Statistica Sinica **10**(2000), 1-50

# HYBRID RESAMPLING METHODS FOR CONFIDENCE INTERVALS

Chin-Shan Chuang and Tze Leung Lai

*Carnegie-Mellon University and Stanford University*

*Abstract:* This paper considers the problem of constructing confidence intervals for a single parameter $\theta$ in a multiparameter or nonparametric family. Hybrid resampling methods, which "hybridize" the essential features of bootstrap and exact methods, are proposed and developed for both parametric and nonparametric situations. In particular, we apply such methods to construct confidence regions, whose coverage probabilities are nearly equal to the nominal ones, for the treatment effects associated with the primary and secondary endpoints of a clinical trial whose stopping rule, specified by a group sequential test, makes the approximate pivots in the nonsequential bootstrap method highly "non-pivotal". We also apply hybrid resampling methods to construct second-order correct confidence intervals in possibly non-ergodic autoregressive models and branching processes.

*Key words and phrases:* Bootstrap confidence intervals, empirical likelihood, group sequential tests, hybrid resampling, nonparametric tilting.

## 1. Introduction

The past two decades have witnessed important developments in group sequential methods for interim analysis of clinical trials. Although these methods allow for early termination while preserving the overall significance level of the test, they introduce substantial difficulties in constructing confidence intervals for parameters of interest following the test. Under strong distributional assumptions and for relatively simple parametric models, exact confidence intervals in group sequential settings have been developed in the literature. For samples of fixed size, an important methodology for constructing confidence intervals without distributional assumptions is the bootstrap method. Although the stopping rule makes approximate pivots in the nonsequential bootstrap method highly "non-pivotal", we have recently shown in Chuang and Lai (1998) that it is possible to "hybridize" the bootstrap and exact methods for constructing confidence intervals following a group sequential test.

In Sections 2, 3 and 6, we give a comprehensive development of the hybrid resampling approach, extending the methodology beyond the group sequential setting considered in Chuang and Lai (1998) and in Section 4 of the present paper.

In particular, in Section 5, we show how the methodology can be used to construct second-order correct confidence intervals for the autoregressive parameter $\theta$ of a possibly nonstationary AR(1) model, for which the bootstrap method has been shown to be inconsistent when $|\theta| = 1$. Since the bootstrap method is a special case of hybrid resampling as shown in Section 2, our development of hybrid resampling also addresses certain basic issues concerning the bootstrap method, such as choice of root, difficulties with the infinitesimal jackknife and linearization, influential observations, calibration and bootstrap iteration. Moreover, in connection with the choice of a resampling family for the hybrid approach, we also discuss certain basic issues concerning Owen's (1988, 1990) empirical likelihood and Efron's (1981, 1987) nonparametric tilting. Some concluding remarks are given in Section 7.

## 2. Exact, Bootstrap and Hybrid Resampling Methods for Confidence Intervals

We begin with the use of exact, bootstrap and hybrid resampling methods for constructing confidence intervals. Let $\mathbf{X} = (X_{11}, \ldots, X_{1p}, \ldots, X_{n1}, \ldots, X_{np})$ be a vector of observations from some family of distributions $\{F : F \in \mathcal{F}\}$. For nonparametric problems, $\mathcal{F}$ is the family of distributions on $\mathbf{R}^p$ satisfying certain prespecified regularity conditions and $(X_{i1}, \ldots, X_{ip})$ are i.i.d. rrandom vectors having common distribution $F \in \mathcal{F}$. For parametric models with parameter $\eta \in \Gamma$, we can denote $\mathcal{F}$ by $\{F_\eta : \eta \in \Gamma\}$, and $(X_{i1}, \ldots, X_{ip})$ may be i.i.d. or may form a time series. The problem of interest is to find a confidence interval for the real-valued parameter $\theta = \theta(F)$. We let $\Theta$ denote the set of all possible values of $\theta$.

*Exact method*: If $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is indexed by a real-valued parameter $\theta$, an exact equal-tailed confidence region can always be found by using the well known duality between hypothesis tests and confidence regions (cf. Rosner and Tsiatis (1988), Schenker (1987)). Suppose one would like to test the null hypothesis that $\theta$ is equal to $\theta_0$. Let $R(\mathbf{X}, \theta_0)$ be some real-valued test statistic. Let $u_\alpha(\theta_0)$ be the $\alpha$-quantile of the distribution of $R(\mathbf{X}, \theta_0)$ under the distribution $F_{\theta_0}$. The null hypothesis is accepted if $u_\alpha(\theta_0) < R(\mathbf{X}, \theta_0) < u_{1-\alpha}(\theta_0)$. An exact equal-tailed confidence region with coverage probability $1 - 2\alpha$ consists of all $\theta_0$ not rejected by the test and is therefore given by

$$\{\theta : u_\alpha(\theta) < R(\mathbf{X}, \theta) < u_{1-\alpha}(\theta)\}. \tag{2.1}$$

*Bootstrap method*: The exact method applies only when there are no nuisance parameters and this assumption is rarely satisfied in practice. The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ by the approximate quantiles

$u^*_\alpha$ and $u^*_{1-\alpha}$ obtained in the following manner. Based on $\mathbf{X}$, construct an estimate $\widehat{F}$ of $F \in \mathcal{F}$. The quantile $u^*_\alpha$ is defined to be the $\alpha$-quantile of the distribution of $R(\mathbf{X}^*, \widehat{\theta})$ with $\mathbf{X}^*$ generated from $\widehat{F}$ and $\widehat{\theta} = \theta(\widehat{F})$; see Efron (1981, 1987). Thus, the bootstrap method yields the following confidence region for $\theta$ with approximate coverage probability $1 - 2\alpha$:

$$\{\theta : u^*_\alpha < R(\mathbf{X}, \theta) < u^*_{1-\alpha}\}. \tag{2.2}$$

In particular, when $\widehat{F}$ is the empirical distribution of i.i.d. $X_1, \ldots, X_n$ and the root $R(\mathbf{X}, \theta)$ is equal to $(\widehat{\theta} - \theta)/\widehat{\sigma}$ for some estimate $\widehat{\sigma}$ of the standard error of $\widehat{\theta}$, the bootstrap confidence interval (2.2) is called the bootstrap-$t$ interval. It is well known that the bootstrap-$t$ interval has coverage error $O(n^{-1})$ at both endpoints when $\theta$ is a smooth function of means; see Hall (1988, 1992).

*Hybrid resampling method*: The hybrid confidence region is based on reducing the family of distributions $\mathcal{F}$ to another family of distributions $\{\widehat{F}_\theta : \theta \in \Theta\}$, where $\theta$ is the unknown parameter of interest. We call this family the "resampling family". This reduction depends on $\mathbf{X}$, and ways for carrying it out are explored in the rest of the paper. Let $\widehat{u}_\alpha(\theta)$ be the $\alpha$-quantile of the sampling distribution of $R(\mathbf{X}, \theta)$ under the assumption that $\mathbf{X}$ has distribution $\widehat{F}_\theta$. The hybrid confidence region results from applying the exact method to $\{\widehat{F}_\theta : \theta \in \Theta\}$ and is given by

$$\{\theta : \widehat{u}_\alpha(\theta) < R(\mathbf{X}, \theta) < \widehat{u}_{1-\alpha}(\theta)\}. \tag{2.3}$$

The construction of (2.3) typically involves simulations to compute the quantiles as in the bootstrap method and is elaborated below. We call this the "hybrid resampling" method because it "hybridizes" the exact method (that uses test inversion) with the bootstrap method (that uses the observed data to determine the resampling distribution). Note that hybrid resampling is a generalization of the bootstrap method, which uses the singleton $\{\widehat{F}\}$ as the resampling family $\{\widehat{F}_\theta\}$. The following two examples, which will be discussed in Sections 4 and 5, illustrate difficulties with the bootstrap method when the sampling distribution of $R(\mathbf{X}, \theta)$ may vary substantially with $\theta$.

**Example 1.** Consider a group sequential trial with Pocock's (1977) boundary and a maximum of 5 groups, as in Chuang and Lai (1998). Let $X_1, X_2, \ldots$ be independent with mean $\theta$ and variance 1, $S_n = X_1 + \cdots + X_n$, $\bar{X}_n = S_n/n$, and let $J = \{15j : j = 1, 2, 3, 4, 5\}$. The stopping rule is $\tau = \min\{n \in J : |S_n| \geq 2.413n^{1/2}\}$, where we define the minimum of $\emptyset$ to be $75 = 15 \times 5$. The choice 2.413 ensures that when $\theta = 0$, $P\{\max_{n \in J} |S_n/n^{1/2}| \geq 2.413\} \doteq 0.05$; see Pocock (1977). Figure 1 in Chuang and Lai (1998) shows that the sampling distribution of $\sqrt{\tau}(\bar{X}_\tau - \theta)$ varies markedly with $\theta$ even for normal observations, and Table 1 there reports poor performance of the bootstrap method.

**Example 2.** Consider the first-order autoregressive AR(1) model given by the following. Let $x_0 = 0$ and $x_i = \theta x_{i-1} + \epsilon_i$, where $\epsilon_i$ are i.i.d. with mean 0 and variance 1. Let $\widehat{\theta}$ be the least squares estimate of $\theta$ based on $(x_1, \ldots, x_n)$, given by $\widehat{\theta} = \sum_{i=1}^n x_i x_{i-1} / \sum_{i=1}^n x_{i-1}^2$. It is well known that the limiting distribution of $(\sum_{i=1}^n x_{i-1}^2)^{1/2}(\widehat{\theta} - \theta)$ is standard normal if $|\theta| < 1$. However, when $|\theta| = 1$, the limiting distribution is given by $\frac{1}{2}(B_1^2 - 1)/(\int_0^1 B_t^2 dt)^{1/2}$, where $B_t$ denotes standard Brownian motion. Basawa, Mallik, McCormick, Reeves and Taylor (1991) showed that bootstrapping the least squares estimate is inconsistent when $|\theta| = 1$.

In practice, it is often desirable to express a confidence set for $\theta$ as an interval. Although the sets (2.1), (2.2) and (2.3) may not be intervals, it often suffices to give only the upper and lower limits of the confidence set. We now describe an algorithm, based on the method of successive secant approximations, to find the upper limit of (2.3). Let $f(\theta) = R(\mathbf{X}, \theta) - \widehat{u}_\alpha(\theta)$ and consider solving the equation $f(\theta) = 0$. First we find $a_1 < b_1$ such that $f(a_1) > 0$ and $f(b_1) < 0$. Let $f_1(\theta)$ be linear in $\theta \in [a_1, b_1]$ with $f_1(a_1) = f(a_1)$ and $f_1(b_1) = f(b_1)$, and let $\theta_1$ be the root of $f_1(\theta) = 0$. If $f(\theta_1) > 0$, set $a_2 = \theta_1$ and $b_2 = b_1$. If $f(\theta_1) < 0$, set $b_2 = \theta_1$ and $a_2 = a_1$. Proceeding inductively in this manner, we let $f_k(\theta)$ linearly interpolate $f(a_k)$ and $f(b_k)$ for $a_k \le \theta \le b_k$, and let $\theta_k \in (a_k, b_k)$ be the root of $f_k(\theta) = 0$. This procedure terminates if $\theta_k$ differs little from $\theta_{k-1}$ or if $k$ reaches some upper bound, and the terminal value $\theta_k$ is taken to be the upper limit of (2.3). Typically $f(\widehat{\theta}) > 0$, so $\widehat{\theta}$ can be chosen as $a_1$. To find $b_1$, one can start with $b_1' = \widehat{\theta} + 2\widehat{\sigma}$, where $\widehat{\sigma}$ is an estimate of the standard error of $\widehat{\theta}$. If $f(b_1') < 0$, set $b_1 = b_1'$; otherwise let $b_2' = b_1' + \widehat{\sigma}/2$ and check whether $f(b_2') < 0$. This procedure is repeated until one arrives at $f(b_h') < 0$ and sets $b_1 = b_h'$. The total number of iterations, $h + k$, is kept no more than some prescribed upper bound $m$. If there are already $m$ iterations before one arrives at $b_1$ with $f(b_1) < 0$ (so $h = m$), take $b_m'$ as the default value of the upper limit of (2.3). For the simulations in Example 3 and those used to produce Tables 4 and 5, we took $m = 8$. For the simulations in Example 7 of Section 6, we took $m = 4$ to ease the computational burden. The quantiles $\widehat{u}_\alpha(\theta_j)$ can be computed from independent samples from $\widehat{F}_{\theta_j}$, as was done in these examples. It is sometimes possible to try to reuse the same random sample for all $\theta$ values, as in the mean and regression models considered in Tables 4 and 5, but there is not much computational saving since we typically do not use a large value of $m$.

## 3. Choice of Root and Implementation of Hybrid Resampling Methods

As the framework of Section 2 suggests, there are two issues that must be addressed for hybrid resampling methods to be used successfully in practice.

First, one must choose an appropriate root $R(\mathbf{X}, \theta)$ that is used in resampling. Second, one needs to find a suitable reduction of the original family $\mathcal{F}$ to the resampling family $\{\widehat{F}_\theta\}$. The second issue for nonparametric problems is quite complicated and is deferred to Section 6; in the examples considered in Section 4 and 5, there is a simple reduction.

### 3.1. Choice of root and resampling family in parametric models

One natural root $R(\mathbf{X}, \theta)$ that can be used in parametric models is the signed root of the log likelihood ratio statistic. Let $\mathbf{X}$ represent a vector of observations from a parametric family $F_\eta$ with joint density $f(\mathbf{x}; \eta)$, and let $\theta = g(\eta)$ be a real-valued parameter of interest. Consider testing the null hypothesis that $\theta = \theta_0$. Let $\widehat{\eta}$ be the (unrestricted) maximum likelihood estimate of $\eta$ based on $\mathbf{X}$ and let $\widehat{\eta}(\theta_0)$ be the maximum likelihood estimate of $\eta$ subject to the constraint $g(\eta) = \theta_0$. The likelihood ratio test rejects the null hypothesis for large values of

$$l(\theta_0) = 2\{\log f(\mathbf{X}; \widehat{\eta}) - \log f(\mathbf{X}; \widehat{\eta}(\theta_0))\}. \tag{3.1}$$

Equivalently, one rejects the null hypothesis for large absolute values of the signed root

$$l^{\pm}(\theta_0) = \operatorname{sgn}(\widehat{\theta} - \theta_0) l^{1/2}(\theta_0). \tag{3.2}$$

Here, $\widehat{\theta}$ represents the maximum likelihood estimate of $\theta$ based on $\mathbf{X}$, i.e. $\widehat{\theta} = g(\widehat{\eta})$. With $R(\mathbf{X}, \theta) = l^{\pm}(\theta)$, the quantiles $u^*_\alpha$ and $u^*_{1-\alpha}$ used in the bootstrap confidence region (2.2) are determined from the distribution of $l^{\pm}(\theta)$ under the assumption that $\mathbf{X}$ is generated from $F_{\widehat{\eta}}$. The quantiles $\widehat{u}_\alpha(\theta)$ and $\widehat{u}_{1-\alpha}(\theta)$ used in the hybrid confidence region (2.3) are determined from the distribution of $l^{\pm}(\theta)$ under the assumption that $\mathbf{X}$ is generated from $F_{\widehat{\eta}(\theta)}$. Note that the hybrid region (2.3) is obtained by reducing the family $F_\eta$ to the family $F_{\widehat{\eta}(\theta)}$, whereas the bootstrap confidence region (2.2) is based on the single distribution $F_{\widehat{\eta}}$.

**Example 3.** Consider the following Galton-Watson branching process with immigration (BPI). Let $\xi_1, \xi_2, \ldots$ be i.i.d. Poisson random variables with mean $\theta$, and let $\psi_1, \psi_2, \ldots$ be i.i.d. Poisson random variables with mean $\lambda$. Here, the mean $\theta$ of the offspring distribution is of interest, whereas $\lambda$ is regarded as a nuisance parameter. The Galton-Watson BPI is defined as follows. Assume that $X_0 = x_0$ for some positive integer $x_0$. Let $X_1 = \xi_1 + \cdots + \xi_{x_0} + \psi_1$, which represents the size of the first generation. The size $X_n$ of the $n$th generation, for $n = 2, 3, \ldots$, is given by

$$X_n = \xi_{x_0 + X_1 + \cdots + X_{n-2} + 1} + \cdots + \xi_{x_0 + X_1 + \cdots + X_{n-1}} + \psi_n.$$

Let $\mathbf{X} = (X_0, \ldots, X_n, \psi_1, \ldots, \psi_n)$ be the vector of observations. Let $N_{n-1} = x_0 + X_1 + \cdots + X_{n-1}$ and $\eta = (\theta, \lambda)$. Then as shown by Bhat and Adke (1981),

$$\begin{aligned} \log f(\mathbf{X}; \eta) &= \sum_{i=1}^{N_{n-1}} \xi_i \log\theta - N_{n-1}\theta + \sum_{j=1}^{n} \psi_j \log\lambda - n\lambda + c \qquad (3.3) \\ &= \sum_{i=1}^{n} (X_i - \psi_i) \log\theta - N_{n-1}\theta + \sum_{j=1}^{n} \psi_j \log\lambda - n\lambda + c, \end{aligned}$$

where $c$ is a constant that depends only on $\mathbf{X}$ but not on $\eta$. Therefore, the (unrestricted) maximum likelihood estimate $\widehat{\eta} = (\widehat{\theta}, \widehat{\lambda})$ is given by

$$\widehat{\theta} = \sum_{i=1}^{N_{n-1}} \xi_i / N_{n-1}, \quad \widehat{\lambda} = \sum_{j=1}^{n} \psi_j / n,$$

while the constrained maximum likelihood estimate $\widehat{\eta}(\theta)$ is equal to $(\theta, \widehat{\lambda})$, so the log likelihood ratio statistic $l(\theta)$ can be simplified to

$$l(\theta) = 2N_{n-1}\{(\theta - \widehat{\theta}) + \widehat{\theta}\log(\widehat{\theta}/\theta)\}. \qquad (3.4)$$

The function $l^{\pm}(\theta) = \mathrm{sgn}(\widehat{\theta} - \theta) l^{1/2}(\theta)$ is decreasing in $\theta$ and therefore the bootstrap confidence region (2.2) with $R(\mathbf{X}, \theta) = l^{\pm}(\theta)$ is an interval, the endpoints of which are obtained by solving the equations $l^{\pm}(\theta) = u^*_{\alpha}$ and $l^{\pm}(\theta) = u^*_{1-\alpha}$.

