Extrapolation of subsampling distribution
estimators: the i.i.d. and strong mixing cases

Patrice BERTAIL and Dimitris N. POLITIS

Key words and phrases: Bootstrap; confidence limits; extrapolation; interpolation; large sample inference; subsampling.

MSC 2000: Primary 62M20; secondary 62G05.

Abstract: Politis & Romano (1994) proposed a general subsampling methodology for the construction of large-sample confidence regions for an arbitrary parameter under minimal conditions. Nevertheless, the subsampling distribution estimators may sometimes be inefficient (in the case of the sample mean of i.i.d. data, for instance) as compared to alternative estimators such as the bootstrap and/or the asymptotic normal distribution (with estimated variance). The authors investigate here the extent to which the performance of subsampling distribution estimators can be improved by interpolation and extrapolation techniques, while at the same time retaining the robustness property of consistent distribution estimation even in nonregular cases; both i.i.d. and weakly dependent (mixing) observations are considered.

1. INTRODUCTION

Let \( \{X_1, \ldots, X_n\} \) be an observed stretch of a (strictly) stationary, strong mixing sequence of real-valued random variables \( \{X_t, t \in \mathbb{Z}\} \); the probability measure generating the observations is denoted by \( \mathbb{P} \). The strong mixing condition means that the sequence \( \alpha_X(k) = \sup_{A,B} |P(A \cap B) - P(A)P(B)| \) tends to zero as \( k \) tends to infinity, where \( A \) and \( B \) are events in the \( \sigma \)-algebras generated by \( \{X_t, t < 0\} \) and \( \{X_t, t \geq k\} \) respectively; the case where \( X_1, \ldots, X_n \) are independent, identically distributed (i.i.d.) is an important special case where \( \alpha_X(k) = 0 \) for all \( k > 0 \).

In Politis & Romano (1994), a general subsampling methodology has been put forth for the construction of large-sample confidence regions based on statistics \( T_n = T_n(\mathbf{X}_n) \) estimating a general unknown parameter \( \theta = \theta(P) \), under very minimal conditions. In the case of stationary data (time series or random fields), subsampling is closely related to the blocking methods introduced by Hall (1985), Carlstein (1986), Künsch (1989), and Liu & Singh (1992); see Shao & Tu (1995, ch 9), as well as Wu (1990), Sherman (1992) and Sherman & Carlstein (1994) for related ideas.

Let us now make the simplifying assumption that \( T_n \) and \( \theta \) are real-valued. To obtain asymptotically pivotal (or at least, scale-free) statistics, a standardization (also known as "studentization" when the norming is data-based and random) is often required. Since we will later discuss the influence of the studentization, we introduce a statistic \( S_n = S_n(\mathbf{X}_n) > 0 \) converging in probability to some constant \( \sigma > 0 \); heuristically, \( \sigma^2 \) may stand for the asymptotic variance of \( \tau_n(T_n - \theta) \), but this is not necessarily always the case. Without loss of generality, the unstuden-
In the present paper, we explore the asymptotic performance of extrapolation to seek the desired improvement. More specifically, we show that Richardson extrapolation considered by Bickel & Yahav (1988), Bertail (1997) and, in the specialized case, corresponds to $S_n = 1$.

Although i.i.d. data can be seen as a special case of stationary strong mixing data, the construction of the subsampling distribution can take advantage of the i.i.d. structure when such a structure exists:

(i) General case (strong mixing data). Define $Y_i$ to be the subsequence $(X_{i1}, \ldots, X_{i+b-1})$, for $i = 1, \ldots, q$, and $q = n - b + 1$; note that $Y_i$ consists of $b$ consecutive observations from the $X_1, \ldots, X_n$ sequence, and the order of the observations is preserved.

(ii) Special case (i.i.d. data). Let $Y_1, \ldots, Y_q$ be equal to the $q = \binom{n}{b}$ subsets of size $b$ chosen from $\{X_1, \ldots, X_n\}$, and then ordered in any fashion; here the subsets $Y_i$ consist of unordered observations.

In either case, let $T_{b,i}$ and $S_{b,i}$ be the values of statistics $T_b$ and $S_b$ calculated from just subsample $Y_i$. The subsampling distribution of the root $\tau_n S_n^{-1} (T_n - \theta)$, based on a subsample size $b$, is defined by

$$K_b(x) \equiv \frac{1}{q} \sum_{i=1}^{q} \mathbb{1}\{\tau_{b,i} S_{b,i}^{-1} (T_{b,i} - T_n) \leq x\}.$$ 

If there is a non-degenerate distribution $K(x, P)$, continuous in $x$, such that

$$K_n(x, P) \equiv \mathbb{P}_P \{\tau_n S_n^{-1} (T_n - \theta) \leq x\} \to K(x, P)$$

as $n \to \infty$, for any real number $x$, the subsampling methodology was shown to 'work' asymptotically provided that the integer "subsample size" $b$ satisfies $b \to \infty$ and as $n \to \infty$, $\max(b/n, \tau_b/\tau_n) \to 0$.

The subsampling distribution turns out to be a relatively low-accuracy approximation to the true sampling distribution $K_n(x, P)$ and is actually worse than the asymptotic normal distribution (with estimated variance). Indeed it was proved by Bertail (1997) that the subsampling distribution admits, for suitable $b$, the same Edgeworth expansion as $K_n(x, P)$—when such an expansion exists—but in powers of $b$ instead of $n$. This result has a straightforward consequence when there exists a standardization $S_b$ such that the asymptotic distribution is pivotal and known, i.e., if $K(x, P) = K(x)$ does not depend on $P$. If the rate of the first term in the Edgeworth expansion $f_1(n)$ is known (typically $f_1(n) = \sqrt{n}$ in the regular case) then it is possible to improve the subsampling distribution by considering a linear combination of that distribution with the asymptotic distribution

$$K_n^{\text{inf}}(x) = \left\{1 - \frac{f_1(b)}{f_1(n)}\right\} K(x) + \frac{f_1(b)}{f_1(n)} K_b(x).$$

This type of linear (convex) combination with positive coefficients may be seen as an interpolation in that $K_n^{\text{inf}}(x)$ is an intermediate point on the straight line segment joining $K_b(x)$ to the asymptotic $K(x)$, in the same way that sample size $n$ is an intermediate point between sample sizes $b$ and $\infty$; note the ordering $b < n < \infty$ and recall that we are interested in obtaining an estimate of the ordinate (sampling distribution) at sample size $n$ (based on the ordinates at sample sizes $b$ and $\infty$). This interpolation idea was first considered in Bickel & Yahav (1988) and generalized in Bertail (1997).

