

The bagged median and the bragged mean

José R. Berrendero *

Departamento de Matemáticas

Universidad Autónoma de Madrid

28049 Madrid, Spain

June 18, 2007

Abstract

Bagging (bootstrap aggregating) is a procedure that aims at reducing the prediction error of a classifier or the mean square error of an estimate by averaging its values across bootstrap samples. We illustrate some of the effects of bagging on point estimation using only averages and medians. Our examples show that when we compute the bagged version of a robust estimate, the size of the bootstrap samples can be viewed as a tuning constant that controls the trade-off between efficiency and robustness. To quantify the robustness properties of bagged estimates we introduce a new concept of breakdown point that is useful in situations when resampling is needed. Finally, a robust version of bagging applied to the average leads to generalizations of previous results about the Hodges-Lehmann estimate.

KEY WORDS: Bagging, median, robustness, bootstrap, breakdown point, Hodges-Lehmann estimate.

MSC: 62F35, 62F40.

*This work has been partially supported by grant MTM2004-00098 and grant S-0505/ESP/0158 of Comunidad de Madrid. The author thanks Antonio Cuevas for some helpful comments

1 Introduction

This paper explores some links between robustness and *bagging* (an acronym for **bootstrap aggregating**). We limit ourselves to very simple elements, namely, averages, sample medians and their population counterparts, the expectation and the population median. We will show that the combination of these elementary blocks results in a variety of connections among resampling, non-parametric estimation and robustness. Since the effects of bagging are particularly clear in the examples that we present, they could be useful as exercises or examples in an introduction to this technique. They could also be used to address the advantages and drawbacks of some popular measures of robustness, such as the breakdown point or the influence function, and to illustrate the techniques to compute them.

Breiman (1996) introduced bagging as a technique for reducing the prediction error in classification methods. At a simpler level, bagging can also be applied for reducing the mean square error of point estimates. Generally speaking, given an estimate $\hat{\theta} = T(X_1, \dots, X_n)$ based on i.i.d. observations X_1, \dots, X_n , the bagged version of $\hat{\theta}$ is defined as $\hat{\theta}_B := E_{F_n}(\hat{\theta}^*)$, where $\hat{\theta}^* = T(X_1^*, \dots, X_m^*)$ and X_1^*, \dots, X_m^* is a sample of i.i.d. observations drawn from F_n , the empirical distribution function corresponding to the original sample. Thus, $\hat{\theta}_B$ is the average of the bootstrap values $\hat{\theta}^*$ across all the possible bootstrap samples. Usually this average can be computed neither analytically nor exactly, but can be approximated by drawing a large enough number of samples from F_n .

The effects of bagging on the mean square error of an estimate are uncertain. On the one hand averaging will usually decrease the variance but, on the other hand, one can also expect an increase of the bias. Indeed, we have that

$$\hat{\theta}_B = E_{F_n}(\hat{\theta}^*) = \hat{\theta} + (E_{F_n}(\hat{\theta}^*) - \hat{\theta}),$$

where $E_{F_n}(\hat{\theta}^*) - \hat{\theta}$ is a bootstrap bias estimate. Thus the bias of the bagged estimate is roughly double the bias of the original estimate.

Some recent papers are devoted to clarifying the effects of bagging for different classes of estimates. Bühlmann and Yu (2002) study regression and classification trees, Chen and Hall (2003) consider M-estimates and Buja and Stuetzle

(2006) analyze U-statistics. In a nice survey on bagging by Bühlmann (2004), a robust modification of bagging is described and termed *bragging* (**bootstrap robust aggregating**). The difference between bragging and bagging is just that the median is used instead of the expectation to aggregate the results obtained across the bootstrap samples.

