Statistica Sinica **13**(2003), 97-109

EFFECTS OF BAGGING AND BIAS CORRECTION ON ESTIMATORS DEFINED BY ESTIMATING EQUATIONS

Song Xi Chen and Peter Hall

National University of Singapore and Australian National University

Abstract: Bagging an estimator approximately doubles its bias through the impact of bagging on quadratic terms in expansions of the estimator. This difficulty can be alleviated by bagging a suitably bias-corrected estimator, however. In these and other circumstances, what is the overall impact of bagging and/or bias correction, and how can it be characterised? We answer these questions in the case of general estimators defined by estimating equations, including for example maximum likelihood and method of moments estimators. It is shown that, despite the considerable variety of estimators that can be constructed by bagging and bias correction, the number of modes of behaviour is very small. In particular, bagging a bias-corrected estimator produces a new estimator that is second-order equivalent to the original, unadjusted estimator. Furthermore, the conventional bagged estimator, and the standard bias-corrected estimator, represent virtually equal but opposite adjustments of the conventional estimator. In particular, bagging adds back the adjustment provided by bias correction. If we bag a doubly bias corrected estimator, constructed so as to counteract the tendency of bagging to exacerbate bias, then the result is an estimator that is second-order equivalent to the standard bias-corrected estimator. These results do not depend on the manner of bias correction; that procedure may be implemented using the jackknife, the parametric bootstrap or the nonparametric bootstrap. They show that, when bagging is applied to relatively conventional statistical problems, it cannot reliably be expected to improve performance. Its domain is, in effect, restricted to problems such as regression trees, where variability is so high that it cannot be plausibly modelled using the approach taken here.

Key words and phrases: Bootstrap, estimating function, jackknife, maximum likelihood, mean square error, parametric bootstrap.

1. Introduction

Bagging was introduced by Breiman (1996, 1999) as a means of improving performance of statistical methods. In relatively conventional problems where an estimator can be represented as a smooth function of the data, for example as a smooth function of sums of independent random variables, it is known that bagging a conventional estimator generally tends to increase bias. In such cases, any improvements in performance will arise principally through the impact that

bagging has on variability. However, one does not need to bag the "raw" estimator $\hat{\theta}$; the deleterious effects of bagging on bias can be reduced by bagging bias-reduced forms of $\hat{\theta}$, constructed using either jackknife or bootstrap methods to effect bias corrections. In these general circumstances, what will be the overall impact of bagging, for example on mean squared error? If bagging is applied in familiar statistical problems, in particular those where estimators are defined by estimating equations, can its impact be characterised in a simple manner?

In the present paper we answer these questions, detailing the effects of bagging applied to either conventional or bias-reduced estimators. Despite the considerable variety of estimators that can be constructed in this way, the number of different modes of behaviour is surprisingly small. To summarise them, let us take the basic estimator $\hat{\theta}$ to be determined by an estimating equation as a function of a random dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$, and its bagged form $\tilde{\theta} = E(\hat{\theta}^*|\mathcal{X})$ to be the conditional mean of the value $\hat{\theta}^*$ of $\hat{\theta}$ computed from a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ drawn by sampling randomly, with replacement, from $\mathcal{X}$. We show that bagging a jackknife or bootstrap bias-corrected version $\hat{\theta}_{\rm bc}$ of $\hat{\theta}$ produces a new estimator that is second-order equivalent to the original $\hat{\theta}$. If, on the other hand, we bag an estimator that is computed by subtracting twice the conventional bias correction from $\hat{\theta}$, then the result is second-order equivalent to the standard bias-corrected, unbagged estimator $\hat{\theta}_{\rm bc}$.

These results, and those discussed below, continue to hold if either the parametric or non-parametric bootstrap is used to construct bias corrections. Thus in a parametric setting we may estimate bias by $E(\hat{\theta}^\star|\mathcal{X}) - \hat{\theta}$, where $\hat{\theta}^\star$ denotes the value the estimator $\hat{\theta}$ would take if it were computed instead from a random sample drawn from the distribution with density $f(\cdot|\hat{\theta})$.

Any one of the three estimators $\hat{\theta}$, $\hat{\theta}_{\rm bc}$ and $\tilde{\theta}$ may have asymptotically least mean squared error to second order. Particular interest centres on $\hat{\theta}_{\rm bc}$ and $\tilde{\theta}$, which represent "corrections" of $\hat{\theta}$ by approximately equal amounts in opposite directions. In fact, to second order, $\tilde{\theta}$ equals $\hat{\theta} + (\hat{\theta} - \hat{\theta}_{\rm bc})$, and so bagging adds back rather than subtracts away the bias estimate.

Of course, the impact of empirical bias correction on mean squared error can be substantially greater than just its effect on bias. The bias correction introduces another term, arising from the correlation between the correction and the estimator, which in the formula for mean squared error is of the same order as squared bias. Depending on its sign and size this correlation can more than compensate for the increased bias suffered by $\tilde{\theta}$; that property characterises occasions where $\tilde{\theta}$ outperforms $\hat{\theta}_{\rm bc}$. However, while there exist problems for which this result holds, they do not seem to arise commonly. In particular, in the setting of univariate maximum likelihood estimation of location, neither $\hat{\theta}$ nor $\tilde{\theta}$ ever bests $\hat{\theta}_{\rm bc}$ in terms of second-order properties of mean squared error. And in that context, whenever second-order bias does not vanish, $\hat{\theta}_{\rm bc}$ has strictly smaller mean squared error than either $\hat{\theta}$ or $\tilde{\theta}$ for all sufficiently large n. In this

important class of problems the potential performance advantages of bagging are never realised.