Nevertheless the generality of the subsampling methodology lies in the fact that $K(x, P)$ does not have to be known in order for subsampling to work. Therefore, it is of interest to seek improvements upon the subsampling distribution estimators that do not explicitly involve $K(x, P)$.

In the present paper, we explore the asymptotic performance of extrapolation similar to the notion of Richardson extrapolation considered by Bickel & Yahav (1988), Bertail (1997) and Bickel, Götze & van Zwet (1997) to seek the desired improvement. More specifically, we show
that the extrapolation of two undersampling distribution \( K_{b_1}(x) \) and \( K_{b_2}(x) \) can be used to provide us with the linear combination effecting the aforementioned extrapolation of subsampling distribution estimators, and we quantify the improvement achieved by the extrapolation.

In Section 2, we focus on the i.i.d. case and show that since in this case subsampling amounts to sampling without replacement from a finite population, the finite population correction \( 1 - f \), with \( f = b/n \), should necessarily be taken into account to build an accurate approximation of the true distribution. Then extrapolation of subsampling distributions for the sample mean or non degenerate U-V statistics of i.i.d. data achieves second order accuracy.

The strong mixing case studied in Section 3 is more complicated because of inaccurate variance estimation. Thus, we include an Appendix where a simple variance estimator is proposed based on the ideas of Politis & Romano (1995) with the property of being almost \( \sqrt{n} \) consistent; this accurate variance estimator can then be used in the construction of confidence intervals for the sample mean of strong mixing observations achieves second order accuracy; this finding extends the i.i.d. result of Booth & Hall (1993). In particular, this finding applies to many econometric models with stationary data. Section 4 presents some finite-sample simulations in the context of an ARMA model.

2. I.I.D. DATA

2.1. Finite population correction.

Bertail (1997) noticed that subsampling may be seen as a particular case of the weighted bootstrap considered in Barbe & Bertail (1995) with some exchangeable weights \( W_n = (w_{i,n})_{1 \leq i \leq n} \). The proper normalizing factor for these weights is none other than the finite population correction factor \( 1 - f \) with \( f = b/n \), a result foreshadowed by Shao & Wu (1989) in the case of variance estimation (see also Booth & Hall 1993). This suggests that the adequate renormalization factor in the subsampling distribution is \( \tau_r \) (instead of \( \tau_b \)), where \( r \) is defined by \( r = b/(1 - f) \). This also leads to a more general definition of the corrected subsampling distribution as

\[
\tilde{K}_b(x) = \frac{1}{q} \sum_{i=1}^{q} \{ \tau_r, S_{b,i}^{-1}(T_{b,i} - T_{n}) \leq x \}.
\]

Clearly, the factor \( 1 - f \) has no first-order asymptotic effect on the subsampling which remains consistent under very weak assumptions provided that \( f \to 0 \). However, the factor \( 1 - f \) is of great importance for second-order properties as shown in the sequel.

2.2. The studentized sample mean.

Consider the problem of estimating the mean \( \theta(P) = E_P X_1 \). In the following we assume that \( E_P X_1^4 < \infty \), and we take

\[
T_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]
as usual. From Bhattacharya & Ghosh (1978), under the usual Cramér condition, we have the Edgeworth expansion

\[
K_n(x, P) = P \left[ n^{1/2} S_n^{-1} \{ \bar{X}_n - \theta(P) \} \leq x \right] = \Phi(x) + n^{-1/2} p_1(x, P) \phi(x) + O(n^{-1}) \quad \text{(1)}
\]
with

\[
p_1(x, P) = \frac{k_3}{6} (2x^3 + 1),
\]
where \( k_3 \) is the skewness.
Following Booth & Hall (1993), under the Cramér condition and assuming that $E|X_i|^6 < \infty$, for some $\eta > 0$ we have from Babu & Singh (1985) an Edgeworth expansion for sampling without replacement from a finite population, with $b/n \to 0$ as $n \to \infty$; so for any $\varepsilon > 0$,

$$\tilde{K}_b(x) = \Phi(x) + b^{-1/2} p_1(x, P) \phi(x) + \frac{1}{4} k_3 \phi(x) + O_P(b^{-1/2} n^{-1/2+\varepsilon} + b^{3/2} n^{-2}) + o(b^{-1}),$$

(2)

where

$$p_2(x, P) = \frac{1}{12} k_4(x^3 - x) - \frac{1}{18} k_3(x^5 + 2x^3 - 3x) - \frac{1}{4} (x^3 + 3x),$$

and $k_4$ is the kurtosis. Notice first that for $b$ such that $b/n \to 0$, the right-hand side of (2) cannot be made smaller than $O_P(n^{-1/2})$. Thus even with the finite population correction, the subsampling distribution will not be second-order correct.