Let $\text{ave}\{X_1, \dots, X_m\}$ and $\text{med}\{X_1, \dots, X_m\}$ be, respectively, the average and the median of X_1, \dots, X_m , and let $E_F[T(X_1, \dots, X_m)]$ and $M_F[T(X_1, \dots, X_m)]$ be, respectively, the expectation and population median of $T(X_1, \dots, X_m)$ under F . The average and the median can play two different roles in the bootstrap aggregating process: both may be the estimate that is aggregated, and both can be used as the operator for aggregating. Accordingly, we will study some properties of the *bagged median*

$$E_{F_n}[\text{med}\{X_1^*, \dots, X_m^*\}] \tag{1}$$

and the *bragged mean*

$$M_{F_n}[\text{ave}\{X_1^*, \dots, X_m^*\}]. \tag{2}$$

It is easy to see that the bagged average amounts to the average and that the bragged median coincides with the median. Therefore, (1) and (2) are the only two combinations that deserve further attention. In both cases, we allow the bootstrap sample size m to be different from the original sample size n . This flexibility helps to interpret the estimates arising after the bagging process, which is our main goal.

The paper is organized as follows: Section 2 is devoted to studying the properties of the bagged median as a function of the bootstrap sample size m . In particular, a new measure of robustness, the *breakdown probability profile* is defined in this section. In Section 3 we study the properties of the bragged mean. Section 4 contains some further remarks and comments on possible generalizations. More technical material is contained in a final Appendix.

2 The bagged median

2.1 Representation of bagged order statistics as L-estimates

Instead of the median we consider first a slightly more general situation, the application of bagging to any order statistics. Although approximations of bagged estimates can be easily obtained using a Monte Carlo scheme, explicit expressions for the bootstrap distribution of order statistics are available so that simulation is not needed.

Henceforth the k th order statistic of X_1, \dots, X_n will be denoted by $X_{k:n}$. Fix $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, m\}$. Notice that $X_{k:m}^* > X_{i:n}$ if and only if fewer than k bootstrap observations X_j^* are less than or equal to $X_{i:n}$. As a consequence,

$$P_{F_n}(X_{k:m}^* > X_{i:n}) = \sum_{j=0}^{k-1} \binom{m}{j} \left(\frac{i}{n}\right)^j \left(1 - \frac{i}{n}\right)^{m-j}.$$

Therefore, if we denote $p_{n,m,k,i} := P_{F_n}(X_{k:m}^* = X_{i:n})$ and use the relationship between binomial probabilities and the incomplete beta function, we have

$$\begin{aligned} p_{n,m,k,i} &= P_{F_n}(X_{k:m}^* > X_{i-1:n}) - P_{F_n}(X_{k:m}^* > X_{i:n}) \\ &= k \binom{m}{k} \int_{(i-1)/n}^{i/n} t^{k-1} (1-t)^{m-k} dt. \end{aligned} \quad (3)$$

and

$$E_{F_n}(X_{k:m}^*) = \sum_{i=1}^n p_{n,m,k,i} X_{i:n}. \quad (4)$$

This means that bagged order statistics are linear combinations of order statistics, that is, L-estimates.

Expressions (3) and (4) for $m = n$ already appear in Efron (1979) and, for the special case of the median, in Maritz and Jarrett (1978). The use of (3) and (4) for estimating quantiles was proposed by Harrell and Davis (1982). The expression for $m \neq n$ presented above may have some additional interest. For instance, one may be interested in estimating the expectation of the maximum among m observations, $E_F(X_{m:m})$, using a sample of size n . These expectations are useful to characterize some stochastic orders (see de la Cal and Cárcamo, 2006). Using

(3) and (4) with $k = m$, an estimate is

$$E_{F_n}(X_{m:m}^*) = \sum_{i=1}^n \left[\binom{i}{n}^m - \binom{i-1}{n}^m \right] X_{i:n}.$$

Still, the main advantage of considering $m \neq n$ is the possibility of modulating the degree of bagging, as we discuss in the next subsection.