Table 1 reports a simulation study comparing the bootstrap and hybrid methods for finding confidence intervals for $\theta$ with $l^{\pm}(\theta)$ as the root $R(\mathbf{X}, \theta)$. In the simulation study, $x_0 = 10, n = 10$ and $\lambda = 0.25$. We simulated 2000 sets of data for values of $\theta$ ranging from 0.8 to 1.5, and used the secant method described at the end of Section 2 to compute the hybrid and bootstrap confidence intervals explicitly. Also given for comparison in Table 1 are the coverage errors of the normal and bootstrap-$t$ confidence limits. The normal limits are based on normal approximation for $l^{\pm}(\theta)$ so that the lower confidence limit is obtained by solving the equation $l^{\pm}(\theta) = \Phi^{-1}(1 - \alpha)$, where $\Phi$ denotes the standard normal distribution function. The bootstrap-$t$ confidence limits use the Studentized root

$$R(\mathbf{X}, \theta) = (\widehat{\theta} - \theta)/(\widehat{\theta}/N_{n-1})^{1/2} \qquad (3.5)$$

since $l''(\widehat{\theta})/2 = N_{n-1}/\widehat{\theta}$. An advantage of using the linear function (3.5) of $\theta$ as the root is that (2.2) reduces simply to $\widehat{\theta} - (\widehat{\theta}/N_{n-1})^{1/2} u^*_{\alpha} \geq \theta \geq \widehat{\theta} - (\widehat{\theta}/N_{n-1})^{1/2} u^*_{1-\alpha}$, which is the usual bootstrap-$t$ interval. The nominal coverage error $\alpha$ for the upper and lower confidence bounds is 5%. With this choice of $\alpha$, we used resample sizes of 999 to compute $u^*_{\alpha}, u^*_{1-\alpha}, \widehat{u}_{\alpha}(\theta), \widehat{u}_{1-\alpha}(\theta)$ for the bootstrap, bootstrap-$t$ and hybrid confidence limits, following a suggestion by Davison and Hinkley (1997).

Table 1. Coverage errors in % for lower (L) and upper (U) confidence limits based on 2000 simulations for the offspring mean $\theta$ in a Galton-Watson BPI with $x_0 = 10$, $n = 10$ and an immigration rate of $\lambda = 0.25$. Nominal coverage errors are 5%.

| | Normal | | Bootstrap | | Bootstrap-$t$ | | Hybrid | |
|---|---|---|---|---|---|---|---|---|
| $\theta$ | L | U | L | U | L | U | L | U |
| 0.80 | 3.10 | 7.45 | 4.60 | 5.00 | 4.00 | 7.95 | 4.75 | 5.35 |
| 0.95 | 2.50 | 9.50 | 3.85 | 6.45 | 3.25 | 8.90 | 3.90 | 5.65 |
| 1.00 | 3.30 | 10.50 | 4.60 | 6.95 | 4.15 | 9.35 | 4.95 | 5.70 |
| 1.05 | 4.45 | 8.15 | 5.75 | 5.75 | 5.35 | 7.35 | 5.95 | 4.55 |
| 1.20 | 3.80 | 7.70 | 4.55 | 5.15 | 4.30 | 5.55 | 4.90 | 5.15 |
| 1.50 | 3.65 | 5.90 | 4.40 | 4.20 | 4.30 | 4.65 | 4.35 | 4.55 |

Table 1 shows that the hybrid confidence limits have coverage errors close to the nominal value of 5% for all cases considered. The bootstrap confidence limits have coverage errors reasonably close to the nominal value of 5% except for the upper limit when $\theta = 0.95$ or 1. In contrast, the bootstrap-$t$ confidence limits based on the Studentized root have coverage errors that differ substantially from the nominal value of 5% in about half the cases considered. The normal confidence limits have poor coverage except in the case $\theta = 1.5$. An explanation for some of these results is provided by Lemma 1, whose proof is given in the Appendix and in which $Y_t$ is the highly non-Gaussian "square-root diffusion process" in mathematical finance (cf. Cox, Ingersoll and Ross (1985)). In fact, since the distribution of $l^{\pm}(\theta)$ changes drastically for $\theta$ near 1 in view of Lemma 1, the bootstrap and normal confidence limits are not valid for $\theta$ near 1, but the hybrid resampling method can be used to overcome the difficulty.

**Lemma 1.** *As $n \to \infty$, $l^{\pm}(\theta)$ has a limiting standard normal distribution if $\theta \neq 1$. If $\theta = 1$, then $l^{\pm}(\theta)$ converges weakly to $(Y_1 - \lambda)/(\int_0^1 Y_t dt)^{1/2}$, where $Y_t$ satisfies the stochastic differential equation $dY_t = \lambda dt + Y_t^{1/2} dB_t$, $Y_0 = 0$ and $B_t$ is standard Brownian motion.*

The preceding construction of the hybrid confidence region (2.3) uses the obvious choice $\widehat{F}_\eta = \widehat{F}_{\widehat{\eta}(\theta)}$ for the resampling family. By conditioning on $\widehat{\lambda}$ to construct an alternative resampling family and by modifying slightly $l^{\pm}(\theta)$ as the choice of the root, we obtain below a hybrid confidence region (2.3) with exact coverage probability $1 - 2\alpha$. Because $l^{\pm}(\theta)$ is discrete, its distribution function may not assume the values $\alpha$ and $1 - \alpha$. However, adding an independent uniform random variable on $(-n^{-1}, n^{-1})$ to $l^{\pm}(\theta)$ yields a continuous distribution function. This randomized version of $l^{\pm}(\theta)$ is used as the root in the following lemma, whose proof is given in the Appendix.

**Lemma 2.** *Let $R(\mathbf{X},\theta) = l^{\pm}(\theta) + n^{-1}U$, where $U$ is uniformly distributed on $(-1,1)$ and is independent of $\mathbf{X}$.*

(i) *Let $\widehat{F}_\theta$ be the distribution of the Galton-Watson BPI in which $\xi_1^*, \xi_2^*, \ldots$ are i.i.d. Poisson with mean $\theta$ and $(\psi_1^*, \ldots, \psi_n^*)$ has the multinomial distribution $M(n\widehat{\lambda}; n^{-1}, \ldots, n^{-1})$ corresponding to $n\widehat{\lambda}$ independent trials, with $n$ equally likely outcomes in each trial. With $R(\mathbf{X},\theta)$ and $\widehat{F}_\theta$ thus defined, the hybrid confidence region (2.3) has coverage probability $1-2\alpha$.*

(ii) *The conclusion of* (i) *still holds if $\widehat{F}_\theta$ is the distribution function under which $\psi_i^* = \psi_i$ for $1 \le i \le n$ and the $\xi_i^*$ are i.i.d. Poisson with mean $\theta$.*

In Example 3, since $l(\widehat{\theta}) = l'(\widehat{\theta}) = 0$, the leading term in the Taylor expansion of (3.1) is the square of (3.5). More generally, for smooth parametric models, (3.2) is asymptotically equivalent to the Studentized root

$$t(\theta) = (\widehat{\theta} - \theta)/\widehat{\sigma}. \tag{3.6}$$

Here $\widehat{\sigma}$ is the estimated standard error of $\widehat{\theta} = g(\widehat{\eta})$ given by

$$\widehat{\sigma}^2 = -\widehat{\nabla}^T(\nabla^2 \log f(\mathbf{X},\eta))^{-1}\Big|_{\eta=\widehat{\eta}}\,\widehat{\nabla}, \quad \widehat{\nabla} = \nabla g(\eta)|_{\eta=\widehat{\eta}}, \tag{3.7}$$

in which $\nabla g$ represents the column vector of partial derivatives (gradient vector) of $g$, while $\nabla^2 g$ represents the matrix of second partial derivatives (Hessian matrix). Table 1 shows that replacing $l^{\pm}(\theta)$ by its linear approximation $t(\theta)$ even when $\widehat{\theta}$ and $\theta$ are many standard errors apart may result in inferior performance. On the other hand, for nonlinear multiparameter problems, $\widehat{\eta}(\theta)$ and therefore $l^{\pm}(\theta)$ also may be difficult to compute. To reduce the computational task in simulating the distribution of $l^{\pm}(\theta)$ under $\widehat{\eta}(\theta)$, we replace $l^{\pm}(\theta)$ by $t(\theta)$ when $\theta$ is reasonably close to $\widehat{\theta}$ and arrive at the root

$$R(\mathbf{X},\theta) = \begin{cases} t(\theta), & \text{if } |t(\theta)| \le M, \\ l^{\pm}(\theta), & \text{if } |t(\theta)| > M. \end{cases} \tag{3.8}$$

Likewise, to construct the resampling family, we replace $\widehat{\eta}(\theta)$ by its linear approximation

$$\widetilde{\eta}(\theta) = \widehat{\eta} - \widehat{\sigma}^{-2}(\theta - \widehat{\theta})(\nabla^2 \log f(\mathbf{X},\eta))^{-1}\Big|_{\eta=\widehat{\eta}}\,\widehat{\nabla} \tag{3.9}$$

when $|t(\theta)| \le \widetilde{M}$, or equivalently, when $\theta^l \le \theta \le \theta^u$, where $\theta^l = \widehat{\theta} - \widehat{\sigma}\widetilde{M}$ and $\theta^u = \widehat{\theta} + \widehat{\sigma}\widetilde{M}$. When $t(\theta)$ is an approximate pivot, we can choose the resampling family $\{\widehat{F}_\theta\}$ to be the singleton $\{F_{\widehat{\eta}}\}$, which is tantamount to the bootstrap method. When $t(\theta)$ is not an approximate pivot, we define

$$\widehat{F}_\theta = \begin{cases} F_{\widetilde{\eta}(\theta)}, & \text{if } \theta^l \le \theta \le \theta^u, \\ F_{\widetilde{\eta}(\theta^u)}, & \text{if } \theta > \theta^u, \\ F_{\widetilde{\eta}(\theta^l)}, & \text{if } \theta < \theta^l. \end{cases} \tag{3.10}$$

This means that the quantiles $\widehat{u}_\alpha(\theta)$ and $\widehat{u}_{1-\alpha}(\theta)$ in the hybrid confidence region (2.3) are obtained from the distribution of (3.8) under $\widehat{F}_\theta$ if $\theta^l \le \theta \le \theta^u$, and are obtained by extrapolation for $\theta$ outside this range. Choosing $M$ and $\widetilde{M}$ to be of the order of $\log n$ in the preceding gives asympotically correct confidence limits, as will be explained in Section 3.2 and Section 6, where the root (3.8) and the resampling family (3.10) will be extended to nonparametric problems.

Besides equal-tailed confidence regions, the Studentized root (3.6) has been used via its absolute value to construct symmetric bootstrap confidence intervals (cf. Section 3.6 of Hall (1992)). A refinement of this approach along the lines of (3.8) yields the nonnegative root

$$R(\mathbf{X},\theta) = \begin{cases} t^2(\theta) \text{ if } |t(\theta)| \le M, \\ l(\theta) \quad \text{if } |t(\theta)| > M, \end{cases} \tag{3.11}$$

and a hybrid confidence region associated with this root is $\{\theta : R(\mathbf{X},\theta) \le \widehat{u}_{1-2\alpha}(\theta)\}$, with nominal two-sided coverage error $2\alpha$.

### 3.2. Choice of root for nonparametric models and a correlation coefficient example

Let $X_1,\ldots,X_n$ be i.i.d. random variables having a common unknown distribution $F$. Without assuming that $F$ belongs to some parametric family, the nonparametric maximum likelihood estimate of $F$ is the empirical distribution $\widehat{F} = n^{-1}\sum_{i=1}^n \delta_{X_i}$. Letting $\widehat{F}^{(\theta_0)}$ be the distribution that maximizes the empirical likelihood $f(\mathbf{X};F) = \prod_{i=1}^n F(\{X_i\})$ among all distributions $F \ll \widehat{F}$ subject to the constraint $\theta(F) = \theta_0$, Owen (1988, 1990) extended the log likelihood ratio statistic (3.1) to the nonparametric case and obtained the empirical likelihood ratio statistic

$$l(\theta_0) = 2\{\log f(\mathbf{X};\widehat{F}) - \log f(\mathbf{X},\widehat{F}^{(\theta_0)})\}. \tag{3.12}$$

Under certain conditions on the functional $\theta(F)$, Owen (1988, 1990) showed that for fixed sample sizes, $l(\theta_0)$ has a limiting chi-square distribution with 1 degree of freedom.

Let $p_i = F(\{X_i\})$. To maximize $\prod_{i=1}^n p_i$ subject to the constraint $\theta(F) = \theta_0$, one can introduce a Lagrange multiplier $\lambda$ associated with the constraint. Differentiation with respect to $p_i$ and $\lambda$ leads to $n+1$ equations which are the first-order conditions for the constrained optimization problem. When $\theta(F)$ is the mean of $F$, Owen (1990) combined these equations into an equation

$$\sum_{i=1}^n (X_i - \theta_0)/\{1 + \lambda(X_i - \theta_0)\} = 0 \tag{3.13}$$

for $\lambda$, with $p_i^{-1} = n\{1 + \lambda(X_i - \theta_0)\}$. Section 6 of Owen (1988) and Section 10.2 of Davison and Hinkley (1997) give numerical results on the use of (3.12) as the root $R(\mathbf{X}, \theta)$ in forming bootstrap confidence regions $\{\theta : R(\mathbf{X}, \theta) \le u^*_{1-2\alpha}\}$ for the mean of $F$. We can clearly extend the idea to construct equal-tailed bootstrap confidence regions (2.2) and hybrid confidence regions (2.3) for other functionals $\theta(F)$, with $R(\mathbf{X}, \theta) = \text{sgn}(\widehat{\theta} - \theta) l^{1/2}(\theta)$. When $\theta(F)$ is a smooth function of several population means, Owen (1990) developed a nested algorithm to maximize $\prod_{i=1}^n p_i$ subject to $\theta(F) = \theta_0$. However, this algorithm is computationally intensive and it is impractical to repeat many times such computations in simulating the distribution of $l(\theta)$ or the signed root $l^{\pm}(\theta)$ for the bootstrap and hybrid resampling methods. For more complicated nonlinear functionals $\theta(F)$, one does not even have nested algorithms to reduce dimensionality in finding $p_1, \ldots, p_n$.

A much simpler root than the empirical likelihood ratio statistic is the Studentized statistic $t(\theta)$ defined by (3.6) with

$$\widehat{\sigma}^2 = n^{-2} \sum_{i=1}^n U_i^2(\widehat{F}), \quad U_i(F) = \lim_{\epsilon \to 0} \epsilon^{-1} \{\theta((1-\epsilon)F + \epsilon \delta_{X_i}) - \theta(F)\}. \tag{3.14}$$

Note that $U_i(F)$ is the influence function associated with $\theta(F)$ and $\widehat{\sigma}^2$ is the infinitesimal jackknife variance estimate. When $\theta(F)$ is a linear functional of $F$, $\theta(\widehat{F}) = \theta(F) + n^{-1} \sum_{i=1}^n U_i(F)$ and Owen (1990, p.102) shows that $l(\theta_0) = t^2(\theta_0) + o_p(1)$ under the hypothesis $\theta(F) = \theta_0$. For nonlinear but differentiable $\theta(\cdot)$, although consistency of $\widehat{F}$ and the functional delta method yield the linear approximation $\theta(\widehat{F}) \doteq \theta(F) + n^{-1} \sum_{i=1}^n U_i(F)$, this approximation may be highly inadequate in samples of moderate size, where the infinitesimal jackknife variance estimate $\widehat{\sigma}^2$ has been found to be unreliable and unstable. This instability has been linked to unsatisfactory performance of bootstrap-$t$ confidence intervals based on the Studentized root defined by (3.6) and (3.14) despite the attractive asymptotic properties of these intervals.