The extrapolation of two subsampling distributions (one with subsample size $b_1$ and the other with $b_2$) is given by

$$\tilde{K}_b^{(2)}(x) = \lambda_1 \tilde{K}_{b_1}(x) + \lambda_2 \tilde{K}_{b_2}(x),$$

(3)

where $\lambda_1$ and $\lambda_2$ are chosen to solve

$$\lambda_1 + \lambda_2 = 1 \quad \text{and} \quad \lambda_1 b_1^{-1/2} + \lambda_2 b_2^{-1/2} = n^{-1/2}$$

in order to match the second order term in (1) and get a second order valid approximation. We then get

**Proposition 1.** Let $b_1$ and $b_2$ be such that $b_i/n \to 0$ for $i = 1, 2$ and $b_2/b_1 \to C \in [0, 1]$ when $n \to \infty$. Then the extrapolation of the two finite-population-corrected subsampling distributions of the studentized mean is actually second order correct with the best choice given by $b_1 = C_1 n^{2/3}$ and $b_2 = C_2 n^{2/3}$ with $C_2 < C_1$. We also have

$$\sup_x |\tilde{K}_b^{(2)}(x) - K_n(x, P)| = O_P(n^{-2/3}).$$

**Proof.** From equation (2) and the definition of the $b_i$'s, we get

$$\tilde{K}_b^{(2)}(x) - K_n(x, P) = -b_1^{-1/2} b_2^{-1/2} p_2(x, P) \phi(x) + b_1^{-1/2} (1 + C^{1/2} n^{-1} + \frac{1}{4} k_3 \phi(x) + O_P(b_1^{-1/2} n^{-1/2+\varepsilon} + b_1^{3/2} n^{-2}) + o(b_1^{1/2} n^{-1}).$$

(4)

Minimizing the order of the right-hand side of (4) leads us to choose $b_1$ and $b_2$ such that $b_1 b_2^{1/2}$ is proportional to $n$. This yields the result.

**Remark 1.** Note that the order of the whole approximation is worse than that which is obtained by considering the interpolation of one subsampling distribution with its asymptotic distribution when the latter is known; see Booth & Hall (1993) and Bertail (1997), who obtain an error of size $n^{-5/6}$.

**Remark 2.** It is important to point out that if we do not take into account the finite population correction factor, then the second order validity of the extrapolated version of the two distributions fails; in the case of interpolation, the second order property still holds but with a loss in term of coverage probability; see, e.g., Bertail (1997) and Bickel, Götze & van Zwet (1997, p. 17) in which the importance of this correction factor was recognized but not really exploited. Indeed in
the absence of the finite population correction, if we let $K_n^{(2)}(x)$ be the extrapolation of $K_{b_1}(x)$ and $K_{b_2}(x)$, we have

$$K_n^{(2)}(x) = \Phi(x) + n^{-1/2}p_1(x, P)\phi(x) + O_P\left(\frac{b_1}{n}\right) + O_P\left(b_1^{3/2}n^{-1/2+\varepsilon} + b_2^{1/2}n^{-2/2+\varepsilon} + b_1^{-1/2}b_2^{-1/2} + b_1^{1/2}n^{-1}\right).$$

Once again, to obtain the second order correctness we would have to choose $n^{1/2} = o(b_1)$ and the loss induced by the sampling-without-replacement scheme (typically of order $b_1/n$) implies that second order correctness cannot be attained.

2.3. General extrapolation result in the i.i.d. case.

Of course, in the case of the mean, interpolation and/or extrapolation can be thought to give no advantage since we already know that the usual bootstrap (Efron 1979) gives an approximation up to $O_P(n^{-1})$; see Hall (1992). It is quite obvious that the same holds for smooth functions of means. Even though the usual bootstrap works in these mean-like cases, subsampling may still be useful for computational reasons, since constructing two subsampling distributions requires less simulation than constructing the usual bootstrap distribution. Moreover, in contrast to the interpolation schemes studied in Booth & Hall (1993) and Bertail (1997), the extrapolation does not depend on the asymptotic approximation and is thus more robust. Indeed, if the original assumptions break down, e.g., the assumptions leading to the Edgeworth expansion, or even if the statistic $T_n$ is not asymptotically Gaussian in which case interpolation is not applicable and the bootstrap fails, the extrapolation of the (corrected or uncorrected) subsampling distributions remains a consistent distribution estimator.

We conjecture, however, that in a great number of situations, extrapolation together with a finite population correction will yield second order accuracy, provided that one takes $\tau_n$ instead of $\tau_{n}^\ast$ in the definition of $K_b$ for $b$ such that $f^{-1}_b(h) = o\{f_1(n)\}^{-1}$ by analogy to our Sections 2.1 and 2.2. Using recent Edgeworth expansions results for finite population as proposed by Bloznelis & Götze (2000), it is easy to see that this conjecture holds for U-statistics of degree 2 with non-degenerate first gradient (influence function) and as a consequence it holds for any smooth statistical functional, differentiable according to some nice metric (see Barbe & Bertail 1995). In that case, $\tau_n = \sqrt{n}$ and $\tau_n = \sqrt{b/(1-f)}$ is the adequate normalization which makes the extrapolation second order correct. In any case, even if second order accuracy is not achieved, the extrapolated distribution will always improve over each individual subsampling distribution, i.e., an extrapolated distribution will always improve over a single distribution.

2.4. When the convergence rate to the asymptotic approximation is unknown.

A case of interest in some practical applications occurs when the order of the difference between the asymptotic and the true distribution is unknown. Indeed the knowledge of the rate $f_1$ of the asymptotic approximation is implicit in the construction of both the interpolation and extrapolation. Consider for instance the simple case of estimating the mean. If $E_P(X_i - E_PX_i)^3 \neq 0$, then (3) yields a second order correct approximation and improves over the asymptotic. But if $E_P(X_i - E_PX_i)^3 = 0$, then the asymptotic distribution is already second order correct and (3) is less accurate; this follows from the fact that when the skewness is zero, the correct extrapolation (which will indeed improve upon the asymptotic distribution) should be built with $f_1(n) = n$ and not $\sqrt{n}$. Other problematic situations occur when the distribution of the $X_i$ is lattice. These examples suggest that we either have to make a preliminary test on some parameter appearing in the Edgeworth expansion (depending on the statistic and the underlying distribution), or we have to directly construct and employ an accurate estimator of $f_1$, which may then be used in forming the extrapolation. The first suggestion is unsatisfactory, however, because it is highly problem-dependent, and the Edgeworth expansion may be quite complicated.

Under very general conditions on the statistic and under some conditions on $b$, the order of the subsampling distribution is $f_1(b)$. Thus if we study a collection of subsampling distributions,
when \( b \) varies in its domain, we should be able to observe their convergence to the asymptotic distribution in connection with \( f_1 \). When \( f_1(n) = n^{-\alpha} \), where \( \alpha \) is unknown, the following proposition shows that it is possible to estimate the accuracy rate \( f_1 \) by a simple regression.