If the bootstrap rather than the jackknife is employed to construct $\hat{\theta}_{\rm bc}$ then the operation of bagging involves the double bootstrap. We use tilde notation to denote bagged estimators, so in either the jackknife or the bootstrap case the bagged "version" of $\hat{\theta}_{\rm bc}$ is represented by $\tilde{\theta}_{\rm bc}$. However, we argue that $\tilde{\theta}_{\rm bc}$ should not be computed simply by bagging $\hat{\theta}_{\rm bc}$, since that would not adequately compensate for bias. The main effects of bias arise from quadratic terms, and bagging virtually doubles their size. This can be seen from the fact that if $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed data pairs, and (X_i^*, Y_i^*) are the corresponding resampled values, then

$$\frac{1}{n^2} E\Big(\sum_{i=1}^n X_i^* \sum_{i=1}^n Y_i^*\Big) - E(X_1)E(Y_1) = (2 - n^{-1})\Big\{\frac{1}{n^2} E\Big(\sum_{i=1}^n X_i \sum_{i=1}^n Y_i\Big) - E(X_1)E(Y_1)\Big\}.$$

That is why, as noted earlier, $\tilde{\theta}_{\rm bc}$ should be calculated by bagging an unconventional version of $\hat{\theta}_{\rm bc}$ in which bias is corrected by subtracting twice the usual amount.

In Section 2 we establish these properties for any estimator that is defined in terms of conventional estimating equations and is computed from a sample of random data. Particular examples are given in Section 3. Results from a simulation study are reported in Section 4, and technical arguments are outlined in Section 5. The conclusion of our analysis is that when bagging is applied to relatively conventional statistical problems, it cannot reliably be expected to improve performance. Its domain is, in effect, restricted to problems such as regression trees, where variability is so high that it cannot be plausibly modelled using the approach taken here.

Recent studies of the performance of bagging include those of Bühlmann and Yu (2000), who describe properties of a relatively sophisticated mathematical model for bagging regression trees; Buja and Stuetzle (2000a, b), who address theoretical features of bagging statistical functionals related to quadratics; and Friedman and Hall (2000), who discuss variance and mean squared error properties of bagging under an abstract model for the objective function. The work of Buja and Stuetzle is most closely related, in that it uses techniques based on Taylor expansion to explore properties of bias, variance and mean squared error of the estimator.

2. Main Results

2.1. Estimators and their bias-corrected forms

Let $\sum_i \psi_0(X_i|\theta) = 0$ denote an estimating equation defining an estimator $\hat{\theta}$ of the parameter θ. We may take θ to be p-variate, for a general $p \geq 1$, in which

case $\psi_0(\cdot|\theta)$ is a p-variate function; our main conclusions do not depend on p. For simplicity, however, we discuss detailed results only for $p = 1$, leaving until Section 2.4 an account of the general case. Likewise we confine attention here to bias-reduction methods based in the nonparametric bootstrap, giving in Section 2.5 a summary of the parametric case. Main conclusions are the same in both cases, although there are algebraic differences.

Examples of estimators defined by estimating equations include maximum likelihood estimators, where $\psi_0(x) = \dot{f}(x|\theta)/f(x|\theta)$ with $\dot{f}(x|\theta) = (\partial/\partial\theta)f(x|\theta)$. More generally, put $\psi_j(x|\theta) = (\partial/\partial\theta)^j\psi_0(x|\theta)$ and let θ_0 denote the true value of θ. Taylor expanding the estimating equation we obtain

$$S_0 + (\theta - \theta_0)\, S_1 + \cdots + \frac{1}{j!}\,(\theta - \theta_0)^j\, S_j + \cdots = 0\,, \tag{2.1}$$

where $S_j = n^{-1}\sum_i \psi_j(X_i|\theta_0)$. Solving (2.1) for $\hat\theta$, and expanding the solution in successively higher powers of S_0/S_1, we deduce that

$$\hat\theta - \theta_0 = -\Big\{\frac{S_0}{S_1} + \frac{1}{2}\Big(\frac{S_0}{S_1}\Big)^2\frac{S_2}{S_1}\Big\} + \Big(\frac{S_0}{S_1}\Big)^3\Big\{\frac{1}{6}\,\frac{S_3}{S_1} - \frac{1}{2}\Big(\frac{S_2}{S_1}\Big)^2\Big\} + \cdots\,. \tag{2.2}$$

Without loss of generality, $\theta_0 = 0$. Then the expected value of $\hat\theta$ may be expressed as

$$E(\hat\theta) = n^{-1}\,(\alpha_1 - \tfrac{1}{2}\,\rho_2\,\sigma_0^2) + O(n^{-2})\,, \tag{2.3}$$

where $\alpha_j = \mu_1^{-2}\,E\{\psi_0(X)\,\psi_j(X)\}$, $\mu_j = E\{\psi_j(X|\theta_0)\}$, $\rho_j = \mu_j/\mu_1$ and $\sigma_j^2 = \mu_1^{-2}\mathrm{var}\{\psi_j(X|\theta_0)\}$. It is assumed throughout that $\mu_1 \neq 0$, which implies that $\hat\theta$ is root-n consistent for θ_0.

The standard jackknife and bootstrap estimators of the bias of $\hat\theta$ are, respectively,

$$\widehat{\mathrm{bias}}_{\mathrm{jack}} = \sum_{i=1}^n \hat\theta_i - n\,\hat\theta\,, \quad \widehat{\mathrm{bias}}_{\mathrm{boot}} = E(\hat\theta^*|\mathcal{X}) - \hat\theta\,, \tag{2.4}$$

where $\hat\theta_i$ denotes the version of $\hat\theta$ computed from the sample $\mathcal{X}\backslash\{X_i\}$. The standard jackknife bias-corrected and bootstrap bias-corrected estimators are $\hat\theta_{\mathrm{jack}} = \hat\theta - \widehat{\mathrm{bias}}_{\mathrm{jack}}$ and $\hat\theta_{\mathrm{boot}} = \hat\theta - \widehat{\mathrm{bias}}_{\mathrm{boot}}$, respectively. They completely correct for all bias of order n^{-1}, and hence satisfy $E(\hat\theta_{\mathrm{jack}}) = O(n^{-2})$ and $E(\hat\theta_{\mathrm{boot}}) = O(n^{-2})$; compare (2.3).

