Observations on Bagging

Andreas Buja[1]
Statistics Department, The Wharton School, University of Pennsylvania, USA

Werner Stuetzle[2]
Department of Statistics, University of Washington, Seattle, USA

December 28, 2002

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.

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. We extend the definition of bagging from statistics (defined on samples) to statistical functionals (defined on distributions), and we study the von Mises expansion of bagged statistical functionals. We show that a bagged functional is smooth in the sense that the von Mises expansion is *finite* of length 1 + resample size M. The resample size may be different from the original sample size N; it acts as a smoothing parameter, where smaller M means more smoothing.

We then study the effects of bagging on U-statistics. 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.

[1] Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302; www-stat.wharton.upenn.edu/~buja. Part of this work was performed while the first author was with AT&T Labs.

[2] Department of Statistics, University of Washington, Seattle, WA 98195-4322; wxs@stat.washington.edu. Research partially supported by NSF grant DMS-9803226 and NSA grant 62-1942. Part of this work was performed while the second author was on sabbatical leave at AT&T Labs.

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.

In this article 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 statistics $\theta(X_1, \ldots, X_N)$ is defined as averaging it over bootstrap samples $X_1^*, \ldots, X_N^*$:

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

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

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

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.

In the literature one sometimes finds a slight generalization of bagging in which the bootstrap resample size is permitted to be different from the sample size N. That is, one permits averaging over resamples $\{\, X_1^*, \ldots, X_M^*\}$ of arbitrary size M. We call the resulting procedure M-bagging and denote the M-bagged statistic by

$$\theta_M^B(X_1, \ldots, X_N) \;=\; \text{ave}\; \theta^B(X_1^*, \ldots, X_M^*) \;.$$

The resample size M is typically N or less. We will, however, have reason to consider resampling sizes M greater than N as well.

A goal of the present article is to contribute to the theoretical understanding of bagging. In a first step, we extend the notion of bagging to statistical functionals, that is, functions of distributions. Most statistics can be derived from statistical functionals, for example, means of samples from expectations, or medians of samples from medians of distributions. We will denote statistical functionals by $\theta(F)$, where the argument F is a general distribution. The link between statistics $\theta(X_1, \ldots, X_N)$ and statistical functionals $\theta(F)$ is provided by the interpretation of i.i.d. samples $X_1, \ldots, X_N$ as empirical distributions:

$$F_N \;=\; \frac{1}{N} \sum_{i=1,\ldots,N} \delta_{X_i} \;,$$

where δ_x denotes a unit point mass distribution at x. The specialization of statistical functionals to empirical distributions yields statistics:

$$\theta(F_N) \;=\; \theta(X_1, \ldots, X_N) \;.$$

In order to extend bagging to statistical functionals, we cast it in terms of empirical distributions. Just like F_N is the empirical distribution corresponding to an i.i.d. sample $X_1, \ldots, X_N$ from F, we let F^*_M be the empirical distribution of an i.i.d. bootstrap sample $X^*_1, \ldots, X^*_M$ from F_N. We write $F_N \sim F$ and $F^*_M \sim F_N$. With these conventions, an M-bagged statistic can be written as

$$\theta^B_M(F_N) \;=\; \mathbf{E}_{F_N} \theta(F^*_M) \;,$$

where $\mathbf{E}_{F_N}$ is the expectation referring to $F^*_M \sim F_N$. The extension of bagging to statistical functionals is now obtained by permitting arbitrary distributions F where empirical distributions F_N have appeared so far. We define the bagged functional $\theta^B_M(F)$ by

$$\theta^B_M(F) \;=\; \mathbf{E}_F \theta(F^*_M) \;.$$

Alternatively, the right hand side can also be written as an intergral if one writes $\theta(F^*_M)$ as $\theta(x^*_1, \ldots, x^*_M)$:

$$\theta^B_M(F) \;=\; \int \theta(x^*_1, \ldots, x^*_M) \,\mathrm{d}F(x^*_1) \ldots \mathrm{d}F(x^*_M)$$

The primary motivation for formulating a definition of bagged functionals is the availability of a potent technical tool for the analysis of statistical functionals, namely, the von Mises expansion. This is a kind of Taylor expansion and allows a similar interpretation.

It turns out that (i) bagged functionals have finite von Mises expansions, and (ii) the length of the expansions is given by the resample size M. This simple finding may well bolster the case in favor of the above definition of bagged functionals. For one thing, it yields an extremely simple characterization of the role of the resample size M as a smoothing parameter: the smaller M, the smoother the bagged functional.

The von Mises expansion of a functional is typically used to derive asymptotic properties of the corresponding statistic. The von Mises expansion of an M-bagged functional, however, leads to asymptotics that are open to criticism: Asymptotics with fixed resample size M and $N \to \infty$ is unrealistic; M should be a fixed fraction of M, or at least $M \to \infty$ albeit at a slower rate than N. While these types of asymptotics should certainly be pursued (as we will later in this article), asymptotics based on fixed resample size M should not be discarded out of hand. We will, in fact, give an example where this type of asymptotics produces excellent approximations to finite sample variances for realistic values of M and N.

The above definition of a bagged statistical functional has a blind spot: It would be interesting to consider both conventional bootstrap sampling with replacement as well as subsampling without replacement (as in Friedman and Hall (2000), and Buhlmann and Yu (2000), for example). If the bootstrap is extended to infinite populations, however, the difference between sampling with and without replacement disappears. Thus, in order to study both sampling modes, we have to work with finite samples and statistics as opposed to distributions and statistical functionals. This is in fact what we do in the second part of the article.

The class of statistics we consider is that of finite sums of U-statistics. We obtain such sums by applying functionals with finite von Mises expansion to empirical distributions. 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 (2003).)

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 resampling without replacement is done with a fraction $\alpha_{w/o} = M_{w/o}/N$ of the sample size N, and resampling with replacement with a multiple $\alpha_{with} = M_{with}/N$ of the sample size, where

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

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 expansions, 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 (2002) and Knight and Bassett (2002), we obtain effects that are only of order $O(N^{-2})$. 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 the 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 values $\alpha_{with} = 1$ and $\alpha_{w/o} = 1/2$, improvements in MSE are obtained only if the resample sizes falls in the respective critical intervals. However, there can arise odd situations in which the MSE is improved only for $\alpha_{with} > 1$ and $\alpha_{w/o} > 1/2$. Details are given in Section 8.3.

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 the notion of bias for simple statistics differs from the notion of bias for fitted functions:

$$\mathbf{E}\,\theta(F_N) - \theta(F) \qquad \text{versus} \qquad \mathbf{E}\,\theta(x|F_N) - f(x) \;,$$

where, as usual in non-parametric fitting, $\theta_N(x|F_N)$ is a function of N not only through F_N. This point applies to the present theory of bagged U-statistics, Chen and Hall's (2002) 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 (2003) 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 fitting 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 fitting bias when the underlying $f(x)$ is smooth.

2 Preliminaries 1: The von Mises Expansion of a Statistical Functional

The von Mises expansion of a functional θ around a distribution F is an expansion of the form

$$\theta(G) \;=\; \theta(F) + \int \psi_1(x)\, d(G-F)(x) + \frac{1}{2}\int \psi_2(x_1,x_2)\, d(G-F)^{\otimes 2} + \cdots$$

$$= \theta(F) + \sum_{k=1}^{\infty} \frac{1}{k!} \int \psi_k(x_1,\ldots,x_k)\, d(G-F)^{\otimes k} \ .$$

It can be interpreted as the Taylor expansion of $\theta((1-s)F+sG) = \theta(F+s(G-F))$ evaluated at $s = 1$. The first term in the sum is a linear functional, the second term is a quadratic functional, etc. There is of course no guarantee that the expansion exists. Reeds (1976) gives a discussion of conditions under which this expansion is meaningful in terms of remainders and convergence. We are not concerned with technical difficulties because the expansions we encounter below are finite and exact. See also Serfling (1980, chap. 6).

