}_{\mathcal{L}}(T_\alpha^{w/o}(X;\mathcal{L}) - f(X)))^2] \\
&= \text{Var}(T_\alpha^{w/o}) + \text{Bias}^2_{est}(T_\alpha^{w/o})\,.
\end{aligned}$$

The MSE of T_α^{wi} is defined analogously.

Recall that the definition of bias used here is *estimation bias* — expected difference between the estimated regression function for a finite sample size and the true regression function. As pointed out in the introduction, this is different from *plug-in bias* — expected difference between the value of a statistic for a finite sample size and its value for infinite sample size — which was analyzed in the earlier sections of the article. CART trees with a fixed number of leaves and their bagging averages are not in general consistent estimates of the true regression function, and in cases where they are not, as in scenarios (2) and (4) above, the two notions of bias differ.

Operational details of the experiment. We estimated plug-in bias, estimation bias, variance, and MSE for $\alpha = 0.1, 0.2, \ldots, 0.9, 0.95, 0.99, 1$; $\alpha = 1$ corresponds to unbagged CART. Estimates were obtained by averaging over 100 training samples and 10,000 test observations.

We approximated the bagged trees $T_\alpha^{w/o}(\cdot; \mathcal{L})$ and $T_\alpha^{wi}(\cdot; \mathcal{L})$ by averaging over 50 resamples. A finite number of resamples adds a significant variance component to the Monte Carlo estimates of $\mathrm{Var}(T_\alpha^{w/o})$ and $\mathrm{Var}(T_\alpha^{wi})$. This component can be easily estimated and subtracted out, thus adjusting for the finite number of resamples. There is no influence of the number of resamples on bias.

To calculate the plug-in bias we need to know the CART tree for infinite training sample size. In Scenarios 1 and 3 this is not a problem because the trees are consistent estimates for the true regression functions. In Scenarios 2 and 4 we approximated the tree for infinite training sample size by a tree grown on a training sample of size $n = 100,000$.

Simulation results. Figure 5.2 summarizes the simulation results for Scenario 1. The top panels show variance, squared plug-in bias, and squared estimation bias as functions of the resampling fraction α, for resampling with and without replacement. The bottom panel shows the MSE for both resampling modes, and variance and squared estimation bias for sampling with replacement only. To make the tick mark labels more readable, vertical scales in all the panels are relative to the MSE of the unbagged tree.

We note that variance decreases monotonically with decreasing resampling fraction, which confirms the intuition that smaller resample size means more averaging. Estimation bias and plug-in bias agree because a tree with two leaves is a consistent estimate for the true regression function, which in this scenario is a step function. Squared plug-in bias increases with decreasing resampling fraction, as predicted by the theory presented in Sections 4.1 through 4.3.

Figure 5.3 shows the corresponding results for Scenario 2. Again, variance is decreasing with decreasing resampling fraction, and squared plug-in bias is increasing, as predicted by the theory. Squared estimation bias, however, is *de-*

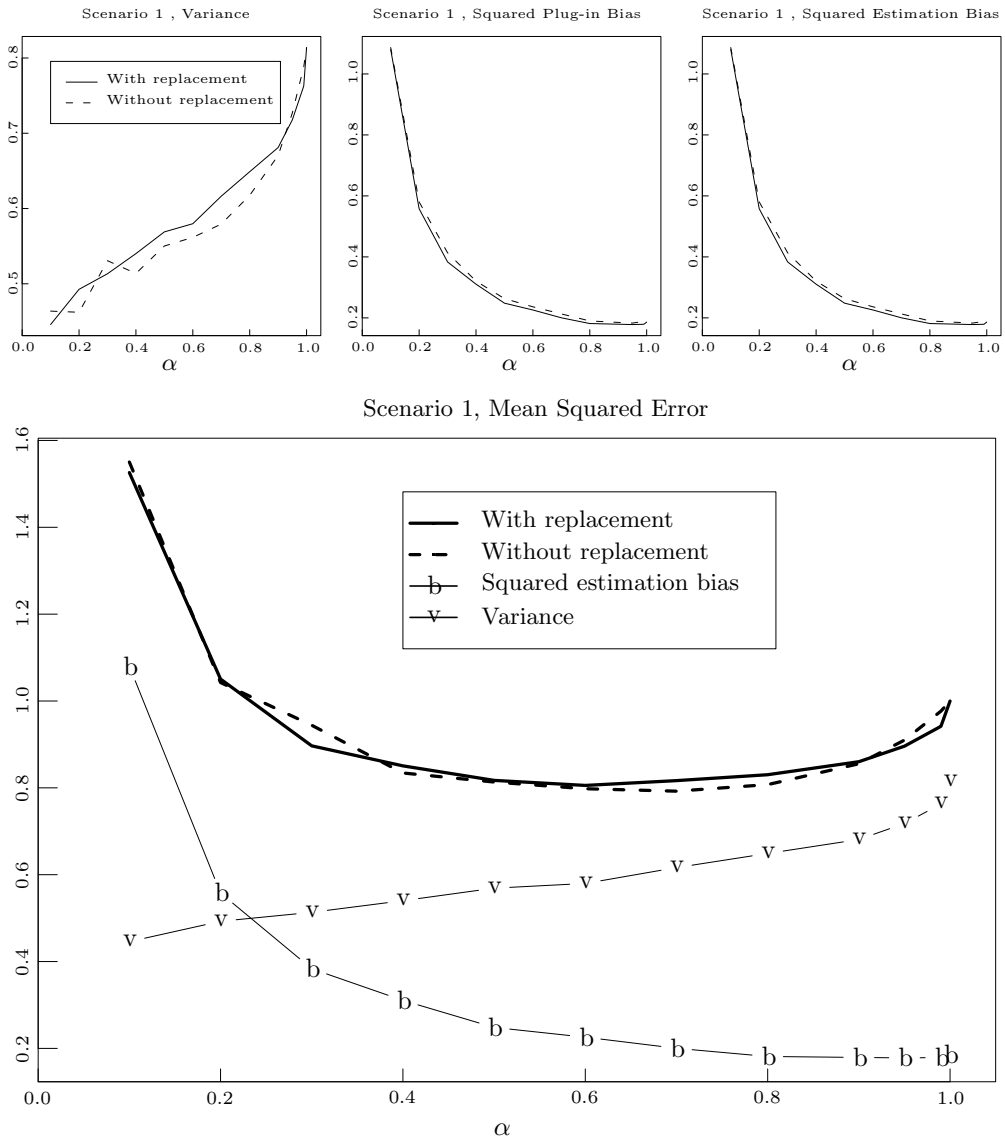

Figure 5.2. Simulation results for Scenario 1. Top panels: Variance, squared plug-in bias, and squared estimation bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared estimation bias for resampling with replacement.

creasing with decreasing resampling fraction. Bagging therefore conveys a double benefit, decreasing both variance and squared (estimation) bias. The explanation is simple: a bagged CART tree is smoother than the corresponding unbagged tree, because bagging smoothes out the discontinuities of a piecewise constant model.

If the true regression function is smooth, smoothing the estimate can be expected to be beneficial. Admittedly, the scenario considered here is highly unrealistic, but the beneficial effect can also be expected in more realistic situations, like Scenario 4 discussed below.