**Proposition 2.** Let \( b^{(i)}_1, b^{(i)}_2 \) for \( i = 1, \ldots, I \) (for simplicity \( b^{(i)}_2 = b_2 \) may be chosen to be the same for all \( i \)) be several pairs of different subsampling sizes, satisfying the assumptions of Bertail (1997) A1–A5, with \( b^{(i)}_1 = n^{\beta_1}, 1/2 > \beta_1 > \cdots > \beta_I \), such that \( b^{(i)}_1/b^{(i)}_2 \to 0 \) when \( n \to \infty \). Assume in addition that \( f_1(n) = n^{-\alpha} \), where \( \alpha \) is unknown. Then we have, uniformly in \( x \),

\[
\log(|K_{b^{(i)}_1}(x) - K_{b^{(i)}_2}(x)|) = -\alpha \log(b^{(i)}_1) + \log(|p(x, P)|) + o_P(1), \quad i = 1, \ldots, I.
\]

(5)

Let \( \hat{\alpha} \) be the least square estimator obtained by regressing \( \log(|K_{b^{(i)}_1}(x) - K_{b^{(i)}_2}(x)|) \) on \( -\log b^{(i)}_1 \). Then we have

\[
\hat{\alpha} = \alpha + o_P\{\log(n)^{-1}\}.
\]

(6)

As a consequence, interpolation and extrapolation of the finite-population-corrected subsampling distributions with estimated rate \( f_1(n) = n^{-\hat{\alpha}} \) are second-order correct.

**Proof.** Under our assumptions, \( K_b(x) \) has the same Edgeworth expansion as \( K_n(x, P) \) but on functions of \( b \) instead of \( n \). If we choose \( b_1 \) and \( b_2 \) such that \( b_1/b_2 \to 0 \) when \( n \to \infty \), then it is easy to see that both

\[
|K_{b_1}(x) - K_{b_2}(x)| = f_1(b_1)^{-1} \left\{ 1 + o(1) \right\} |p(x, P)|;
\]

(7)

and (5) follow when we take the logarithm. Now, since

\[
\sum_{i=1}^I \left\{ \log(b^{(i)}_1) - I^{-1} \sum_{i=1}^I \log(b^{(i)}_1) \right\}^2 = C_0(\log n)^2
\]

for some constant \( C_0 \), we thus obtain (6). Finally, it is easy to see that if we now use \( \hat{f}_1(n) = f_1(n) \left\{ 1 + o_P(1) \right\} \) in place of \( f_1(n) \), then the extrapolation and the interpolation remain second-order correct.

We observe that our idea here differs from but is in the same spirit as results of Bertail, Politis & Romano (1999), who were trying to estimate the rate of the statistics \( T_n \) itself. In the more general case, when the functional form of \( f_1 \) is unknown, we may consider (7), or rather, its logarithm, as a nonparametric regression for \( f_1 \). Under a monotonicity constraint and the assumption that \( f_1(b) \to \infty \) as \( b \to \infty \), consistent estimation of \( f_1 \) may still be possible, albeit more complicated and slower. We will not pursue this approach here.

### 3. Strong Mixing Data

The case of strong mixing data is complicated by the fact that an adequate standardization is needed; see Götze & Künsch (1996). In Hall & Jing (1996), interpolation was used in the context of stationary data. However, their results are weakened by the fact that their hypothesis on the Edgeworth expansion with a remainder of size \( O(n^{-1}) \) only hold in very particular circumstances (typically i.i.d. data with an adequate standardization). Indeed, for dependent data, the bias of the variance estimator may be so important that the second-order validity of the interpolation may not even hold, a possibility that might explain the bad results of their simulation. Thus, an accurate simple variance estimator is proposed in our Appendix. Even with this estimator, a close study of the improvements of interpolations and extrapolations is needed.
3.1. The studentized sample mean.

Let $T_n = \overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ be the sample mean, and let $\theta = E X_0$ be the mean. Also let $R(s) = E(X_0 - \theta)(X_0 + a - \theta)$, for $s = 0, \pm 1, \pm 2, \ldots$ be the autocovariance sequence. Both $\theta$ and $R$ are generally unknown, and the objective is to obtain interval estimates for $\theta$ based on the data in a nonparametric fashion. Following Götze & Künsch (1996), we assume that

$$\alpha X(k) \leq d^{-1}e^{-dk}$$

(8)

for some $d > 0$ and that

$$E|X_0|^s < \infty$$

(9)

for some $s \geq 5$. We will also assume the Cramér-type regularity conditions A3, A5, and A6 of Götze & Künsch (1996).

Let $s_n^2$ be an estimator of $\sigma^2_\infty$ based on $X_1, \ldots, X_n$ and accurate enough so that $s_n^2 = \sigma^2_\infty + O_P(\sqrt{\log n}/n)$ under conditions (8) and (9); for instance, we can let $s_n^2 = \hat{\sigma}^2_{\hat{\theta}, M, M, n}$ with $M = A \log n$, where the estimator $\hat{\sigma}^2_{\hat{\theta}, M, M, n}$ is defined as in the Appendix.

Consider now the subsampling distribution of the studentized sample mean, which is defined by

$$L_0(x) = \frac{1}{q} \sum_{i=1}^{q} \left( \frac{T_{0,i} - T_n}{s_{b,i}} \right) \leq x,$$

where $T_{b,i} = b^{-1} \sum_{k=i}^{i+b-1} X_k$ and $s_{b,i}$ is the statistic $s_b$ computed on block $\{X_1, \ldots, X_{i+b-1}\}$.

The following proposition states that interpolation of an undersampling distribution with the adequate standardization is second order correct. We give some rate of convergence. Interpolation is not second order correct but improves over only one subsampling distribution.