The functions ψ_k are not uniquely determined. We can choose them such that all the integrals w.r.t. F vanish, that is,

$$\begin{aligned} 0 &= \int \psi_1(x)\,\mathrm{d}F \\ 0 &= \int \psi_2(x_1,x_2)\,\mathrm{d}F(x_1) = \int \psi_2(x_1,x_2)\,\mathrm{d}F(x_2) \ , \end{aligned}$$

and so on. The von Mises expansion then simplifies to

$$\begin{aligned} \theta(G) &= \theta(F) + \mathbf{E}_G\, \psi_1(X) + \frac{1}{2}\mathbf{E}_G\, \psi_2(X_1,X_2) + \ldots \\ &= \theta(F) + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{E}_G\, \psi_k(X_1,\ldots,X_k) \ . \end{aligned}$$

The function $\psi_1(x)$ is also known as the influence function of θ, but we will similarly call $\psi_k(x_1,\ldots,x_k)$ the k-th order influence function. Influence functions of any order are permutation symmetric in their arguments.

Assuming sufficient smoothness of the functional, ψ_k can be obtained by differentiation:

$$\psi_k(x_1,\ldots,x_k) = \left.\frac{d}{ds_1}\right|_{s_1=0} \cdots \left.\frac{d}{ds_k}\right|_{s_k=0} \theta((1-\sum s_i)F + \sum s_i \delta_{x_i}) \ .$$

3 Preliminaries 2: The ANOVA Expansion of a Statistic

Efron and Stein (1981) introduced an ANOVA-type expansion for statistics that are functions of independent random variables $X_1,\ldots,X_M$. Because we are only interested in symmetric functions of i.i.d. data as they arise from evaluating statistical functionals on empirical distributions, we use an earlier simplified version of the expansion which can be found for example in Serfling (1980). Define partial expectations

$$\mu_0 = \mathbf{E}_F\, \theta(X_1,\ldots,X_M)$$

$$\begin{aligned} \mu_1(x_1) &= \mathbf{E}_F\,\theta(x_1, X_2 \ldots, X_M) \\ \mu_2(x_1, x_2) &= \mathbf{E}_F\,\theta(x_1, x_2, X_3 \ldots, X_M) \\ \ldots & \\ \mu_k(x_1, \ldots, x_k) &= E_F\,\theta(x_1, \ldots, x_k, X_{k+1}, \ldots, X_M) \\ \ldots & \\ \mu_M(x_1, \ldots, x_M) &= \theta(x_1, \ldots, x_M)\ . \end{aligned}$$

Permutation symmetry of $\theta(x_1, \ldots, x_M)$ implies that the free arguments x_j could be in any position, a fact that will be used extensively below.

Define ANOVA terms

$$\begin{aligned} \alpha_0 &= \mu_0 \\ \alpha_1(x_1) &= \mu_1(x_1) - \mu_0 \\ \alpha_2(x_1, x_2) &= \mu_2(x_1, x_2) - \mu_1(x_1) - \mu_1(x_2) + \mu_0 \\ \ldots & \\ \alpha_k(x_1, \ldots, x_k) &= \sum_{\nu=0}^{k} (-1)^{k-\nu} \sum_{1 \le i_1 < \ldots < i_\nu \le k} \mu_\nu(x_{i_1}, \ldots, x_{i_\nu}) \\ \ldots & \\ \alpha_M(x_1, \ldots, x_M) &= \sum_{\nu=0}^{M} (-1)^{M-\nu} \sum_{1 \le i_1 < \ldots < i_\nu \le M} \mu_\nu(x_{i_1}, \ldots, x_{i_\nu})\ . \end{aligned}$$

Then the ANOVA expansion of $\theta(x_1, \ldots, x_M)$ is

$$\begin{aligned} \theta(x_1, \ldots, x_M) = &\quad \alpha_0 + \sum_{j=1}^{M} \alpha_1(x_j) + \sum_{1 \le j_1 < j_2 \le M} \alpha_2(x_{j_1}, x_{j_2}) + \ldots \\ = &\quad \sum_{k=0}^{M} \sum_{1 \le j_1 < \ldots < j_k \le M} \alpha_k(x_{j_1}, \ldots, x_{j_k})\ . \end{aligned}$$

This expansion is tautological and holds without assumptions other than permutation symmetry of $\theta(x_1, \ldots, x_M)$ in its arguments. The proof is by showing that the partial expectations implicit in the ANOVA terms cancel each other except for $\mu_M = \theta(x_1, \ldots, x_M)$.

If one assumes that the variables $X_1, \ldots, X_M$ are i.i.d., then the terms α_k have vanishing marginals in all arguments:

$$\mathbf{E}_F\,\alpha_k(x_1, \ldots, x_{j-1}, X_j, x_{j+1}, \ldots, x_k) = 0\ .$$

As a consequence, all terms in the ANOVA expansion are pairwise uncorrelated.

Note that all functions μ_k and α_k are implicitly dependent on M because they derive from a statistic of M arguments, $\theta(x_1, \ldots, x_M)$. If necessary we make the dependence explicit by writing μ_k^M and α_k^M. By contrast, the influence functions ψ_k in the von Mises expansion are independent of any sample size because this expansion is centered at F as opposed to F_M.

The zero-th term $\alpha_0^M = \mu_0^M$ is also called the "grand mean", and the first term $\alpha_1^M(x)$ the "main effect" function or the "additive component" of θ. Correspondingly we call $\alpha_k^M(x_1, \ldots, x_k)$ the k-th order "interaction" function.

4 The von Mises Expansion of a Bagged Functional

4.1 A Warm-Up Exercise: The Additive Term

Before deriving a general formula for the terms of the von Mises expansion of θ_M^B, we calculate the additive term to illustrate the idea. The influence function will be denoted $\psi_1^B(x)$ as a reminder that it belongs to the bagged functional:

$$\begin{aligned} \psi_1^B(x) &= \left.\frac{d}{ds}\right|_{s=0} \theta_M^B((1-s)F + s\delta_x) \\ &= \left.\frac{d}{ds}\right|_{s=0} \mathbf{E}_{(1-s)F+s\delta_x}\, \theta(X_1, \ldots, X_M) \ . \end{aligned}$$

The expectation $\mathbf{E}_{(1-s)F+s\delta_x}\, \theta(X_1, \ldots, X_M)$ is a polynomial of degree M in s and hence arbitrarily differentiable. We expand it by applying the mixture $(1-s)F + s\delta_x$ to each argument X_i, resulting in 2^M terms. These terms in turn can be bundled according to the number of times δ_x occurs:

$$\begin{aligned} &\mathbf{E}_{(1-s)F+s\delta_x}\, \theta(X_1, \ldots, X_M) \\ &= \ (1-s)^M \ \mathbf{E}_F\, \theta(X_1, \ldots, X_M) \\ &\ \ + \ (1-s)^{M-1}\, s\ M\, \mathbf{E}_F\, \theta(x, X_2, \ldots, X_M) \\ &\ \ + \ (1-s)^{M-2}\, s^2\ \frac{M(M-1)}{2}\, \mathbf{E}_F\, \theta(x, x, X_3, \ldots, X_M) \\ &\ \ + \ O(s^3) \ . \end{aligned}$$

In this rearrangement we also used permutation symmetry; it implies, for example, that

$$\mathbf{E}_F\, \theta(\ldots, X_{j-1}, x, X_{j+1}, \ldots) = \mathbf{E}_F\, \theta(x, X_2, \ldots, X_M) \ .$$

When we differentiate w.r.t. s at $s = 0$, only the first two terms make a contribution:

$$\psi_1^B(x) = M\ \left[-\mathbf{E}_F\, \theta(X_1, \ldots, X_M) + \mathbf{E}_F\, \theta(X_1, \ldots, X_{M-1}, x) \right] = M\, \alpha_1^M(x) \ ,$$

where as above α_1^M is the main effect function in the ANOVA expansion of the unbagged statistic $\theta(F_M)$.