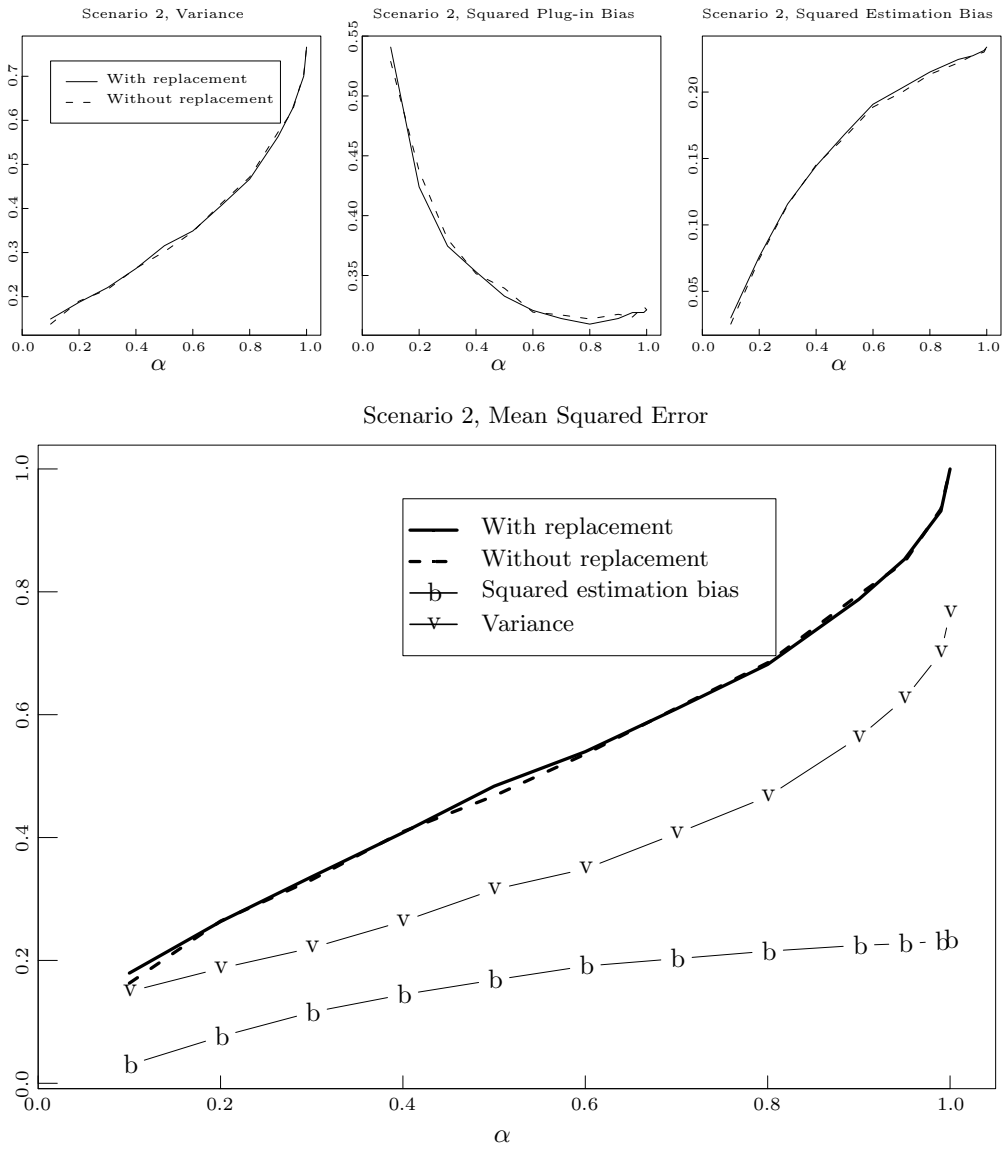

Figure 5.3. Simulation results for Scenario 2. Top panels: Variance, squared plug-in bias, and squared estimation bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared estimation bias for resampling with replacement.

Scenario 3 is analogous to Scenario 1, with 10-dimensional instead of one-dimensional predictor space. The true regression function is piecewise constant and can be consistently estimated by a CART tree with 50 leaves. The results, shown in Figure 5.4, parallel those for Scenario 1.

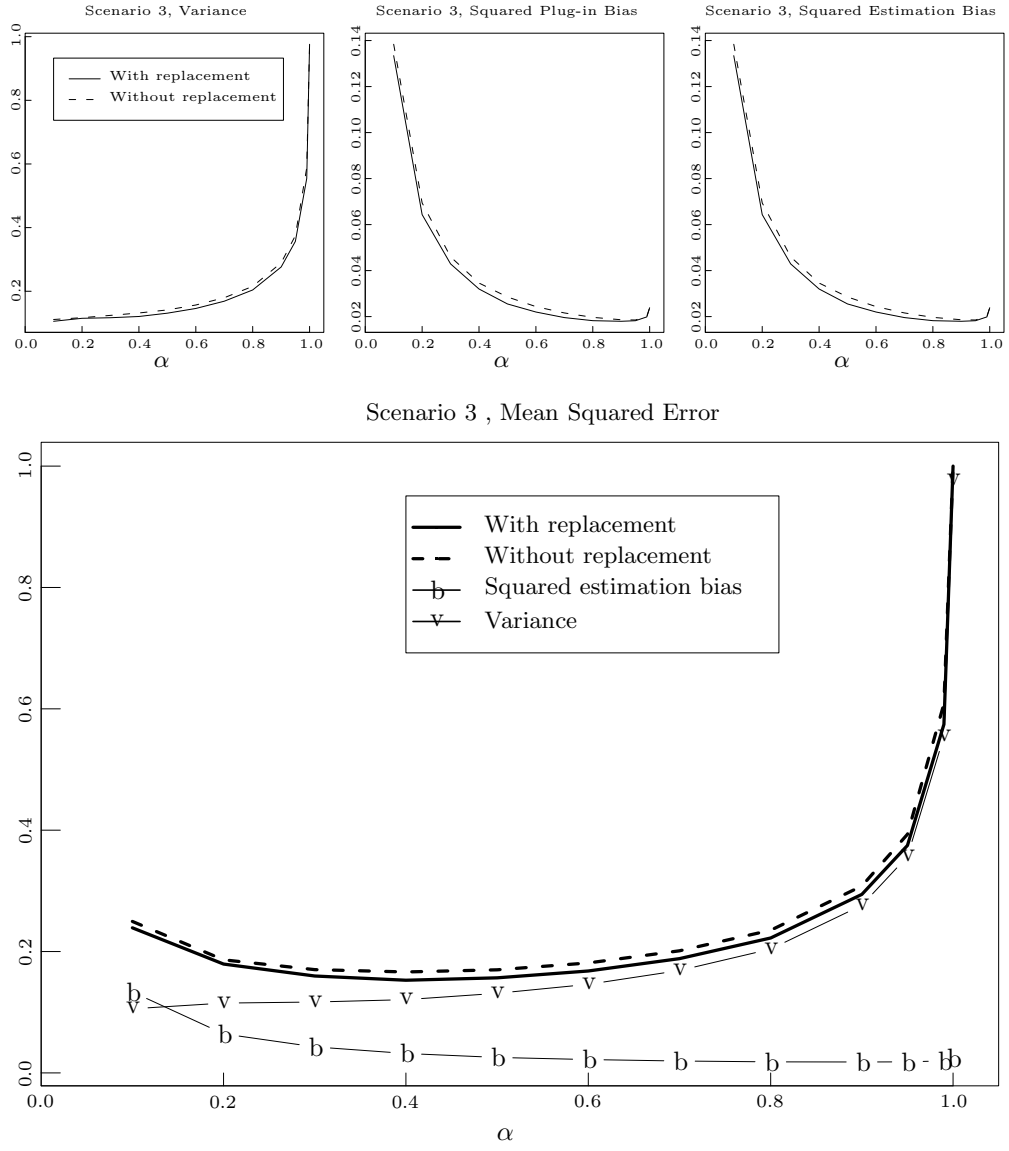

Figure 5.4. Simulation results for Scenario 3. Top panels: Variance, squared plug-in bias, and squared estimation bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared estimation bias for resampling with replacement.