**Proposition 3.** If $b = \Omega(n^{s/(3s-4)} \log n^{-1})$, then under the preceding assumptions, the interpolation of $L_0$ with $\Phi$ is second order correct with an error rate

$$O_P \left( \frac{\log n}{n^{(3s-4)/(3s-4)}} \right)$$

that is close to $O_P(n^{-2/3})$ when $s$ is large. Let $b_i = c_i^b b_i$, $i = 1, 2$, with $b = \Omega \left( (n \log^2 n)^{s/(3s-4)} \right)$. Then the extrapolation of two undersampling distributions satisfies

$$\tilde{L}_2(x) = \Phi(x) + O_P \left( n^{(3-s)/(3s-4)} \log(n^{s/(3s-4)}) \right),$$

which for large $s$ becomes close to $n^{-1/3}$, thereby improving upon the $n^{-1/4}$ rate of $L_0(x)$, but not achieving second order correctness.

**Proof.** In the following we use the notation $b = \Omega(m)$ to mean that $b/m \to \text{constant} \neq 0$. Define the Edgeworth expansion

$$Q(x) = \Phi(x) + n^{-1/2}k_\infty \sigma^{-3/2} \left\{ \frac{1}{1} \phi(3)(x) - \frac{1}{3} \phi(x) \right\},$$

where $k_\infty = \sum_{i,j} E \left\{ (X_i - \theta)(X_j - \theta)(X_i - \theta)(X_j - \theta) \right\}$ is finite because of (8) and (9). Here, $\Phi(x), \phi(x), \phi^{(k)}(x)$ denote the standard normal distribution, density, and the $k$th derivative of its density, respectively. Now under the assumed conditions, we can employ the results of Götze & Künsch (1996) to infer that

$$\sup_x \left| P \left( \sqrt{\frac{n}{\sum_{i} X_i - \theta}} \leq x \right) - Q(x) \right| = O \left( \frac{M}{n^{1-2/s}} \right) + O(\beta_n) = O \left( \frac{\log n}{n^{1-2/s}} \right)$$

(10)
where $\beta_n = \mathbb{E} s_n^2 - \sigma_n^2$, and where $M$ is the equivalent width of the autocovariance window; see the Appendix for more details. Using the choice $M = A \log n$ for a sufficiently large constant $A$ implies that $\beta_n = O(n^{-1})$ and yields the rate in (10).

Now, similarly to the proof of Theorem 3.1 in Politis & Romano (1994), it can be shown that $\text{var} \{ I_b(x) \} = O(b/n)$ due to the geometric mixing rate (8). Now, using (10), we obtain

$$
\frac{1}{q} \sum_{i=1}^{q} \mathbb{P} \left( \sqrt{\frac{b}{s_{b,i}}} \left( \frac{T_{b,i} - \theta}{s_{b,i}} \right) \leq x \right) = \mathbb{E} \left\{ \mathbb{P} \left( \sqrt{\frac{b}{s_{b,i}}} \left( \frac{T_{b,i} - \theta}{s_{b,i}} \right) \leq x \right) \right\} + O_P \left( \frac{\log b}{b^{1/2}} \frac{\log n}{b^{1/2}} \right),
$$

The above together with the fact that

$$
\sqrt{\frac{b}{s_{b,i}}} \left( \frac{T_{b,i} - \theta}{s_{b,i}} \right) = O_P \left( \frac{\log b}{b^{1/2}} \frac{\log n}{b^{1/2}} \right)
$$

yield

$$
I_b(x) = \Phi(x) + \frac{k_{\infty} p(x)}{b^{1/2} \sigma_{\infty}^2} + O_P \left( \frac{\log b}{b^{1/2}} \frac{\log n}{b^{1/2}} \right) + O_P \left( \frac{\log n}{b^{1/2}} \right). \tag{11}
$$

It follows that taking $b = \Omega \left( \frac{n}{\sqrt{s}} \right)$, we minimize the Mean Squared Error (MSE) of $I_b(x)$, thus having $I_b(x) = \Phi(x) + O_P \left( n^{-1/4} \right)$. The remainder of the proof is straightforward: we use (11) and minimize the remainders in the interpolation and the extrapolation, respectively.

**Remark 3.** The choice of $b$ is nearly optimal up to a log factor. However the rate of the interpolation is still worse than the best rate of the block-bootstrap obtained by Götte & Künsch (1996), which can be made close to $O_P \left( n^{-3s/4} \right)$ that gives $O_P \left( n^{-3s/4} \right)$ when $s$ is infinite. Moreover, we note that for the interpolation to be correct we need only $s \geq 5$ whereas at least $s \geq 24$ is needed for the block bootstrap to be second-order correct; see Götte & Künsch (1996).

**Remark 4.** The fact that the second order correctness is not attained may be explained by the need of some finite population correction factor as in the i.i.d. case. Recall that the finite population correction factor in the i.i.d. case in particular was due to the fact that

$$
\text{var} \left\{ \sqrt{\frac{b}{s_n}} \left( \bar{X}_{b,i} - \bar{X}_n \right) \right\} = \left( 1 - \frac{b}{s_n} \right) \sigma^2,
$$

where $\bar{X}_{b,i}$ is the sample mean of subsample $Y_i$. In the strong mixing case, it is interesting to see that a similar relation holds, thus indicating that the same finite population correction factor (surprisingly of same form as in the i.i.d. case) may be appropriate. A straightforward calculation shows that we have

$$
\text{var} \left\{ \sqrt{\frac{b}{s_n}} \left( \bar{X}_{b,i} - \bar{X}_n \right) \right\} = \left( 1 - \frac{b}{n} \right) \sigma^2 + O \left( \frac{b}{n^{2s}} + \frac{b^2}{n^3} \right).
$$

So it is obvious that taking into account the finite population correction factor is generally advisable, as it will reduce the error of the subsampling distribution in the mixing case as well. However, it is not clear that the correction will improve the order of the extrapolation.