Whereas $l^{\pm}(\theta)$ may be difficult to compute in practice, it is more stable than $t(\theta)$ since it does not require an auxiliary variance estimate because of its "self-normalizing" property. When $\widehat{\theta}$ is close to $\theta$, $l^{\pm}(\theta)$ should be close to $t(\theta)$ as in the parametric case. This suggests that, as in (3.8), we replace $l^{\pm}(\theta)$ by $t(\theta)$ when $|t(\theta)| \le M$ (so that $\widehat{\theta}$ is not too far from $\theta$). When $|t(\theta)| > M$, $\widehat{\sigma}$ may be erratic and we should use some better-behaved alternative to $t(\theta)$. Because of the computational complexity, it is not practical to use $l^{\pm}(\theta)$ even though $|t(\theta)| > M$ may not occur frequently in the simulations. We therefore modify (3.8) as

$$R(\mathbf{X}, \theta) = \begin{cases} t(\theta), & \text{if } |t(\theta)| \le M, \\ d(\theta), & \text{if } |t(\theta)| > M, \end{cases} \tag{3.15}$$

where the choice of $M$ and $d(\theta)$ is discussed below.

Methods in the bootstrap literature that bypass estimation of the standard error of $\widehat{\theta} = \theta(\widehat{F})$ include Efron's (1987) $\mathrm{BC}_a$ method, and bootstrap calibration of the percentile intervals proposed by Beran (1987) and Loh (1987). As pointed out by Bickel (1987), the $\mathrm{BC}_a$ method uses an approximate pivot of the form $\widehat{S}^{-1}(\Phi^{-1}(G_{\widehat{F}}(\theta)))$, where $G_F$ denotes the distribution of $\widehat{\theta}$ under $F$ (so the bootstrap distribution of $\widehat{\theta}$ under $\widehat{F}$ is given by $G_{\widehat{F}}$) and

$$\widehat{S}(z) = \widehat{z}_0 + (\widehat{z}_0 + z)/\{1 - \widehat{a}(\widehat{z}_0 + z)\}, \tag{3.16}$$

in which $\widehat{z}_0 = \Phi^{-1}(G_{\widehat{F}}(\widehat{\theta}))$ and $\widehat{a} = \frac{1}{6}\{\sum_{i=1}^n U_i^3(\widehat{F})\}/\{\sum_{i=1}^n U_i^2(\widehat{F})\}^{3/2}$ are the bias-correction and acceleration constants. The calibrated percentile method starts with Efron's (1981) percentile interval based on percentiles of the bootstrap distribution to define the upper and lower confidence limits for each nominal coverage error $\lambda$, and then uses the bootstrap method to calibrate the actual coverage error $\widehat{p}(\lambda)$ and chooses the upper/lower confidence limit with $\widehat{p}(\lambda) = \alpha$. Therefore the equal-tailed calibrated percentile interval is a special case of (2.2) with $R(\mathbf{X}, \theta) = G_{\widehat{F}}(\theta)$. Because this confidence region is invariant under monotone transformations, we can also take $(1-\Phi)^{-1}(G_{\widehat{F}}(\theta))$ as the root. When $\widehat{\theta}$ is a smooth function of sample means, standard results in Hall (1988, 1992) show that the bootstrap distribution has an Edgeworth-type expansion

$$G_{\widehat{F}}(\theta) = P\{\widehat{\theta}^* \le \theta | \widehat{F}\} = \Phi(-t(\theta)) + n^{-1/2} p(-t(\theta))\phi(-t(\theta)) + O_p(n^{-1}), \tag{3.17}$$

where $p(\cdot)$ is a polynomial and $\phi(\cdot)$ is the standard normal density. Since $\Phi(-x) = 1 - \Phi(x)$, it follows from (3.17) that $(1 - \Phi)^{-1}(G_{\widehat{F}}(\theta)) = t(\theta) + O_p(n^{-1/2})$. In view of this connection between $t(\theta)$ and $(1-\Phi)^{-1}(G_{\widehat{F}}(\theta))$, we propose to choose $d(\theta)$ in (3.15) to be

$$d(\theta) = (1 - \Phi)^{-1}(G_{\widehat{F}}(\theta)). \tag{3.18}$$

To avoid infinite values in (3.18), redefine $d(\theta)$ as $(1-\Phi)^{-1}(1/(2n^*))$ if $G_{\widehat{F}}(\theta) = 0$, and as $(1 - \Phi)^{-1}(1 - 1/(2n^*))$ if $G_{\widehat{F}}(\theta) = 1$, for some $n^* \ge n$. With $d(\theta)$ thus defined, choosing $M$ to be some constant times of $\log n$ in (3.15) yields

$$P\{|t(\theta)| > M\} = O(n^{-a}) \text{ for any } a > 0, \tag{3.19}$$

when $\widehat{\theta}$ is an infinitely differentiable function of means of i.i.d. random vectors whose common moment generating function is finite in some neighborhood of the origin.

In view of (3.19), $t(\theta)$ is asymptotically equivalent to the root $R(\mathbf{X}, \theta)$ defined by (3.15) and (3.18). The latter, however, is considerably more stable than

$t(\theta)$ for nonlinear functionals of $\widehat{F}$ in samples of small to moderate size, since we use percentiles of the bootstrap distribution in (3.18) rather than the infinitesimal jackknife standard error estimate of $\widehat{\theta}$ when $\widehat{\theta}$ is not near $\theta$. Although the root $\widehat{S}^{-1}(\Phi^{-1}(G_{\widehat{F}}(\theta)))$ in Efron's $\mathrm{BC}_a$ method is an attractive alternative to $t(\theta)$, the bias-correction and acceleration constants in (3.16) involve linearization arguments which may not be appropriate when $\widehat{\theta}$ is not near $\theta$, so we do not choose it for $d(\theta)$. Using the root (3.15) with $d(\theta)$ given by (3.18) in (2.2) yields a confidence region whose infimum L and supremum U are given by

$$\begin{aligned} \mathrm{L} &= \mathrm{L}'1\{\mathrm{L}' < \widehat{\theta} - \widehat{\sigma}M\} + \max(\widehat{\theta} - \widehat{\sigma}u^*_{1-\alpha}, \widehat{\theta} - \widehat{\sigma}M)1\{\mathrm{L}' \geq \widehat{\theta} - \widehat{\sigma}M\}, \\ \mathrm{U} &= \mathrm{U}'1\{\mathrm{U}' > \widehat{\theta} + \widehat{\sigma}M\} + \min(\widehat{\theta} - \widehat{\sigma}u^*_{\alpha}, \widehat{\theta} + \widehat{\sigma}M)1\{\mathrm{U}' \leq \widehat{\theta} + \widehat{\sigma}M\}, \end{aligned} \quad (3.20)$$

where $\mathrm{L}' = G^{-1}_{\widehat{F}}(1-\Phi(u^*_{1-\alpha}))$, $\mathrm{U}' = G^{-1}_{\widehat{F}}(1-\Phi(u^*_{\alpha}))$ and $1\{\cdot\}$ is the usual indicator function.

Analogous to (3.11), we can extend the preceding idea to obtain a nonnegative root of the form

$$R(\mathbf{X}, \theta) = \begin{cases} |t(\theta)|, & \text{if } |t(\theta)| \leq M, \\ (1-\Phi)^{-1}(\max\{1/(2n^*), \min(G_{\widehat{F}}(\theta), 1 - G_{\widehat{F}}(\theta))\}), & \text{if } |t(\theta)| > M, \end{cases} \quad (3.21)$$

noting that $(1-\Phi)^{-1}(\min\{G_{\widehat{F}}(\theta), 1 - G_{\widehat{F}}(\theta)\}) = |t(\theta)| + O_p(n^{-1/2})$ by (3.17). A bootstrap confidence region associated with this root is $\{\theta : R(\mathbf{X}, \theta) \leq u^*_{1-2\alpha}\}$, with nominal (two-sided) coverage error $2\alpha$. When $\widehat{\theta}$ is a smooth function of sample means, Hall (1988, 1992) showed that for the equal-tailed bootstrap confidence interval based on $t(\theta)$ as the root, both the upper and lower confidence bounds have coverage error $\alpha + O(n^{-1})$, and that the symmetric bootstrap confidence interval based on $|t(\theta)|$ as the root has coverage error $2\alpha + O(n^{-2})$. When $M$ in (3.15) and (3.21) is chosen to be some constant times $\log n$, the modified bootstrap-$t$ confidence region with root given by (3.15) and (3.18), or by (3.21), also has coverage error differing from the nominal value by $O(n^{-1})$, or by $O(n^{-2})$ in the two-sided case, in view of (3.19). However, the modified bootstrap-$t$ confidence region has much more stable finite-sample behavior, as illustrated below in the problem of constructing a confidence interval for a correlation coefficient, called by Hall (1992, p.152) a "smoking gun" of bootstrap methods. Note that choosing $M = \infty$ in (3.15) and (3.18) (or (3.21)) gives the equal-tailed (or symmetric) bootstrap-$t$ interval, while choosing $M = 0$ in (3.15) and (3.18) (or (3.21)) gives the equal-tailed (or symmetric) calibrated percentile interval. Using the root (3.15) or (3.21) with suitably chosen $M > 0$ combines the computational simplicity of the bootstrap-$t$ method and the stability of the much more computationally expensive calibrated percentile method. Note that it involves a second

layer of bootstrapping only for a small fraction of the resamples, namely, those with $|t(\theta)| > M$. When we have a stable estimate $\widehat{\sigma}$, as in mean and regression problems, we can simply take $M = \infty$.

**Example 4.** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. bivariate vectors from some unknown distribution $F$ with correlation coefficient $\theta$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i, \bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$, $\widehat{\sigma}^2_{x,n} = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, $\widehat{\sigma}^2_{y,n} = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, and let $\widehat{\theta}_n = (n\widehat{\sigma}_{x,n}\widehat{\sigma}_{y,n})^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$ be the sample correlation coefficient. The infinitesimal jackknife variance estimate is $\widehat{\sigma}^2 = n^{-2}\sum_{i=1}^n U_i^2$, where

$$U_i = X_i' Y_i' - \widehat{\theta}(X_i'^2 + Y_i'^2)/2, \quad X_i' = (X_i - \bar{X}_n)/\widehat{\sigma}_{x,n}, \quad Y_i' = (Y_i - \bar{Y}_n)/\widehat{\sigma}_{y,n}. \tag{3.22}$$

Since $\theta$ is a smooth function of the means $EX_i, EY_i, EX_i^2, EY_i^2$ and $EX_iY_i$, the bootstrap-$t$ upper and lower confidence bounds are second-order accurate. However many authors, notably Efron (1982, 1987) have observed that they can be very long, with endpoints larger than 1 in absolute value, and that $\widehat{\sigma}$ can be erratic and grossly inaccurate. Tables 2 and 3 give the results of a simulation study comparing the performance of the equal-tailed bootstrap-$t$ interval with its modification (3.20) with $n = 30$, $M = 1.5$ and $\alpha = 5\%$. Also included for comparison are the interval $\widehat{\theta} \pm 1.645\widehat{\sigma}$ based on the normal approximation, the parametric interval involving Fisher's transformation (cf. Efron and Tibshirani (1993, pp.54, 163)) which is nearly exact in the bivariate normal case, and Efron's (1987) BC$_a$ interval (see (3.16) above). The first three rows of the tables consider the bivariate normal population with correlation coefficient 0, 0.5, and 0.9 respectively. The fourth row of each table considers the regression model $Y_i = X_i/2 + \sqrt{3}\epsilon_i/2$, in which $X_1, X_2, \ldots, \epsilon_1, \epsilon_2, \ldots$ are independent and have a common double exponential distribution with mean 0 and variance 1 so that $\mathrm{Var}(Y_i) = 1$ and the correlation coefficient between $X_i$ and $Y_i$ is 0.5. The fifth row of each table considers the case where $(\log X_i, \log Y_i)$ has a bivariate normal distribution with zero means, unit variances and correlation coefficient 0.5, so that $(X_i, Y_i)$ has the bivariate lognormal distribution with correlation coefficient $(1 + \sqrt{e})^{-1}$, as considered in the simulation study in Hall, Schucany and Martin (1989). We used bootstrap resamples of size 999 to compute the quantiles $u^*_\alpha$ and $u^*_{1-\alpha}$ in (3.20), the bootstrap-$t$ limits and $\widehat{z}_0 = \Phi^{-1}(G_{\widehat{F}}(\widehat{\theta}))$ for the BC$_a$ limits (3.16). To compute $u^*_\alpha$ and $u^*_{1-\alpha}$ in (3.20), whenever $t(\widehat{\theta}) = (\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*$ exceeded $M = 1.5$ for a given resample, a second layer of 299 resamples was used to compute $d(\widehat{\theta})$, taking $n^* = 300$. The estimates of the coverage errors are summarized in Table 2, whereas the average confidence limits are summarized in Table 3.

Table 2. Coverage errors in % for lower (L) and upper (U) confidence limits based on 2000 simulations for the correlation coefficient $\theta$. The first three distributions considered are bivariate normal distributions. The fourth row indicated by 0.5 Exp uses a regression model with double exponential regressors and errors so that $\theta = 0.5$. The fifth row indicated by 0.5 Log uses a bivariate lognormal distribution with $\theta = (1 + \sqrt{e})^{-1}$ (whose underlying bivariate normal distribution has correlation 0.5). The sample size is 30. Nominal coverage errors are 5%.

| | Normal | | Bootstrap-$t$ | | Modified $t$ | | $BC_a$ | | Fisher | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | L | U | L | U | L | U | L | U | L | U |
| 0.0 | 8.05 | 7.30 | 4.05 | 3.85 | 4.75 | 4.20 | 5.50 | 5.25 | 4.20 | 4.25 |
| 0.5 | 10.80 | 4.80 | 5.05 | 3.60 | 5.30 | 4.20 | 5.95 | 6.00 | 5.65 | 4.40 |
| 0.9 | 13.65 | 2.45 | 5.30 | 4.95 | 4.85 | 5.25 | 5.95 | 6.10 | 5.50 | 4.35 |
| 0.5 Exp | 13.60 | 6.80 | 6.05 | 5.40 | 6.10 | 5.65 | 7.55 | 7.00 | 8.50 | 6.50 |
| 0.5 Log | 18.50 | 13.30 | 6.40 | 11.40 | 6.05 | 6.50 | 6.75 | 9.20 | 18.40 | 4.25 |

Table 3. Average lower (L) and upper (U) confidence limits based on 2000 simulations for the correlation coefficient. The distributions used are the same as those used for Table 2.

| | Normal | | Bootstrap-$t$ | | Modified $t$ | | $BC_a$ | | Fisher | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | L | U | L | U | L | U | L | U | L | U |
| 0.0 | −0.27 | 0.28 | −0.34 | 0.34 | −0.31 | 0.31 | −0.29 | 0.29 | −0.29 | 0.30 |
| 0.5 | 0.28 | 0.71 | 0.18 | 0.71 | 0.20 | 0.70 | 0.22 | 0.68 | 0.23 | 0.69 |
| 0.9 | 0.84 | 0.95 | 0.80 | 0.94 | 0.80 | 0.94 | 0.81 | 0.94 | 0.81 | 0.94 |
| 0.5 Exp | 0.28 | 0.72 | 0.13 | 0.74 | 0.18 | 0.71 | 0.21 | 0.70 | 0.24 | 0.69 |
| 0.5 Log | 0.18 | 0.64 | −0.05 | 0.80 | 0.08 | 0.69 | 0.09 | 0.65 | 0.14 | 0.63 |

As Table 2 shows, both the bootstrap-$t$ interval and its modified version (3.20) have coverage errors reasonably close to the nominal value of 5% for bivariate normal distributions and the regression model considered. In contrast, the interval based on normal approximation has rather poor coverage. However, (3.20) is shorter on average than the bootstrap-$t$ interval. Figure 1 shows a boxplot of the lower and upper confidence limits for the 2000 sets of simulated data from the regression model with double exponential errors. Here, the occasional erratic behavior of the bootstrap-$t$ limits becomes apparent, whereas its modification (3.20) is considerably more stable. The $BC_a$ confidence limits are also stable, and the interval tends to be shorter than (3.20) but undercovers the true value $\theta = 0.5$ in this regression model. The advantage of the modified version (3.20) over the bootstrap-$t$ method is also apparent when the population is lognormal, where the bootstrap-$t$ upper limit has coverage error exceeding twice the nominal coverage error of 5%.

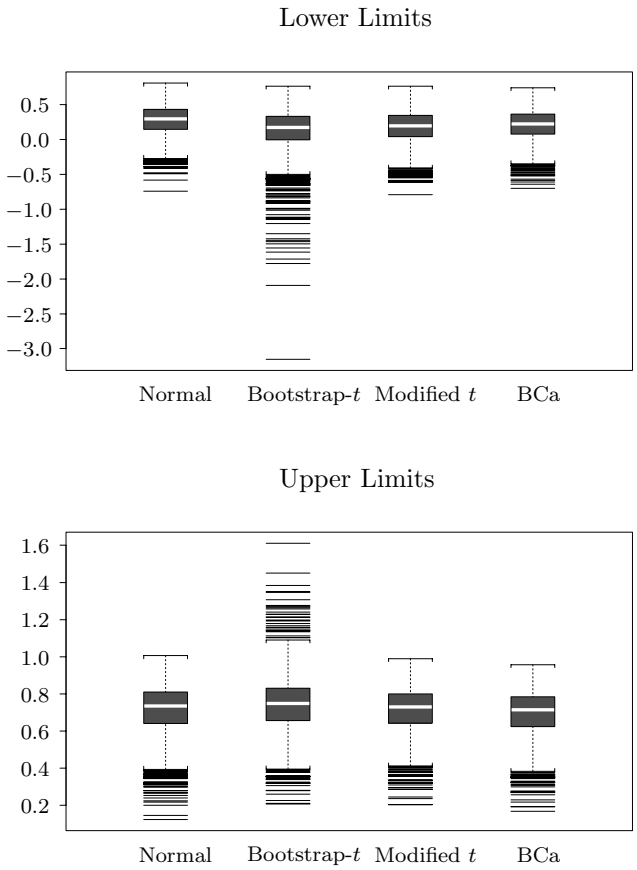


Figure 1. Boxplots of lower and upper confidence limits for the correlation coefficient based on 2000 simulated data sets from a double exponential regression family with correlation 0.5.