The results for Scenario 4, shown in Figure 5.5, closely parallel those for Scenario 2. Again, both variance and squared (estimation) bias decrease with decreasing resampling fraction.

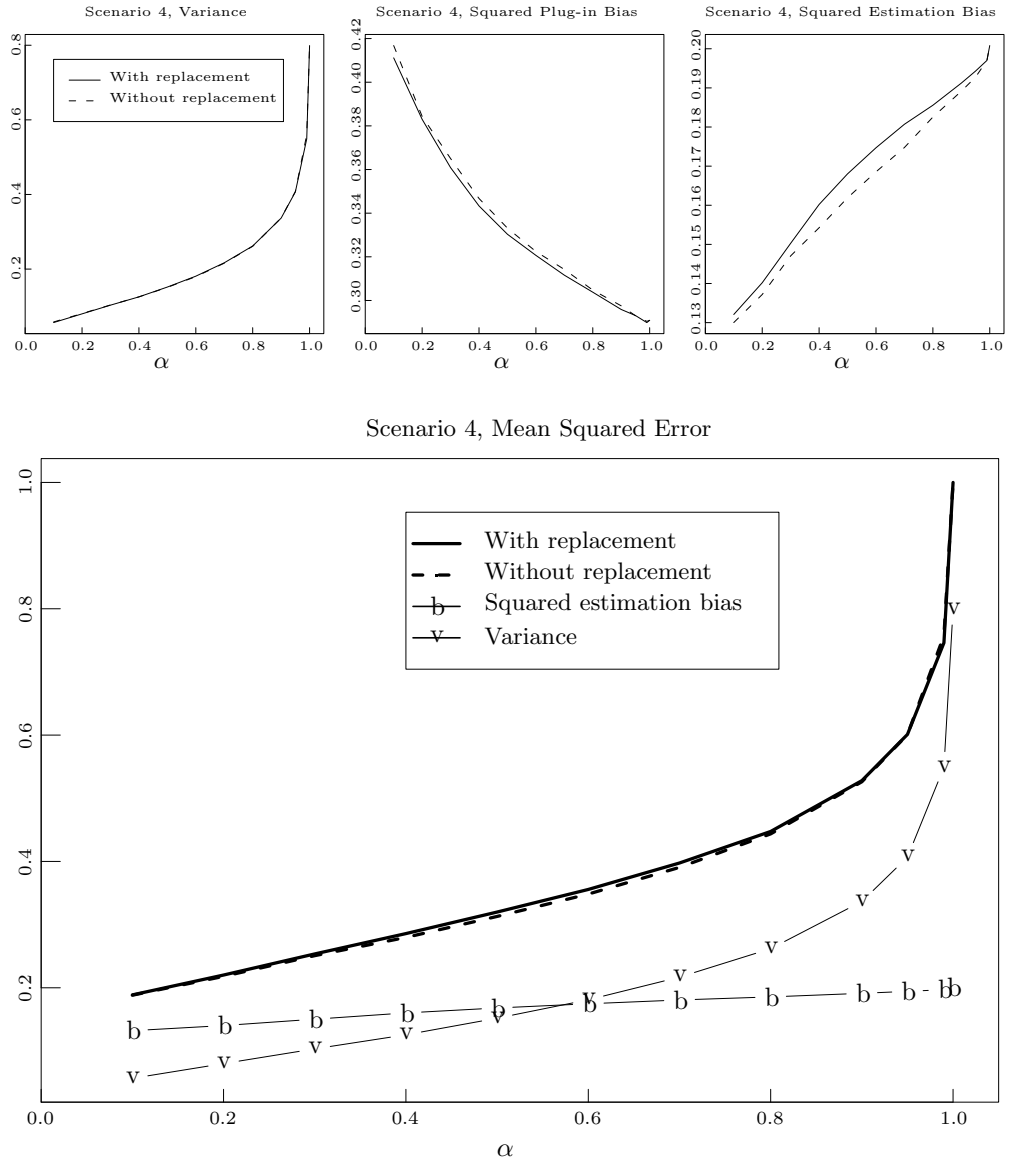

Figure 5.5. Simulation results for Scenario 4. Top panels: Variance, squared plug-in bias, and squared estimation bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared estimation bias for resampling with replacement.

The experiments confirm the agreement between bagging with and without replacement predicted by the theory developed in Section 3: Bagging without replacement with resample size $N\,\alpha$ gives almost the same results in terms of bias, variance, and MSE as bagging with replacement with resample size $N\,\alpha/(1-\alpha)$.

The experiments also confirm that bagging does increase squared plug-in bias. However, the relevant quantity in a regression context is estimation bias: if the true regression function is smooth, bagging can in fact *reduce* estimation bias as well as variance and therefore yield a double benefit.

6. Summary and Conclusions

We studied the effects of bagging for U-statistics of any order and finite sums thereof. U-statistics of high order can describe complex data dependencies and yet they admit a rigorous asymptotic analysis. The findings are as follows.

- The effects of bagging on variance, squared plug-in bias and mean squared error are of order N^{-2}. This may not seem to explain the sometimes considerable improvements due to bagging seen in trees; on the other hand, pointwise tree-based function estimates often rely on small terminal nodes, implying a small N in terms of our theory and hence allowing sizable effects even for order N^{-2}. (The following statements are all valid to second order.)
- If one allows boostrap samples *with or without* replacement and arbitrary resample sizes, then bagging based on "sampling with" for resample size M_{with} is equivalent to "sampling without" for resample size $M_{w/o}$ if $N/M_{with} = N/M_{w/o} - 1 = g$ $(> 0,\ < \infty)$. While our derivation is limited to U-statistics and sums thereof, the equivalence seems to hold more widely, as illustrated by our experiments with bagged CART trees.
- Var(bagged) − Var(raw) is a linear function of g; bagging improves variance if the slope is negative.
- Bias^2(bagged) − Bias^2(raw) is a positive quadratic function of g; bagging hence always increases squared bias.
- MSE^2(bagged) − MSE^2(raw) is a quadratic function of g; bagging may or may not improve mean squared error, and if it does, it is for sufficiently small g, that is, sufficiently large resample sizes M.

Even though CART trees and U-statistics are quite different, there is *qualitative* agreement between our theoretical findings and the experimental results for trees. Plug-in bias increases and variance decreases with decreasing resample size, as predicted by the theory, and the theoretically predicted equivalence between bagging with and without replacement is indeed observed in the experiments.