Suppose we have an i.i.d. sample $x_1, \ldots, x_N$ of size N from F with empirical distribution $F_N = \frac{1}{N} \sum \delta_{x_i}$. The first order von Mises approximation to the plug-in estimate $\theta_M^B(F_N)$ of $\theta_M^B(F)$ is

$$\theta_M^B(F_N) \;\approx\; \theta_M^B(F) + \frac{1}{N} \sum_{i=1}^{N} \psi_1^B(x_i) \;=\; \mu_0^M + \frac{M}{N} \sum_{i=1}^{N} \alpha_1^M(x_i) \; .$$

For $M = N$ this is exactly the first order ANOVA expansion of $\theta(F_N)$.

4.2 The Full von Mises Expansion of a Bagged Functional

More generally we have the following theorem whose proof can be found in Subsection 11.1 of the appendix.

Theorem: *The k-th order influence function ψ_k^B of an M-bagged functional $\theta_M^B(F)$ is proportional to the k-th order interaction function α_k^M of the statistic $\theta(F_M)$:*

$$\psi_k^B(x_1, \ldots, x_k) \;=\; \begin{cases} \dfrac{M!}{(M-k)!} \, \alpha_k^M(x_1, \ldots, x_k) & \text{for } k \le M \; , \\ 0 \quad \text{for } k > M \; . \end{cases}$$

It is now a simple matter to write down the full von Mises expansion of an M-bagged functional:

$$\begin{aligned} \theta_M^B(G) \;&=\; \theta_M^B(F) \,+\, \sum_{k \ge 1} \frac{1}{k!} \, \mathbf{E}_G \, \psi_k(X_1, \ldots, X_k) \\ &=\; \alpha_0^M \,+\, \sum_{k=1}^{M} \binom{M}{k} \, \mathbf{E}_G \, \alpha_k^M(X_1, \ldots, X_k) \; . \end{aligned}$$

We summarize:

Theorem: *Bagged functionals are smooth in the sense that the von Mises expansion exists and is of finite length $M+1$:*

$$\theta_M^B(G) \;=\; \sum_{k=0}^{M} \binom{M}{k} \, \mathbf{E}_G \, \alpha_k^M(X_1, \ldots, X_k) \; .$$

Because the von Mises expansion is effectively a Taylor expansion, it is natural for exact finite expansions to use their length as an inverse measure of smoothness: the shorter the expansion the smoother the functional. With this interpretation and in light of the theorem, bagging performs more smoothing for smaller M.

Suppose now we have an i.i.d. sample $y_1, \ldots, y_N$ of size N from the distribution F.

The von Mises expansion of θ_M^B around F evaluated at $F_N = \frac{1}{N}\sum_1^N \delta_{y_j}$ is

$$\theta_M^B(F_N) \;=\; \sum_{k=0}^{M} \binom{M}{k} \frac{1}{N^k} \sum_{1 \le j_1, \ldots, j_k \le N} \alpha_k^M(y_{j_1}, \ldots, y_{j_k}) \;.$$

The bagging parameter M is independent of the sample size N, which raises the question of criteria for its choice. This is just another instance of the problem of smoothing parameter selection.

For the conventional choice $M = N$ one obtains an interesting comparison with the ANOVA expansion of $\theta(F_N)$:

Theorem: *The terms in the von Mises expansion of the conventional N-bagged statistic $\theta_N^B(F_N)$ form a superset of the terms in the ANOVA expansion of $\theta(F_N)$.*

$$\begin{aligned} \theta_N^B(F_N) \;&=\; \sum_{k=0}^{N} \binom{N}{k} \frac{1}{N^k} \sum_{1 \le j_1, \ldots, j_k \le N} \alpha_k^N(y_{j_1}, \ldots, y_{j_k}) \;, \\ \theta(F_N) \;&=\; \sum_{k=0}^{N} \sum_{1 \le j_1 < \ldots < j_k \le N} \alpha_k^N(y_{j_1}, \ldots, y_{j_k}) \;. \end{aligned}$$

The inner sums in the first and the second line have N^k and $\binom{N}{k}$ terms, respectively, the difference being that the first inner sum runs over unconstrained indices, the second over strictly ordered indices. The ratio $\binom{N}{k}/N^k$ downweights the inner sum in the first line to match the smaller number of terms in the second line. The difference between the unbagged and the N-bagged statistic is that the latter includes “diagonal” terms such as $\alpha_2^N(y_1, y_1)$, arising from sampling with replacement in the bootstrap procedure.

4.3 Asymptotic Variances of Bagged Medians

The first order von Mises expansion of a statistical functional is frequently used to estimate the variance of the corresponding statistic: if

$$\theta(F_N) \approx \theta(F) + \frac{1}{N}\sum \psi_1(X_i)$$

then

$$\mathrm{Var}_F(\theta(F_N)) \approx \frac{1}{N}\mathbf{E}_F(\psi_1^2(X))\,.$$

We use this approach to estimate the variances of bagged versions of the sample median for a range of resample sizes. The purpose of the exercise is to show that these estimates, obtained using the von Mises expansions of bagged medians, are close to the true variances, thereby supporting our claim that such expansions indeed can accurately reflect reality.

For our experiment we chose standard gaussian F, sample size $N = 50$ and resample sizes $M = 5$, 20, and 80. For each resample size we computed the influence function of the M-bagged median, i.e., M times the main effect function in the ANOVA decomposition of median$(X_1, \ldots, X_M)$, on a grid, using Monte Carlo quadrature. We then evaluated $\mathbf{E}_F \psi_1^2(X)$ using Simpson's rule. The table below gives the estimated variances of the bagged median for the three resample sizes and the "true" variances calculated by simple Monte Carlo. The results show that the variance of a bagged median decreases with decreasing resample size, and that this fact is accurately reflected in the asymptotic results.

M	5	20	80
True variance	0.0212	0.0247	0.0270
Est. variance	0.0215	0.0254	0.0277

5 From von Mises Expansions to U-Statistics

If $\theta(F)$ is a statistical functional with a finite von Mises expansion of length K,

$$\begin{aligned}\theta(G) &= \theta(F) + \sum_{k=1}^{K} \frac{1}{k!} \mathbf{E}_G \, \psi_k(X_1, \ldots, X_k) \\ &= \theta(F) + \int \psi_1(x) \, \mathrm{d}G(x) + \frac{1}{2} \int \psi_2(x_1, x_2) \, \mathrm{d}G^{\otimes 2}(x_1, x_2) \\ &+ \frac{1}{6} \int \psi_3(x_1, x_2, x_3) \, \mathrm{d}G^{\otimes 3}(x_1, x_2, x_3) + \cdots \ ,\end{aligned}$$

then the plug-in estimate for θ from an i.i.d. sample $X_1, X_2, \ldots, X_N$ is a U-statistic:

$$\begin{aligned}\theta(F_N) &= \theta(F) + \frac{1}{N} \sum_{i=1}^{N} \psi_1(X_i) \ + \ \frac{1}{2} \frac{1}{N^2} \sum_{i,j=1}^{N} \psi_2(X_i, X_j) \\ &+ \ \frac{1}{6} \frac{1}{N^3} \sum_{i,j,k=1}^{N} \psi_3(X_i, X_j, X_k) \ + \ \cdots \ .\end{aligned}$$

We now adopt the conventional notation for U-statistics:

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

where the constant corresponding to $\theta(F)$ is absorbed in the other terms. 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.

Strictly speaking, only the off-diagonal sums such as $\sum_{i<j} B_{X_i,X_j}$ and $\sum_{i<j<k} C_{X_i,X_j,X_k}$ are proper U-statistics. Unrestricted sums that include the diagonal terms are usually called V-statistics or von Mises statistics (Serfling 1980, Sec. 5.1.2). However, we use the better known term "U-statistics" anyway. Our problem with U-statistics in the traditional sense is that they are not plug-in estimates of statistical functionals. — In another slight deviation from common usage, we often refer not only to the terms in U as U-statistics, but to the sum U as well.

6 Resampling U-Statistics

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.

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 is for $M = N$, 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*, the distribution of $\mathbf{W}$ is Hypergeometric(M, N). Half-sampling, for example, is for $M = N/2$, 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.