**Remark 5.** The preceding discussion has made apparent that perhaps second order accuracy may be too much to hope for from extrapolated distributions in the case of strong mixing data because of the “bad” effect of the standardization. Thus it may be interesting in that case to use unstandardized distributions. Using the conditions (2.3), (2.5) and (2.6) of Götte & Hipp (1983) as well as our conditions (8) and (9)—with $s > 3$—it is easy to see that regular statistics (functions of moments) admits an Edgeworth expansion uniformly in $x$. Using the results in Bertail (1997) (conditions $A2[i]$, $\alpha 3[i]$, and $\alpha 4[i]$), the extrapolation of $I$ undersampling distribution is clearly
first-order correct and improves over one undersample distribution. Moreover in regular cases (function of means) if one chooses \( b = n^{1/2}/\log n \), then as \( n \) grows we come close to the first order rate \( O(n^{-1/2}) \) (without achieving it); see Bertail & Politis (1996) for further details. Obviously, extrapolated subsampling will definitely not be second-order correct in the unstudentized case; therefore, it will be inferior to the studentized block-bootstrap when the block-bootstrap applies as well; see Künsch (1989), Liu & Singh (1992), Götze & Künsch (1996). The possibility, however, that extrapolated subsampling distributions may provide a robust and more accurate asymptotic approximation under very weak assumptions is rather remarkable.

4. SOME SIMULATION RESULTS FOR ARMA PROCESSES

A straightforward application of our result concerns confidence intervals for parameters of ARMA processes using pseudo-maximum likelihood or robust methods (see, e.g., Künsch 1984). In this section, we give some simulations results on ARMA processes estimated by pseudo-maximum likelihood assuming that the underlying likelihood is normal.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( \alpha % )</th>
<th>2.5</th>
<th>5.0</th>
<th>95.0</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>( K^{-1}_n(\alpha) )</td>
<td>-7.127</td>
<td>-4.603</td>
<td>1.564</td>
<td>2.511</td>
</tr>
<tr>
<td>Asymptotic</td>
<td>( \Phi^{-1}(\alpha) )</td>
<td>-1.960</td>
<td>-1.645</td>
<td>1.645</td>
<td>1.900</td>
</tr>
<tr>
<td>( (\beta %) )</td>
<td></td>
<td>12.5</td>
<td>15.4</td>
<td>94.9</td>
<td>96.3</td>
</tr>
<tr>
<td>Interpolation</td>
<td>Mean</td>
<td>-5.174</td>
<td>-2.332</td>
<td>1.907</td>
<td>4.920</td>
</tr>
<tr>
<td>( b_n = 7 )</td>
<td>Median</td>
<td>-3.636</td>
<td>-2.270</td>
<td>1.672</td>
<td>2.436</td>
</tr>
<tr>
<td>( (\beta %) )</td>
<td></td>
<td>10.8</td>
<td>14.6</td>
<td>94.6</td>
<td>96.8</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>Mean</td>
<td>-7.393</td>
<td>-3.776</td>
<td>3.244</td>
<td>9.425</td>
</tr>
<tr>
<td>( b_1 = 7, b_2 = 13 )</td>
<td>Median</td>
<td>-4.132</td>
<td>-2.632</td>
<td>1.672</td>
<td>2.813</td>
</tr>
<tr>
<td>( (\beta %) )</td>
<td></td>
<td>11.4</td>
<td>13.7</td>
<td>94.7</td>
<td>97.1</td>
</tr>
</tbody>
</table>

The model considered here is an ARMA(2,1) model

\[
X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} .
\]

We are interested in confidence intervals for \( \phi_1 \). In the following tables, we compare the asymptotic distribution with the interpolated distribution and our extrapolation technique, taking into account the “finite population correction”. The exact quantiles of the distribution of the maximum likelihood are computed by Monte-Carlo replications by generating 100,000 processes. The mean, the median of the bounds and estimated coverage probability of the interpolated and extrapolated distributions are calculated over 10,000 iterations of the procedure.

In Tables 1 and 2, the true parameters are \( \phi_1 = 0.5, \phi_2 = 0.3 \) and \( \theta_1 = 0.6 \) and the residuals are \( \mathcal{N}(0, 1) \). Tables 1 and 2 give the results for an observed stretch of size 50 and 100, respectively.

The first striking feature of these simulation results is how far the asymptotic quantiles are from the true quantile. Both the extrapolation and the interpolation succeed in catching the asymmetry of the true distribution. Curiously the extrapolation gives better results in terms of average estimation of the quantile than the interpolation. However in term of coverage probability, the interpolation gives a better result as predicted by the theoretical results. Both in terms of quantile estimation and coverage probability, the extrapolation and the interpolation outperform the asymptotic distribution.
TABLE 2: Confidence intervals for $\phi_2$, normal residuals, $n = 100$.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\alpha%$</th>
<th>2.5</th>
<th>5.0</th>
<th>95.0</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>$K_n^{-1}(\alpha)$</td>
<td>$-3.853$</td>
<td>$-2.312$</td>
<td>$1.754$</td>
<td>$2.306$</td>
</tr>
<tr>
<td>Asymptotic</td>
<td>$\Phi^{-1}(\alpha)$</td>
<td>$-1.960$</td>
<td>$-1.645$</td>
<td>$1.645$</td>
<td>$1.960$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$7.8$</td>
<td>$10.2$</td>
<td>$95.0$</td>
<td>$96.9$</td>
</tr>
<tr>
<td>Interpolation</td>
<td>Mean</td>
<td>$-2.753$</td>
<td>$-1.994$</td>
<td>$1.669$</td>
<td>$2.211$</td>
</tr>
<tr>
<td>$b_n = 10$</td>
<td>Median</td>
<td>$-2.848$</td>
<td>$-2.008$</td>
<td>$1.618$</td>
<td>$2.135$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$5.2$</td>
<td>$8.4$</td>
<td>$95.5$</td>
<td>$97.9$</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>Mean</td>
<td>$-3.424$</td>
<td>$-2.279$</td>
<td>$1.704$</td>
<td>$2.646$</td>
</tr>
<tr>
<td>$b_1 = 10$, $b_2 = 21$</td>
<td>Median</td>
<td>$-3.300$</td>
<td>$-2.422$</td>
<td>$1.549$</td>
<td>$2.106$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$5.9$</td>
<td>$8.3$</td>
<td>$93.5$</td>
<td>$97.1$</td>
</tr>
</tbody>
</table>

In comparison to the previous simulations, the model is actually AR(1), thus corresponding to $\phi_1 = 0.9$, $\phi_2 = 0$ and $\theta_1 = 0$. Moreover, the true residuals have a lognormal distribution recentered at 0. These simulations are used to test the effect of the asymmetry of the distribution of the residuals on the pseudo-likelihood estimator and the robustness of the extrapolations.