7. Appendix

7.1. Summation patterns for U-statistics

The calculations for U-statistics in this and the following sections are reminiscent of those found in Hoeffding (1948). We introduce notation for statistical functionals that are interactions of order J and K, respectively:

$$\mathbf{B} = \frac{1}{N^J} \sum_{\mu} B_\mu, \quad \mathbf{C} = \frac{1}{N^K} \sum_{\nu} C_\nu,$$

where

$$\mu = (\mu_1, \ldots, \mu_J) \in \{1, \ldots, N\}^J, \quad B_\mu = B_{X_{\mu_1}, \ldots, X_{\mu_J}},$$

$$\nu = (\nu_1, \ldots, \nu_K) \in \{1, \ldots, N\}^K, \quad C_\nu = C_{X_{\nu_1}, \ldots, X_{\nu_K}}.$$

We assume the functions $B_{x_1,\ldots,x_J}$ and $C_{y_1,\ldots,y_K}$ to be permutation symmetric in their arguments, the random variables $X_1, \ldots, X_N$ to be i.i.d., and the second moments of B_μ and C_ν to exist for all μ and ν. As is usual in the context of von Mises expansions, we do not limit the summations to distinct indices as is usual in the context of U-statistics. One reason is that we wish $\mathbf{B}$ and $\mathbf{C}$ to be plug-in estimates of the functionals $\mathbf{E}\, B_{1,\ldots,J}$ and $\mathbf{E}\, C_{1,\ldots,K}$. Another reason is that bagging produces lower order interactions from higher order, as we will see.

In what follows we will need to partition sums such as σ_μ according to how many indexes appear multiple times in $\mu = (\mu_1, \ldots, \mu_J)$. To this end, we introduce $t(\mu)$ as the numbers of "essential ties" in μ:

$$t(\mu) = \#\{(i,j) | i < j, \mu_i = \mu_j, \quad \mu_i \neq \mu_1, \ldots, \mu_{i-1}\}.$$

The sub-index i marks the first appearance of the index μ_i, and all other μ_j equal to μ_i are counted relative to i. For example, $\mu = (1, 1, 2, 1, 2)$ has three essential ties: $\mu_1 = \mu_2$, $\mu_1 = \mu_4$, and $\mu_3 = \mu_5$; the tie $\mu_2 = \mu_4$ is inessential because it can be inferred from the essential ties.

An important observation concerns the counts of indexes with a given number of essential ties. The following will be used repeatedly:

$$\begin{aligned}
\#\{\mu \mid t(\mu) = 0\} &= \begin{bmatrix} N \\ J \end{bmatrix} = O(N^J) \ , \\
\#\{\mu \mid t(\mu) = 1\} &= \begin{bmatrix} N \\ J \end{bmatrix} \binom{J}{2} = O(N^{J-1}) \ , \\
\#\{\mu \mid t(\mu) = 0\} &= O(N^{J-2}) \ .
\end{aligned}$$

Another notation we need is for the number $c(\mu,\nu)$ of essential cross-ties between μ and ν:

$$c(\mu,\nu) = \#\{\ (i,j) \mid \mu_i = \nu_j\ ,\ \ \mu_i \neq \mu_1,\ldots,\mu_{i-1}\ ,\nu_j \neq \nu_1,\ldots,\nu_{j-1}\ \}\ .$$

We exclude inessential cross-ties that can be inferred from the ties within μ and ν. For example, for $\mu = (1,2,1)$ and $\nu = (3,1)$ the only essential cross-tie is $\mu_1 = \nu_2 = 1$; the remaining inessential cross-tie $\mu_3 = \nu_2$ can be inferred from the essential tie $\mu_1 = \mu_3$ within μ.

With these definitions we have the following fact for the number of essential ties of the concatenated sequence (μ,ν):

$$t((\mu,\nu)) = t(\mu) + t(\nu) + c(\mu,\nu)\ .$$

7.2. Covariance of general interactions

In expanding the covariance between $\mathbf{B}$ and $\mathbf{C}$, we note that the terms with zero cross-ties between μ and ν vanish due to independence. Thus:

$$\mathrm{Cov}\,(\mathbf{B},\mathbf{C}) = \frac{1}{N^{J+K}} \sum_{c(\mu,\nu)>0} \mathrm{Cov}\,(B_\mu, C_\nu)\ .$$

Because $\#\{(\mu,\nu) \mid c(\mu,\nu) > 0\ \}$ is of order $O(N^{J+K-1})$ (a crude upper bound is JKN^{J+K-1}), it follows that $\mathrm{Cov}\,(\mathbf{B},\mathbf{C})$ is of order $O(N^{-1})$, as it should.

We now show that in order to capture terms of order N^{-1} and N^{-2} in $\mathrm{Cov}\,(\mathbf{B},\mathbf{C})$ it is sufficient to limit the summation to those (μ,ν) that satisfy either

- $t(\mu) = 0$, $t(\nu) = 0$ and $c(\mu,\nu) = 1$, or
- $t(\mu) = 1$, $t(\nu) = 0$ and $c(\mu,\nu) = 1$, or
- $t(\mu) = 0$, $t(\nu) = 1$ and $c(\mu,\nu) = 1$,

or $t(\mu) + t(\nu) = 0,1$ and $c(\mu,\nu) = 1$ for short. To this end, we note that the number of terms with $t(\mu) + t(\nu) \geq 2$ and $c(\mu,\nu) \geq 1$ is of order N^{J+K-3}. This is seen from the following crude upper bound:

$$\begin{aligned}
&\#\{\ (\mu,\nu) \mid t(\mu) + t(\nu) \geq 2\ ,\ c(\mu,\nu) \geq 1\ \} \\
&\leq \#\{\ (\mu,\nu) \mid t((\mu,\nu)) \geq 3\ \} \\
&\leq \left(\binom{K+J}{4, K+J-4} + \binom{K+J}{3,2,K+J-5} + \binom{J+K}{2,2,2,J+K-6} \right) \cdot N^{J+K-3}\ ,
\end{aligned}$$

where the "choose" terms arise from choosing the index patterns $(1,1,1,1)$, $(1,1,1,2,2)$ and $(1,1,2,2,3,3)$ in all possible ways in a sequence (μ,ν) of length

$K + J$; these three patterns are necessary and sufficient for $t((\mu, \nu)) \geq 3$. Using N^{J+K-3} instead of $N(N-1)\cdots(N-(J+K-4))$ makes this an upper bound.

With the assumption of finite second moments of B_μ and C_ν for all μ and ν, it follows that the sum of terms with $t(\mu) + t(\nu) \geq 2$ and $c(\mu, \nu) \geq 1$ is of order $O(N^{-3})$. Abbreviating

$$\begin{bmatrix} N \\ L \end{bmatrix} = \frac{N!}{(N-L)!} = N(N-1)\cdots(N-(L-1))$$

we have:

$$\begin{aligned}
\mathrm{Cov}\,(\mathbf{B}, \mathbf{C}) &= \frac{1}{N^{J+K}} \sum_{t(\mu)+t(\nu)=0,1; c(\mu,\nu)=1} \mathrm{Cov}\,(B_\mu, C_\nu) + O(N^{-3}) \\
&= \frac{1}{N^{J+K}} \sum_{t(\mu)=0,\ t(\nu)=0, c(\mu,\nu)=1} \mathrm{Cov}\,(B_\mu, C_\nu) \\
&\quad + \frac{1}{N^{J+K}} \sum_{t(\mu)=1, t(\nu)=0, c(\mu,\nu)=1} \mathrm{Cov}\,(B_\mu, C_\nu) \\
&\quad + \frac{1}{N^{J+K}} \sum_{t(\mu)=0, t(\nu)=1, c(\mu,\nu)=1} \mathrm{Cov}\,(B_\mu, C_\nu) + O(N^{-3}) \\
&= \frac{1}{N^{J+K}} JK \begin{bmatrix} N \\ J+K-1 \end{bmatrix} \cdot \mathrm{Cov}\,(B_{(1,\ldots)}, C_{(1,\ldots)}) \\
&\quad + \frac{1}{N^{J+K}} \binom{J}{2} KN \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \cdot \big(\mathrm{Cov}\,(B_{(1,1,\ldots)}, C_{(1,\ldots)}) \\
&\quad + \mathrm{Cov}\,(B_{(1,1,2,\ldots)}, C_{(2,\ldots)})\big) \\
&\quad + \frac{1}{N^{J+K}} J \binom{K}{2} N \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \cdot \big(\mathrm{Cov}\,(B_{(1,\ldots)}, C_{(1,1,\ldots)}) \\
&\quad + \mathrm{Cov}\,(B_{(2,\ldots,J)}, C_{(1,1,2,\ldots)})\big) + O(N^{-3}),
\end{aligned}$$