2.2 The bagged median and the size of the resamples

The bagged median based on bootstrap samples of size m is defined in (1), where the sample median is defined as usual,

$$\text{med}\{X_1, \dots, X_n\} = \begin{cases} X_{(n+1)/2:n}, & \text{if } n \text{ is odd} \\ (X_{n/2:n} + X_{n/2+1:n})/2, & \text{if } n \text{ is even} \end{cases}.$$

When $m = 1$, the bagged median is just the sample mean (so that this case corresponds to a maximum degree of bagging). For very large values of m , approximately m/n of the bootstrap values will be equal to each original observation X_i so that the bagged median equals the median (the limiting case $m = \infty$ corresponds to no bagging at all). For intermediate values of m , (3) and (4) allow us to represent the bagged median as an L-estimate:

$$E_{F_n}[\text{med}\{X_1^*, \dots, X_m^*\}] = \sum_{i=1}^n w_{n,m,i} X_{i:n}, \quad (5)$$

where $w_{n,m,i} = p_{n,m,(m+1)/2,i}$ if m is odd, and $w_{n,m,i} = (p_{n,m,m/2,i} + p_{n,m,m/2+1,i})/2$ if m is even.

To understand the role played by m it is illustrative to display the weights $w_{n,m,i}$ versus the rank of each observation i varying m for a fixed sample size n . In Figure 1 we have plotted $w_{n,m,i}$ versus i for $n = 51$, and $m = 1, 11, 21, \dots, 101$. As we have mentioned above, when $m = 1$ the bagged median coincides with the sample mean. However, as m increases, the weights progressively concentrate about the median of the original sample. For different choices of m we obtain estimates that represent different levels of compromise between the median and the mean, maybe closer to the mean (small m) or to the median (large m). We

can therefore understand m as a tuning constant analogous to those that appear in the definitions of many robust estimates. Thus the effect of bagging the median is closely related to the robustness-efficiency trade-off that usually arises in robust estimation.

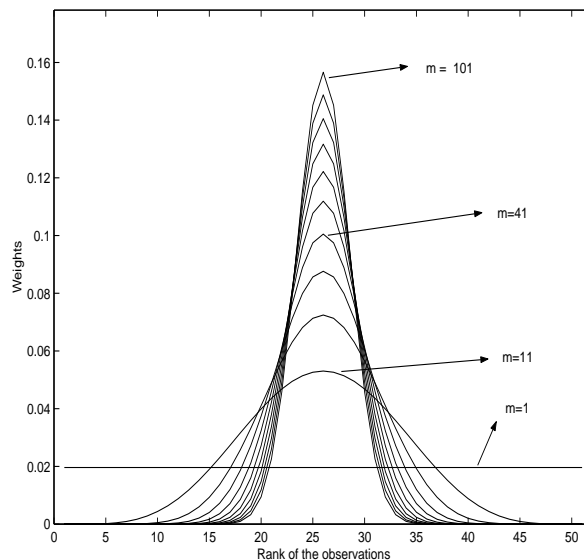


Figure 1: Weights corresponding to each observation as a function of its rank for several values of m ($n = 51$).

The shape of the weight curves in Figure 1 immediately suggests a relationship with the normal density curve. Indeed, it can be shown (see Appendix) that the following representation is valid for large n ,

$$E_{F_n}[\text{med}\{X_1^*, \dots, X_m^*\}] \approx \frac{1}{nh_m} \sum_{i=1}^n K\left(\frac{i/n - 1/2}{h_m}\right) X_{i:n}, \quad (6)$$

where $K(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$ is a Gaussian kernel and $h_m = [4(m+1)]^{-1/2}$ is a smoothing parameter. Accordingly, the effects of bagging the median could also be studied in the context of the bias-variance trade-off that appears in non-parametric estimation problems. For $m = n$, the approximation (6) was derived by Sheather and Marron (1990).