As argued in Section 1, in order to counteract the effect of bias on the bagged form of $\hat\theta$ we should bag an estimator that has twice the bias estimate subtracted, i.e., $\hat\theta_{\mathrm{bjack}} \equiv \hat\theta - 2\,\widehat{\mathrm{bias}}_{\mathrm{jack}}$ or $\hat\theta_{\mathrm{bboot}} \equiv \hat\theta - 2\,\widehat{\mathrm{bias}}_{\mathrm{boot}}$ instead of $\hat\theta_{\mathrm{jack}}$ and $\hat\theta_{\mathrm{boot}}$. The subscript notation "bjack" here denotes "jackknife bias-corrected estimator suitable for bagging", with a similar interpretation for "bboot". The respective bagged forms are

$$\tilde\theta_{\mathrm{jack}} \equiv E(\hat\theta^*_{\mathrm{bjack}}|\mathcal{X}) = 3\tilde\theta - 2E(\hat\theta_1^*|\mathcal{X}),\ \ \tilde\theta_{\mathrm{boot}} \equiv E(\hat\theta^*_{\mathrm{bboot}}|\mathcal{X}) = 3\tilde\theta - 2E(\hat\theta^{**}|\mathcal{X}), \tag{2.5}$$

where $\tilde{\theta} = E(\hat{\theta}^*|\mathcal{X})$ is the standard bagged estimator, $\hat{\theta}^*_{\rm bjack}$, $\hat{\theta}^*_{\rm bboot}$ and $\hat{\theta}^*_i$ denote the values $\hat{\theta}_{\rm bjack}$, $\hat{\theta}_{\rm bboot}$ and $\hat{\theta}_i$ would assume if they were calculated from a bootstrap resample $\mathcal{X}^*$ instead of the original sample $\mathcal{X}$, and $\hat{\theta}^{**}$ is the value $\hat{\theta}$ would have if it were computed from a double-bootstrap resample, i.e., a bootstrap resample derived from a bootstrap resample derived from $\mathcal{X}$. We denote $\tilde{\theta}_{\rm jack}$ and $\tilde{\theta}_{\rm boot}$ generically by $\tilde{\theta}_{\rm bc}$.

Of course, we can nevertheless bag the conventional bias-corrected estimators $\hat{\theta}_{\rm jack}$ and $\hat{\theta}_{\rm boot}$, obtaining respectively

$$\tilde{\theta}_{\rm pjack} \equiv E(\hat{\theta}^*_{\rm jack}|\mathcal{X})\,, \quad \tilde{\theta}_{\rm pboot} \equiv E(\hat{\theta}^*_{\rm boot}|\mathcal{X})\,, \tag{2.6}$$

where $\hat{\theta}^*_{\rm jack}$ and $\hat{\theta}^*_{\rm boot}$ denote the versions of $\hat{\theta}_{\rm jack}$ and $\hat{\theta}_{\rm boot}$ computed from $\mathcal{X}^*$. The subscripts "pjack" and "pboot" denote "partial jackknife" and "partial bootstrap", respectively. We denote $\tilde{\theta}_{\rm pjack}$ and $\tilde{\theta}_{\rm pboot}$ generically by $\tilde{\theta}_{\rm pbc}$. The subscript "pbc" denotes "partial bias-correction".

2.2. Approximations to bias-corrected and bagged estimators

First we discuss the unbagged, bias-corrected estimators $\hat{\theta}_{\rm jack}$ and $\hat{\theta}_{\rm boot}$, which we denote generically by $\hat{\theta}_{\rm bc}$ where the subscript stands for "bias corrected". It is shown in Section 5 that, assuming $\theta_0 = 0$, we have for either choice of $\hat{\theta}_{\rm bc}$,

$$\hat{\theta}_{\rm bc} = \hat{\theta}\,(1 - n^{-1}\,\hat{\gamma}_1) - n^{-1}\,\hat{\gamma}_2 + O_p(n^{-2})\,, \tag{2.7}$$

where $\hat{\gamma}_1 = \hat{\gamma}_3 - 3\,\hat{\rho}_2\,\hat{\gamma}_2$, $\hat{\gamma}_2 = \hat{\alpha}_1 - \frac{1}{2}\,\hat{\rho}_2\,\hat{\sigma}_0^2$, $\hat{\gamma}_3 = \hat{\sigma}_1^2 + \hat{\alpha}_2 - \frac{1}{2}\,\hat{\rho}_3\,\hat{\sigma}_0^2$, and $\hat{\alpha}_j$, $\hat{\rho}_j$ and $\hat{\sigma}_j^2$ are empirical versions of α_j, ρ_j and σ_j^2: $\hat{\rho}_j = S_j/S_1$,

$$\hat{\alpha}_j = \frac{1}{nS_1^2}\,\sum_{i=1}^n \{\psi_0(X_i) - S_0\}\,\{\psi_j(X_i) - S_j\}\,, \quad \hat{\sigma}_j^2 = \frac{1}{nS_1^2}\,\sum_{i=1}^n \{\psi_j(X_i) - S_j\}^2\,.$$

It will be proved too that the standard bagged estimator $\tilde{\theta}$ admits the expansion

$$\tilde{\theta} = \hat{\theta}\,(1 + n^{-1}\,\hat{\gamma}_1) + n^{-1}\,\hat{\gamma}_2 + O_p(n^{-2})\,. \tag{2.8}$$

Comparing (2.7) and (2.8) we see that, while $\hat{\theta}_{\rm bc}$ and $\tilde{\theta}$ both equal $\hat{\theta}$ to first order, to second order they represent adjustments in completely different directions:

$$\hat{\theta}_{\rm bc} - \hat{\theta} = -(\tilde{\theta} - \hat{\theta}) + O_p(n^{-2})\,. \tag{2.9}$$

In particular to second order, $\tilde{\theta}$ equals $\hat{\theta} + (\hat{\theta} - \hat{\theta}_{\rm bc})$.

Next we address the bagged bias-corrected estimator $\tilde{\theta}_{\rm bc}$. Again we assume for convenience that $\theta_0 = 0$. It is shown in Section 5 that, for either choice of $\tilde{\theta}_{\rm bc}$,

$$\tilde{\theta}_{\rm bc} = \tilde{\theta} - 2\,n^{-1}\,(\hat{\theta}\,\hat{\gamma}_1 + \hat{\gamma}_2) + O_p(n^{-2})\,. \tag{2.10}$$

Combining (2.7)−(2.10) we deduce that

$$\tilde{\theta}_{\rm bc} = \hat{\theta}\,(1 - n^{-1}\,\hat{\gamma}_1) - n^{-1}\,\hat{\gamma}_2 + O_p(n^{-2}) = \hat{\theta}_{\rm bc} + O_p(n^{-2})\,. \tag{2.11}$$

For the sake of definiteness we interpret $\tilde{\theta}_{\rm bc}$ as either $\tilde{\theta}_{\rm jack}$ or $\tilde{\theta}_{\rm boot}$ in the same respective order that $\hat{\theta}_{\rm bc}$ denotes $\hat{\theta}_{\rm jack}$ or $\hat{\theta}_{\rm boot}$, although (2.11) holds even if the order is switched.