As we have pointed out after (3.21), the symmetric bootstrap interval based on using $|t(\theta)|$ as the root and the calibrated symmetric percentile interval both have coverage errors that differ from the nominal value by $O(n^{-2})$. Consequently, we expect the two-sided coverage error to be closer to the nominal value than the one-sided coverage errors of equal-tailed intervals. Indeed, based on 2000 simulations when the root (3.21) with $M = 1.5$ was used to obtain the modified symmetric $t$ interval with samples of size 30 generated from the lognormal distribution, the simulated coverage error is 9.35% when the nominal coverage error is 10%. This agrees qualitatively with results in Hall, Martin and Schucany (1989), who reported excellent coverage properties of the calibrated symmetric percentile interval for the correlation coefficient in small sample sizes. Lee and Young (1995) use analytic approximations to reduce the computational task of

the calibrated symmetric percentile interval, but their intervals appear to substantially undercover the true correlation for this lognormal distribution. Their Table 4 reports that the estimated coverage error is about 15% when the nominal coverage error is 10% and the sample size is 30. Section 6 contains some further discussion of this lognormal example.

As has been pointed out by Efron and Tibshirani (1993), the bootstrap variance estimate $\widehat{\sigma}_b^2$ is usually more stable and reliable than the infinitesimal variance estimate $\widehat{\sigma}^2$. However, bootstrapping the root $(\widehat{\theta}-\theta)/\widehat{\sigma}_b$ is computationally expensive. The computational burden can be substantially reduced by using the root (3.15) with $d(\theta) = (\widehat{\theta}-\theta)/\widehat{\sigma}_b$. We call this modified root the $t_*$–root. In a simulation study for the correlation coefficient with $n = 30$ and $M = 1.5$, we found the $t_*$–root gave results similar to those in Tables 2 and 3 for the bootstrap-$t$ interval when a second layer of 100 resamples was used to compute $\widehat{\sigma}_b^*$ whenever $d(\theta)$ was used. We have focused so far only on the choice of root for nonparametric models. In Section 6, the choice of the resampling family $\{\widehat{F}_\theta\}$ for implementing the hybrid resampling method in nonparametric problems will be discussed systematically. For the mean and regression problems considered in Sections 4 and 5, there is a natural and simple resampling family. For highly nonlinear functionals $\theta = \theta(F)$, such as the correlation coefficient, Section 6 addresses complications and provides further discussion of Example 4 and also improvements of hybrid over bootstrap confidence intervals for the correlation coefficient based on sequential samples.

## 4. Hybrid Confidence Regions Following Group Sequential Tests

We have found hybrid resampling methods particularly useful in statistical inference after group sequential tests. When the testing procedure is fully sequential, in that one decides whether to stop or continue collecting data based on review after each new observation has been collected, Woodroofe (1986, 1992) and Coad and Woodroofe (1996) have developed techniques based on "very weak expansions". These expansions, however, depend on the structure of exponential families and are difficult to extend to nonparametric problems. In this section we show how the hybrid resampling method approach can be used to construct confidence intervals for population means following group sequential tests. Possible extensions are given at the end of the section.

### 4.1. Hybrid resampling methods for population mean

In a group sequential test with $k$ interim analyses, the number $\tau$ of observations may be either $n_1$, or $n_2, \ldots,$ or $n_k$, where $n_j$ is the number of observations available at the $j$th analysis; see Example 1 for an example of a group sequential stopping rule. Let $X_1, X_2, \ldots$ be i.i.d. random variables with unknown mean $\theta$

and known variance 1, and suppose the stopping rule $\tau$ depends on the sample sums $S_n = \sum_{i=1}^n X_i$ up to the stopping time. In the notation of Section 2, $\mathbf{X} = (X_1, \ldots, X_\tau, \tau)$ and $R(\mathbf{X}, \theta) = \sqrt{\tau}(\bar{X}_\tau - \theta)$. If $X_i$ are assumed to be standard normal, then the exact confidence region (2.1) corresponds to the method proposed by Rosner and Tsiatis (1988). An obvious way of extending Efron's bootstrap method to the present situation is the following. Use the empirical distribution $\widehat{F} = \widehat{F}_\tau$ of $X_i, 1 \le i \le \tau$, to generate $X_1^*, \ldots, X_{\tau^*}^*$, where $\tau^*$ is the stopping rule $\tau$ applied to $X_1^*, X_2^*, \ldots$. Let $u_\alpha^*$ and $u_{1-\alpha}^*$ denote the $\alpha$- and $(1-\alpha)$-quantiles of the distribution of $\sqrt{\tau^*}(\bar{X}_{\tau^*}^* - \bar{X}_\tau)$. The bootstrap confidence interval (2.2) is given by $\bar{X}_\tau - u_{1-\alpha}^*/\sqrt{\tau} \le \mu \le \bar{X}_\tau - u_\alpha^*/\sqrt{\tau}$. As pointed out in Chuang and Lai (1998), $\sqrt{\tau}(\bar{X}_\tau - \theta)$ fails to be an approximate pivot and the bootstrap method does not work well.

The hybrid resampling method uses the location family given by $\theta + \widehat{G}$ as the resampling family, where $\widehat{G}$ is the empirical distribution of $(X_i - \bar{X}_\tau)/\widehat{\sigma}_{x,\tau}, 1 \le i \le \tau$, and $\widehat{\sigma}_{x,\tau}^2 = \tau^{-1}\sum_{i=1}^\tau (X_i - \bar{X}_\tau)^2$. Let $\epsilon_1, \epsilon_2, \ldots$ be i.i.d. from $\widehat{G}$ and let $X_i(\theta) = \theta + \epsilon_i$. Let $\tau(\theta)$ be the stopping rule $\tau$ applied to $X_1(\theta), X_2(\theta), \ldots$. Let $\widehat{u}_\alpha(\theta)$ and $\widehat{u}_{1-\alpha}(\theta)$ be the $\alpha$- and $(1-\alpha)$-quantiles of the distribution of $(\tau(\theta))^{-1/2}(\epsilon_1 + \cdots + \epsilon_{\tau(\theta)})$. The hybrid confidence region (2.3) is then given by

$$\{\theta : \widehat{u}_\alpha(\theta) \le \sqrt{\tau}(\bar{X}_\tau - \theta) \le \widehat{u}_{1-\alpha}(\theta)\}. \tag{4.1}$$

Chuang and Lai (1998, Tables 1 and 2) report a simulation study comparing the exact, bootstrap, and hybrid methods for finding equal-tailed 90% confidence intervals for the stopping rule of Example 1 with normal observations. The hybrid method was found to yield results very similar to the exact method. Theorem 1 of Chuang and Lai (1998) also establishes the second-order accuracy of the hybrid confidence region (2.3), i.e. (4.1) has coverage error $2\alpha + O(n_k^{-1})$ under certain regularity conditions.

### 4.2. Bivariate confidence regions for two population means

The stopping rule $\tau$ in Section 4.1 involves the partial sum $S_n$ of i.i.d. random variables $X_i$ whose common mean $\theta$ is unknown and is to be estimated by a confidence interval. Suppose that one is also interested in estimating the common mean $\mu$ of i.i.d. random variables $Y_1, Y_2, \ldots$ which are observed up to the stopping time $\tau$. For example, consider a group sequential trial with multiple endpoints. Here $X_i$ represents a linear combination of the components, representing various endpoints, of a response vector (cf. Tang, Genecco and Geller (1989)) and $Y_i$ represents one of these components, corresponding to a given endpoint. We shall first assume that $X_i$ and $Y_i$ have common known variance 1. Let $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. Although $\sqrt{n}(\bar{Y}_n - \mu)$ is an asymptotic pivot having a limiting standard normal distribution, $\sqrt{\tau}(\bar{Y}_\tau - \mu)$ is no longer an asymptotic

pivot since its limiting distribution depends on $\theta$ that determines the distribution of $\tau$. We therefore consider a joint confidence region for both $\mu$ and $\theta$ in lieu of a confidence interval for $\mu$ alone.

First, suppose that $(X_i, Y_i)$ is bivariate normal with known correlation coefficient $\rho$. Let $V$ denote the covariance matrix of $(X_i, Y_i)$ having 1 as its diagonal entries and $\rho$ as the other entries. In the notation of Section 2, $\mathbf{X} = (X_1, \ldots, X_\tau; Y_1, \ldots, Y_\tau; \tau)$ and an exact $1-2\alpha$ confidence region for $(\theta, \mu)$ is

$$\{(\theta, \mu) : R(\mathbf{X}, \theta, \mu) \le u_{1-2\alpha}(\theta)\}, \tag{4.2}$$

where $R(\mathbf{X}, \theta, \mu) = \tau(\bar{X}_\tau - \theta, \bar{Y}_\tau - \mu)V^{-1}(\bar{X}_\tau - \theta, \bar{Y}_\tau - \mu)^T$, and $u_{1-2\alpha}(\theta)$ is the $(1-2\alpha)$-quantile of $R(\mathbf{X}, \theta, \mu)$ whose distribution depends on $\theta$ (that determines the distribution of $\tau$) but not on $\mu$ (since $Y_i - \mu$ is standard normal).

Without assuming $(X_i, Y_i)$ to be standard normal and $\rho$ to be known, we can replace $\rho$ in $V$ by the sample correlation coefficient $\widehat{\rho}_\tau$. Letting $\widehat{V}$ denote the matrix with 1 as its diagonal entries and $\widehat{\rho}_\tau$ elsewhere, we modify (4.2) as

$$\widehat{R}(\mathbf{X}, \theta, \mu) = \tau(\bar{X}_\tau - \theta, \bar{Y}_\tau - \mu)\widehat{V}^{-1}(\bar{X}_\tau - \theta, \bar{Y}_\tau - \mu)^T. \tag{4.3}$$

Define the sample variances $\widehat{\sigma}^2_{x,n}$ and $\widehat{\sigma}^2_{y,n}$ as in Example 4, and let $\widehat{G}$ be the empirical distribution of $((X_i - \bar{X}_\tau)/\widehat{\sigma}_{x,\tau}, (Y_i - \bar{Y}_\tau)/\widehat{\sigma}_{y,\tau}), 1 \le i \le \tau$. Let $(\epsilon_1, \eta_1), (\epsilon_2, \eta_2), \ldots$ be i.i.d. with common distribution $\widehat{G}$ and let $X_i(\theta) = \theta + \epsilon_i$. Let $\tau(\theta)$ be the stopping rule $\tau$ applied to $X_1(\theta), X_2(\theta), \ldots$. Using $\widetilde{\rho}_{\tau(\theta)}$ to denote the sample correlation coefficient of the $(\epsilon_i, \eta_i), 1 \le i \le \tau(\theta)$, we let $\widehat{V}_{\tau(\theta)}$ denote the matrix with 1 as the diagonal entries and $\widetilde{\rho}_{\tau(\theta)}$ elsewhere. Defining $\widehat{u}_{1-2\alpha}(\theta)$ as the $(1-2\alpha)$-quantile of

$$(\tau(\theta))^{-1}\Big(\sum_{i=1}^{\tau(\theta)} \epsilon_i, \sum_{i=1}^{\tau(\theta)} \eta_i\Big)\widehat{V}^{-1}_{\tau(\theta)}\Big(\sum_{i=1}^{\tau(\theta)} \epsilon_i, \sum_{i=1}^{\tau(\theta)} \eta_i\Big)^T, \tag{4.4}$$

the hybrid confidence region for $(\mu, \theta)$ with nominal coverage error $2\alpha$ is given by

$$\{(\theta, \mu) : \widehat{R}(\mathbf{X}, \theta, \mu) \le \widehat{u}_{1-2\alpha}(\theta)\}. \tag{4.5}$$

Without assuming known unit variance of $Y_i$, we can replace $\widehat{V}$ in (4.3) and $\widehat{V}_{\tau(\theta)}$ in (4.4) by

$$\widetilde{V} = \begin{pmatrix} 1 & \widehat{\rho}_\tau \widehat{\sigma}_{y,\tau} \\ \widehat{\rho}_\tau \widehat{\sigma}_{y,\tau} & \widehat{\sigma}^2_{y,\tau} \end{pmatrix}, \quad \widetilde{V}_{\tau(\theta)} = \begin{pmatrix} 1 & \widetilde{\rho}_{\tau(\theta)} \widetilde{\sigma}_{\eta,\tau(\theta)} \\ \widetilde{\rho}_{\tau(\theta)} \widetilde{\sigma}_{\eta,\tau(\theta)} & \widetilde{\sigma}^2_{\eta,\tau(\theta)} \end{pmatrix}, \tag{4.6}$$

where $\widetilde{\sigma}^2_{\eta,m} = m^{-1}\sum_{i=1}^m (\eta_i - \bar{\eta}_m)^2$. Let $\widetilde{u}_{1-2\alpha}(\theta)$ be the $(1-2\alpha)$-quantile of (4.4) with $\widehat{V}_{\tau(\theta)}$ replaced by $\widetilde{V}_{\tau(\theta)}$. The hybrid confidence region in this case is

$$\{(\theta, \mu) : \widetilde{R}(\mathbf{X}, \theta, \mu) \le \widetilde{u}_{1-2\alpha}(\theta)\}, \tag{4.7}$$

where $\widetilde{R}(\mathbf{X}, \theta, \mu)$ is defined by (4.3) with $\widehat{V}$ replaced by $\widetilde{V}$. The confidence region (4.7) can be inverted to yield a more explicit region as follows. Let $\widehat{a} = \tau(1 - \widehat{\rho}_\tau^2)^{-1}, \widehat{b} = -\tau\widehat{\rho}_\tau/\{(1 - \widehat{\rho}_\tau^2)\widehat{\sigma}_{y,\tau}\}$, and $\widehat{c} = \tau/\{(1 - \widehat{\rho}_\tau^2)\widehat{\sigma}_{y,\tau}^2\}$. Then $\widetilde{R}(\mathbf{X}, \theta, \mu) = \widehat{a}(\bar{X}_\tau - \theta)^2 + 2\widehat{b}(\bar{X}_\tau - \theta)(\bar{Y}_\tau - \mu) + \widehat{c}(\bar{Y}_\tau - \mu)^2$. For fixed $\theta$ with

$$(\widehat{b}^2 - \widehat{a}\widehat{c})(\bar{X}_\tau - \theta)^2 + \widehat{c}\widetilde{u}_{1-\alpha}(\theta) \geq 0, \tag{4.8}$$

$(\theta, \mu)$ belongs to the confidence region (4.7) if $\mu$ lies in the interval

$$\bar{Y}_\tau + \widehat{b}\widehat{c}^{-1}(\bar{X}_\tau - \theta) \pm \widehat{c}^{-1}\{(\widehat{b}^2 - \widehat{a}\widehat{c})(\bar{X}_\tau - \theta)^2 + \widehat{c}\widetilde{u}_{1-\alpha}(\theta)\}^{1/2}. \tag{4.9}$$

In Example 5 below, we use the following algorithm to compute the hybrid confidence region explicitly. First find the two values of $\underline{\theta} < \overline{\theta}$ for which equality holds in (4.8). Then partition the interval $[\underline{\theta}, \overline{\theta}]$ by $\underline{\theta} = \theta_1 < \cdots < \theta_k = \overline{\theta}$. For each $\theta_j$, find the endpoints of the interval (4.9). The hybrid confidence region (4.7) is then computed by linearly interpolating the upper endpoints to form the upper boundary, and the lower endpoints to form the lower boundary of the quasi-elliptical confidence region. A similar method can be used to compute the hybrid confidence region (4.5) and the exact confidence region (4.2).

**Example 5.** Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be independent bivariate normal with unit variances and correlation $\rho$, and consider the stopping rule $\tau$ defined from the partial sums of the $X_i$ given in Example 1. A random sample $\{(X_1, Y_1), \ldots, (X_\tau, Y_\tau)\}$ was generated with $\rho = 0.8, \sqrt{15}\theta = 0.5$, and $\mu = 0$. For this sample, $\sqrt{15}(\bar{X}_\tau, \bar{Y}_\tau)$ $= (0.18, -0.37)$, and $\widehat{\rho}_\tau = 0.81$. Figure 2 shows the 90% hybrid bivariate confidence region (4.7) for $\sqrt{15}(\theta, \mu)$ computed for this sample, using 21 evenly spaced points $\sqrt{15}\theta_i$ in $[\underline{\theta}, \overline{\theta}]$ to obtain the boundary of the region. Also shown in the figure are the exact confidence region $\{(\theta, \mu) : R(\mathbf{X}, \theta, \mu) \leq u_{0.9}(\theta)\}$ defined by (4.2), and the naive chi-square region $\{(\theta, \mu) : R(\mathbf{X}, \theta, \mu) \leq 4.61\}$ that assumes $R(\mathbf{X}, \theta, \mu)$ has an approximate chi-square distribution with 2 degrees of freedom and 90th percentile 4.61. Table 4 gives the simulated coverage errors of the hybrid confidence regions (4.5) and (4.7), and of the naive chi-square region, at various values of $\rho$ and $\theta$. Each simulation was based on 2000 sets of simulated data, and we computed $\widehat{u}_{0.9}(\theta)$ and $\widetilde{u}_{0.9}(\theta)$ with 999 resamples. Table 4 indicates that the two hybrid regions have coverage errors close to the nominal coverage error but that the chi-square approximation is not appropriate even when the true $\rho$ and $\mathrm{Var}(Y_i)$ are used in $R(\mathbf{X}, \theta, \mu)$.