2.3 Robustness: breakdown probability profile

One of the most popular robustness measures is the breakdown point, that is, the minimum proportion of outliers that can arbitrarily determine the value of an estimate. As we have seen, the bagged median converges to the sample median as $m \rightarrow \infty$, however this convergence is not reflected in an improvement of the breakdown point, which is zero for all values of m , since the weights $w_{n,m,i}$ are always strictly positive. In this section, we introduce a probabilistic concept of breakdown point which is able to capture the increase in robustness as m gets larger.

Assume that we approximate the bagged median using resampling, that is, drawing many bootstrap samples and averaging the corresponding bootstrap medians. There is a positive probability that this approximate estimate does not break down, regardless of the fraction and the configuration of the outliers. The reason is that the fraction of outliers that appear in the bootstrap samples, which is random, might not attain the breakdown point of the median and this event could happen for each of the bootstrap samples, so that the average would not break down. We take advantage of this fact and define the *breakdown probability profile* (BPP) of a bagged estimate as the probability of breakdown of its resampling approximation for each fraction of outliers ϵ in the original sample.

If the original sample size is n and we use B bootstrap samples of size m to compute the bagged median, each of the bootstrap samples does not break down if and only if it contains fewer than $[m + 1]/2$ outliers, where $[x]$ stands for the integer part of x . Therefore, the corresponding BPP is given by

$$\text{BPP}_m(\epsilon) = 1 - \left[\text{P} \left(\text{Bin}(m; \epsilon) < \left[\frac{m + 1}{2} \right] \right) \right]^B,$$

where ϵ takes the values $1/n, 2/n, \dots, 1$. In Figure 2 we display the BPP of the bagged median for $n = 50$, $B = 1000$, and several values of m ranging from 10 to 100.

An increase of m shifts the position of the BPP toward the right, meaning that the breakdown behavior of the estimate approaches that of the median, whose

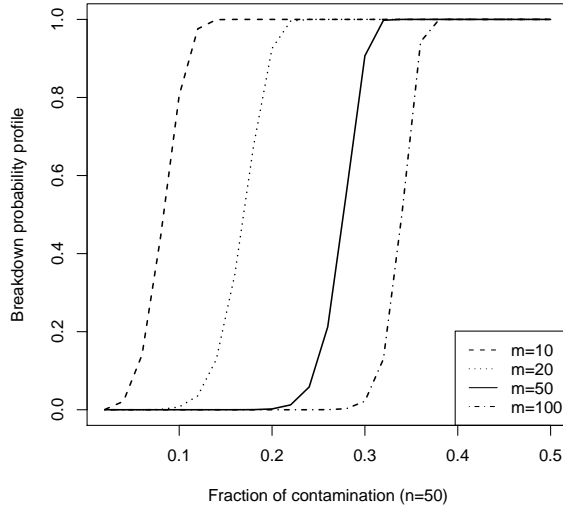


Figure 2: Breakdown probability profiles of the bagged median for several values of m ($n = 50$ and $B = 1000$).

BPP is 0 for $\epsilon < [(n + 1)/2]/n$, and 1 for $\epsilon \geq [(n + 1)/2]/n$.

Observe that for each value of m , BPP_m has a sigmoidal shape. This implies that for each m there exists a threshold value of ϵ beyond which the probability of breakdown grows abruptly. For practical purposes, we could consider such a threshold value as the breakdown point. For instance, given a small risk of breakdown α that we are willing to assume, we could define the α *resampling breakdown point* as

$$\epsilon_{\alpha, m, B}^* := \sup\{\epsilon : \text{BPP}_m(\epsilon) \leq \alpha\}. \quad (7)$$

For $n = 50$, $B = 1000$, $\alpha = 0.001$ and several values of m , the numerical results are displayed in Table 1. Notice how this weaker concept of breakdown point reflects the convergence of the bagged median to the median as $m \rightarrow \infty$.

m	2	10	20	50	100	∞
$\epsilon_{\alpha, m, B}^*$	0	0.02	0.08	0.18	0.26	0.5

Table 1: α resampling breakdown point for $\alpha = 0.001$, $n = 50$, $B = 1000$ and several values of m .