Analogously to (2.11) it may be proved that the bagged and partially bias-corrected estimators $\tilde{\theta}_{\rm pbc}$ are second-order equivalent to $\hat{\theta}$:

$$\tilde{\theta}_{\rm pbc} = \hat{\theta} + O_p(n^{-2})\,. \tag{2.12}$$

2.3. Bias, variance and mean squared error

We continue to assume $\theta_0 = 0$. Since, as noted in Section 2.1, $\hat{\theta}_{\rm bc}$ has bias of order $O(n^{-2})$, then (2.11) implies that the same is true of $\tilde{\theta}_{\rm bc}$:

$$E(\tilde{\theta}_{\rm bc}) = O(n^{-2})\,. \tag{2.13}$$

This confirms the wisdom of subtracting twice the bias estimator when computing jackknife or bootstrap bias-corrected estimators prior to bagging. Combining (2.11) and (2.13) we deduce that

$$\mathrm{var}(\tilde{\theta}_{\rm bc}) = \mathrm{var}(\hat{\theta}_{\rm bc}) + O(n^{-3})\,. \tag{2.14}$$

Together, (2.13) and (2.14) imply there is no first- or second-order advantage, in terms of bias or variance or mean squared error, in using a bagged bias-corrected estimator rather than an unbagged bias-corrected estimator. To derive (2.14) we have used the well-known result that expansions of moments are power series in n^{-1}, rather than simply $n^{-1/2}$. This property will be employed below without further comment.

Analogously we deduce from (2.12) that

$$E(\tilde{\theta}_{\rm pbc}) = E(\hat{\theta}) + O(n^{-2})\,, \quad \mathrm{var}(\tilde{\theta}_{\rm pbc}) = \mathrm{var}(\hat{\theta}) + O(n^{-3})\,. \tag{2.15}$$

Hence, there are no first- or second-order differences between the bagged partially bias-corrected estimators $\tilde{\theta}_{\rm pjack}$ and $\tilde{\theta}_{\rm pboot}$, and the original estimator $\hat{\theta}$. Note that in the context of variance, "second order" means terms of size n^{-3}; in the case of bias, "second order" terms are of size n^{-2}. Changes of orders n^{-2} and n^{-3} to bias and variance, respectively, both introduce adjustments of order n^{-3} to mean squared error.

Next we relate the mean squared error of the non-bagged, bias-corrected estimator $\hat{\theta}_{\rm bc}$ to that of the conventional estimator $\hat{\theta}$. Define $\beta_j = \mu_1^{-3}\,E\{\psi_0(X)^2 \times \psi_j(X)\}$, $\gamma_1 = \gamma_3 - 3\,\rho_2\,\gamma_2$, $\gamma_2 = \alpha_1 - \frac{1}{2}\,\rho_2\,\sigma_0^2$, $\gamma_3 = \sigma_1^2 + \alpha_2 - \frac{1}{2}\,\rho_3\,\sigma_0^2$, $\gamma_4 =$

$\alpha_1\,(3\gamma_2 - \alpha_1) + \sigma_0^2\,(1 + \frac{1}{2}\,\alpha_2) + \frac{1}{2}\,\rho_2\,\beta_0 - \beta_1$ and $\gamma_0 = \sigma_0^2\,\gamma_1 + \gamma_4$. In particular, $\gamma_1, \gamma_2, \gamma_3$ are the population counterparts of $\hat\gamma_1, \hat\gamma_2, \hat\gamma_3$ respectively. We show in Section 5 that, assuming $\theta_0 = 0$,

$$E(\hat\theta\,\hat\gamma_2) = n^{-1}\,(\gamma_2^2 + \gamma_4) + O(n^{-2})\,. \tag{2.16}$$

It follows from (2.7), (2.16) and the fact that $E(\hat\theta^2) = n^{-1}\,\sigma_0^2 + O(n^{-2})$, that

$$\begin{aligned} E(\hat\theta_{\rm bc}^2) &= E(\hat\theta^2)\,(1 - 2\,n^{-1}\,\gamma_1) - 2\,n^{-1}\,E(\hat\theta\,\hat\gamma_2) + n^{-2}\gamma_2^2 + O(n^{-3}) \\ &= E(\hat\theta^2) - 2\,n^{-2}\,(\gamma_0 + \tfrac{1}{2}\gamma_2^2) + O(n^{-3})\,. \end{aligned} \tag{2.17}$$

Similarly, using (2.8) in place of (2.7) in this argument,

$$E(\tilde\theta^2) = E(\hat\theta^2) + 2\,n^{-2}\,(\gamma_0 + \tfrac{3}{2}\,\gamma_2^2) + O(n^{-3})\,. \tag{2.18}$$

We know from (2.3) that $E(\hat\theta) = n^{-1}\gamma_2 + O(n^{-2})$, and of course $E(\hat\theta_{\rm bc}) = O(n^{-2})$. Therefore by (2.17), writing $\mathrm{MSE}(Z)$ to denote the mean squared error of a random variable Z, we have

$$\mathrm{MSE}(\hat\theta_{\rm bc}) = \mathrm{MSE}(\hat\theta) - n^{-2}\,(2\,\gamma_0 + \gamma_2^2) + O(n^{-3})\,, \tag{2.19}$$

$$\mathrm{MSE}(\tilde\theta) = \mathrm{MSE}(\hat\theta) + n^{-2}\,(2\,\gamma_0 + 3\gamma_2^2) + O(n^{-3})\,. \tag{2.20}$$

This result makes it clear that $\hat\theta_{\rm bc}$ in a sense "overcorrects" $\hat\theta$ for the effects of bias on mean squared error; it reduces mean squared error by subtracting off twice the squared bias contribution to mean squared error, rather than one lot of squared bias. The origin of the extra component is clear from (2.16): it arises from the correlation between the empirical bias correction and the estimator $\hat\theta$, and to this extent at least that correlation works in favour of improved performance. By way of comparison, (2.20) shows that the empirical adjustment offered by bagging works in the wrong direction.