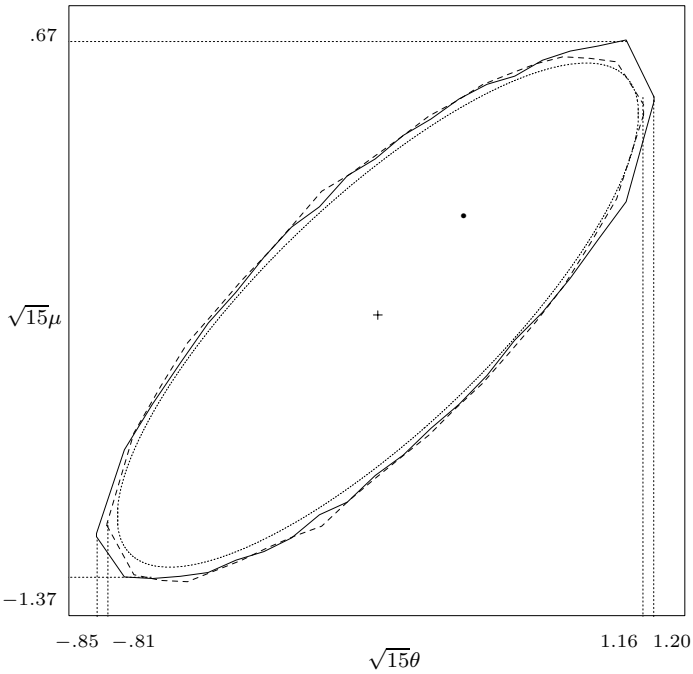


Figure 2. Comparison of bivariate confidence regions for two population means after optional stopping for a simulated data set. The true population mean vector is $\sqrt{15}(\theta, \mu) = (0.5, 0.0)$, which is indicated by $\bullet$. The observed stopped mean vector is given by $\sqrt{15}(\bar{X}_\tau, \bar{Y}_\tau) = (.18, -.37)$, which is indicated by $+$. The dotted ellipsoid is based on the naive chi-square approximation. The confidence region drawn by the solid line is the hybrid confidence region which does not assume $\mathrm{Var}(Y_i)$ to be known. The exact confidence region computed under the assumption of bivariate normality with known correlation and variances is indicated by the broken line.

Table 4. Coverage errors in % for joint confidence regions based on 2000 simulations for two population means. The hybrid region H1 assumes both variances to be known, whereas the hybrid region H2 allows for $Y_i$ to have unknown variance. Also included is the region obtained by a naive chi-square approximation. Nominal coverage errors are 10%.

| $\rho$ | $\sqrt{15}\theta$ | H1 | H2 | Chisq |
|---|---|---|---|---|
| 0.0 | 0.0 | 10.75 | 10.35 | 13.40 |
| 0.0 | 0.5 | 9.90 | 10.00 | 13.40 |
| 0.5 | 0.0 | 9.85 | 10.00 | 12.85 |
| 0.8 | 0.0 | 9.35 | 9.50 | 12.15 |
| 0.8 | 0.5 | 10.20 | 10.10 | 13.05 |

The following theorem, whose proof is similar to that of Theorem 1 of Chuang and Lai (1998), states that the hybrid confidence regions (4.5) and (4.7) are second-order accurate.

**Theorem 1.** *Suppose that the stopping rule $\tau$ is of the form*

$$\tau = \inf\Big\{n_j : \sum_{i=1}^{n_j} X_i \geq \gamma_j \quad or \quad \sum_{i=1}^{n_j} X_i \leq \lambda_j\Big\}, \tag{4.10}$$

*in which $\lambda_j < \gamma_j$ are real numbers and $n_1 < \cdots < n_k = n$ are positive integers such that*

$$\liminf_{n\to\infty}(n_j - n_{j-1})/n > 0 \text{ for } 1 \leq j \leq k \quad (n_0 = 0). \tag{4.11}$$

*Suppose there exist $r > 18$ and $C > 0$ such that*

$$E|X_1 - \mu|^r \leq C, \;\; E|Y_1 - \theta|^r \leq C \text{ and } \limsup_{|t|+|s|\to\infty} |E\exp\{\sqrt{-1}(tX_1 + sY_1)\}| < 1. \tag{4.12}$$

*Then both* (4.5) *and* (4.7) *have coverage probability* $1 - 2\alpha + O(n^{-1})$.

The hybrid confidence region (4.7) assumes the variance of $X_i$ to be known. If $\sigma_x^2 = \mathrm{Var}(X_i)$ is unknown, then the stopping rule would also involve some estimate of $\sigma_x^2$ instead of being based only on the partial sums of $X_i$. In this case the hybrid resampling method can be used to construct a bivariate confidence region for $(\theta/\sigma_x, \mu)$. More generally, we can extend the hybrid resampling method to construct bivariate confidence regions for $(g(\mu_1, \ldots, \mu_k), h(\mu_1, \ldots, \mu_k))$ for smooth functions of means $\mu_1, \ldots, \mu_k$, where we replace $\sum_{i=1}^{n_j} X_i$ in (4.10) by $n_j g(\bar{\mathbf{X}}_{n_j})$, in which $\mathbf{X}_i$ are i.i.d. $1 \times k$ vectors with common mean $(\mu_1, \ldots, \mu_k)$, and $g$ and $h$ are smooth real-valued functions. We can further extend the idea to nonparametric statistics that are more general than smooth functions of sample means. The technical details, however, are considerably more complicated and will be treated elsewhere. Some of the issues that need to be addressed are discussed in Section 6, where the correlation coefficient is discussed as an example, and in Chuang and Lai (1998).

## 5. Confidence Intervals in Possibly Nonstationary First-Order Autoregressive Models

We now apply the hybrid resampling method to solve the long-standing problem of interval estimation of the autoregressive parameter in a possibly nonstationary AR(1) model $x_i = \theta x_{i-1} + \epsilon_i$, where $x_0 = 0$ and $\epsilon_i$ are i.i.d. with mean 0 and variance $v$. This Markov chain has a stationary distribution when $|\theta| < 1$, is a random walk when $\theta = 1$ and has similar asymptotic behavior when $\theta = -1$, and is explosive in the sense that $\theta^{-n} x_n$ converges a.s. when $|\theta| > 1$. The least

squares estimate $\widehat{\theta}$ of $\theta$ based on $x_1, \ldots, x_n$ is given by $\sum_{i=1}^n x_i x_{i-1} / \sum_{i=1}^n x_{i-1}^2$ and is well known to be consistent. Moreover, $\widehat{v} = n^{-1} \sum_{i=1}^n (x_i - \widehat{\theta} x_{i-1})^2$ is a consistent estimate of $v$. In the notation of Section 2, let $\mathbf{X} = (x_1, \ldots, x_n)$ and $R(\mathbf{X}, \theta) = (\widehat{\theta} - \theta)/\widehat{\sigma}$, where $\widehat{\sigma}^2 = \widehat{v} / \sum_{i=1}^n x_{i-1}^2$. The following lemma, which follows from the results of White (1958) and Anderson (1959), shows that $R(\mathbf{X}, \theta)$ has a limiting distribution as $n \to \infty$ in all cases.

**Lemma 3.**

(i) *If* $|\theta| < 1$, *the limiting distribution of* $R(\mathbf{X}, \theta)$ *is standard normal and* $n^{-1} \sum_{i=1}^n x_{i-1}^2 \to v/(1-\theta^2)$ *a.s.*

(ii) *If* $|\theta| = 1$, $R(\mathbf{X}, \theta)$ *converges in distribution to* $\frac{1}{2}(B_1^2 - 1)/(\int_0^1 B_u^2 du)^{1/2}$, *where* $B_t$ *is standard Brownian motion, and* $n^{-2} \sum_{i=1}^n x_{i-1}^2$ *converges in distribution to* $v \int_0^1 B_t^2 dt$.

(iii) *Suppose* $|\theta| > 1$. *Then* $R(\mathbf{X}, \theta)$ *converges in distribution to* $v^{-1/2}(1 - \theta^{-2})^{1/2} YZ/|Z|$, *which is standard normal if the* $\epsilon_i$ *are normal, where* $Y$ *and* $Z$ *are independent and have the same distribution as* $\sum_{i=0}^\infty \theta^{-i} \epsilon_{i+1}$. *Moreover,* $|\theta|^{-2(n-1)} \sum_{i=1}^n x_{i-1}^2 \to (\sum_{i=0}^\infty \theta^{-i} \epsilon_{i+1})^2/(\theta^2 - 1)$ *a.s.*

In view of Lemma 3, it is difficult to apply standard large-sample techniques to construct confidence intervals for $\theta$ unless it is known a priori that $|\theta| < 1$. When $|\theta| < 1$, Bose (1988) proposed to use the bootstrap confidence interval (2.2) with the preceding choice of the root $R(\mathbf{X}, \theta)$. Let $\widehat{\epsilon}_i = x_i - \widehat{\theta} x_{i-1}, 1 \le i \le n$, and define the centered residuals to be $\widetilde{\epsilon}_i = \widehat{\epsilon}_i - n^{-1} \sum_{i=1}^n \widehat{\epsilon}_i$. Consider $x_i^* = \widehat{\theta} x_{i-1}^* + \epsilon_i^*$, where $\epsilon_i^*$ are i.i.d. from the empirical distribution $\widehat{G}$ of $\widetilde{\epsilon}_i, 1 \le i \le n$. Let $\mathbf{X}^* = (x_1^*, \ldots, x_n^*)$ and let $u_\alpha^*$ be the $\alpha$-quantile of $R(\mathbf{X}^*, \widehat{\theta})$. The bootstrap-$t$ confidence interval (2.2) for $\theta$ is $\widehat{\theta} - \widehat{\sigma} u_{1-\alpha}^* \le \theta \le \widehat{\theta} - \widehat{\sigma} u_\alpha^*$. Bose (1988) showed that the maximum difference of the bootstrap distribution function and the actual distribution function of $R(\mathbf{X}, \theta)$ is of the order $o_p(n^{-1/2})$, and Fuh and Lai (1998) recently sharpened this result to yield the $O_p(n^{-1})$ order.

When $|\theta| = 1$, Basawa *et al.* (1991) showed that the bootstrap method is not asymptotically valid and that the bootstrap distribution function converges to a random distribution function when $\epsilon_n^*$ are generated from a normal distribution (assuming $\epsilon_n$ are normal). To address the difficulties of the bootstrap method, Heimann and Kreiss (1996) proposed an $m$-out-of-$n$ bootstrap, with $m \to \infty$ but $m = o(n)$ as $n \to \infty$. Their idea is to resample $\epsilon_1^*, \ldots, \epsilon_m^*$ from $\widehat{G}$ and to form $x_j^* = \widehat{\theta} x_{j-1}^* + \epsilon_j^*, 1 \le j \le m$, for computing the least squares estimate $\widehat{\theta}_m^* = \sum_{i=1}^m x_i^* x_{i-1}^* / \sum_{i=1}^m x_i^{*2}$. They showed that the difference between the distribution function of $(\sum_{i=1}^m x_i^*)^{1/2}(\widehat{\theta}_m^* - \widehat{\theta}_m)$ and that of $R(\mathbf{X}, \theta)$ converges in probability to 0 for any fixed value of $\theta$. However, no rates of convergence have been established when $|\theta| \ge 1$, and the problem of constructing confidence intervals for $\theta$ without a priori stability assumptions has been relatively unexplored.

The hybrid resampling approach provides a unified solution to this problem, irrespective of the (unknown) value of $\theta$. Consider the resampling family $\{\widehat{F}_\theta\}$ given by $x_i(\theta) = \theta x_{i-1}(\theta) + \epsilon_i^*, 1 \le i \le n$, where $x_0(\theta) = 0$ and $\epsilon_1^*, \ldots, \epsilon_n^*$ are i.i.d. from $\widehat{G}$. The approximate equal-tailed $1-2\alpha$ confidence region (2.3) is then given by

$$\{\theta : \widehat{u}_\alpha(\theta) < (\widehat{\theta} - \theta)/\widehat{\sigma} < \widehat{u}_{1-\alpha}(\theta)\}, \tag{5.1}$$

which is the set of $\theta_0$ for which the "bootstrap test" using the test statistic $(\widehat{\theta} - \theta_0)/\widehat{\sigma}$ accepts the null hypothesis $\theta = \theta_0$. When $\theta_0 = 1$, Ferreti and Romo (1996) considered such bootstrap tests and showed that $\widehat{u}_\alpha(1)$ and $\widehat{u}_{1-\alpha}(1)$ indeed converge in probability to $u_\alpha(1)$ and $u_{1-\alpha}(1)$ but provided no convergence rates. The following theorem gives the convergence rates and thereby establishes the second-order correctness of the hybrid confidence region (5.1).

**Theorem 2.** *Let $\theta_0$ denote the true value of the autoregressive parameter.*

(i) *Suppose $|\theta_0| < 1$, $E|\epsilon_1|^6 < \infty$ and $\limsup_{|t|+|s|\to\infty} |E \exp\{\sqrt{-1}(t\epsilon_1 + s\epsilon_1^2)\}| < 1$. Then for any $K > 0$,*

$$\max_{|\theta - \widehat{\theta}| \le Kn^{-1/2}} |\widehat{u}_\alpha(\theta) - u_\alpha(\theta)| = O_p(n^{-1}).$$

(ii) *For any $K > 0$,*

$$\max_{|\theta - \widehat{\theta}| \le Kn^{-1}} |\widehat{u}_\alpha(\theta) - u_\alpha(\theta)| = O_p(n^{-1}), \quad \text{if} \quad |\theta_0| = 1,$$

$$\max_{|\theta - \widehat{\theta}| \le K|\theta_0|^{-n}} |\widehat{u}_\alpha(\theta) - u_\alpha(\theta)| = O_p(n^{-1/2}|\theta_0|^{-n}), \quad \text{if} \quad |\theta_0| > 1.$$

*Similar results hold for $\widehat{u}_{1-\alpha}(\theta) - u_{1-\alpha}(\theta)$.*

Suppose the true quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ are known for every $\theta$. Then an exact equal-tailed $1-2\alpha$ confidence region for $\theta$ is

$$\{\theta : u_\alpha(\theta) < (\widehat{\theta} - \theta)/\widehat{\sigma} < u_{1-\alpha}(\theta)\}. \tag{5.2}$$

In the stable case $|\theta_0| < 1$, it follows from Lemma 3(i) that for all sufficiently large $K$,

$$\lim_{n\to\infty} P\{(5.2) \text{ is contained in the interval } (\widehat{\theta} - Kn^{-1/2}, \widehat{\theta} + Kn^{-1/2})\} = 1. \tag{5.3}$$

Combining this result with Theorem 2(i) (which follows from the Edgeworth expansions for the distributions of $(\widehat{\theta} - \theta)/\widehat{\sigma}$ under the models $x_i = \theta x_{i-1} + \epsilon_i$ and $x_i(\theta) = \theta x_{i-1}(\theta) + \epsilon_i^*$ given in Fuh and Lai (1998)) establishes the second-order correctness of (5.1) in the stable case. When $|\theta_0| = 1$, it follows from Theorem 1 of Chan and Wei (1987) that (5.3) still holds with $\pm Kn^{-1/2}$ replaced

by $\pm Kn^{-1}$ for sufficiently large $K$. Moreover when $|\theta_0| > 1$, (5.3) still holds with $\pm Kn^{-1/2}$ replaced by $\pm K|\theta_0|^{-n}$, in view of Lemma 3(iii). The proof of Theorem 2(ii) uses of a Skorohod-type embedding that puts $\epsilon_i$ and $\epsilon_i^*$ on the same probability space to analyze the difference between $(\sum_{i=1}^n x_{i-1}\epsilon_i, \sum_{i=1}^n x_{i-1}^2)$ and $(\sum_{i=1}^n x_{i-1}(\theta)\epsilon_i^*, \sum_{i=1}^n x_{i-1}^2(\theta))$. The details are quite lengthy and technical and will be presented elsewhere.

Table 5 summarizes a simulation study comparing the performance of the bootstrap-$t$ and hybrid confidence intervals for $\theta$ and also the normal interval given by $\widehat{\theta} \pm z_{1-\alpha}\widehat{\sigma}$. The nominal coverage error is $\alpha = 5\%$ for both upper and lower confidence limits. The length of the time series is $n = 30$ for each case, and the simulated coverage errors are based on 2000 time series. For the bootstrap and hybrid methods, 999 resamples were used, and the secant method described at the end of Section 2 was used to obtain the hybrid interval. Standard normal $\epsilon_i$ were used in the simulations. As Table 5 indicates, the hybrid confidence interval has excellent coverage properties for all values of $\theta$, whereas the actual coverage errors of the normal and bootstrap intervals may be quite far from the nominal coverage errors.