3 The bragged mean

3.1 Interpretation

In this section the roles of the median and the mean are exchanged, which leads to the bragged mean defined in equation (2) above. To give further insight we first consider a few particular values of m . For $m = 1$, the bragged mean is obviously the sample median. For $m = 2$ we have

$$M_{F_n} [\text{ave}\{X_1^*, X_2^*\}] = \text{med} \left\{ \frac{X_i + X_j}{2} : i, j = 1, \dots, n \right\},$$

so that the bragged mean is almost the well-known Hodges-Lehmann estimate. It would be exactly the Hodges-Lehmann estimate if resampling were without replacement. In the limiting case, $m \rightarrow \infty$, approximately m/n of the bootstrap observations will be equal to each original observation X_i . Therefore,

$$M_{F_n} [\text{ave}\{X_1^*, \dots, X_m^*\}] \approx M_{F_n} [\text{ave}\{X_1, \dots, X_n\}] = \text{ave}\{X_1, \dots, X_n\},$$

and the bragged mean is just the mean. We see that m is again a tuning constant giving different levels of compromise between the median and the mean. Comparing with the bagged median, what we obtain now are generalized Hodges-Lehmann estimates instead of L-estimates (see Serfling (1984) for a closely related general class of robust estimates). Note also that there is a role reversal since now efficiency increases and robustness decreases as m gets larger.

It should be noted that for moderate values of m the bragged mean is already very close to the mean. When F is continuous, for fixed n and $m \rightarrow \infty$, it can be shown (see Appendix) that,

$$M_{F_n} [\text{ave}\{X_1^*, \dots, X_m^*\}] = \bar{X}_n - \frac{\gamma_n}{6m} + o\left(\frac{1}{m}\right) \quad \text{a.s.}, \quad (8)$$

conditionally on the sample, where $\gamma_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^3 / s_n^3$ is the skewness coefficient. As a consequence, the bragged mean can also be interpreted as a skewness corrected average.

3.2 Robustness: breakdown probability profile

The breakdown probability profile of the bragged mean can also be easily computed. Let $\epsilon = 1/n, 2/n, \dots, 1$ be the fraction of outliers in the original sample and assume that we use B bootstrap samples of size m to approximate the bragged mean. The average of each bootstrap sample breaks down when this sample contains at least one outlier. This event happens with probability $1 - (1 - \epsilon)^m$. Therefore, the random number of broken down bootstrap samples follows a binomial distribution $\text{Bin}(B; 1 - (1 - \epsilon)^m)$. The bragged mean breaks down when at least $\lceil (B + 1)/2 \rceil$ bootstrap averages break down (since we are using the median for aggregating the averages). Hence the BPP is given by

$$\text{BPP}_m(\epsilon) = \text{P} \left(\text{Bin}(B; 1 - (1 - \epsilon)^m) \geq \left\lceil \frac{B + 1}{2} \right\rceil \right), \quad \epsilon = 1/n, 2/n, \dots, 1.$$

In Figure 3 we have represented the BPP of the bragged mean for $n = 50$ and $m = 2, 3, 4, 5$.

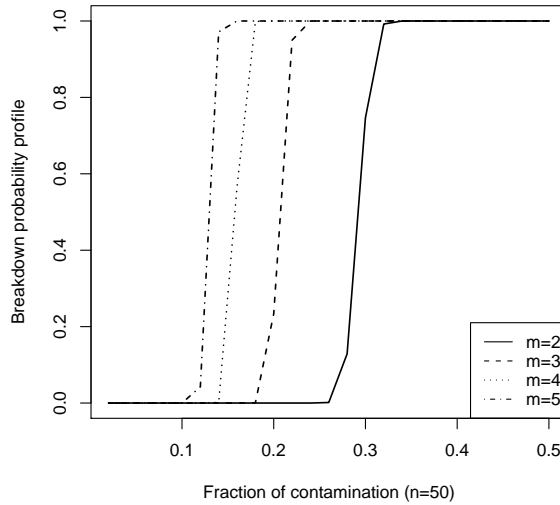


Figure 3: Breakdown probability profiles of the bragged mean for several values of m ($n = 50$).