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 $\mathbf{W}$:

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

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

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

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. In this notation we suppress the dependence on M.

$$\begin{aligned}
U^{with} &= \frac{1}{N}\sum_i \left(A_{X_i} + \frac{1}{M} B_{X_i,X_i} \right) \;+\; \frac{1}{N^2}\sum_{i,j} \left(1 - \frac{1}{M}\right) B_{X_i,X_j} \ , \\
U^{w/o} &= \frac{1}{N}\sum_i \left(A_{X_i} + \left(\frac{1-\frac{M}{N}}{1-\frac{1}{N}}\right) \frac{1}{M}\, B_{X_i,X_i} \right) \;+\; \frac{1}{N^2}\sum_{i,j} \left(\frac{1-\frac{1}{M}}{1-\frac{1}{N}}\right) B_{X_i,X_j} \ .
\end{aligned}$$

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

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 $B_{x,y}$ as follows:*

$$\begin{aligned}
A_x^{with} &= A_x + \frac{1}{M} B_{x,x} \ , & B_{x,y}^{with} &= \left(1 - \frac{1}{M}\right) B_{x,y} \ , \\
A_x^{w/o} &= A_x + \left(\frac{1-\frac{M}{N}}{1-\frac{1}{N}}\right) \frac{1}{M}\, B_{x,x} \ , & B_{x,y}^{w/o} &= \left(\frac{1-\frac{1}{M}}{1-\frac{1}{N}}\right) B_{x,y} \ .
\end{aligned}$$

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 (Section 11).

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 additive part. Similar effects take place in higher orders where variability is shifted to lower orders.

7 Equivalence of Resampling With and Without Replacement in Bagging

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 $O(N^{-\frac{2}{p}})$, where $\frac{1}{p} + \frac{1}{q} = 1$.

The proof is in Subsection 11.5 of the appendix.

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:

$$\alpha_{with} = \frac{M_{with}}{N} \;\; (> 0, \;\; < \infty) \quad \text{and} \qquad \alpha_{w/o} = \frac{M_{w/o}}{N} \;\; (> 0, \;\; < 1).$$

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.

8 The Effect of Bagging on Variance, Bias and MSE

We need some notation: For U-statistics $C_{X,Y,Z,...}$ of any order we denote partial conditional expectations by

$$C_X = \mathbf{E}\,[\,C_{X,Y,Z,W,...}|\,X\,]\ ,\quad C_{X,Y} = \mathbf{E}\,[\,C_{X,Y,Z,W,...}|\,X,Y\,]\ ,\quad C_{X,Y,Z} = \mathbf{E}\,[\,C_{X,Y,Z,W,...}|\,X,Y,Z\,]\ .$$

Under the assumption of independent variables, one can re-express them as partial marginal expectations (Serfling, 1980, Sec. 5.1.5):

$$C_x = \mathbf{E}\,[\,C_{x,Y,Z,W,...}]\ ,\quad C_{x,y} = \mathbf{E}\,[\,C_{x,y,Z,W,...}]\ ,\quad C_{x,y,z} = \mathbf{E}\,[\,C_{x,y,z,W,...}]\ .$$

It will turn out that for variance and bias calculations to order N^{-2} these three partial conditional expectations are the only information needed about a U-statistic of any order. We will use simple facts such as the following without further mention:

$$\begin{aligned}\mathrm{Cov}(B_{X,Y,Z,...}, C_{X,Y',Z',...}) &= \mathrm{Cov}(B_X, C_X)\ ,\\ \mathrm{Cov}(C_{X,Y,Z,...}, C_{X,Y',Z',...}) &= \mathrm{Var}(C_X)\ ,\end{aligned}$$

where X, Y, Z, Y', Z' are independent.

8.1 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

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

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

Proposition 3: *Let $g = \frac{N}{M}$ for sampling with replacement and $g = \frac{N}{M} - 1$ 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:*

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

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 effect of bagging on variance is of order $O(N^{-2})$. The proof is in Subsection 11.7 of the appendix; Subsection 11.6 shows how to calculate S_{Var} for statistics with U-terms of any order.

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 the 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.

8.2 Bias

We show that bagging U-statistics *always* increases squared bias. 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

$$S_{\mathrm{Bias}} \;=\; \left(\mathbf{E}\,B_{X,X} - \mathbf{E}\,B_{X,Y}\right)^2 \;.$$

Subsection 11.8 of the appendix has proofs and a general formula for S_{Bias} for statistics with U-terms of any order.

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.

8.3 Mean Squared Error

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

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

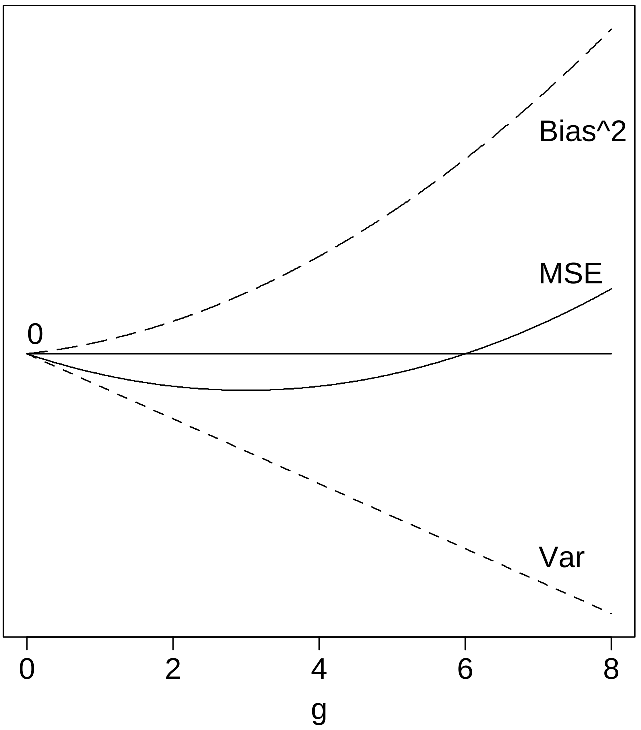

Figure 1: Dependence of Variance, Squared Bias and MSE on g. The graph shows the situation for $S_{\mathrm{Var}}/S_{\mathrm{Bias}} = -4$. Bagging is beneficial for $g < 6$, that is, for resample sizes $M_{with} > N/6$ and $M_{w/o} > N/7$. Optimal is $g = 3$, that is, $M_{with} = N/3$ and $M_{w/o} = N/4$.

The 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:*

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

for both sampling with and without replacement.

8.4 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:

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

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

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

The resample size with optimal MSE improvement is

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

Conventional bootstrap, $M_{with} = N$, and half-sampling, $M_{w/o} = N/2$, (both characterized by $g = 1$) are beneficial iff

$$\frac{S_{\text{Var}}}{S_{\text{Bias}}} < -\frac{3}{2} \ ,$$

and they are optimal iff

$$\frac{S_{\text{Var}}}{S_{\text{Bias}}} = -2 \ .$$

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 1 for an illustration.

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. This follows from the fact that S_{Bias} is always positive, hence bagging always increases bias, but if the variance dips sufficiently strongly, an overall benefit results.

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.

8.5 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

$$(\kappa-1)\,\frac{a}{b} \;+\; 1 \;<\; 0\;.$$

Since $\kappa > 1$ for all distributions except a balanced 2-point mass, the condition becomes

$$\frac{a}{b} \;<\; -\frac{1}{\kappa-1}\;.$$

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.

9 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, and 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	$f(x) = x$
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, till 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 $\mathcal{L} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, 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

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

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

Note that the definition of bias used here — expected difference between the estimated regression function for a finite sample size and the true regression function — is different from the definition used in the earlier sections of the article, where we took bias to mean the expected difference between the value of a statistic for a finite sample size and its value for infinite sample size, i.e., for the underlying distribution. We refer to the former as *regression bias* and to the latter as *estimation bias.* 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 regression 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 $\mathrm{Var}(T_\alpha^{wi})$. We adjusted the estimates by removing this component. The influence on bias is of smaller order.

To calculate the estimation 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 2 summarizes the simulation results for Scenario 1. The top panels show variance, squared estimation bias, and squared regression 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 regression 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. Regression bias and estimation 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 estimation bias increases with decreasing resampling fraction, as predicted by the theory presented in Sections 8.1 through 8.3.