where "..." inside a covariance stands for as many *distinct other* indices as necessary. Using

$$\begin{bmatrix} N \\ L \end{bmatrix} = N^L - \binom{L}{2} N^{L-1} + O(N^{L-2})$$

we obtain

$$\begin{aligned}\operatorname{Cov}(\mathbf{B},\mathbf{C}) = &\left(N^{-1} - \binom{J+K-1}{2} N^{-2} + O(N^{-3})\right) JK \cdot \operatorname{Cov}\left(B_{(1,\ldots)}, C_{(1,\ldots)}\right) \\ &+ \left(N^{-2} + O(N^{-3})\right) \binom{J}{2} K \cdot \big(\operatorname{Cov}\left(B_{(1,1,\ldots)}, C_{(1,\ldots)}\right) \\ &+ \operatorname{Cov}\left(B_{(1,1,2,\ldots)}, C_{(2,\ldots)}\right)\big) \\ &+ \left(N^{-2} + O(N^{-3})\right) J \binom{K}{2} \cdot \big(\operatorname{Cov}\left(B_{(1,\ldots)}, C_{(1,1,\ldots)}\right) \\ &+ \operatorname{Cov}\left(B_{(2,\ldots)}, C_{(1,1,2\ldots)}\right)\big) + O(N^{-3}).\end{aligned}$$

Collecting terms $O(N^{-3})$, the above can be written in a more sightly manner as

$$\begin{aligned}\operatorname{Cov}(\mathbf{B},\mathbf{C}) = &\left(N^{-1} - \binom{J+K-1}{2} N^{-2}\right)\ JK \cdot \operatorname{Cov}\left(B_X, C_X\right) \\ &+N^{-2}\ \binom{J}{2} K \cdot \left(\operatorname{Cov}\left(B_{X,X}, C_X\right) + \operatorname{Cov}\left(B_{X,X,Y}, C_Y\right)\right) \\ &+N^{-2}\ J \binom{K}{2} \cdot \left(\operatorname{Cov}\left(B_X, C_{X,X}\right) + \operatorname{Cov}\left(B_X, C_{X,Y,Y}\right)\right) + O(N^{-3}) \\ = &\, a \cdot N^{-1} + b \cdot N^{-2} + O(N^{-3})\ .\end{aligned}$$

7.3. Moments of resampling coefficients

We consider sampling in terms of M draws from N objects $\{1,\ldots,N\}$ with and without replacement. The draws are M exchangeable random variables $\S_1,\ldots,\S_M$, where $\S_i \in \{1,\ldots,N\}$. Each draw is equally likely: $P[\S_i = n] = N^{-1}$, but for sampling with replacement the draws are independent; for sampling w/o replacement they are dependent and the joint probabilities are $P[\S_1 = n_1, \S_2 = n_2, \ldots, \S_J = n_J] = \begin{bmatrix} M \\ J \end{bmatrix} \Big/ \begin{bmatrix} N \\ J \end{bmatrix}$ for distinct n_i's, and $= 0$ if ties exist among the n_i's.

For resampling one is interested in the count variables

$$W_{n,M,N} = W_n = \sum_{\mu=1,\ldots,M} 1_{[\S_\mu = n]}\ ,$$

where we drop M and N from the subscripts if they are fixed. We let $\mathbf{W} = \mathbf{W}_{M,N} = (W_1,\ldots,W_N)$ and recall:

- For resampling *with* replacement: $\mathbf{W} \sim \text{Multinomial}(1/N,\ldots,1/N; M)$.
- For resampling *w/o* replacement: $\mathbf{W} \sim \text{Hypergeometric}(M,N)$.

For bagging one needs the moments of $\mathbf{W}$. Because of exchangeability of $\mathbf{W}$ for fixed M and N, it is sufficient to consider moments of the form

$$\mathbf{E}\,[W_{n=1,M,N}^{i_1}\ W_{n=2,M,N}^{i_2}\cdots W_{n=L,M,N}^{i_L}]\ .$$

The following recursion formulae hold for $i_l \geq 1$:

$$\begin{aligned}
&\mathbf{E}\,[W_{n=1,M,N}^{i_1}\ W_{n=2,M,N}^{i_2}\cdots W_{n=L,M,N}^{i_L}] \\
&= \begin{cases} \text{with}: \frac{M}{N}\ \mathbf{E}\,[(W_{n=1,M-1,N}+1)^{i_1-1}\ W_{n=2,M-1,N}^{i_2}\cdots W_{n=L,M-1,N}^{i_L}]\ , \\ \text{w/o}: \frac{M}{N}\ \mathbf{E}\,[W_{n=2,M-1,N-1}^{i_2}\cdots W_{n=L,M-1,N-1}^{i_L}]\ . \end{cases}
\end{aligned}$$

From these we derive the moments that will be needed below. Recall $\alpha = M/N$, and $g = 1/\alpha$ for resampling with, $g = (1/\alpha) - 1$ for resampling without, replacement. Using repeatedly approximations such as

$$\begin{bmatrix} N \\ L \end{bmatrix} = N^L - \binom{L}{2}\ N^{L-1} + O(N^{L-2})\ ,$$

we obtain:

$$\mathbf{E}\,[W_1^{i_1}\ W_2^{i_2}\cdots W_L^{i_L}] = O(1)$$

$$\begin{aligned}
\mathbf{E}\,[W_1\ W_2\cdots W_L] &= \begin{cases} \text{with}: \begin{bmatrix} M \\ L \end{bmatrix} \Big/ N^L \\ \text{w/o}: \begin{bmatrix} M \\ L \end{bmatrix} \Big/ \begin{bmatrix} N \\ L \end{bmatrix} \end{cases} \\
&= \begin{cases} \text{with}: \alpha^L - \alpha^L \binom{L}{2} \frac{1}{\alpha} N^{-1} + O(N^{-2}) \\ \text{w/o}:\ \alpha^L - \alpha^L \binom{L}{2} \left(\frac{1}{\alpha} - 1\right) N^{-1} + O(N^{-2}) \end{cases} \\
&= \alpha^L\ \left(1 - \binom{L}{2}\ g\ N^{-1}\right) + O(N^{-2}) \\
\mathbf{E}\,[W_1^2\ W_2\cdots W_{L-1}] &= \begin{cases} \text{with}: \begin{bmatrix} M \\ L \end{bmatrix} \Big/ N^L + \begin{bmatrix} M \\ L-1 \end{bmatrix} \Big/ N^{L-1} \\ \text{w/o}: \begin{bmatrix} M \\ L-1 \end{bmatrix} \Big/ \begin{bmatrix} N \\ L-1 \end{bmatrix} \end{cases} \\
&= \begin{cases} \text{with}: \alpha^L + \alpha^{L-1} + O(N^{-1}) \\ \text{w/o}:\ \alpha^{L-1} + O(N^{-1}) \end{cases} \\
&= \alpha^L(g+1) + O(N^{-1}).
\end{aligned}$$