The interpretation of Figure 3 is similar to that of Figure 2. In this case,

a decrease of m displaces the BPP to the right. The sigmoidal shape appears again so that the definition of α resampling breakdown point given in (7) is also meaningful in this case. For $n = 50$, $B = 1000$, $\alpha = 0.001$ and several values of m , the numerical results are displayed in Table 2.

m	2	3	4	5
$\epsilon_{\alpha, m, B}^*$	0.240	0.375	0.120	0.100

Table 2: α resampling breakdown point for $\alpha = 0.001$, $n = 50$, $B = 1000$ and several values of m .

3.3 Robustness: the influence function

Most estimates can be viewed as the value that a functional $T(\cdot)$, defined on the set of probability distribution functions, takes at the empirical distribution corresponding to the sample. In these cases, the *influence function* (IF), defined as the derivative

$$\text{IF}(x; T, F) := \lim_{\epsilon \rightarrow 0} \frac{T(F_{\epsilon, x}) - T(F)}{\epsilon},$$

where $F_{\epsilon, x} = (1 - \epsilon)F + \epsilon\Delta_x$, provides information about how the estimate changes when we add a new observation located at x to the sample. The computation of the IF requires determination of the functional T from which the estimate arises. In this section we derive such a functional for the braggged mean and then we use a standard argument to derive the corresponding IF.

The key point is that the median M_G of a distribution G satisfies the equation $E_G\psi(X - M_G) = 0$, for the score function $\psi(x) = \text{sgn}(x)$. Hence, if $M_{n, m}$ stands for the braggged mean we have $E_{F_n}\psi(\text{ave}\{X_1^*, \dots, X_m^*\} - M_{n, m}) = 0$. Now we just have to replace F_n with an arbitrary distribution G to obtain the desired functional version: $E_G\psi[\text{ave}\{X_1, \dots, X_m\} - T_m(G)] = 0$. When we put $G = F_{\epsilon, x}$, we get

$$\begin{aligned} & (1 - \epsilon)^m E_{F_n}\psi[\text{ave}\{X_1, \dots, X_m\} - T_m(F_{\epsilon, x})] \\ & + m\epsilon(1 - \epsilon)^{m-1} E_{F_n}\psi[\text{ave}\{X_1, \dots, X_m\} - T_m(F_{\epsilon, x})] + o(\epsilon) = 0 \end{aligned}$$

The standard procedure to compute the influence function of M-estimates is differentiating the expression above w.r.t. ϵ and evaluating at $\epsilon = 0$. Notice that the terms gathered in $o(\epsilon)$ are not relevant for the computation. Following this method we end up with this expression for the IF of the bragged mean:

$$IF(x; T_m, F) = \frac{mE_F\psi(x + X_2 + \cdots + X_m)}{-\left[\frac{\partial}{\partial t}E_F\psi(X_1 + X_2 + \cdots + X_m - mt)\right]_{t=0}} \quad (9)$$

Although expression (9) is rather general, it can be worked out under more particular conditions. For instance, if we assume that F has an even density f , then it can be shown using a standard induction argument that (9) reduces to

$$IF(x; T_m, F) = \frac{2F^{(m-1)}(x) - 1}{2 \int f(u)f^{(m-1)}(u)du}, \quad (10)$$

where $F^{(m)}$ is the m th convolution of F and $f^{(m)}$ is its corresponding density. In particular, for $m = 2$, we obtain the influence function of the Hodges-Lehmann estimate. For the normal distribution $F = \Phi$, we have that $\Phi^{(m)}(u) = \Phi(u/\sqrt{m})$ so that a still more explicit expression is available:

$$IF(x; T_m, \Phi) = \sqrt{\frac{\pi m}{2}} \left[2\Phi\left(\frac{x}{\sqrt{m-1}}\right) - 1 \right]. \quad (11)$$

Figure 4 exhibits the IF of the bragged mean for several values of m under the normal model. We observe that the part in the middle of the functions, $x \in (-2, 2)$ say, is fairly similar for all the values of m . However the influence of larger values of x increases quickly with m .