TABLE 3: Confidence intervals for $\phi_1$, lognormal residuals, $n = 50$.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\alpha%$</th>
<th>2.5</th>
<th>5.0</th>
<th>95.0</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>$K_n^{-1}(\alpha)$</td>
<td>$-1.802$</td>
<td>$-1.521$</td>
<td>$2.978$</td>
<td>$4.404$</td>
</tr>
<tr>
<td>Asymptotic</td>
<td>$\Phi^{-1}(\alpha)$</td>
<td>$-1.960$</td>
<td>$-1.645$</td>
<td>$1.645$</td>
<td>$1.960$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$1.7$</td>
<td>$4.3$</td>
<td>$89.5$</td>
<td>$92.0$</td>
</tr>
<tr>
<td>Interpolation</td>
<td>Mean</td>
<td>$-1.851$</td>
<td>$-1.499$</td>
<td>$2.238$</td>
<td>$3.332$</td>
</tr>
<tr>
<td>$b = 7$</td>
<td>Median</td>
<td>$-1.751$</td>
<td>$-1.406$</td>
<td>$2.259$</td>
<td>$3.301$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$2.7$</td>
<td>$6.5$</td>
<td>$94.2$</td>
<td>$96.6$</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>Mean</td>
<td>$-1.836$</td>
<td>$-1.503$</td>
<td>$2.448$</td>
<td>$3.561$</td>
</tr>
<tr>
<td>$b_1 = 7$, $b_2 = 13$</td>
<td>Median</td>
<td>$-1.635$</td>
<td>$-1.270$</td>
<td>$2.464$</td>
<td>$3.372$</td>
</tr>
<tr>
<td></td>
<td>($\beta%$)</td>
<td>$3.2$</td>
<td>$6.7$</td>
<td>$92.9$</td>
<td>$94.7$</td>
</tr>
</tbody>
</table>

In comparison to the previous simulations, the true distribution exhibits an opposite asymmetric behaviour. Once again the interpolation and the extrapolation capture this asymmetry. For $n = 100$, the coverage probability are very close to the nominal level but the interpolation which is known to be second-order correct in that case clearly exhibits a better behaviour in terms of coverage probability.
Here, the full-overlap case corresponding to \( \alpha | = 0 \) between the starting points of block \( \alpha | = 0 \) and block \( \alpha | = 0 \). Different names and variations, see Politis & Romano 1995) is condition (8)) have been proposed in the literature; probably the most popular one (under many different names and variations, see Politis & Romano 1995) is
\[
\hat{\sigma}^2_{M,n} = \frac{M}{Q} \sum_{i=1}^{Q} \left( \bar{X}_{i,M,L} - \bar{X}_n \right)^2,
\]
where \( \bar{X}_{i,M,L} = \frac{1}{M} \sum_{s=L(i-1)+1}^{L(i-1)+M} X_s \) is the mean of the block \( \{ X_{L(i-1)+1}, \ldots, X_{L(i-1)+M} \} \) of the data, the numbers \( L, M \) are integers depending on the sample size \( n \), and \( Q = \left\lfloor \frac{n - M}{L} \right\rfloor + 1 \), with \( \lfloor \cdot \rfloor \) being the integer part. Here, \( M \) is the block’s size, \( L \) is the amount of ‘lag’ between the starting points of block \( i \) and block \( i+1 \), and \( Q \) is the total number of such blocks that can be extracted from the data. If \( L = M \), there is no overlap between block \( i \) and block \( i+1 \). The full-overlap case corresponding to \( L = 1 \) is recommended (see, e.g., Künsch 1989); thus we set \( L = 1 \) in what follows.

Under regularity conditions, \( \hat{\sigma}^2_{M,n} \) is a consistent and asymptotically normal estimator. The regularity conditions are moment and mixing conditions and conditions on the design parameters; typically \( M \to \infty \), but with \( M/n \to 0 \). Consistency is immediate if we consider the first two moments of \( \hat{\sigma}^2_{M,n} \) that can be asymptotically calculated to be
\[
\text{bias} \left( \hat{\sigma}^2_{M,n} \right) = \mathbb{E} \left( \hat{\sigma}^2_{M,n} - \sigma^2_{\infty} \right) = O(1/M) + O(M/n),
\]
\[
\text{var} \left( \hat{\sigma}^2_{M,n} \right) \approx 2\epsilon \frac{M}{n} \sigma^4_{\infty} + o(M/n).
\]

Realizing that the poor rate of convergence of \( \hat{\sigma}^2_{M,n} \) is due to its bias, Politis & Romano (1995) proposed a bias-corrected version in the more general case of estimation of the spectral density \( g(w) \). In the present setting, this bias-corrected version is given by
\[
\hat{\sigma}^2_{m,M,n} \equiv (h+1)\hat{\sigma}^2_{M,n} - h\hat{\sigma}^2_{m,n}.
\]
Here, \( h \) is some chosen positive constant, and \( m \) is chosen as \( m = hM/(1+h) \). The choice \( h = 1 \), leading to \( m = M/2 \) is proposed as a simple solution, and an empirical data-driven method for choosing \( M \) is presented in Politis & Romano (1995). Equation (13) can be interpreted as an