7.4. Equivalence of resampling with and without replacement

We show the equivalence of resampling with and without replacement to order N^{-2}. To this end we need to distinguish between the resampling sizes M_{with} and $M_{w/o}$, and the corresponding resampling fractions $\alpha_{with} = M_{with}/N$ and $\alpha_{w/o} = M_{w/o}/N$. The equivalence holds under the condition

$$\frac{1}{\alpha_{with}} = \frac{1}{\alpha_{w/o}} - 1 \ (=: g) \ .$$

The two types of bagged U-statistics are denoted, respectively, by

$$\mathbf{B}^{with} = \frac{1}{M_{with}^J} \sum_{\mu} \ \mathbf{E}\left[W_{\mu_1}^{with} \cdots W_{\mu_J}^{with}\right] \cdot B_\mu \ ,$$

$$\mathbf{B}^{w/o} = \frac{1}{M_{w/o}^J} \sum_{\mu} \ \mathbf{E}\left[W_{\mu_1}^{w/o} \cdots W_{\mu_J}^{w/o}\right] \cdot B_\mu \ .$$

Bagging differentially reweights the parts of a general interaction in terms of moments of the resampling vector $\mathbf{W}$. The result of bagging is no longer a pure interaction but a general U-statistic because bagging creates lower-order interactions from higher orders.

Recall two facts about the bagging weights, that is, the moments of $\mathbf{W}$: 1) They depend on the structure of the ties in the index vectors $\mu = (\mu_1, \ldots, \mu_J)$ only; for example, $\mu = (1, 1, 2)$ and $\mu = (3, 2, 3)$ have the same weights, $\mathbf{E}\,[W_1^2 W_2]$ $= \mathbf{E}\,[W_3^2 W_2]$ due to exchangeability. 2) The moments of $\mathbf{W}$ are of order $O(1)$ in N (Subsection 7.3) and hence preserve the orders $O(N^{-1})$, $O(N^{-2})$, $O(N^{-3})$ of the terms considered in Subsection 7.1.

We derive a crude bound on their difference using $B_{bound} = \max_\mu |B_\mu|$. We assume the above condition on α_{with} and $\alpha_{w/o}$ and obtain:

$$\begin{aligned}
&|\mathbf{B}^{with} - \mathbf{B}^{w/o}| \\
&\le \sum_{\mu} \Big| \frac{1}{M_{with}^J} \mathbf{E}\left[W_{\mu_1}^{with} \cdots W_{\mu_J}^{with}\right] - \frac{1}{M_{w/o}^J} \mathbf{E}\left[W_{\mu_1}^{w/o} \cdots W_{\mu_J}^{w/o}\right] \Big| \cdot B_{bound} \\
&= \Big(\sum_{t(\mu)=0} + \sum_{t(\mu)=1} + \sum_{t(\mu)>1} \Big) |\cdots| \cdot B_{bound} \\
&= \sum_{t(\mu)=0} \Big| \frac{1}{M_{with}^J} \Big[\alpha_{with}^J \Big(1 - \binom{J}{2} g N^{-1} \Big) + O(N^{-2}) \Big] \\
&\quad - \frac{1}{M_{w/o}^J} \Big[\alpha_{w/o}^J \Big(1 - \binom{J}{2} g N^{-1} \Big) + O(N^{-2}) \Big] \Big| \cdot B_{bound} \\
&\quad + \sum_{t(\mu)=1} \Big| \frac{1}{M_{with}^J} \Big[\alpha_{with}^J (g+1) + O(N^{-1}) \Big]
\end{aligned}$$

$$
\begin{aligned}
&\quad -\frac{1}{M^J_{w/o}}\Big[\alpha^J_{w/o}(g+1)+O(N^{-1})\Big]\Big|\cdot B_{bound}\\
&\quad +\sum_{t(\mu)>1}\Big|\frac{1}{M^J_{with}}[O(1)]-\frac{1}{M^J_{w/o}}[O(1)]\Big|\cdot B_{bound}\\
&=\frac{1}{N^J}\Big(\sum_{t(\mu)=0}O(N^{-2})+\sum_{t(\mu)=1}O(N^{-1})+\sum_{t(\mu)>1}O(1)\Big)\cdot B_{bound}\\
&=\frac{1}{N^J}\Big[\begin{bmatrix}N\\J\end{bmatrix}O(N^{-2})+\begin{bmatrix}N\\J-1\end{bmatrix}\binom{J}{2}O(N^{-1})\\
&\quad +\Big(N^J-\begin{bmatrix}N\\J\end{bmatrix}-\begin{bmatrix}N\\J-1\end{bmatrix}\binom{J}{2}\Big)O(1)\Big]\cdot B_{bound}\\
&=\frac{1}{N^J}\Big[O(N^J)O(N^{-2})+O(N^{J-1})O(N^{-1})+O(N^{J-2})O(1)\Big]\cdot B_{bound}\\
&=O(N^{-2})\cdot B_{bound}
\end{aligned}
$$

This proves the per-sample equivalence of bagging based on resampling with and without replacement up to order $O(N^{-2})$. The result is somewhat unsatisfactory because the bound depends on the extremes of the U-terms B_μ, which tend to infinity for $N\to\infty$, unless B_μ is bounded. Other bounds at a weaker rate can be obtained with the Hölder inequality:

$$
|\mathbf{B}^{with}-\mathbf{B}^{w/o}|\ \le\ O\Big(N^{-\frac{2}{p}}\Big)\Big(\frac{1}{N^J}\sum_\mu|B_\mu|^q\Big)^{\frac{1}{q}}\qquad\text{for}\ \ \frac{1}{p}+\frac{1}{q}=1\ .
$$

This specializes to the previously derived bound when $p=1$ and $q=\infty$, for which the best rate of $O(N^{-2})$ is obtained, albeit under the strongest assumptions on B_μ.

7.5. Covariances of bagged interactions

Resuming calculations begun in Subsection 7.2 for covariances of unbagged interaction terms, we now derive the covariance of their M-bagged versions:

$$
\mathbf{B}^{bag}=\frac{1}{M^J}\sum_\mu\mathbf{E}\left[W_{\mu_1}\cdots W_{\mu_J}\right]\cdot B_\mu,\quad \mathbf{C}^{bag}=\frac{1}{M^K}\sum_\nu\mathbf{E}\left[W_{\nu_1}\cdots W_{\nu_K}\right]\cdot C_\nu.
$$