The behavior of the IF is sometimes summarized by the so called *gross error sensitivity* (GES) given by $GES = \sup_x |IF(x; T, F)|$. Using (11) we see that $GES_m = \sqrt{\pi m/2}$ under the normal model. Although we have bounded inference for all the values of m , it also holds that $\lim_{m \rightarrow \infty} GES_m = \infty$, which is not surprising since the estimates approach the sample average.

A well-known heuristic argument (see e.g. Huber (1981), p. 14) shows that the asymptotic variance (AV) of an estimate can be computed by integrating the square of its influence function, $AV(T, F) = \int IF(x; T, F)^2 dF(x)$. In Table 3 we

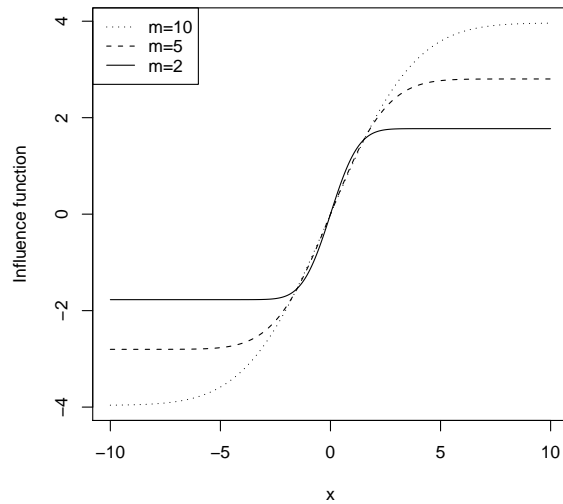


Figure 4: Influence functions of the bagged mean for several values of m under the normal model.

have displayed both the asymptotic efficiency and the GES of the bagged mean for several values of m under the normal model. Notice that the efficiencies are very close to 100% in agreement with the fact that the bagged mean is quite close to the sample mean even for moderate values of m . Table 3 shows the usual trade-off between efficiency and robustness.

m	2	3	4	5
eff_m	95.5%	98.1%	98.9%	99.3%
GES_m	1.77	2.17	2.51	2.80

Table 3: GES and asymptotic efficiency of the bagged mean for several values of m under the normal model.

4 Conclusions

When we compute the bagged version of an estimate, the size of the bootstrap samples can be viewed as a tuning constant that controls the trade-off between

efficiency and robustness. In this paper, we have illustrated this fact using the median and the average. Of course, the same ideas can also be applied to other estimates. For instance, we could have considered dispersion estimates such as the sample standard deviation or the median of absolute deviations (MAD). The *bagged standard deviation* or the *bragged MAD* would yield new families of dispersion estimates whose robustness and efficiency properties as a function of m could be a matter of further study.

The concept of breakdown probability profile is useful to describe the breakdown properties of any estimate obtained through resampling, not just the bagged median or the bragged mean. In contrast with previous measures of robustness (see e.g. Singh, 1998) the breakdown probability profile focuses on the Monte Carlo approximation of the estimates rather than on the ideal bootstrap estimates. This could be relevant in practice since Monte Carlo approximations are certainly the only estimates that can be effectively computed in most cases. Among other issues, the breakdown probability profile takes into account the number B of bootstrap samples that we are drawing or the probability α of breakdown that we are willing to assume.