Of course it is the total correlation, proportional to $\gamma_2^2 + \gamma_4$, between the empirical bias correction and $\hat\theta$ that is important, not just the term γ_2^2. However, we show in Section 5 that when $\hat\theta$ is a maximum likelihood location estimator, γ_0 vanishes. In that case the performance advantages of $\hat\theta_{\rm bc}$ over $\tilde\theta$ are starkly clear from (2.19) and (2.20).

More generally the relationships among mean squared errors of $\hat\theta$, $\hat\theta_{\rm bc}$ and $\tilde\theta$ are demonstrated by (2.19) and (2.20). In particular, in terms of second-order properties of mean squared error, $\hat\theta$ outperforms $\hat\theta_{\rm bc}$, $\hat\theta_{\rm bc}$ outperforms $\tilde\theta$, and $\tilde\theta$ outperforms $\hat\theta$ if and only if $\gamma_0 + \gamma_2^2 < 0$, $\gamma_0 + \frac{5}{4}\,\gamma_2^2 > 0$ and $\gamma_0 + \frac{3}{2}\,\gamma_2^2 < 0$, respectively. This triangle of inequalities allows any one of $\hat\theta$, $\hat\theta_{\rm bc}$ and $\tilde\theta$ to outperform the other two, to second order. To appreciate why, note that if $\gamma_0 + \frac{3}{2}\,\gamma_2^2$ is strictly negative then so too will be $\gamma_0 + \gamma_2^2$, and so the best-performing among $\hat\theta$, $\hat\theta_{\rm bc}$ and $\tilde\theta$ will be $\tilde\theta$. On the other hand, if $\gamma_0 + \frac{3}{2}\,\gamma_2^2$ is strictly positive

then the best-performing estimator will be either $\hat\theta$ or $\hat\theta_{\rm bc}$; it will be $\hat\theta$ if and only if $\gamma_0+\gamma_2^2$ is strictly negative.

2.4. Multivariate case

When θ is p-variate and $p\ge1$, $\psi_0(\cdot|\theta)$ is a p-variate function. For brevity we confine our treatment of this setting to a brief account of the bootstrap-based biased corrected estimator. Other cases are similar.

Formula (2.4) for the bias estimators $\widehat{\rm bias}_{\rm jack}$ and $\widehat{\rm bias}_{\rm boot}$ is applicable as it stands in multivariate settings, and leads directly to multivariate versions of $\tilde\theta_{\rm bc}$ and $\tilde\theta_{\rm pbc}$, defined by (2.5) and (2.6) respectively. The principal result in Section 2 relating $\hat\theta$, $\hat\theta_{\rm bc}$, $\tilde\theta$, $\tilde\theta_{\rm bc}$ and $\tilde\theta_{\rm pbc}$ is (2.9), and it holds in multivariate settings. This leads in turn to the following analogues of (2.11) and (2.12) in multivariate problems: $\tilde\theta_{\rm bc}=\hat\theta_{\rm bc}+O_p(n^{-2})$ and $\tilde\theta_{\rm pbc}=\hat\theta+O_p(n^{-2})$. As a result, formulae (2.13)–(2.15) connecting the biases and variances of estimators $\tilde\theta_{\rm bc}$ and $\hat\theta_{\rm bc}$, and $\tilde\theta_{\rm pbc}$ and $\hat\theta$, hold almost without change; the only alteration required is that "var" should be replaced by "cov", denoting the covariance matrix, in (2.14) and (2.15).

2.5. Parametric bootstrap

For brevity we consider only the univariate case, the multivariate setting is virtually identical. Let $X_1^\star,\dots,X_n^\star$ be a resample drawn by sampling randomly, conditional on $\mathcal{X}$, from the distribution with density $f(\cdot|\hat\theta)$, let $\theta=\hat\theta^\star$ denote the solution of $\sum_i\psi(X_i^\star|\theta)=0$, and write $\theta=\hat\theta_1$ for the solution of $E\{\psi(X_1^\star|\theta)|\mathcal{X}\}=0$.

The quantity γ_j, defined in Section 2.3, is of course a function of the unknown θ. Let $\gamma_j(\hat\theta)$ denote the same function evaluated at $\hat\theta$. It may be shown that (2.7) and (2.8) continue to hold if we replace $\hat\gamma_j$ by $\gamma_j(\hat\theta)$: $\hat\theta_{\rm bc}=2\,\hat\theta-\hat\theta_1\{1+n^{-1}\,\gamma_1(\hat\theta)\}-n^{-1}\,\gamma_2(\hat\theta)+O_p(n^{-2})$, $\tilde\theta=\hat\theta_1\,\{1+n^{-1}\,\gamma_1(\hat\theta)\}+n^{-1}\,\gamma_2(\hat\theta)+O_p(n^{-2})$. It follows that (2.9)–(2.11) remain true if we replace $\hat\theta$ and $\hat\gamma_j$ by $\hat\theta_1$ and $\gamma_j(\hat\theta)$, respectively. Results (2.19) and (2.20) also continue to hold, provided we replace γ_0 there by $\gamma_0+3\gamma_5\mu_1^{-2}$, where

$$\gamma_5=\frac12\,\mu_1^{-1}\Big(\int\frac{\ddot f^2}{f^2}-\int\frac{\ddot f\,\dot f^2}{f^2}\Big)-\frac34\,\rho_2\,\Big(\rho_2+\frac13\int\frac{\ddot f\,\dot f}{f}\Big)\,.$$

Therefore, the discussion of (2.19) and (2.20), immediately below those formulae, applies equally to the parametric case.