Figure 3 shows the corresponding results for Scenario 2. Again, variance is decreasing with decreasing resampling fraction, and squared estimation bias is increasing, as predicted by the theory. Squared regression bias, however, is *decreasing* with decreasing resampling fraction. Bagging therefore conveys a double benefit, decreasing both variance and squared (regression) 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.

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 4, parallel those for Scenario 1.

The results for Scenario 4, shown in Figure 5, closely parallel those for Scenario 2. Again, both variance and squared (regression) bias decrease with decreasing resampling fraction.

The experiments confirm the agreement between bagging with and without replacement predicted by the theory developed in Section 7: 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 estimation bias. However, the relevant quantity in a regression context is regression bias — the expected difference between the estimated and the true regression function. If the true regression function is smooth, bagging can in fact *reduce* regression bias as well as variance and therefore yield a double benefit.

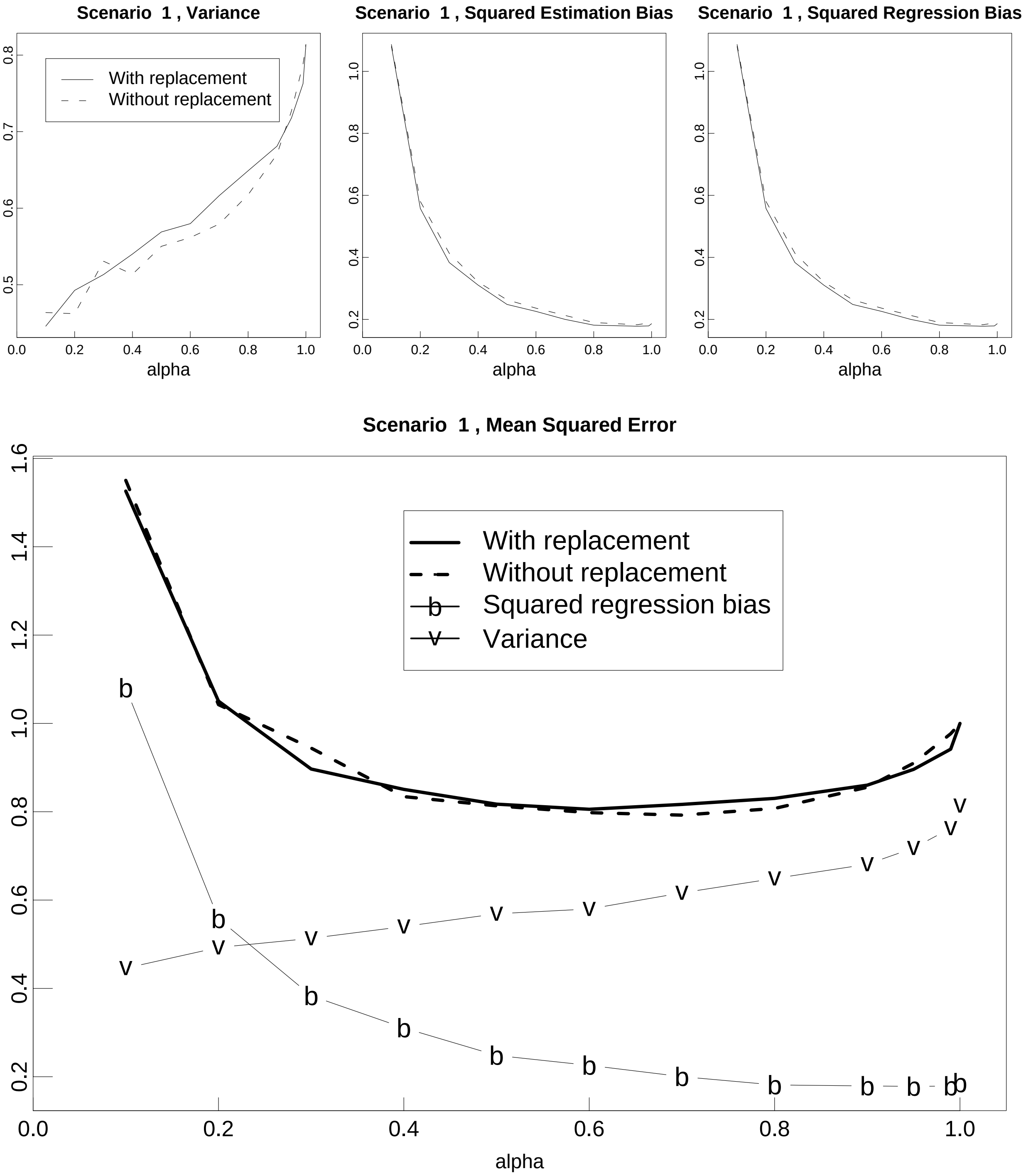

Figure 2: Simulation results for Scenario 1. Top panels: Variance, squared estimation bias, and squared regression bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared regression bias for resampling with replacement.

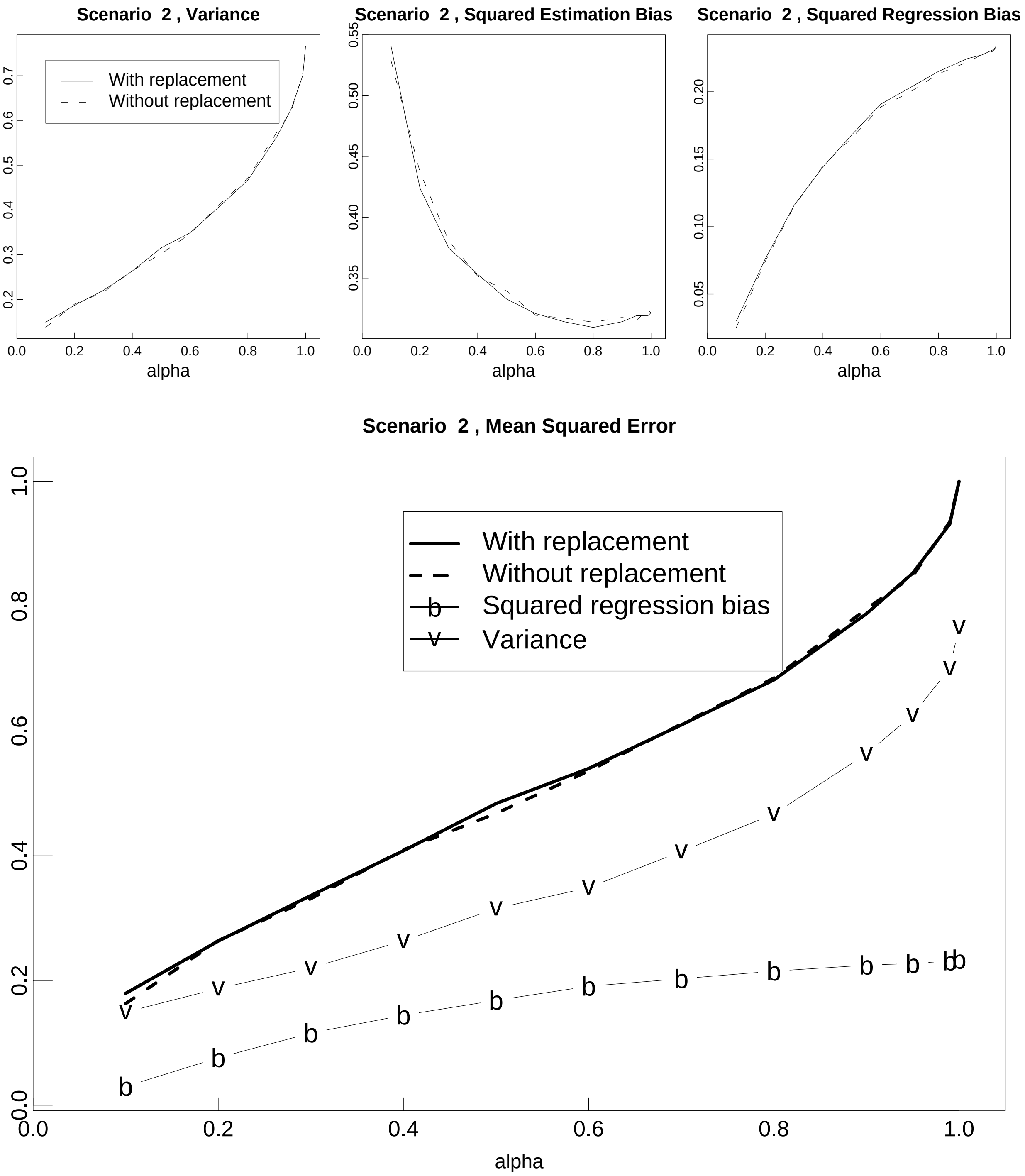

Figure 3: Simulation results for Scenario 2. Top panels: Variance, squared estimation bias, and squared regression bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared regression bias for resampling with replacement.