Table 5. Coverage errors and mean values for lower (L) and upper (U) confidence limits based on 2000 simulations for autoregressive parameter. The errors are standard normal, and the length of the time series is 30. Nominal coverage errors are 5%.

Coverage Errors (in %)

| | Normal | | Bootstrap | | Hybrid | |
|---|---|---|---|---|---|---|
| $\theta$ | L | U | L | U | L | U |
| 0.00 | 5.60 | 5.40 | 5.70 | 5.85 | 5.40 | 5.45 |
| 0.50 | 4.15 | 6.40 | 5.75 | 5.40 | 5.20 | 5.10 |
| 0.80 | 3.80 | 6.70 | 7.20 | 5.45 | 6.10 | 4.65 |
| 0.95 | 3.60 | 9.30 | 5.15 | 6.65 | 5.45 | 5.50 |
| 1.00 | 3.75 | 11.05 | 5.40 | 7.50 | 5.90 | 5.25 |
| 1.05 | 3.75 | 15.60 | 5.35 | 8.25 | 5.85 | 5.20 |
| 1.20 | 4.30 | 8.00 | 4.10 | 4.40 | 4.30 | 4.75 |
| 1.50 | 5.90 | 6.15 | 4.95 | 5.05 | 5.20 | 5.15 |

Average Values

| Normal | | Bootstrap | | Hybrid | |
|---|---|---|---|---|---|
| L | U | L | U | L | U |
| -0.30 | 0.30 | -0.30 | 0.31 | -0.31 | 0.32 |
| 0.21 | 0.74 | 0.22 | 0.76 | 0.21 | 0.77 |
| 0.56 | 0.95 | 0.58 | 0.98 | 0.57 | 1.00 |
| 0.76 | 1.03 | 0.78 | 1.06 | 0.77 | 1.08 |
| 0.84 | 1.06 | 0.85 | 1.08 | 0.85 | 1.10 |
| 0.93 | 1.08 | 0.94 | 1.10 | 0.94 | 1.11 |
| 1.18 | 1.21 | 1.18 | 1.21 | 1.19 | 1.21 |
| 1.50 | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 |

## 6. Choice of Resampling Family in Nonparametric Models

We have discussed in Section 3.2 the choice of the root in nonparametric models and applied this choice to construct bootstrap confidence regions (2.2) in cases where $(\widehat{\theta} - \theta)/\widehat{\sigma}$ is an asymptotic pivot. For situations where $(\widehat{\theta} - \theta)/\widehat{\sigma}$ is not an asymptotic pivot, we use the hybrid resampling method to overcome the difficulties with the bootstrap method. For the population mean problem in Section 4 and the regression problem in Section 5, the resampling family

$\{\widehat{F}_\theta\}$ in the hybrid resampling method is defined by using the distribution of the residuals and the location/regression parameter $\theta$, from which the simulated data are generated. For more complicated nonparametric problems, under the assumption that $(X_{i1}, \ldots, X_{ip})$ in Section 2 are i.i.d. random vectors, a natural choice of $\widehat{F}_{\theta_0}$ is the nonparametric maximum likelihood estimate of $F$ subject to the constraint $\theta(F) = \theta_0$.

A simpler alternative that is asymptotically equivalent to the computationally intensive constrained nonparametric maximum likelihood estimate of $F$ has been developed by Efron (1981, 1987) when $p = 1$. He proposed a one-parameter "tilting family" of distributions $\widehat{H}_\delta$ such that

$$\widehat{H}_\delta(\{X_i\}) = \exp(\delta U_i(\widehat{F}))/\sum_{j=1}^{n} \exp(\delta U_j(\widehat{F})), \quad i = 1, \ldots, n, \tag{6.1}$$

where $U_i(\widehat{F})$ is defined in (3.14). Efron (1981) proposed to use (6.1) in his "nonparametric tilting method" to construct confidence intervals for the mean $\theta$ of $F$, for which $U_i(\widehat{F}) = X_i - \bar{X}_n$. Let $\widehat{\theta}_\delta$ be the mean of the distribution $\widehat{H}_\delta$ in (6.1) with $U_i(\widehat{F}) = X_i - \bar{X}_n$. Then the tilted upper $1 - \alpha$ confidence bound for the mean $\theta$ is given by $\theta_{\delta'}$, with $\delta'$ being the value of $\delta$ for which

$$P(\bar{X}_n^* \le \bar{X}_n | \widehat{H}_{\delta'}, \bar{X}_n) = \alpha, \tag{6.2}$$

and the lower endpoint of the tilted interval can be expressed similarly. Let $G_F$ denote the distribution of $\widehat{\theta}$ ($= \bar{X}_n$ in the present case) under $F$, as in Section 3.2. Then

$$G_{\widehat{H}_\delta}(\bar{X}_n) = P(\bar{X}_n^* \le \bar{X}_n | \widehat{H}_\delta, \bar{X}_n) = E\Big\{ e^{n\delta \bar{X}_n^*} \Big( n^{-1} \sum_{j=1}^{n} e^{\delta X_j^*} \Big)^{-n} I_{\{\bar{X}_n^* < \bar{X}_n\}} | \widehat{F} \Big\}. \tag{6.3}$$

An attractive feature of the tilting family (6.1) is that we need only generate $X_1^*, \ldots, X_n^*$ from $\widehat{F}$ and then apply (6.3) to evaluate the probability under $\widehat{H}_\delta$, without the need to resample again under $\widehat{H}_\delta$ for each trial value of $\delta$ to solve (6.2). Since the left-hand side of (6.2) is decreasing and continuous in $\delta$ and converges to 0 as $\delta \to \infty$, as can be seen from (6.3), (6.2) has a unique solution $\delta'$. For more general functionals $\theta(F)$, the analogue of the left-hand side of (6.2) is $G_{\widehat{H}_\delta}(\widehat{\theta})$, which need not be monotone in $\delta$. DiCiccio and Romano (1990) therefore defined the tilted upper confidence bound for $\theta(F)$ to be $\sup\{\theta(\widehat{H}_\delta) : G_{\widehat{H}_\delta}(\theta) \ge \alpha\}$ and the corresponding lower confidence bound for $\theta(F)$ to be $\inf\{\theta(\widehat{H}_\delta) : G_{\widehat{H}_\delta}(\theta) \le 1 - \alpha\}$.

The tilting family (6.1) provides a choice of the resampling family for the hybrid resampling method. Specifically, let $\widehat{F}_{\theta(\widehat{H}_\delta)} = \widehat{H}_\delta$. With this choice of

the resampling family, if we restrict the values of $\theta$ in the hybrid confidence set (2.3) to $\{\theta(\widehat{H}_\delta) : -\infty < \delta < \infty\}$ and choose the root to be $R(\mathbf{X}, \theta) = \widehat{\theta}$, then the supremum and infimum of the set (2.3) are the same as the Efron-DiCiccio-Romano tilted upper and lower confidence bounds. Under certain regularity conditions, DiCiccio and Romano (1990) showed that these confidence bounds are second-order accurate for smooth functions of mean vectors, and that second-order accuracy also holds if instead of (6.1) we choose $\widehat{F}_{\theta_0}$ to be the nonparametric maximum likelihood estimate of $F$ subject to the constraint $\theta(F) = \theta_0$.

In Section 3.1 dealing with parametric models, quadratic approximation to the log likelihood ratio statistic $l(\theta)$ when $\theta$ is near $\widehat{\theta}$ has led us to the root (3.8) and to the resampling family (3.10). The bootstrap corresponds to taking $\theta^l = \theta^u = \widehat{\theta}$ in (3.10). For nonparametric problems, an analogue of (3.9) and (3.10) is (6.1) with $\delta = n^{-1}(\theta - \widehat{\theta})/\widehat{\sigma}^2$, which leads to the resampling family

$$\widehat{F}_\theta = \begin{cases} \widehat{H}_{n^{-1}(\theta-\widehat{\theta})/\widehat{\sigma}^2}, & \text{if } \theta^l \le \theta \le \theta^u, \\ \widehat{H}_{n^{-1}(\theta^u-\widehat{\theta})/\widehat{\sigma}^2}, & \text{if } \theta > \theta^u, \\ \widehat{H}_{n^{-1}(\theta^l-\widehat{\theta})/\widehat{\sigma}^2}, & \text{if } \theta < \theta^l, \end{cases} \tag{6.4}$$

where $\theta^l = \widehat{\theta} - \widehat{\sigma}\widetilde{M}$, $\theta^u = \widehat{\theta} + \widehat{\sigma}\widetilde{M}$ and $\widehat{\sigma}$ is defined in (3.14). Since $\theta(\widehat{H}_\delta) = \theta(\widehat{F}) + n^{-1}\delta\sum_{i=1}^n U_i^2(\widehat{F}) + O(\delta^2) = \widehat{\theta} + n\delta\widehat{\sigma}^2 + O(\delta^2)$, for $\theta_0$ near $\widehat{\theta}$ the equation $\theta(\widehat{H}_\delta) = \theta_0$ has solution $\delta \doteq n^{-1}(\theta_0 - \widehat{\theta})/\widehat{\sigma}^2$; see Davison and Hinkley (1997, p.452) and DiCiccio and Romano (1990, p.72, in which $\widehat{\sigma}_n^2 = n\widehat{\sigma}^2$). The resampling family (6.4) is based on this linear approximation to the equation $\theta(\widehat{H}_\delta) = \theta_0$. Alternatively we can solve this equation numerically for $\delta^l \le \delta \le \delta^u$, where $\delta^l = -(n\widehat{\sigma})^{-1}\widetilde{M}$ and $\delta^u = (n\widehat{\sigma})^{-1}\widetilde{M}$ by using, for example, Brent's method (cf. Press, Teukolsky, Vetterling and Flannery (1992, Section 9.3)). Letting $\widetilde{\theta}^l = \theta(\delta^l)$ and $\widetilde{\theta}^u = \theta(\delta^u)$, this yields the alternative resampling family

$$\widehat{F}_\theta = \begin{cases} \widehat{H}_\delta \text{ if } \widetilde{\theta}^l \le \theta \le \widetilde{\theta}^u, \text{ where } \theta(\widehat{H}_\delta) = \theta, \\ \widehat{H}_{\delta^u} \text{ if } \theta > \widetilde{\theta}^u, \\ \widehat{H}_{\delta^l} \text{ if } \theta < \widetilde{\theta}^l. \end{cases} \tag{6.5}$$

The $n$ appearing in (6.4) (and in (6.5) via the $\widetilde{\theta}^l$ and $\widetilde{\theta}^u$) refers to the sample size, which can be a random variable, as in the case of group sequential trials. In general, for a resampling family $\{\widehat{F}_\theta\}$ to yield second-order accuracy of the associated hybrid confidence region (2.3), we need $\widehat{F}_{\theta(F)}$ to differ from the true distribution $F$ by $O_p(n^{-1/2})$. As noted by Efron and Tibshirani (1993, p.322), it is desirable to have in addition to accuracy the "correctness" of a confidence limit, which refers to how closely the confidence limit matches an ideal or exact confidence limit. To achieve second-order correctness for the hybrid confidence

region (2.3) for samples of fixed size $n$, we require $\widehat{F}_\theta$ to differ from $F$ by $O_p(n^{-1/2})$ not only at $\theta = \widehat{\theta}$ by also at the other values of $\theta$ that are included in (2.3). This explains why we use extrapolation beyond the interval $[\theta^l, \theta^u]$ in (6.4) to ensure that the $\widehat{F}_\theta$ thus defined will not give spurious estimates of the quantiles $\widehat{u}_\alpha(\theta)$ and $\widehat{u}_{1-\alpha}(\theta)$.

A disadvantage of the tilting family (6.1) is that it is very sensitive to influential observations. Whereas $\widehat{F}$ puts mass $1/n$ at each $X_i$, $\widehat{H}_\delta$ can put most of its mass at a few influential observations. Consequently, such $\widehat{H}_\delta$ would differ substantially from $\widehat{F}$ and not yield reasonably accurate and correct confidence limits, in view of the discussion of the preceding paragraph.

**Example 6.** Consider interval estimation of the correlation coefficient for the bivariate lognormal distribution with the correlation $(1 + \sqrt{e})^{-1} = 0.3775$ used in the simulation study in Example 4. Because $X_i$ and $Y_i$ are exponentials of normal variables, they tend to have a few outlying values, showing a pattern as in Figure 3 for a typical data set of 30 pairs $(X_i, Y_i)$. With $U_i$ defined by (3.22), the standardized $\widetilde{U}_i = U_i/(n^{-1}\sum_{i=1}^n U_i^2)^{1/2}$ of six outlying points in Figure 3 are indicated. These six points include three with the largest $|\widetilde{U}_i|$ values, whose sum of $U_i^2$ values makes up about 88% of $\sum_{i=1}^n U_i^2$. In particular the most influential point, labeled A, contributes 59% to $\sum_{i=1}^n U_i^2$. The sample correlation coefficient is 0.42 with all 30 points, but increases to 0.62 when point A is omitted. The infinitesimal jackknife standard error estimate $\widehat{\sigma}$ for the data set is 0.18. The lower and upper 5th percentiles of $(\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*$ are $-2.02$ and 6.29 based on 999 resamples, so the bootstrap-$t$ interval is $[-0.71, 0.78]$. The lower and upper 5th percentiles of the sampling distribution of $t(\theta)$ obtained using 999 samples from the true lognormal distribution are $-2.88$ and 3.83. Thus the bootstrap distribution of the Studentized root has a much heavier right tail than the actual sampling distribution, which translates into a bootstrap-$t$ interval that extends too far to the left. For the modified-$t$ root with $M = 2$, $u^*_{0.05}$ and $u^*_{0.95}$ are given by $-1.70$ and 2.17, which are quite close to the lower and upper 5th percentiles of the modified-$t$ root under the true lognormal distribution, given by $-1.86$ and 1.96 respectively. The bootstrap confidence region based on the modified-$t$ root is $[0.03, 0.73] \cup (0.78, 0.83]$, where $0.73 = \widehat{\theta} - u^*_{0.05}\widehat{\sigma}$, $0.78 = \widehat{\theta} + 2\widehat{\sigma}$, $0.83 =$ U$'$ and $0.03 =$ L$'$; see (3.20). When the bootstrap confidence region based on the modified-$t$ root is further refined by using the resampling family (6.4) with $\widetilde{M} = 2$, we obtain the hybrid interval $[0.15, 0.71]$. The lower limit of the hybrid interval is the correlation coefficient of $\widehat{H}_{-0.3}$, which puts about 11% of its mass at point A, in comparison with 3.3% that $\widehat{F}$ puts at each of the 30 points. The upper limit of the hybrid interval is the correlation coefficient of $\widehat{H}_{0.38}$, which puts about 8% of its mass at point B. However, without extrapolation beyond

$[\theta^l, \theta^u]$ as in (6.4), we have not been able to apply hybrid resampling to refine the bootstrap-$t$ interval that uses the Studentized root instead of the modified-$t$ root. In particular, it is basically impossible to reweight the data so that the correlation coefficient of $\widehat{H}_\delta$, for some $\delta$, is equal to $-0.71$, which is the lower limit of the bootstrap-$t$ interval.

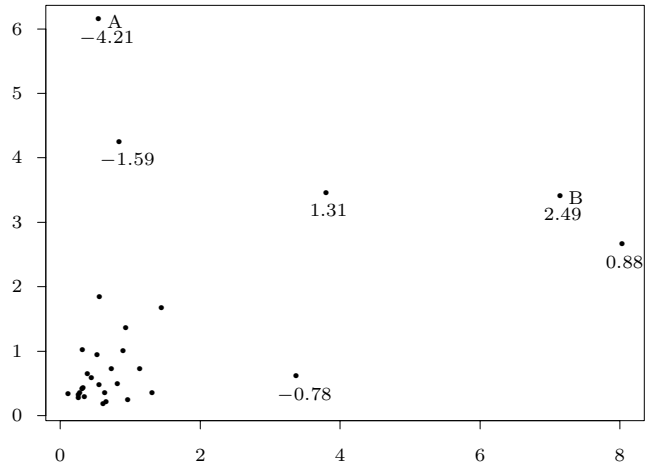


Figure 3. Scatterplot of a simulated data set of size 30 from a lognormal distribution with correlation equal to $(1+\sqrt{e})^{-1}$. Six outlying points with their standardized influences are marked in the plot.

We next illustrate how the resampling family (6.4) or (6.5) can be used to construct hybrid confidence regions following sequential tests by considering interval estimation of the correlation coefficient $\theta$ of a bivariate vector $(X, Y)$ after a group sequential test of independence.