3. Examples from Maximum Likelihood Estimation

Let $\ddot f$ and $f^{[3]}$ denote the second and third derivatives, respectively, of f with respect to θ, evaluated at θ_0. We prove in Section 5 that

$$\gamma_2\Big(\int\frac{\dot f^2}{f}\Big)^2=-\frac12\int\frac{\dot f\,\ddot f}{f}\,, \tag{3.1}$$

$$\gamma_0\Big(\int \frac{\dot f^2}{f}\Big)^4 = \frac{1}{2}\Big(\int \frac{\dot f^2}{f}\Big)\int\Big(\frac{\dot f^2 \ddot f}{f^2} - \frac{\dot f f^{[3]}}{f} - \frac{\ddot f^2}{f}\Big) + 2\Big(\int \frac{\dot f \ddot f}{f}\Big)^2 - \Big(\int \frac{\dot f \ddot f}{f}\Big)\Big(\int \frac{\dot f^3}{f^2}\Big). \quad (3.2)$$

To appreciate the implications of these formulae, consider the location estimation problem where $f(x|\theta) = f(x-\theta)$ and f is a smooth density. Then

$$\int \frac{f'f''}{f} = \frac{1}{2}\int \frac{(f')^3}{f^2}, \qquad \int \frac{(f')^2 f''}{f^2} = \frac{2}{3}\int \frac{(f')^4}{f^3}, \qquad \int \frac{f'f'''}{f} = \frac{2}{3}\int \frac{(f')^4}{f^3} - \int \frac{(f'')^2}{f}.$$

It follows from the latter formulae and (3.2) that $\gamma_0 = 0$, and that γ_2 vanishes if and only $\tau \equiv \int\{(f')^3/f\} = 0$. Therefore, by (2.17) and (2.18), $\hat\theta_{\rm bc}$ strictly outperforms $\tilde\theta$, to second order in terms of mean squared error, unless $\tau = 0$; and $\hat\theta_{\rm bc}$ is never inferior to either $\tilde\theta$ or $\hat\theta$, to second order.

This result does not extend to other related problems. Consider for example the problem of location estimation in one component of a generalised mixture, where $f(x|\theta) = p\,g(x-\theta) + (1-p)\,h(x)$, $p \neq 0$ is known and satisfies $-\infty < p < \infty$, and g and h are known densities. There is no difficulty in taking p small and negative provided the resulting f is nonnegative for values of θ in a neighbourhood of θ_0. Examples are easily found, for example by taking h and g to be rescaled Student's t densities with equal degrees of freedom. Assume for simplicity that g and h are both symmetric about 0, and $\theta_0 = 0$, in which case $\gamma_2 = 0$. Again for simplicity, suppose h is close to g. Then if we evaluate γ_0 for $g = h$, the error we commit can be made arbitrarily small by sufficiently reducing the distance between g and h. Making this approximation we obtain:

$$\gamma_0 = \tfrac{2}{3}\,p^{-1}\,(1-p)\,\Big(\int \frac{(f')^2}{f}\Big)^{-3}\int \frac{(f')^4}{f^3}\,, \quad (3.3)$$

which has the same sign as p.

It follows from (2.17), (2.18) and (3.3) that by choosing $p > 0$ we ensure that $\hat\theta_{\rm bc}$ outperforms $\tilde\theta$, to second order in terms of mean squared error; and that the position is reversed if $p < 0$. Therefore either the standard bagged estimator, or the conventional unbagged bias-corrected estimator, can outperform the other to second order.

4. Simulation Results

We have found the theoretical properties reported above to be accurately reflected in numerical work. Specifically, we generated data from two models, one an exponential distribution with mean θ^{-1} and the other a Binomial distribution with mean $e^{-\theta}$. In neither case was θ the mean of the distribution. We computed $\hat\theta$ using maximum likelihood, employed 500 bootstrap simulations to calculate $\hat\theta_{\rm bc}$, and used 500 and 200 simulations at the first and second levels, respectively, to calculate $\tilde\theta_{\rm bc}$. Mean squared errors were calculated using 1000 replications, and sample sizes ranged from 40 to 200.

The simulations showed that $\hat{\theta}_{\rm bc}$ and $\tilde{\theta}_{\rm bc}$ had almost identical performance, as implied by (2.11). For each sample size and parameter value, the averages of the simulated values of $\hat{\theta}_{\rm bc}$ and $\tilde{\theta}$ were almost equidistant from $\hat{\theta}$ but on opposite sides of $\hat{\theta}$, as predicted by (2.9). Bagging increased both variance and mean square error, and the bias-corrected estimator had lowest bias, variance and mean square error. Figure 1 shows typical results.

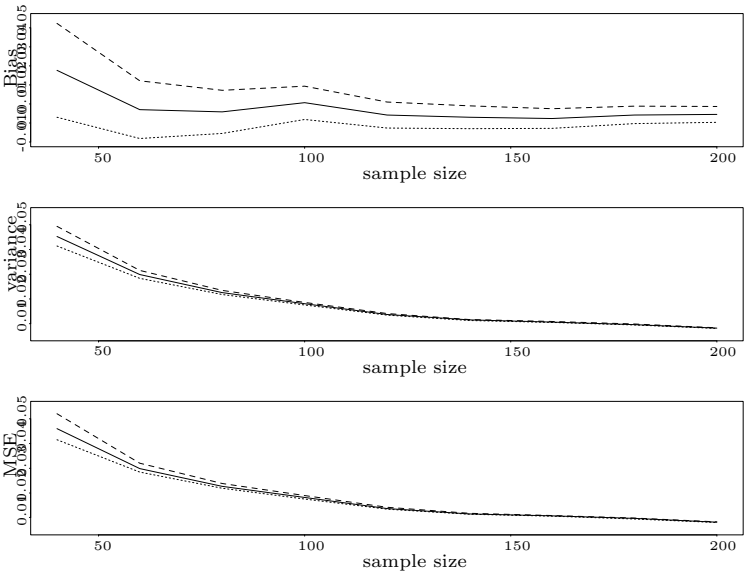

Figure 1. *Bias, variance and mean squared error in Binomial case.* The problem illustrated is that of estimating θ using data from a Binomial $(n, e^{-\theta})$ distribution with $\theta = 1$. The value of sample size, n, is graphed along the horizontal axis. Results for $\hat{\theta}$, for the bias-corrected estimated $\hat{\theta}_{\rm bc}$, and for the bagged estimator $\tilde{\theta}$, are represented by solid, dotted and dashed lines, respectively.