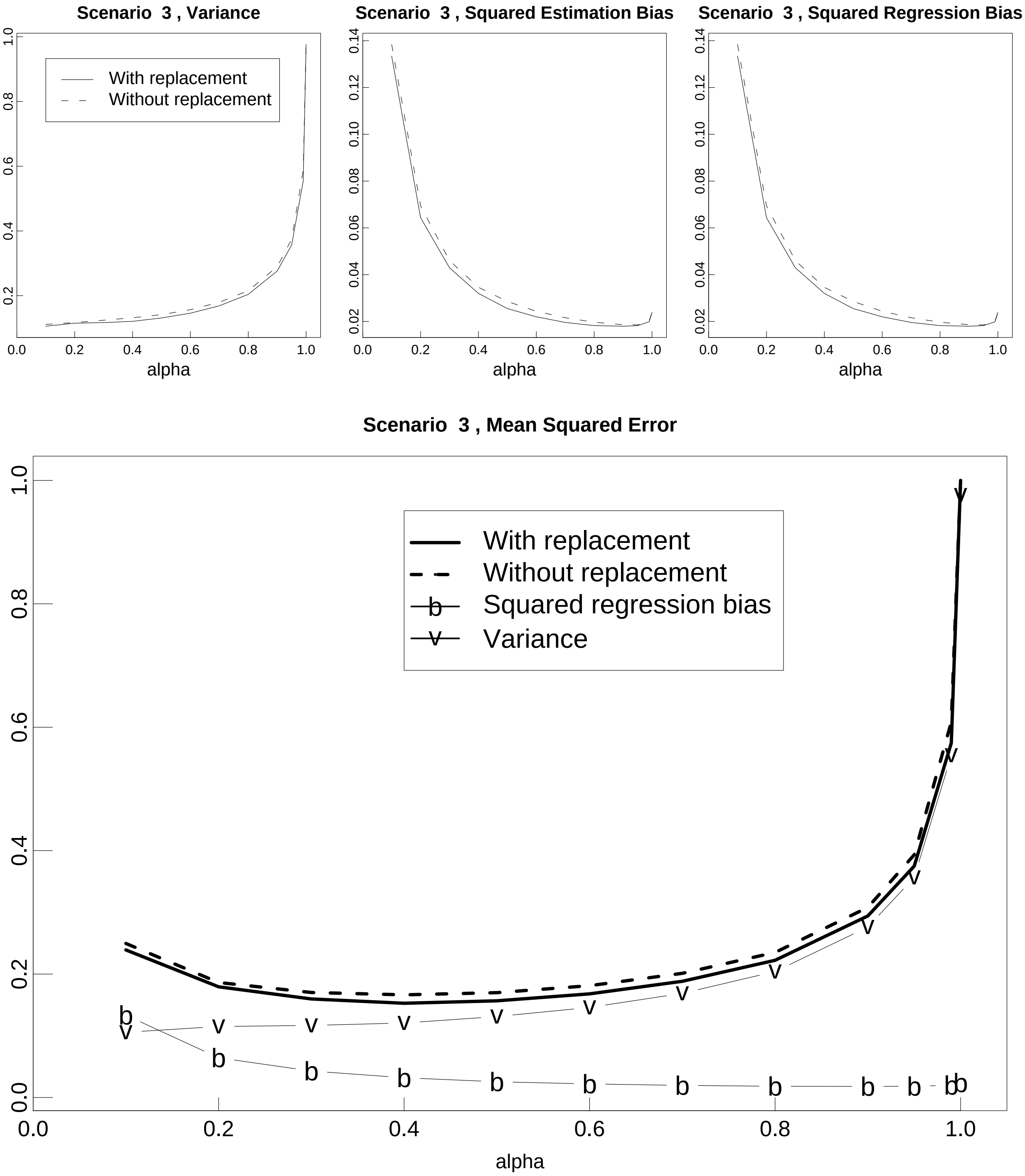

Figure 4: Simulation results for Scenario 3. Top panels: Variance, squared estimation bias, and squared regression bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared regression bias for resampling with replacement.

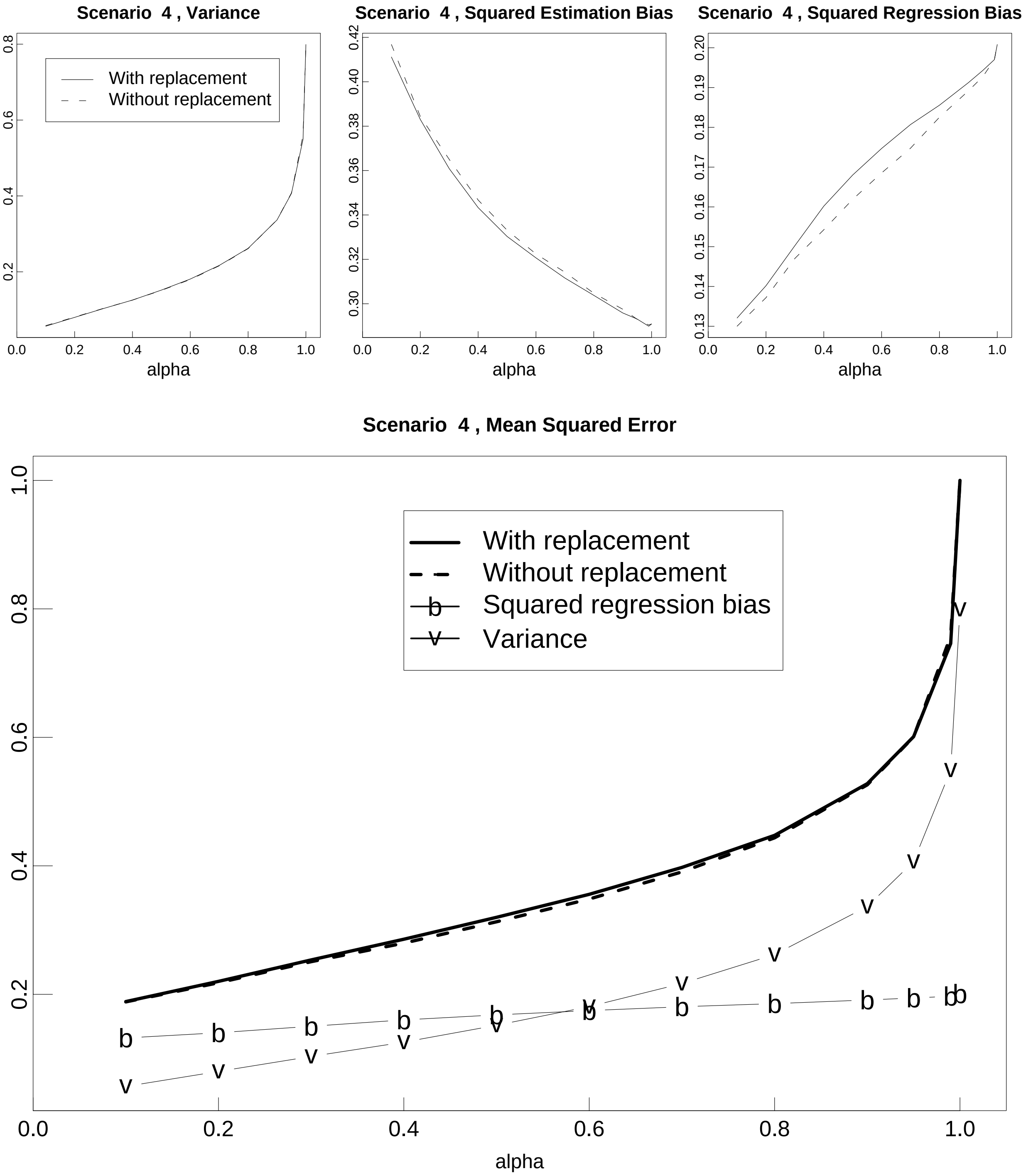

Figure 5: Simulation results for Scenario 4. Top panels: Variance, squared estimation bias, and squared regression bias for resampling with and without replacement. Bottom panel: MSE for both resampling modes, and variance and squared regression bias for resampling with replacement.