**Example 7**. Using the same notation as in Example 4, let $\sigma_x^2 = \mathrm{Var}(X_i)$ and $\sigma_y^2 = \mathrm{Var}(Y_i)$. Under independence between $X_i$ and $Y_i$, the sample correlation coefficient has the representation $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n (X_i - \mu_x)(Y_i - \mu_y)/(\sigma_x\sigma_y) + o_p(n^{-1/2})$, which behaves like an average of i.i.d. random variables with mean 0 and variance 1. Therefore we can use a group sequential test for a normal mean with known variance to test the null hypothesis of independence between $X$ and $Y$ based on the sequential sample correlation coefficients, provided the group size is large enough to justify the normal approximation. In particular, Pocock's stopping rule in Example 1 leads to $\tau = \min\{n \in J : \sqrt{n}|\widehat{\theta}_n| \geq 2.413\}$, and under the null hypothesis of independence, $P\{\max_{n\in J}\sqrt{n}|\widehat{\theta}_n| \geq 2.413\} \doteq 0.05$. In the notation of Section 2, $\mathbf{X} = (X_1, Y_1, \ldots, X_\tau, Y_\tau, \tau)$ and we let $\widehat{F}$ be the empirical distribution of $\{(X_1, Y_1), \ldots, (X_\tau, Y_\tau)\}$. Defining $U_i$ by (3.22) (with $n$ replaced

by $\tau$), $\widehat{\sigma}^2(\tau) = \tau^{-2}\sum_{i=1}^{\tau} U_i^2$, and letting $\widehat{\sigma}_b^2(\tau)$ be the variance of $\widehat{\theta}_\tau^*$ generated from the distribution $\widehat{F}_\tau$ with the sample size fixed at $\tau$, we consider the $t_*$-root, introduced in the last paragraph of Section 3.2, of the form

$$R(\mathbf{X},\theta) = \begin{cases} (\widehat{\theta}_\tau - \theta)/\widehat{\sigma}(\tau) & \text{if } |\widehat{\theta}_\tau - \theta| \le M\widehat{\sigma}(\tau), \\ (\widehat{\theta}_\tau - \theta)/\widehat{\sigma}_b(\tau) & \text{if } |\widehat{\theta}_\tau - \theta| > M\widehat{\sigma}(\tau). \end{cases} \tag{6.6}$$

Define $\widehat{H}_\delta$ by (6.1) with $n$ replaced by $\tau$, and replace $(n, \widehat{\sigma}, \widehat{\theta})$ by $(\tau, \widehat{\sigma}(\tau), \widehat{\theta}_\tau)$ in the resampling family (6.4) or (6.5). With this choice for $\mathbf{X}, R(\mathbf{X},\theta)$ and $\widehat{F}_\theta$, the hybrid resampling method is used to construct an equal-tailed $1-2\alpha$ confidence region for the correlation coefficient following a group sequential test of independence.

In particular, suppose $(X, Y)$ is bivariate normal with correlation coefficient $\theta = (15)^{-1/2}(0.5)$. If one ignores the effects of optional stopping, one can use Fisher's transformation to construct a naive parametric confidence interval for $\theta$ in the present bivariate normal case. We performed a simulation study based on 1000 replications, for a nominal coverage error of 5%, to compare this interval with the nonparametric hybrid confidence interval based on the root (6.6). The hybrid interval was computed by using the secant method at the end of Section 2. We set $M = 2$ and $\widetilde{M} = 1.5$ and used bootstrap resample sizes of 999, and 100 for an additional layer of bootstrapping to compute $\widehat{\sigma}_b(\tau)$. The naive parametric interval for the 1000 replications averages to be $[-.06, .33]$, compared to $[-.07, .32]$ for the hybrid method using the resampling family (6.4). The simulated coverage errors for the lower and upper endpoints of the hybrid interval are 5.5% and 5.7%, while those of the other interval are 9% and 5.5%. In the same simulations, we also computed the hybrid confidence limits using (6.5) as the resampling family. The average hybrid confidence limits are the same (to two decimal places) as those using (6.4) instead, and the simulated coverage errors for the lower and upper limits are 5.9% and 5.8%.

## 7. Conclusion

As pointed out in Section 2, the hybrid resampling method generalizes the bootstrap method by incorporating a resampling family $\{\widehat{F}_\theta\}$, instead of using a single estimate of $F$ as in the bootstrap method. Young (1994, p.385) has commented that choosing a bootstrap procedure involves not only choice of the root to be bootstrapped but also choice of the estimate of $F$ to resample from, which can be a difficult issue for the general user. Sections 3.1, 4, 5 and 6 above have discussed in detail the choice of the resampling family, while Sections 3.1 and 3.2 address the issue of choosing the root in parametric and nonparametric models. The root (3.8) or (3.15) is closely related to (parametric or empirical) likelihood

ratio statistics which are, however, often difficult to compute repeatedly in simulating the distribution of the root. Analogously, the resampling family (3.10) or (6.4) is used to circumvent the computational difficulties of constrained maximum likelihood estimates of $F$. Instead of relying on normal/chi-square approximations or higher-order asymptotics in (parametric or empirical) likelihood confidence regions, the hybrid resampling approach simulates the sampling distribution of the root under the resampling family. It can therefore handle complex situations where such approximations fail, as in estimation following group sequential tests or in possibly non-ergodic autoregressive models and branching processes, where the bootstrap also fails. If the likelihood ratio statistic is readily computable, and if accurate and not too complicated analytic approximations to its sampling distribution are available for the problem at hand, then perhaps the most natural and attractive way to mimic (2.1) is a likelihood confidence region, which is in fact a special case of the hybrid confidence region $\{\theta : R(\mathbf{X}, \theta) \le \widehat{u}_{1-2\alpha}(\theta)\}$ with $R(\mathbf{X}, \theta)$ chosen as the log likelihood ratio statistic and $\widehat{u}_{1-2\alpha}(\theta)$ given by an analytic approximation instead of by simulations.

Influential observations are troublesome for the bootstrap and other resampling methods. Note the deterioration in the coverage accuracy in Table 2 for the lognormal case compared to the other cases. When highly influential observations are present, there are inherent difficulties in estimating the sampling distribution of $\theta(\widehat{F})$, on which the bootstrap and the more general hybrid resampling methods are based. The need to think critically about the answers provided by the bootstrap method in these situations has been pointed out by Efron (1992). On the other hand, for data of the type plotted in Figure 3, which shows a highly unstable correlation pattern between $X$ and $Y$, it is questionable whether the correlation coefficient has any value in describing the underlying bivariate distribution, and therefore interval or point estimates of the correlation coefficient would be of little practical relevance for such data. In this connection, the family $\mathcal{F}$ in Section 2 for nonparametric problems should satisfy certain regularity conditions related to $\theta(F)$ for the bootstrap and hybrid resampling methods to work. For the case where $\theta(F)$ is the mean of a univariate distribution $F$, Bahadur and Savage (1956) have shown that if $\mathcal{F}$ is a convex family of distributions having finite means such that $\{\theta(F) : F \in \mathcal{F}\}$ is the whole real line, and if $I$ is a confidence set such that $P_F\{\theta(F) \in I\} \ge 1-\alpha$ for all $F \in \mathcal{F}$, then $P_F\{x \in I\} \ge 1-\alpha$ for all $x \in \mathbf{R}$ and $F \in \mathcal{F}$. This difficulty arises because $\theta(F)$ is sensitive to the tails of $F \in \mathcal{F}$, and can be circumvented by putting further restrictions on $\mathcal{F}$ (for example, that the support of every $F \in \mathcal{F}$ lies in a given compact set). Indeed, for a heavy-tailed distribution $F$, the mean is not a useful summary of $F$ and one should use other measures of location that are more outlier-resistant. When $\theta(F)$ is robust for $F \in \mathcal{F}$, we have shown good performance of the hybrid resampling approach for constructing confidence intervals.

## Acknowledgement

This research was supported by the National Science Foundation and the National Security Agency.

## Appendix

**Proof of Lemma 1.** Theorem 2.1 of Bhat and Adke (1981) shows that $\widehat{\theta}$ and $\widehat{\lambda}$ are strongly consistent estimates of $\theta$ and $\lambda$. Since $l'(\widehat{\theta}) = 0$ and $l''(\widehat{\theta})/2 = N_{n-1}/\widehat{\theta}$, it then follows that $l^{\pm}(\theta) = t(\theta)+o_p(1)$, where $t(\theta)=(\widehat{\theta}-\theta)/(\widehat{\theta}/N_{n-1})^{1/2}$. The proof of Lemma 2.3 in Bhat and Adke (1981) shows that if $\theta \neq 1$, then $t(\theta)$ has a limiting standard normal distribution. Suppose that $\theta = 1$. Then

$$t(1) = \{\sum_{i=1}^{n}(X_i - \psi_i) - N_{n-1}\}/(\widehat{\theta}N_{n-1})^{1/2} = (X_n - x_0 - \sum_{i=1}^{n}\psi_i)/(\widehat{\theta}N_{n-1})^{1/2}.$$

Since $(n^{-1}X_n, n^{-2}N_{n-1}, n^{-1}\sum_{i=1}^{n}\psi_i)$ converges in distribution to $(Y_1, \int_0^1 Y_t dt, \lambda)$ (cf. Wei and Winnicki (1989, Remark 2.4)) and since $\widehat{\theta} \to \theta = 1$ and $\widehat{\lambda} \to \lambda$ a.s., it follows from the continuous mapping theorem that $t(1)$ converges to $(Y_1 - \lambda)/(\int_0^1 Y_t dt)^{1/2}$ in distribution.

**Proof of Lemma 2.** To prove (i), note that the conditional distribution of $\xi_1, \xi_2, \ldots, \psi_1, \ldots, \psi_n$ given $n\widehat{\lambda}$ is the same as the distribution of i.i.d. Poisson $\xi_i^*$ having mean $\theta$ and independent of $(\psi_1^*, \ldots, \psi_n^*)$ that has the $M(n\widehat{\lambda}; n^{-1}, \ldots, n^{-1})$ distribution. Let $l_*^{\pm}(\theta)$ denote the signed root of the log likelihood ratio statistic based on $\xi_1^*, \ldots, \xi_{N_{n-1}^*}^*, \psi_1^*, \ldots, \psi_n^*$. Since $P\{l_*^{\pm}(\theta)+n^{-1}U \leq \widehat{u}_\alpha(\theta)|\widehat{\lambda}\} = \alpha$, it then follows that

$$P\{R(\mathbf{X}, \theta) \leq \widehat{u}_\alpha(\theta)\} = E[P\{l_*^{\pm}(\theta) + n^{-1}U \leq \widehat{u}_\alpha(\theta)|\widehat{\lambda}\}] = \alpha.$$

Similarly, $P\{l_*^{\pm}(\theta)+n^{-1}U \geq \widehat{u}_{1-\alpha}(\theta)|\widehat{\lambda}\} = \alpha$. To prove (ii), condition on $(\psi_1, \ldots, \psi_n)$ instead of on $\widehat{\lambda}$ alone.

## References

Anderson, T. W. (1959). On asymptotic distribution of estimates of parameters of stochastic difference equations. *Ann. Math. Statist.* **30**, 676-687.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27**, 1115-1122.

Basawa, I. V., Mallik, A. K., McCormick, W. P., Reeves, J. H. and Taylor, R. L. (1989). Bootstrapping explosive autoregressive processes. *Ann. Statist.* **17**, 1479-1486.

Basawa, I. V., Mallik, A. K., McCormick, W. P., Reeves, J. H. and Taylor, R. L. (1991). Bootstrapping unstable first-order autoregressive processes. *Ann. Statist.* **19**, 1098-1101.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457-468.

Bhat, B. R. and Adke, S. R. (1981). Maximum likelihood estimation for branching processes with immigration. *Adv. Appl. Probab.* **13**, 498-509.

Bickel, P. J. (1987). Comment on "Better Bootstrap Confidence Intervals". *J. Amer. Statist. Assoc.* **82**, 191.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Ann. Statist.* **16**, 1709-1722.

Chan, N. H. and Wei, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Ann. Statist.* **15**, 1050-1063.

Chuang, C. and Lai, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**, 317-332.

Coad, D. S. and Woodroofe, M. B. (1996). Corrected confidence intervals after sequential testing with applications to survival analysis. *Biometrika* **83**, 763-777.

Cox, J., Ingersoll, J. and Ross, S. A. (1985). A theory of term structure of interest rates. *Econometrica* **53**, 385-408.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications.* Cambridge University Press.

DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *Internat. Statist. Rev.* **58**, 59-76.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9**, 139-172.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82**, 171-200.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 83-127.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Ferretti, N. and Romo, R. (1996). Unit root bootstrap tests for AR(1) models. *Biometrika* **83**, 849-860.

Fuh, C. D. and Lai, T. L. (1998). Edgeworth expansions and bootstrap methods for Markov chains. Technical Report, Department of Statistics, Stanford University.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16**, 927-985.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hall, P., Martin, M. A. and Schucany, W. R. (1989). Better nonparametric bootstrap confidence intervals for the correlation coefficient. *J. Statist. Comput. Simul.* **33**, 161-172.

Heimann, G. and Kreiss, J. P. (1996). Bootstrapping general first order autoregression. *Statist. Probab. Lett.* **30**, 87-98.

Lee, S. M. S. and Young, G. A. (1995). Asymptotic iterated bootstrap confidence intervals. *Ann. Statist.* **23**, 1301-1330.

Loh, W. Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82**, 155-162.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* **75**, 723-729.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing.* 2nd edition. Cambridge University Press.

Schenker, N. (1987). Comment on "Better Bootstrap Confidence Intervals". *J. Amer. Statist. Assoc.* **82**, 192-194.

Tang, D. I., Genecco, C. and Geller, N. L. (1989). Design of group sequential trials with multiple endpoints. *J. Amer. Statist. Assoc.* **84**, 776-779.

Wei, C. Z. and Winnicki, J. (1989). Some asymptotic results for the branching process with immigration. *Stoch. Proc. Appl.* **31**, 261-282.

White, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Statist.* **29**, 1188-1197.

Woodroofe, M. (1986). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14**, 1049-1067.

Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika* **79**, 347-353.

Young, G. A. (1994). Bootstrap: more than a stab in the dark? (with discussion). *Statist. Sci.* **9**, 382-415.

Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.
E-mail: cschuang@stat.cmu.edu

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
E-mail: lait@stat.stanford.edu

(Received June 1998; accepted September 1999)

## COMMENT

Michael Woodroofe and Ruby C. Weng

*University of Michigan*

Professors Chuang and Lai are to be congratulated for their interesting suggestions. The hybrid approach is a very general one, applicable to both parametric and non-parametric problems, to non-ergodic models, and to irregular problems (though the authors do not emphasize the latter). It is also computationally intensive and can be difficult to implement. To paraphrase Efron (1979) and Tukey, it seems powerful enough to blow the head off a problem, but it may

leave a bit of a mess. Here we will compare the hybrid method with an approach based on asymptotic expansions, with special reference to the group sequential example. By way of contrast, expansions often require highly structured models, with some independence and/or arcane technical conditions; but they are much simpler, where applicable. In many cases expansions may by used to produce an approximate pivot from which confidence intervals are easily identified.

To date there has been relatively little work on using asymptotic expansions to set confidence intervals after group sequential tests. In principle expansions should work provided that there are enough groups and a believable parametric model. In practice, expansions may work better when the stopping boundaries consist of straight lines. These points are illustrated by Weng (1999) and Weng and Woodroofe (2000) in the context of a group sequential test for comparing two Poisson means. A triangular test is considered below. As will be seen the number of potential groups is an important design parameter, and we may ask: how many groups are needed for expansions to produce reliable approximations?

As in the paper, suppose that $X_1, X_2, \ldots$ are i.i.d. with unknown mean $\theta$ and variance one, and that there are $K$ potential groups of size $m$ each. Let $N = Km$, $J = \{km : k = 1, \ldots, K\}$, and $S_n = X_1 + \cdots + X_n$. Consider stopping times of the form

$$\tau = \min\{n \in J : ng(\frac{S_n}{n}) \geq c\}, \tag{1}$$

where $g$ is a real valued function for which $g \geq \delta := c/N$ and $c$ is chosen to control error probabilities. Example 1 in the paper is of the form with $K = 5$, $m = 15$, $c = (2.413)^2$, and $g(x) = \max(\delta, x^2)$. For $m = 1$ and a stopping time of the form (1), Woodroofe (1992) found the following:

$$R = \sqrt{\tau}(\bar{X}_\tau - \theta)$$

is approximately normal with mean

$$\mu = \frac{1}{\sqrt{c}}(\sqrt{g})'(\theta),$$

where $'$ denotes derivative, and

$$Z = \frac{R - \widehat{\mu}}{\sqrt{1 + \widehat{\mu}^2}}$$

is approximately standard normal to order $o(1/c)$ in the very weak sense. See also Woodroofe and Coad (1997). Thus, $Z$ serves as an approximate pivot. It is important here that only one derivative is needed for $g$, since many of the more popular choices only have one derivative.

Of course, the group sequenctial problem may be reduced to the case $m = 1$ by letting $\widetilde{X}_k = [X_{m(k-1)+1} + \cdots + X_{mk}]/\sqrt{m}$ for $k = 1, 2, \ldots$. Then $\theta$, $g$, and $c$, are replaced by $\widetilde{\theta} = \sqrt{m}\theta$, $\widetilde{g}(x) = \sqrt{m}g(x/\sqrt{m})$, and $\widetilde{c} = c/\sqrt{m}$. This reduction has some simple consequences. First, normality may be less of an issue, since the $\widetilde{X}_k$ are sums, and especially so if the original $X_k$ were differences between reponses to treatments and controls. Key features of the reduced problem are the horizon $K$ (the number of groups) and the size of $\widetilde{\theta}$. However if $|\widetilde{\theta}|$ is very big, the procedure is likely to stop after the first group and the sequential nature of the problm is lost. So, we are led to a reduced problem with (at least, approximately) normal data, moderate values of $\widetilde{\theta}$, and a small horizon.