The moment calculations of Subsection 7.3 yield the following:

$$
\begin{aligned}
&\operatorname{Cov}(\mathbf{B}^{bag},\mathbf{C}^{bag})\\
&=\frac{1}{M^{J+K}}\sum_{t(\mu)+t(\nu)=0,1,c(\mu,\nu)=1}\mathbf{E}\left[W_{\mu_1}\cdots W_{\mu_J}\right]\mathbf{E}\left[W_{\nu_1}\cdots W_{\nu_K}\right]\operatorname{Cov}(B_\mu,C_\nu)\\
&\quad +O(N^{-3})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{M^{J+K}} \sum_{t(\mu)=0,t(\nu)=0,c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1}\cdots W_{\mu_J}\right]\mathbf{E}\left[W_{\nu_1}\cdots W_{\nu_K}\right]\operatorname{Cov}\left(B_\mu, C_\nu\right) \\
&\quad + \frac{1}{M^{J+K}} \sum_{t(\mu)=1,t(\nu)=0,c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1}W_{\mu_2}\cdots W_{\mu_J}\right]\mathbf{E}\left[W_{\nu_1}\cdots W_{\nu_K}\right]\operatorname{Cov}\left(B_\mu, C_\nu\right) \\
&\quad + \frac{1}{M^{J+K}} \sum_{t(\mu)=0,t(\nu)=1,c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1}W_{\mu_2}\cdots W_{\mu_J}\right]\mathbf{E}\left[W_{\nu_1}W_{\nu_2}\cdots W_{\nu_K}\right] \\
&\quad \cdot \operatorname{Cov}\left(B_\mu, C_\nu\right) + O(N^{-3}) \\
&= \frac{1}{N^{J+K}\alpha^{J+K}} JK \begin{bmatrix} N \\ J+K-1 \end{bmatrix} \mathbf{E}\left[W_1\cdots W_J\right]\mathbf{E}\left[W_1\cdots W_K\right]\operatorname{Cov}\left(B_{(1,\ldots)}, C_{(1,\ldots)}\right) \\
&\quad + \frac{1}{N^{J+K}\alpha^{J+K}} \binom{J}{2} KN \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \mathbf{E}\left[W_1^2 W_2\cdots W_{J-1}\right]\mathbf{E}\left[W_1\cdots W_K\right] \\
&\quad \cdot \left(\operatorname{Cov}\left(B_{(1,1,\ldots)}, C_{(1,\ldots)}\right) + \operatorname{Cov}\left(B_{(1,1,2,\ldots)}, C_{(2,\ldots)}\right)\right) \\
&\quad + \frac{1}{N^{J+K}\alpha^{J+K}} J\binom{K}{2} N \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \mathbf{E}\left[W_1\cdots W_J\right]\mathbf{E}\left[W_1^2 W_2\cdots W_{K-1}\right] \\
&\quad \cdot \left(\operatorname{Cov}\left(B_{(1,\ldots)}, C_{(1,1,\ldots)}\right) + \operatorname{Cov}\left(B_{(2,\ldots)}, C_{(1,1,2,\ldots)}\right)\right) + O(N^{-3}) \\
&= JK\left(N^{-1} - \binom{J+K-1}{2} N^{-2}\right)\left(1 - \binom{J}{2} gN^{-1}\right)\left(1 - \binom{K}{2} gN^{-1}\right) \\
&\quad \cdot \operatorname{Cov}\left(B_X, C_X\right) + \binom{J}{2} KN^{-2}(g+1)\left(\operatorname{Cov}\left(B_{X,X}, C_X\right) + \operatorname{Cov}\left(B_{X,X,Y}, C_Y\right)\right) \\
&\quad + J\binom{K}{2} N^{-2}(g+1)\left(\operatorname{Cov}\left(B_X, C_{X,X}\right) + \operatorname{Cov}\left(B_X, C_{X,Y,Y}\right)\right) + O(N^{-3}) \\
&= \left(N^{-1} - N^{-2}\binom{J+K-1}{2} - N^{-2}\left(\binom{J}{2} + \binom{K}{2}\right) g\right) JK \operatorname{Cov}\left(B_X, C_X\right) \\
&\quad + N^{-2}\binom{J}{2} K(g+1)\left(\operatorname{Cov}\left(B_{X,X}, C_X\right) + \operatorname{Cov}\left(B_{X,X,Y}, C_Y\right)\right) \\
&\quad + N^{-2} J\binom{K}{2} \ (g+1) \ \left(\operatorname{Cov}\left(B_X, C_{X,X}\right) + \operatorname{Cov}\left(B_X, C_{X,Y,Y}\right)\right) + O(N^{-3})
\end{aligned}
$$

The last three lines form the final result of these calculations.

7.6. Difference between variances of bagged and unbagged

Comparing the results of the Sections 7.2 and 7.5, we get:

$$
\begin{aligned}
&\operatorname{Cov}\left(\mathbf{B}^{bag}, \mathbf{C}^{bag}\right) - \operatorname{Cov}\left(\mathbf{B}, \mathbf{C}\right) \\
&= -N^{-2}\left(\binom{J}{2} + \binom{K}{2}\right) \ g \ JK \operatorname{Cov}\left(B_X, C_X\right)
\end{aligned}
$$

$$
\begin{aligned}
&+N^{-2}\binom{J}{2} K\ g\ \left(\mathrm{Cov}\left(B_{X,X}, C_X\right) + \mathrm{Cov}\left(B_{X,X,Y}, C_Y\right)\right) \\
&+N^{-2} J \binom{K}{2}\ g\ \left(\mathrm{Cov}\left(B_X, C_{X,X}\right) + \mathrm{Cov}\left(B_X, C_{X,Y,Y}\right)\right) + O(N^{-3}) \\
= N^{-2}\ g\ &\left(-\left(\binom{J}{2} + \binom{K}{2}\right) JK\, \mathrm{Cov}\left(B_X, C_X\right)\right. \\
&+\binom{J}{2} K \left(\mathrm{Cov}\left(B_{X,X}, C_X\right) + \mathrm{Cov}\left(B_{X,X,Y}, C_Y\right)\right) \\
&\left. +J\binom{K}{2} \left(\mathrm{Cov}\left(B_X, C_{X,X}\right) + \mathrm{Cov}\left(B_X, C_{X,Y,Y}\right)\right)\ \right)\ +\ O(N^{-3}) \\
=\ N^{-2}\ g\ 2\ &S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) + O(N^{-3})\ ,
\end{aligned}
$$

where

$$
\begin{aligned}
S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \Bigg(&\binom{J}{2} K\, \mathrm{Cov}\left(C_X, B_{X,X} + B_{X,Y,Y} - J B_X\right) \\
&+ \binom{K}{2} J\, \mathrm{Cov}\left(B_X,\ C_{X,X} + C_{X,Y,Y} - K C_X\right)\Bigg) .
\end{aligned}
$$

The expression for $S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C})$ remains correct for J and K as low as 1, in which case one interprets $\binom{J}{2} = 0$ and $B_{X,X} = 0$ when $J = 1$, and $B_{X,Y,Y} = 0$ when $J \leq 2$, and similar for C when K =1 or 2. The result generalizes to arbitrary finite sums of interactions

$$
\begin{aligned}
U &= \mathbf{A} + \mathbf{B} + \mathbf{C} + \dots \\
&= \frac{1}{N} \sum_i A_i + \frac{1}{N^2} \sum_{i,j} B_{i,j} + \frac{1}{N^3} \sum_{i,j,k} C_{i,j,k} + \dots\ .
\end{aligned}
$$

Because $S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C})$ is a bilinear form in its arguments, the corresponding constant $S_{\mathrm{Var}}(U)$ for sums of U-statistics can be expanded as follows:

$$
\begin{aligned}
S_{\mathrm{Var}}(U) = &S_{\mathrm{Var}}(\mathbf{A}, \mathbf{A}) + 2\, S_{\mathrm{Var}}(\mathbf{A}, \mathbf{B}) + S_{\mathrm{Var}}(\mathbf{B}, \mathbf{B}) \\
&+ 2\, S_{\mathrm{Var}}(\mathbf{A}, \mathbf{C}) + 2\, S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) + S_{\mathrm{Var}}(\mathbf{C}, \mathbf{C}) + \dots\ ,
\end{aligned}
$$

so that

$$
\mathrm{Var}(U^{bag}) - \mathrm{Var}(U) = N^{-2}\, g\, 2\, S_{\mathrm{Var}}(U) + O(N^{-3})\ .
$$