5. Technical Arguments

5.1. Derivation of (2.8)

Let $\hat{\rho}_j^* = S_j^*/S_1^*$. In this notation, by analogy with (2.2),

$$\hat{\theta}^* = -\{\hat{\rho}_0^* + \tfrac{1}{2}\,(\hat{\rho}_0^*)^2\,\hat{\rho}_2^*\} + (\hat{\rho}_0^*)^3\,\{\tfrac{1}{6}\,\hat{\rho}_3^* - \tfrac{1}{2}\,(\hat{\rho}_2^*)^2\} + \cdots\,. \tag{5.1}$$

Put $\Delta_j^* = (S_j^* - S_j)/S_1$ and note that

$$\hat{\rho}_j^* = (\hat{\rho}_j + \Delta_j^*)/(1 + \Delta_1^*) = \hat{\rho}_j + \Delta_j^* - \hat{\rho}_j\,\Delta_1^* + \hat{\rho}_j\,(\Delta_1^*)^2 - \Delta_j^*\,\Delta_1^* + \cdots\,. \tag{5.2}$$

Substituting into (5.1), Taylor expanding, taking conditional expectation, and noting that $E\{(\Delta_j^*)^2|\mathcal{X}\} = n^{-1}\,\hat\sigma_j^2$ and $E(\Delta_0^*\Delta_j^*|\mathcal{X}) = n^{-1}\,\hat\alpha_j$, we deduce that

$$\begin{aligned}
E(\hat\rho_0^*|\mathcal{X}) &= \hat\rho_0\,(1 + n^{-1}\,\hat\sigma_1^2) - n^{-1}\,\hat\alpha_1 + O_p(n^{-2})\,,\\
E\{(\hat\rho_0^*)^2\hat\rho_2^*|\mathcal{X}\} &= \hat\rho_0^2\hat\rho_2 + n^{-1}\,\{\hat\rho_2\,\hat\sigma_0^2 + 2\,\hat\rho_0\,(\hat\alpha_2 - 3\,\hat\rho_2\,\hat\alpha_1)\} + O_p(n^{-2})\,,\\
E\{(\hat\rho_0^*)^3\hat\rho_3^*|\mathcal{X}\} &= \hat\rho_0^3\hat\rho_3 + 3\,n^{-1}\,\hat\rho_0\,\hat\rho_3\,\hat\sigma_0^2 + O_p(n^{-2})\,,\\
E\{(\hat\rho_0^*)^3\,(\hat\rho_2^*)^2|\mathcal{X}\} &= \hat\rho_0^3\hat\rho_2^2 + 3\,n^{-1}\,\hat\rho_0\,\hat\rho_2^2\,\hat\sigma_0^2 + O_p(n^{-2})\,.
\end{aligned} \tag{5.3}$$

Therefore,

$$\begin{aligned}
\tilde\theta \equiv E(\hat\theta^*|\mathcal{X}) = \hat\theta + n^{-1}\,\Big[&\hat\alpha_1 - \hat\rho_0\,\hat\sigma_1^2 - \tfrac{1}{2}\,\{\hat\rho_2\,\hat\sigma_0^2 + 2\,\hat\rho_0\,(\hat\alpha_2 - 3\,\hat\rho_2\,\hat\alpha_1)\}\\
&+\tfrac{1}{2}\,\hat\rho_0\,(\hat\rho_3 - 3\,\hat\rho_2^2)\,\hat\sigma_0^2\Big] + O_p(n^{-2})\,,
\end{aligned} \tag{5.4}$$

which is equivalent to (2.8).

5.2. Derivation of (2.7)

Equation (2.8) implies the following result for $\widehat{\text{bias}}_{\text{boot}}$, and the analogue for $\widehat{\text{bias}}_{\text{jack}}$ will be derived momentarily:

$$\text{both}\quad \widehat{\text{bias}}_{\text{jack}}\quad\text{and}\quad \widehat{\text{bias}}_{\text{boot}}\quad\text{equal}\quad n^{-1}\,(\hat\theta\,\hat\gamma_1 + \hat\gamma_2) + O_p(n^{-2})\,. \tag{5.5}$$

Result (2.7) follows from (5.5) and the definitions of $\hat\theta_{\text{jack}}$ and $\hat\theta_{\text{boot}}$. Let $\hat\rho_{ji}$ denote the version of $\hat\rho_j$ that arises if we compute the latter from the $(n-1)$-sample $\mathcal{X}\backslash\{X_i\}$ instead of $\mathcal{X}$, and put $\delta_{ji} = (n-1)^{-1}\{\psi_j(X_i) - S_j\}$. Then the analogue of (5.2) is:

$$\hat\rho_{ji} = (\hat\rho_j - \delta_{ji})/(1 - \delta_{j1}) = \hat\rho_j - \delta_{ji} + \hat\rho_j\,\delta_{j1} + \hat\rho_j\,\delta_{j1}^2 - \delta_{ji}\,\delta_{j1} + \cdots\,.$$

Using this analogy we may show that

$$E\{(\hat\rho_0^*)^\ell\,(\hat\rho_2^*)^m|\mathcal{X}\} - \hat\rho_0^\ell\hat\rho_2^m = \sum_{i=1}^n \hat\rho_{0i}^\ell\hat\rho_{2i}^m - n\,\hat\rho_0^\ell\hat\rho_2^m + O_p(n^{-2})$$

for $\ell \ge 1$ and $m \ge 0$. Therefore results (5.3) continue to hold if on the left-hand side each term $E\{(\hat\rho_0^*)^\ell\,(\hat\rho_2^*)^m|\mathcal{X}\}$ is replaced by $\sum_i \hat\rho_{0i}^\ell\hat\rho_{2i}^m - (n-1)\,\hat\rho_0^\ell\hat\rho_2^m$. Result (5.5), in the jackknife case, follows.