10 Summary and Conclusions

Here is a summary of what we take to be the main contributions of our article:

Bagging statistical functionals. We extend the definition of bagging from statistics (defined on samples) to statistical functionals (defined on distributions), and we study the von Mises expansion of bagged statistical functionals. The von Mises expansion of a statistical functional is a generalized Taylor expansion and allows for a similar interpretation. We show the following:

- The von Mises expansion of a bagged functional is related to the Efron-Stein ANOVA expansion of the corresponding unbagged statistic. In particular, the first von Mises term is proportional to the additive term in the Efron-Stein expansion.
- A bagged functional is always smooth in the sense that the von Mises expansion is *finite* of length 1 + resample size M. This holds even if the unbagged functional is rough or unstable.

These results support the intuitive notion that bagging is a smoothing operation and that resample size plays the role of the smoothing parameter — smaller resample size means more smoothing.

Bagging U-statistics U-statistics may be regarded as a generalization of polynomials. They can describe complex dependencies, and yet the effect of bagging on U-statistics is amenable to a rigorous asymptotic analysis. The analysis suggests that bagging always increases (estimation) bias — another fact that is borne out by our experiments — and demonstrates that bagging does *not* always decrease variance and MSE.

Our most striking finding is an equivalence between bagging based on resampling *with* and *without* replacement: resample sizes $M_{with} = \alpha_{with} N$ and $M_{w/o} = \alpha_{w/o} N$ (for resampling with and without replacement, respectively) produce very similar bagged statistics if $\alpha_{with} = \alpha_{w/o}/(1 - \alpha_{w/o})$. This approximate equality holds for each sample, not just on the average over samples. While our derivation is limited to U-statistics, the equivalence seems to hold more widely, as illustrated by our limited experiments with bagging CART trees.

While we believe that our article provides valid, and valuable, insights into the nature and effects of bagging, it does not explain the often striking improvements seen when bagging trees. In fact, our experiments show that bagging can reduce both variance *and* bias. This observation, however, does not contradict the theory: the notion of bias used above for U-statistics is different from the notion of bias used in function fitting.

11 Appendix

11.1 Derivation of the von Mises Expansion of a Bagged Functional

We prove the theorem of Section 4:

Theorem: *The k-th order influence function ψ_k^B of an M-bagged functional $\theta_M^B(F)$ is proportional to the k-th order interaction function α_k^M of the statistic $\theta(F_M)$:*

$$\psi_k^B(x_1, \ldots, x_k) \;=\; \begin{cases} \dfrac{M!}{(M-k)!}\, \alpha_k^M(x_1, \ldots, x_k) \quad \text{for } k \le M \;, \\ 0 \quad \text{for } k > M \;. \end{cases}$$

Proof: We now calculate the k-th order influence function. To this end let

$$\tilde{F}_k = (1 - \sum_1^k s_i) F + \sum_1^k s_i \delta_{x_i} \;.$$

By definition,

$$\psi_k^B(x_1, \ldots, x_k) = \frac{\partial^k}{\partial s_1 \cdots \partial s_k}\bigg|_{s_1, \ldots, s_k = 0} \theta_M^B(\tilde{F}_k) \;.$$

Again we note that $\theta_M^B(\tilde{F}_k) = \mathbf{E}_{\tilde{F}}\, \theta(X_1, \ldots, X_M)$ is effectively a polynomial of degree M in s. Expanding it into $(k+1)^M$ summands, bundling the summands according to the number of δ_{x_i}'s they contain, and using permutation symmetry, we get:

$$\begin{aligned}
\theta_M^B(\tilde{F}_k) \;&=\; \mathbf{E}_{\tilde{F}_k}\, \theta(X_1, \ldots, X_M) \\
=\;& (1 - \sum_{i=1}^k s_i)^M \, \mathbf{E}_F\, \theta(X_1, \ldots, X_M) \\
&+ \sum_{j=1}^k \; (1 - \sum_{i=1}^k \; s_i)^{M-1} \, s_j \, M \, \mathbf{E}_F\, \theta(x_j, X_2, \ldots, X_M) \\
&+ \sum_{1 \le j_1 < j_2 \le k} (1 - \sum_{i=1}^k s_i)^{M-2} \, s_{j_1} \, s_{j_2} \, M\,(M-1)\, \mathbf{E}_F\, \theta(x_{j_1}, x_{j_2}, X_3, \ldots, X_M) \\
&+ \;\ldots \\
&+ \; O(s_1^2, \ldots, s_k^2)
\end{aligned}$$

Terms containing a second or higher power of any s_j have vanishing derivatives at zero and hence will disappear in what follows. This is why the summation on the fourth line can run over index pairs $j_1 \neq j_2$ only, the omitted summands being summarily lumped into $O(s_1^2, \ldots, s_k^2)$. Thus, with the abbreviated notation for partial expectations:

$$\begin{aligned}\theta_M^B(\tilde{F}_k) \;&=\; \sum_{\nu=0}^{\min(k,M)} \sum_{1 \le j_1 < \cdots < j_\nu \le k} (1 - \sum_{i=1}^{k} s_i)^{M-\nu} s_{j_1} \cdots s_{j_\nu} \frac{M!}{(M-\nu)!} \mu_\nu^M(x_{j_1}, \ldots, x_{j_\nu}) \\ &\quad + O(s_1^2, \ldots, s_k^2) \ .\end{aligned}$$

Note that the outer sum extends to $\min(k, M)$ only. As the derivatives can be pulled inside the double sum, we have to calculate

$$\frac{\partial^k}{\partial s_1 \cdots \partial s_k}\bigg|_{s_1,\ldots,s_k=0} \left[(1 - \sum_{i=1}^{k} s_i)^{M-\nu} \, s_{j_1} \cdots s_{j_\nu} \right] \ .$$

We first take partial derivatives w.r.t. $s_{j_1}, \ldots, s_{j_\nu}$ in turn:

$$\begin{aligned}&\frac{\partial}{\partial s_{j_1}}\bigg|_{s_{j_1}=0} \Big[(1 - \sum s_i)^{M-\nu} \, s_{j_1} \cdots s_{j_\nu} \Big] \\ &= \Big[(M-\nu)(1 - \sum s_i)^{M-\nu-1}(-1) \, s_{j_1} \cdots s_{j_\nu} \, + \, (1 - \sum s_i)^{M-\nu} \, s_{j_2} \cdots s_{j_\nu} \Big]\Big|_{s_{j_1}=0} \\ &= (1 - \sum s_i)^{M-\nu} \, s_{j_2} \cdots s_{j_\nu} \ .\end{aligned}$$

Repeating this process we obtain:

$$\frac{\partial^\nu}{\partial s_{j_1} \cdots \partial s_{j_\nu}}\bigg|_{s_1,\ldots,s_k=0} \Big[(1 - \sum s_i)^{M-\nu} \, s_{j_1} \cdots s_{j_\nu} \Big] \;=\; (1 - \sum s_i)^{M-\nu} \ .$$

We still have to take the derivatives w.r.t. indices not among $j_1, \ldots, j_\nu$. Pick one such index l:

$$\frac{\partial}{\partial s_l}\bigg|_{s_l=0} \Big[(1 - \sum s_i)^{M-\nu} \Big] = (M-\nu)(1 - \sum s_i)^{M-\nu-1}(-1)$$