For example a functional consisting of first and second order terms,

$$
U = \mathbf{A} + \mathbf{B} = \frac{1}{N} \sum_i A_i + \frac{1}{N^2} \sum_{i,j} B_{i,j}\ ,
$$

yields

$$\begin{aligned} S_{\mathrm{Var}}(U) &= S_{\mathrm{Var}}(\mathbf{A},\mathbf{A}) + 2\,S_{\mathrm{Var}}(\mathbf{A},\mathbf{B}) + S_{\mathrm{Var}}(\mathbf{B},\mathbf{B}) \\ &= \mathrm{Cov}\,(A_X,\ B_{X,X} - 2B_X) + 2\ \mathrm{Cov}\,(B_X,\ B_{X,X} - 2B_X) \\ &= \mathrm{Cov}\,(A_X + 2B_X, B_{X,X} - 2B_X)\ . \end{aligned}$$

Note that $S_{\mathrm{Var}}(\mathbf{A},\mathbf{A}) = 0$ because bagging leaves additive statistics unchanged.

7.7. Difference between Squared Bias of Bagged and Unbagged

We consider a single K-th order interaction first, with functional and plug-in statistic

$$\begin{aligned} U(F) &= \mathbf{E}\,C_{(1,2,\ldots,K)}, \\ U(F_N) &= \frac{1}{N^K} \sum_{\nu_1,\ldots,\nu_K=1}^{N} C_{(\nu_1,\ldots,\nu_K)}. \end{aligned}$$

[Recall that C_ν and $C_{(\nu_1,\ldots,\nu_K)}$ are short for $C_{X_{\nu_1},\ldots,X_{\nu_K}}$.] The functional $U(F)$ plays the role of the parameter to be estimated by the statistic $U = U(F_N)$, so that the notion of bias applies. We first calculate the bias for the unbagged statistic U and second for the bagged statistic U^{bag}. Note that $\mathbf{E}\,C_X = \mathbf{E}\,C_{1,\ldots,K} = U(F)$.

$$\begin{aligned} &\mathbf{E}\,[U(F_N)] \\ &= \frac{1}{N^K} \sum_{\nu_1,\ldots,\nu_K} \mathbf{E}\ C_{(\nu_1,\ldots,\nu_K)} \\ &= \frac{1}{N^K} \left(\begin{bmatrix} N \\ K \end{bmatrix} \mathbf{E}\,C_{(1,\ldots,K)} + \binom{K}{2} \begin{bmatrix} N \\ K-1 \end{bmatrix} \mathbf{E}\,C_{(1,1,2,\ldots,K-1)} + O(N^{K-2}) \right) \\ &= U(F) + N^{-1} \binom{K}{2} (\mathbf{E}\,C_{X,X} - \mathbf{E}\,C_X) + O(N^{-2}). \end{aligned}$$

Now for the bias of the bagged statistic:

$$\begin{aligned} \mathbf{E}\,U^{bag} &= \frac{1}{M^K} \sum_{\nu_1,\ldots,\nu_k=1}^{N} \mathbf{E}\,[W_{\nu_1}\cdots W_{\nu_K}]\ \mathbf{E}\ C_{(\nu_1,\ldots,\nu_K)} \\ &= \frac{1}{N^K \alpha^K} \left(\sum_{t(\nu)=0} + \sum_{t(\nu)=1} + O(N^{K-2}) \right) \\ &= \frac{1}{N^K \alpha^K} \left(\begin{bmatrix} N \\ K \end{bmatrix} \mathbf{E}\,[W_1\cdots W_K] \mathbf{E}\,C_{(1,\ldots,K)} \right. \\ &\quad \left. + \binom{K}{2} \begin{bmatrix} N \\ K-1 \end{bmatrix} \mathbf{E}\,[W_1^2 W_2 \cdots W_{K-1}] \mathbf{E}\,C_{(1,1,2,\ldots,K-1)} \right) + O(N^{-2}) \end{aligned}$$

$$
\begin{aligned}
&= \left(1 - \binom{K}{2} N^{-1}\right) \left(1 - \binom{K}{2}\ g\ N^{-1}\right)\ \mathbf{E}\ C_{(1,\ldots,K)} \\
&\quad + N^{-1} \binom{K}{2}\ (g+1)\ \mathbf{E}\ C_{(1,1,2,\ldots,K-1)} + O(N^{-2}) \\
&= U(F) - N^{-1} \binom{K}{2}\ (g+1)\ \mathbf{E}\ C_{(1,\ldots,K)} \\
&\quad + N^{-1} \binom{K}{2}\ (g+1)\ \mathbf{E}\ C_{(1,1,2,\ldots,K-1)} + O(N^{-2}) \\
&= U(F) + N^{-1} \binom{K}{2}\ (g+1)\ (\mathbf{E}\ C_{X,X} - \mathbf{E}\ C_X) + O(N^{-2})
\end{aligned}
$$

Thus:

$$
\text{Bias}\,(U^{bag}) = N^{-1} \binom{K}{2}\ (g+1)\ (\mathbf{E}\ C_{X,X} - \mathbf{E}\ C_X) + O(N^{-2})
$$

As for variances, we can consider statistics that are finite sums of interactions:

$$
\begin{aligned}
U &= \mathbf{A} + \mathbf{B} + \mathbf{C} + \ldots \\
&= \frac{1}{N} \sum A_i + \frac{1}{N^2} \sum B_{i,j} + \frac{1}{N^3} \sum C_{i,j,k} + \ldots
\end{aligned}
$$

The final result is:

$$
\begin{aligned}
&\text{Bias}^{\,2}(U^{bag}) - \text{Bias}^{\,2}(U) \\
&= N^{-2} \left((g+1)^2 - 1\right) \left(\binom{2}{2} (\mathbf{E}\, B_{X,X} - \mathbf{E}\, B_X) + \binom{3}{2} (\mathbf{E}\, C_{X,X} - \mathbf{E}\, C_X) + \ldots \right)^2 \\
&\quad + O(N^{-3}).
\end{aligned}
$$

As usual, $g = 1/\alpha$ for sampling with, and $g = (1/\alpha) - 1$ for sampling w/o, replacement.

Acknowledgement

This work was begun while the authors were with AT&T Labs, the first author on the technical staff, the second author on sabbatical from the University of Washington. We thank Daryl Pregibon for his support, and two anonymous referees for insightful comments.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**, 123-140.

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

Buhlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.* **30**, 927-961.

Chen, S. X. and Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statist. Sinica* **13**, 97-109.

Friedman, J. H. and Hall, P. (2000). On bagging and nonlinear estimation. Can be downloaded from http://www.stat.stanford.edu/∼jhf/#reports.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.

Knight, K. and Bassett, Jr. G. W. (2002). Second order improvements of sample quantiles using subsamples. Can be downloaded from http://www.utstat.utoronto.ca/keith/papers/subsample.ps.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, U.S.A.

E-mail: buja@wharton.upenn.edu

Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, U.S.A.

E-mail: wxs@stat.washington.edu

(Received July 2005; accepted October 2005)