**TABLE 4:** Confidence intervals for \( \phi_1 \), lognormal residuals, \( n = 100 \).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5</th>
<th>5.0</th>
<th>95.0</th>
<th>97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>( K^{-1}_\alpha )</td>
<td>-1.820</td>
<td>-1.538</td>
<td>2.177</td>
</tr>
<tr>
<td>Asymptotic</td>
<td>( \Phi^{-1}_\alpha )</td>
<td>-1.960</td>
<td>-1.645</td>
<td>1.645</td>
</tr>
<tr>
<td>(( \beta )%)</td>
<td>1.6</td>
<td>4.4</td>
<td>92.4</td>
<td>94.0</td>
</tr>
<tr>
<td>Interpolation</td>
<td>Mean</td>
<td>-1.843</td>
<td>-1.506</td>
<td>2.095</td>
</tr>
<tr>
<td>( b = 10 )</td>
<td>Median</td>
<td>-1.792</td>
<td>1.452</td>
<td>2.114</td>
</tr>
<tr>
<td>(( \beta )%)</td>
<td>2.6</td>
<td>6.2</td>
<td>94.6</td>
<td>97.4</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>Mean</td>
<td>-1.821</td>
<td>-1.510</td>
<td>2.336</td>
</tr>
<tr>
<td>( b_1 = 10, b_2 = 21 )</td>
<td>Median</td>
<td>-1.683</td>
<td>-1.379</td>
<td>2.254</td>
</tr>
<tr>
<td>(( \beta )%)</td>
<td>2.8</td>
<td>6.4</td>
<td>96.0</td>
<td>97.7</td>
</tr>
</tbody>
</table>
extrapolation of the two subsampling variance estimators \( \hat{\sigma}^2_{m, n} \) and \( \hat{\sigma}^2_{M, n} \) that has an improved asymptotic performance.

The difference between the setup of variance estimation considered here and the setup of spectral density estimation considered in Politis & Romano (1995) is that in the usual spectral density estimation practice, the true mean \( \theta \) is assumed to be known and is used (in place of \( \overline{X}_n \)) in constructing Bartlett’s estimator; the implication is that improper centering leads to some—usually negligible—“edge effects”. For example, the added \( O(M/n) \) bias term in equation (12) above is due to this improper centering in the construction of \( \hat{\sigma}^2_{M, n} \), i.e., centering the data at \( \overline{X}_n \) instead of \( \theta \).

Similarly to what was shown by Politis & Romano (1995), for the bias-corrected Bartlett estimator, it may be shown that not only the bias of \( \hat{\sigma}^2_{m, M, n} \) becomes \( o(1/M) \), but a more spectacular bias correction is achieved: namely, under the exponential strong mixing assumption (8) we have that in the case of the mean

\[
\text{bias} \left( \hat{\sigma}^2_{m, M, n} \right) = O(1/n)
\]

if \( M = A \log n \) for some sufficiently large constant \( A \). In other words, we have (for \( h = 1 \), say) that

\[
\hat{\sigma}^2_{m, A \log n, A \log n, n} = \sigma^2_\infty + O_P \left( \sqrt{\log n/n} \right).
\]

Thus \( \hat{\sigma}^2_{m, A \log n, A \log n, n} \) may be used whenever an accurate variance estimator is needed. For example, it may be used for studentization in the context of subsampling distributions discussed here or block-bootstrap distributions in Götze & Künsch (1996).

Note that if \( a_X(k) = 0 \) for all \( |k| \) bigger than some \( K > 0 \), i.e., if the data are \( K \)-dependent, then it can be shown additionally that \( \text{bias} \left( \hat{\sigma}^2_{m, M, n} \right) = O(1/n) \), even when \( m \) and \( M \) are constants, satisfying \( M \geq m \geq K \). Consequently, taking \( K = m = M/2 \), we obtain that \( \hat{\sigma}^2_{K, 2K, n} = \sigma^2_\infty + O_P \left( 1/\sqrt{n} \right) \); in other words, \( \hat{\sigma}^2_{K, 2K, n} \) achieves the parametric \( \sqrt{n} \) rate in this case!

As a final practical comment, we note that \( \hat{\sigma}^2_{m, M, n} \) is not almost surely nonnegative; this is not a problem with \( \hat{\sigma}^2_{m, M, n} \) in particular but rather a problem with all higher-order accurate variance (or spectral) estimators (see Politis & Romano 1995). Although this problem disappears asymptotically, in finite samples it might pose a real problem, especially if we want to studentize using \( \hat{\sigma}^2_{m, M, n} \). If we happen to compute a \( \hat{\sigma}^2_{m, M, n} < 0 \) and the obvious solution to take zero as our estimate is not acceptable (e.g., in the studentization setup), there are two practical ways out:

(a) try out different choices for \( M \) (or for \( A \), if we take \( M = A \log n \)) and use the corresponding \( \hat{\sigma}^2_{m, M, n} \) if it turns out to be positive, or

(b) use a fraction of \( \hat{\sigma}^2_{m, n} \) (e.g., \( \ell \hat{\sigma}^2_{m, n} \) for some \( \ell \in (0, 1] \)) as our variance estimator, i.e., ‘shrink’ the estimator \( \hat{\sigma}^2_{m, n} \) towards zero.

Regarding (b) note that \( \hat{\sigma}^2_{m, M, n} \) can be interpreted as \( \hat{\sigma}^2_{M, n} - \hat{\text{bias}} \left( \hat{\sigma}^2_{M, n} \right) \); a negative \( \hat{\sigma}^2_{m, M, n} \) indicates that our estimate of the bias of \( \hat{\sigma}^2_{m, n} \) is positive and large (actually, too large). Nevertheless, we might take the hint that \( \hat{\sigma}^2_{m, n} \) has a positive bias and attempt to reduce it by taking \( \ell \hat{\sigma}^2_{M, n} \) as our estimator. The additional difficult question of choosing \( \ell \) actually prompts us to favor method (a) of solving the problem of a negative estimate. Simulations results (see Bertail & Politis 1996) further support suggestion (a), i.e., a negative \( \hat{\sigma}^2_{m, M, n} \) can occur as a result of a poor choice of \( M \).
ACKNOWLEDGEMENTS

Financial support for this research was provided by an AIP contract of the Corela Laboratory at INRA, and by NSF grant DMS-97-03964. We would like to thank the referees and the former Editor, Professor C. Genest, for their comments and suggestions which greatly improved the presentation of this paper.

REFERENCES


Received 27 May 1998
Accepted 14 March 2001

Patrice BERTAIL: Patrice.Bertail@u-paris10.fr
INRA-Corela et MODALX
Université Paris X, UFR Segmi, Batiment G
200, avenue de la République, F-92001 Nanterre, France

Dimitris N. POLITIS: politis@euclid.ucsd.edu
Department of Mathematics
University of California at San Diego
La Jolla, CA 92093, USA