5.3. Derivation of (2.10)

The double-bootstrap analogue of (5.4), in which conditioning is on $\mathcal{X}^*$, can be obtained by replacing $E(\hat\theta^*|\mathcal{X})$ on the left-hand side by $E(\hat\theta^{**}|\mathcal{X}^*)$, and placing asterisks on terms on the right-hand side. Taking expectation of this expansion,

conditional on $\mathcal{X}$, we deduce that $E(\hat\theta^{**}|\mathcal{X}) = E(\hat\theta^*|\mathcal{X}) + n^{-1}\,(\hat\theta\,\hat\gamma_1 + \hat\gamma_2) + O_p(n^{-2})$. Therefore,

$$\tilde\theta_{\rm boot} = 3\,\tilde\theta - 2\,\{E(\hat\theta^{**}|\mathcal{X}) - E(\hat\theta^*|\mathcal{X})\} = \tilde\theta - 2\,n^{-1}\,(\hat\theta\,\hat\gamma_1 + \hat\gamma_2) + O_p(n^{-2})\,.$$

This is equivalent to (2.10) in the case $\tilde\theta_{\rm bc} = \tilde\theta_{\rm boot}$. Similarly we derive the version of (2.10) for $\tilde\theta_{\rm bc} = \tilde\theta_{\rm jack}$, and likewise we establish (2.12).

5.4. Derivation of (2.16)

Given a random variable Z write $(1-E)\,Z$ for $Z - E(Z)$, put $A_j = \mu_1^{-2}\,(1-E)\,n^{-1}\,\sum_i\,\psi_0(X_i)\,\psi_j(X_i)$ and $\Delta_j = (S_j - \mu_j)/\mu_1$, and note that $\hat\gamma_2 = \gamma_2 + \Delta$ where $\Delta = (\alpha_1 - 3\,\gamma_2)\,\Delta_1 + A_1 - \frac12\,\rho_2\,A_0 - \frac12\,\sigma_0^2\,\Delta_2 - \Delta_0 + O_p(n^{-1})$. Observe too that, assuming $\theta_0 = 0$, we have $\hat\theta = -\Delta_0 + \Delta_0\,\Delta_1 - \frac12\,\rho_2\,\Delta_0^2 + O_p(n^{-3/2})$. Thus we obtain (2.16):

$$\begin{aligned} E(\hat\theta\,\hat\gamma_2) &= E(\hat\theta)\,\gamma_2 + E\{(\hat\theta - E\hat\theta)\,(\hat\gamma_2 - E\hat\gamma_2)\} + O(n^{-2}) \\ &= n^{-1}\,\gamma_2^2 - E(\Delta\,\Delta_0) + O(n^{-2}) = n^{-1}\,(\gamma_2^2 + \gamma_4) + O(n^{-2})\,. \end{aligned}$$

5.5. Derivation of (3.1) and (3.2)

It may be proved that

$$\begin{aligned} \mu_1^2\,\gamma_2 &= -\frac12\int\frac{\dot f\,\ddot f}{f}\,, \quad \mu_1^2\,\gamma_3 = \int\Big(\frac{\dot f^2\,\ddot f}{f^2} - \frac{\dot f f^{[3]}}{f} - \tfrac12\,\frac{\ddot f^2}{f}\Big) - \Big(\int\frac{\dot f^2}{f}\Big)^2\,, \\ \mu_1\,\rho_2 &= 2\int\frac{\dot f^3}{f^2} - 3\int\frac{\dot f\,\ddot f}{f}\,, \quad \mu_1^3\,\rho_2\,\gamma_2 = \Big(\int\frac{\dot f\,\ddot f}{f}\Big)\int\Big(\frac32\,\frac{\dot f\,\ddot f}{f} - \frac{\dot f^3}{f^2}\Big)\,. \end{aligned}$$

The first result above implies (3.1). To obtain (3.2) note that since $\gamma_1 = \gamma_3 - 3\,\rho_2\,\gamma_2$ and $|\mu_1| = \int(\dot f^2/f)$, then

$$\begin{aligned} |\mu_1|^3\,\gamma_1 &= \Big(\int\frac{\dot f^2}{f}\Big)\int\Big(\frac{\dot f^2\,\ddot f}{f^2} - \frac{\dot f f^{[3]}}{f} - \frac12\,\frac{\ddot f^2}{f}\Big) + 3\Big(\int\frac{\dot f\,\ddot f}{f}\Big)\int\Big(\frac32\,\frac{\dot f\,\ddot f}{f} - \frac{\dot f^3}{f^2}\Big) - \Big(\int\frac{\dot f^2}{f}\Big)^3\,, \\ \mu_1^3\,\beta_1 &= \int\frac{\dot f^2\,\ddot f}{f^2} - \int\frac{\dot f^4}{f^3}\,, \quad \mu_1^2\,\alpha_2 = \int\frac{\dot f f^{[3]}}{f} - 3\int\frac{\dot f^2\,\ddot f}{f^2} - \int\frac{\dot f^4}{f^3}\,. \end{aligned}$$

Formula (3.2) follows from these results and the relation

$$\mu_1^4\,\gamma_0 = |\mu_1|^3\,\{\gamma_1 - \sigma_0^{-2}\,(\beta_1 - \tfrac12\,\sigma_0^2\,\alpha_2 - \sigma_0^2)\} + \mu_1^4\,\{\alpha_1\,(3\,\gamma_2 - \alpha_1) + \tfrac12\,\rho_2\beta_0\}\,.$$

References

Breiman, L. (1996). Bagging predictors. *Mach. Learning* **24**, 123-140.

Breiman, L. (1999). Using adaptive bagging to debias regressions. Technical Report No. 547, Department of Statistics, University of California, Berkeley.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.* **30**, 927-961.
Buja, A. and Stuetzle, W. (2000a). Bagging does not always decrease mean square error. Preprint.
Buja, A. and Stuetzle, W. (2000b). Smoothing effects of bagging. Preprint.
Friedman, J. H. and Hall, P. (2000). On bagging and nonlinear estimation. Preprint.

Department of Statistics and Applied Probability, National University of Singapore, 117543, Singapore.

E-mail: stacsx@nus.edu.sg

Centre for Mathematics and Its Applications, Australian National University, Canberra 0200, Australia.

E-mail: peter.hall@anu.edu.au

(Received August 2001; accepted July 2002)