Repeating for all such l we get:

$$\begin{aligned}&\frac{\partial^k}{\partial s_1 \cdots \partial s_k}\bigg|_{s_1,\ldots,s_k=0} \Big[(1 - \sum s_i)^{M-\nu} \, s_{j_1} \cdots s_{j_\nu} \Big] \\ &= \begin{cases} (M-\nu)(M-\nu-1) \cdots (M-k+1) \, (-1)^{k-\nu} \;=\; \dfrac{(M-\nu)!}{(M-k)!} (-1)^{k-\nu} & \text{for } k \le M \ , \\ 0 \quad \text{for } k > M \ . \end{cases}\end{aligned}$$

Putting everything together, we get first of all

$$\psi_k^B(x_1, \ldots, x_k) = 0 \quad \text{for } k > M \ .$$

For $k \leq M$ we get

$$\begin{aligned}
\psi_k^B(x_1,\ldots,x_k) &= \frac{M!}{(M-k)!} \sum_{\nu=0}^{k} (-1)^{k-\nu} \sum_{1\leq j_1<\cdots<j_\nu\leq k} \mu_\nu^M(x_{j_1},\ldots,x_{j_\nu}) \\
&= \frac{M!}{(M-k)!}\, \alpha_k^M(x_1,\ldots,x_k)
\end{aligned}$$

This completes the proof.

11.2 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:

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

where

$$\begin{aligned}
\mu = (\mu_1,\ldots,\mu_J) \in \{1,\ldots,N\}^J \ , \qquad B_\mu &= B_{X_{\mu_1},\ldots,X_{\mu_J}} \ , \\
\nu = (\nu_1,\ldots,\nu_K) \in \{1,\ldots,N\}^K \ , \qquad C_\nu &= C_{X_{\nu_1},\ldots,X_{\nu_K}} \ .
\end{aligned}$$

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:

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

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):

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

11.3 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 $N(N-1)\ldots(N-(J+K-4))$ 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

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

we have:

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

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

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

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

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

11.4 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 $R_1,\ldots,R_M$, where $R_i \in \{1,\ldots,N\}$. Each draw is equally likely: $P[R_i = n] = N^{-1}$, but for sampling with replacement the draws are independent; for sampling w/o replacement they are dependent and the joint probabilities are $P[R_1 = n_1, R_2 = n_2, \ldots, R_J = n_J] = \begin{bmatrix} M \\ J \end{bmatrix} / \begin{bmatrix} N \\ J \end{bmatrix}$ for distinct n_i's, and $= 0$ if ties exist among the n_i's.

For resampling one is interested in the count variables

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

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

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

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

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

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

From these we derive the moments that will be needed below. Recall $\alpha = M/N$, and $g = \frac{1}{\alpha}$ for resampling with, $g = \frac{1}{\alpha} - 1$ 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:

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

$$\mathbf{E}\left[W_1\ W_2 \cdots W_L\right]$$

$$= \begin{cases} \text{with}: & \begin{bmatrix} M \\ L \end{bmatrix} / N^L \\ \text{w/o}: & \begin{bmatrix} M \\ L \end{bmatrix} \Big/ \begin{bmatrix} N \\ L \end{bmatrix} \end{cases}$$

$$= \begin{cases} \text{with}: & \alpha^L - \alpha^L \binom{L}{2} \frac{1}{\alpha} N^{-1} + O(N^{-2}) \\ \text{w/o}: & \alpha^L - \alpha^L \binom{L}{2} \left(\frac{1}{\alpha} - 1\right) N^{-1} + O(N^{-2}) \end{cases}$$

$$= \alpha^L \left(1 \; - \; \binom{L}{2} \; g \; N^{-1}\right) \; + \; O(N^{-2})$$

$$\mathbf{E}\,[W_1^2 \; W_2 \cdots W_{L-1}]$$

$$= \begin{cases} \text{with :} & \begin{bmatrix} M \\ L \end{bmatrix} /N^L \; + \; \begin{bmatrix} M \\ L-1 \end{bmatrix} /N^{L-1} \\ & \\ \text{w/o :} & \begin{bmatrix} M \\ L-1 \end{bmatrix} \Big/ \begin{bmatrix} N \\ L-1 \end{bmatrix} \end{cases}$$

$$= \begin{cases} \text{with :} & \alpha^L \; + \; \alpha^{L-1} \; + \; O(N^{-1}) \\ \text{w/o :} & \alpha^{L-1} \; + \; O(N^{-1}) \end{cases}$$

$$= \; \alpha^L \; (g+1) \; + \; O(N^{-1}) \; .$$

11.5 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

$$\begin{aligned} \mathbf{B}^{with} \quad &= \quad \frac{1}{M_{with}^J} \sum_{\mu} \; \mathbf{E}\left[W_{\mu_1}^{with} \cdots W_{\mu_J}^{with}\right] \cdot B_\mu \; , \\ \mathbf{B}^{w/o} \quad &= \quad \frac{1}{M_{w/o}^J} \sum_{\mu} \; \mathbf{E}\left[W_{\mu_1}^{w/o} \cdots W_{\mu_J}^{w/o}\right] \cdot B_\mu \; . \end{aligned}$$

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 $\mu = (\mu_1, ..., \mu_J)$ 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 11.4) and hence preserve the orders $O(N^{-1})$, $O(N^{-2})$, $O(N^{-3})$ of the terms considered in Subsection 11.2.

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:

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

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:

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

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.

11.6 Covariances of Bagged Interactions

Resuming calculations begun in Subsection 11.3 for covariances of unbagged interaction terms, we now derive the covariance of their M-bagged versions:

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

The moment calculations of Subsection 11.4 yield the following:

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

$$+ \binom{J}{2} K\ N^{-2}\ (g+1)\ \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y)\right)$$

$$+ J \binom{K}{2}\ N^{-2}\ (g+1)\ \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y})\right) \quad + \ O(N^{-3})$$

$$= \left(N^{-1} - N^{-2} \binom{J+K-1}{2} \ - \ N^{-2} \left(\binom{J}{2} + \binom{K}{2} \right) g \right)\ JK\ \mathrm{Cov}(B_X, C_X)$$

$$+ N^{-2}\ \binom{J}{2} K\ (g+1)\ \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y)\right)$$

$$+ N^{-2}\ J \binom{K}{2}\ (g+1)\ \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y})\right) \quad + \ O(N^{-3})$$

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

11.7 Difference Between Variances of Bagged and Unbagged

Comparing the results of the Sections 11.3 and 11.6, we get:

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

where

$$S_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) \ = \ \frac{1}{2} \ \left(\binom{J}{2} K\ \mathrm{Cov}(C_X,\ B_{X,X} + B_{X,Y,Y} - JB_X) \right.$$

$$+ \binom{K}{2} J \,\mathrm{Cov}(B_X,\ C_{X,X} + C_{X,Y,Y} - K C_X) \Bigg) \ .$$

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 $J \le 2$, and similar for C when K =1 or 2.

The result generalizes to arbitrary finite sums of interactions

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

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:

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

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.

11.8 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:

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

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 now consider more general statistics that are finite sums of interactions:

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

The final result is:

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

As usual, $g = \frac{1}{\alpha}$ for sampling with, and $g = \frac{1}{\alpha} - 1$ for sampling w/o, replacement.

References

[1] L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

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

[3] P. Buhlmann and B. Yu. Analyzing bagging. *Ann. of Statist.*, 30:927–961, 2002.

[4] S. X. Chen and P. Hall. Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 2003. (to appear).

[5] B. Efron and C. Stein. The jackknife estimate of variance. *Ann. of Statist.*, 9:586–596, 1981.

[6] J.H. Friedman and O. Hall. On bagging and nonlinear estimation. Can be downloaded from http://www-stat.stanford.edu/˜jhf/#reports, May 2000.

[7] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19:293–325, 1948.

[8] K. Knight and Jr. G. W. Bassett. Second order improvements of sample quantiles using subsamples. 2002.

[9] J. A. Reeds. On the definition of von mises functionals. (Ph.D. Dissertation, Harvard University, Cambridge), 1976.

[10] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.