Appendix

Relationship between the bagged median and kernel quantile estimates

In this section we show the approximation (6). For the sake of simplicity, assume that m is odd. From (3), with $k = (m + 1)/2$, we get

$$w_{m,n,i} = \frac{m+1}{2} \binom{m}{(m+1)/2} \int_{(i-1)/n}^{i/n} [t(1-t)]^{(m-1)/2} dt$$

Now, when n is large, the following approximation for the integral applies

$$\int_{(i-1)/n}^{i/n} [t(1-t)]^{(m-1)/2} dt \approx \frac{1}{n} \left[\frac{i}{n} \left(1 - \frac{i}{n} \right) \right]^{(m-1)/2}.$$

Therefore,

$$w_{m,n,i} \approx \frac{m+1}{2n} \binom{m}{(m+1)/2} \left[\frac{i}{n} \left(1 - \frac{i}{n} \right) \right]^{(m-1)/2}.$$

Finally, we just have to use Lemma 1, p. 412, in Sheather and Marron (1990) with $\alpha = \beta = m + 1$ and $p = q = 1/2$ to obtain:

$$w_{m,n,i} \approx \frac{1}{n} \sqrt{\frac{2(m+1)}{\pi}} \exp \left\{ -2(m+1) \left(\frac{i}{n} - \frac{1}{2} \right)^2 \right\}.$$

Asymptotic representation of the bragged mean

First, we justify the asymptotic representation (8). When F is continuous, the empirical distribution F_n is non-lattice a.s. Therefore, we can apply Theorem 4, p. 426, in Hall (1980) to the empirical distribution function, conditionally on the sample, which yields (8) straightforwardly.

If one is not ready to use specialized results, a simple application of Chebychev's inequality shows that the difference between the mean and the bragged mean cannot be large. For i.i.d. observations drawn from a distribution G with mean μ and variance σ^2 we have, for all $\epsilon > 0$,

$$P_G \{ |\bar{X}_m - \mu| > \epsilon \} \leq \frac{\sigma^2}{m\epsilon^2}. \quad (12)$$

Conditionally on X_1, \dots, X_n , we can apply (12) with $G = F_n$ [so that $\mu = \bar{X}_n$ and $\sigma^2 = S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$], and $\epsilon^2 = 2S_n^2/m$. This yields

$$P_{F_n} \left\{ |\bar{X}_m^* - \bar{X}_n| > \sqrt{\frac{2}{m}} S_n \right\} \leq 1/2 \quad \text{a.s.}$$

The last equation implies that

$$|M_{F_n} [\text{ave}\{X_1^*, \dots, X_m^*\}] - \bar{X}_n| \leq \sqrt{\frac{2}{m}} S_n \quad \text{a.s.}$$

Therefore, fixing n and letting $m \rightarrow \infty$, the difference between the bragged mean and the mean is less than $O(1/\sqrt{m})$.

References

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, **26**, 123–140.
- Bühlmann, P. (2004). Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics: Concepts and Methods* (eds. Gentle, J., Härdle, W. and Mori, Y.), pp. 877–907. Springer.
- Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.*, **30**, 927–961.
- Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statist. Sinica*, **16**, 323–351.
- Chen, S. X. and Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statist. Sinica* **13**, 97–109.
- de la Cal, J. and Cárcamo, J. (2006). Stochastic orders and majorization of mean order statistics. *J. Appl. Prob.*, **43**, 704–712.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.
- Hall, P. (1980). On the limiting behaviour of the mode and median of a sum of independent random variables. *Ann. Probab.*, **8**, 419–430.
- Harrell, F.E. and Davis, C.E. (1982). A new distribution-free quantile estimator. *Biometrika*, **69**, 635–640.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
- Sheather, S.J. and Marron, J.S. (1990). Kernel quantile estimators. *J. Amer. Statist. Assoc.*, **85**, 410–416.
- Maritz, J.S. and Jarrett, R.G. (1978). A note on estimating the variance of the sample median. *J. Amer. Statist. Assoc.*, **73**, 194–196.
- Serfling, R.J. (1984). Generalized L-,M-, and R-statistics. *Ann. Statist.*, **12**, 76–86.
- Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Ann. Statist.*, **26**, 1719–1732.