Do the expansions work in this context? They do not work very well for the repeated significance tests in Example 1 of the paper, even for $m = 1$ or, equivalently, a large horizon. With large horizons, however, expansion work very well for triangular tests in which $g(x) = \delta + |x|$ and $N\delta = c$. In a simulation experiment with $m = 1$, $K = 100$, $\theta = 0(.25)1.5$ and 10,000 replications, the simulated distributions of $Z$ were not significantly different from the standard normal, as judged by the Kolmogorov Smirnov Test with $\alpha = .05$. If the number of groups is decreased to five, then the normal approximation to the distribution of $Z$ degenerates, especially in the central part of the distribution. In effect, $\widehat{\mu}$ overestimates the mean of $R$, leaving $Z$ with a negative mean. A plausible reason for this is that the derivations in Woodroofe (1992) neglect the excess over the boundary. To some extent, this can be recovered by replacing $c$ with $c + .583$, as in Siegmund (1985, Section 3.5). This change improves the approximations, but still leaves $Z$ with a negative mean. This may be seen in Table 1 below which reports simulated values of $P_\theta\{Z \le -1.645\}$ and $P_\theta\{Z \ge 1.645\}$ for $K = 5$, $K = 10$, and selected values of $\theta$. Also reported are Monte Carlo estimates of $E_\theta(Z)$, $\sqrt{E_\theta(Z^2)}$, the power $P_\theta\{S_\tau > 0\}$, and the expected number of groups $E_\theta(\tau/m)$. For five groups, the approximations appear to be slightly conservative, with one exception (remember that we are doing multiple comparisons here). For ten groups, the approximations are less conservative, and the overall agreement seems better.

Robustness with respect to the normality assumption may also be investigated by simulation. We compared our approximations, derived assuming normality, to simulations from two non-normal distributions, the centered exponential distribution considered in the paper and a bilateral exponential distribution. The exponential is very different from the normal, and the agreement between the theoretical and simulated values was generally worse than in the normal case and much worse for five groups. However, agreement with the bilateral exponential simulations was very comparable to that reported in Table 1.

Table 1. Group Sequential.
Normal Data

$K = 5$, $\widetilde{c} = 3.8702$

| $\widetilde{\theta}$ | Mean | RMSQ | Lower | Upper | Total | Power | ASN |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.0005 | 0.9761 | 0.0507 | 0.0472 | 0.0979 | 0.5043 | 3.7729 |
| 0.25 | −0.0107 | 0.9861 | 0.0485 | 0.0435 | 0.0920 | 0.7014 | 3.6718 |
| 0.50 | −0.0143 | 0.9953 | 0.0475 | 0.0433 | 0.0908 | 0.8569 | 3.4340 |
| 1.00 | −0.0284 | 0.9967 | 0.0550 | 0.0467 | 0.1017 | 0.9830 | 2.8007 |
| 1.50 | −0.0373 | 1.0102 | 0.0561 | 0.0579 | 0.1140 | 0.9991 | 2.3098 |
| 2.00 | −0.0363 | 1.0264 | 0.0498 | 0.0484 | 0.0982 | 1.0000 | 1.9864 |
| 2.50 | −0.0219 | 1.0328 | 0.0594 | 0.0430 | 0.1024 | 1.0000 | 1.7574 |
| 3.00 | −0.0069 | 1.0112 | 0.0658 | 0.0415 | 0.1073 | 1.0000 | 1.5427 |
| 4.00 | −0.0296 | 0.9456 | 0.0482 | 0.0420 | 0.0902 | 1.0000 | 1.1868 |
| 5.00 | −0.0846 | 0.9874 | 0.0536 | 0.0427 | 0.0963 | 1.0000 | 1.0297 |
| ± | 0.0200 | 0.0141 | 0.0044 | 0.0044 | 0.0060 | | |

$K = 10$, $\widetilde{c} = 5.4733$

| $\widetilde{\theta}$ | Mean | RMSQ | Lower | Upper | Total | Power | ASN |
|---|---|---|---|---|---|---|---|
| 0.00 | 0.0149 | 0.9955 | 0.0463 | 0.0502 | 0.0965 | 0.5037 | 6.9434 |
| 0.25 | 0.0061 | 1.0012 | 0.0506 | 0.0546 | 0.1052 | 0.7743 | 6.5839 |
| 0.50 | −0.0124 | 1.0066 | 0.0515 | 0.0443 | 0.0958 | 0.9321 | 5.7646 |
| 1.00 | −0.0108 | 1.0032 | 0.0534 | 0.0535 | 0.1069 | 0.9973 | 4.2104 |
| 1.50 | −0.0149 | 1.0113 | 0.0495 | 0.0453 | 0.0948 | 1.0000 | 3.2887 |
| 2.00 | −0.0215 | 1.0024 | 0.0507 | 0.0404 | 0.0911 | 1.0000 | 2.7207 |
| 2.50 | −0.0297 | 0.9943 | 0.0614 | 0.0451 | 0.1065 | 1.0000 | 2.3416 |
| 3.00 | −0.0472 | 1.0016 | 0.0480 | 0.0559 | 0.1039 | 1.0000 | 2.1021 |
| 4.00 | −0.0168 | 1.0624 | 0.0637 | 0.0507 | 0.1144 | 1.0000 | 1.8274 |
| 5.00 | −0.0265 | 0.9977 | 0.0597 | 0.0444 | 0.1041 | 1.0000 | 1.4728 |
| ± | 0.0200 | 0.0141 | 0.0044 | 0.0044 | 0.0060 | | |

Notes: Monte Carlo estimates based on 10,000 replications; ± is two standard deviations; $\widetilde{\theta} = \sqrt{m}\theta$; Mean= $E_\theta(Z)$; RMSQ= $\sqrt{E_\theta(Z^2)}$; Lower= $P_\theta\{Z \le -1.645\}$; Upper= $P_\theta\{Z \ge 1.645\}$; Total= $P_\theta\{|Z| \ge 1.645\}$; Power= $P_\theta\{S_\tau > 0\}$; ASN= $E_\theta(\tau/m)$.

Department of Statistics, University of Michigan, Ann Arbor MI 48109, U.S.A.
E-mail: michaelw@umich.edu

# COMMENT

Peter J. Bickel

*University of California, Berkeley*

Chuang and Lai explore an interesting principle for constructing confidence intervals for a real parameter $\theta$ in situations where bootstrapping a "pivot" $R(X, \theta)$ may not work well. The method which can be viewed as an extension of the "standard" pivot method is,

(i) To estimate the parameters other than $\theta$ on which the distribution of the data $X$ depends.

(ii) From these obtain estimates $\widehat{F}_\theta$ of the distribution of the pivot $R(X, \theta)$ as a function of $\theta$.

(iii) Apply univariate inversion methods using $R(X, \theta)$ and the quantiles of $\widehat{F}_\theta$.

They give examples in three principal situations:

**I**. Observations from parametric models such as

(a) Branching processes with immigration and $AR(1)$ schemes where for at least one value of $\theta$ the MLE $\widehat{\theta}$ does not have regular Gaussian behavior.

(b) Data obtained by sequential sampling from a parametric model.

and

**II**. Data $X_1, \ldots, X_n$ i.i.d. from $P$ completely unknown where $\theta(P)$ is "differentiable", i.e.

$$\theta(P_n) = \theta(P) + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i, P) + o_p(n^{-1/2}),$$

where $P_n$ is the empirical df. and $E_p\psi(X_1, P) = 0$, $E_p\psi^2(X_1, P) < \infty$.

In the examples of I and II the difficulties arising are rather different. In I (a) for the special values of $\theta$ (or near them) the authors claim that even the parametric bootstrap distribution is an inconsistent estimate of the true distribution of $R(X, \theta)$. In II (given unstated regularity conditions) the nonparametric bootstrap is consistent as $n \to \infty$ and poor behavior comes from second order effects. Finally in I (b) there are no clear asymptotics which suggest that any particular method is preferable.

In situation I there is an alternative which may give satisfactory behavior generally. A special case is proposed by Heimann and Kreiss (1996). In both examples considered, the pivots have the form $R(X, \theta) = h_n(U_1, \ldots, U_n, \theta)$. Here the $U_i$ are i.i.d. with common distribution $F_{\theta,\eta}$ and $h_n(U_1, \ldots, U_n, \theta) = \lambda_n(X_1, \ldots, X_n)(\widehat{\theta} - \theta)(1 + o_p(1))$, where if $\frac{m}{n} \to 0$, $m \to \infty$, $\frac{\lambda_m(X_1,\ldots,X_m)}{\lambda_n(X_1,\ldots,X_n)} \xrightarrow{P} 0$.

In Example 3 the $U_i$ are $(\epsilon_i, \psi_i)$, and in Example 2 the $\epsilon_i$. It follows then from Theorem 3 of Bickel, Götze and van Zwet (1997) that the bootstrap distribution of $\lambda_m(X_1^*, \ldots, X_m^*)(\widehat{\theta}_m^* - \widehat{\theta}_n)$, where the $\widehat{\theta}_m^*$ are based on $X_1^*, \ldots, X_m^*$ and $X_1^*, \ldots, X_m^*$ are generated from $F_{\widehat{\theta},\widehat{\eta}}$, converges to the same limit as the population distribution of $R(X, \theta)$ if $\theta$ is true. This follows since the variational distance between the joint distribution of $(U_1, \ldots, U_m)$ under $(\widehat{\theta}, \widehat{\eta})$ and under $(\theta, \eta)$ tends to 0 if $|(\widehat{\theta}, \widehat{\eta}) - (\theta, \eta)| = o(\frac{1}{\sqrt{m}})$, which is true in both the critical and supercritical cases in Example 2 in view of Lemma 3. The choice of $m$ is of course an issue. A generally useful rule is studied in Bickel and Sakov (1999).

In situation I (b) it is difficult to suggest alternatives without some asymptotics in the sequential sampling. If the sequential sample size $N$ is governed by a parameter such as the cardinality of $J$ in Example 1 and/or the magnitude of the step (15 in Example 3), call it $n$, and if $\frac{N_n}{n^\alpha} \xrightarrow{P} V(\theta)$, one would expect that the theory of the $m$ out of $n$ bootstrap could be extended to such situations.

In situation II alternatives to the hybrid methods are, of course, the prepivoting approaches of Beran (1987) and Edgeworth-related techniques such as those in Putter and van Zwet (1998).

Which of these is appropriate in particular situations has to be explored by simulation coupled with theory.

Finally, a discussion of the effect of influential observations or failure of the model for studentized estimates may be found in Bickel (1992).

Department of Statistics, University of California, Berkeley, 367 Evans Hall #3860, Berkeley CA 94720-3860, U.S.A.
E-mail: bickel@stat.berkeley.edu

## COMMENT

Peter Hall

*Australian National University*

This very interesting paper raises a range of important issues which, despite the great deal of effort devoted to exploring bootstrap methods over the last two decades, have not really been satisfactorily resolved. Some are as deceptively simple as construction of confidence limits for a simple parameter $\theta$, using a given

random sample; others are substantially more complex, and include bootstrap methods for both confidence intervals and hypothesis tests in sequential analysis.

We address only the former problem, and then only one aspect of it. As Chuang and Lai point out, a natural approach to implementing their ingenious hybrid method in nonparametric settings amounts to using the concept of "nonparametric likelihood", or NPL, employed to such good effect by Owen in his development of empirical likelihood. In the context of the correlation coefficient, Chuang and Lai express a degree of reluctance to use NPL because of the magnitude of the computational task. It should perhaps be noted that the computational cost usually relates more to potential difficulties in choice of a starting point for Newton-Raphson iteration, used to calculate the vector $p$ of NPL weights, than it does to sheer computing time. Algorithms such as those developed by Qin and Lawless (1994, 1995) have removed a significant part of the computational burden.

Nevertheless, it is our experience that the empirical likelihood approach to constructing confidence intervals for the correlation coefficient does not perform particularly well. While the "internal studentisation" feature of empirical likelihood is attractive, and avoids calculation of an explicit variance estimator, it does not seem to substantially overcome difficulties associated with the naively studentised statistic, i.e., with the quantity defined as $t(\theta)$ by Chuang and Lai. In particular, empirical likelihood methods generally perform less well than the calibrated percentile bootstrap in this problem. It seems likely that Chuang and Lai's hybrid method, if implemented using NPL, would perform similarly to empirical likelihood.

Possibly these problems are due at least in part to choice of the loss function, or divergence measure, in traditional NPL. Note that NPL aims to minimise $L_0(p) \equiv -\sum_i \log p_i$, where $p = (p_1, \ldots, p_n)$ is a vector of multinomial weights, subject to constraints. If we try to let $p_i$ tend to 0, for some $i$, then $L_0(p)$ becomes unboundedly large. Therefore, the method strenuously resists downweighting of specific data values. In small to moderate samples, where the "stability" problems associated with NPL are often caused by a small number of outlying data vectors, NPL with loss function $L_0(p)$ attempts to share the pain around by downweighting a relatively large number of data by a relatively small amount each, rather than by heavily downweighting the small number of outliers that really matter.

This difficulty can be alleviated by changing the loss function to one that permits more uneven downweighting. A range of alternatives has been discussed recently by Baggerly (1998), Corcoran (1998) and Hall and Presnell (1999a), and their use in connection with outliers has been addressed by Hall and Presnell (1999b). When employed with empirical likelihood they preserve many of the

properties of that method, in particular the fact that it satisfies Wilks' theorem (i.e., the empirical likelihood-ratio statistic has an asymptotic chi-squared distribution).

For example, we might minimise $L_1(p) \equiv \sum_i p_i \log p_i$, instead of $L_0(p)$; the former loss function permits $p_i$ to be reduced to 0 without incurring more than a finite penalty. Using $L_1$ rather than $L_0$ should make NPL a little more robust against the effects of outlying data values. These advantages may be expected to carry through to methods that are based on NPL, for example to empirical likelihood and to the empirical likelihood version of Chuang and Lai's hybrid bootstrap technique.

Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 0200, Australia.
E-mail: halpstat@durras.anu.edu.au

## COMMENT

Jiahua Chen and Hanfeng Chen

*University of Waterloo and Bowling Green State University*

The interesting idea of hybrid resampling method for the construction of confidence intervals was first introduced by Professors Chuang and Lai in their Biometrika paper (Chuang and Lai (1998)). The method is very effective in the analysis of treatment effects associated with the primary and secondary endpoints of a clinical trial whose stopping rule is specified by a group sequential test. In this paper, Chuang and Lai explore the idea further and develop it into a general resampling method for constructing confidence regions in nonparametric models or multiparameter models in the presence of nuisance parameters. Professors Chuang and Lai are to be congratulated on presenting a solution to a difficult problem and for adding another excellent method to the statistical toolbox for real applications.

The hybrid resampling method hybridizes the exact method of confidence interval estimation and the bootstrap method. Chuang and Lai discuss and illustrate the three methods, i.e., exact, bootstrap and hybrid. Following them, we first would like to comment on the connections between the three methods.

We start with models where a useful pivotal quantity is available. Let $X$ be an observation vector from the population $F$ and $\theta = \theta(F)$ be the parameter of interest. Suppose that $R(X, \theta_0)$ is the test statistic for testing $\theta = \theta_0$. Let $u_\alpha(\theta_0)$ be the $\alpha$-quantile of the distribution of $R(X, \theta_0)$ under $\theta = \theta_0$. If $R(X, \theta)$ is a pivotal quantity and its $\alpha$-quantile is denoted by $u_\alpha$, then a $(1 - 2\alpha)100\%$ confidence region for $\theta$ is given by

$$\{\theta : u_\alpha < R(X, \theta) < u_{1-\alpha}\}.$$

When the distribution of $R(X, \theta_0)$ under the assumption $\theta = \theta_0$ depends on $\theta_0$, the quantile $u_\alpha$ has to be calculated for each $\theta_0$, which leads to the exact methods.

In most applications $u_\alpha$ cannot be obtained explicitly. In these situations a contemporary treatment is to use $u^*_\alpha$ to replace $u_\alpha$, where $u^*_\alpha$ is the $\alpha$-quantile of $R(X^*, \widehat{\theta})$ under the assumption $X^* \sim \widehat{F}$, with $\widehat{\theta} = \theta(\widehat{F})$. An ordinary choice of $\widehat{F}$ in a nonparametric set-up is the empirical distribution. This is the so-called bootstrap method. Note that the replacement of $u_\alpha$ with $u^*_\alpha$ is universal, disregarding $\theta_0$. Therefore, the bootstrap method implicitly assumes that $R(X, \theta)$ is pivotal. If $\theta$ is a smooth function of means, $R(X, \theta)$ is often an asymptotic pivotal quantity within a small neightborhood of the true value of $\theta$. However, Chuang and Lai (1998) warn us that the asymptotic pivotal property of $R(X, \theta)$ should not be exaggerated without limit. They show that even such a benign looking quantity as $R(X, \theta) = \sqrt{\tau}(\widehat{X}_\tau - \theta)$ fails to be pivotal asymptotically in a group sequential setting. Blind use of the bootstrap method can lead to poor coverage rates. It is an important contribution to reveal that $u^*_\alpha$ has to be determined case by case for each $\theta_0$. Consequently the hybrid method is urgently called for.

We are impressed by the delicate choice of resampling distribution $\widehat{F}_\theta$ to determine the hybrid $\alpha$-quantile $\widehat{u}_\alpha(\theta)$ of $R(X, \theta)$, and the studies of various applications. Here we would like to remark on the close connection between the hybrid method and the empirical likelihood method.

Suppose that $X = (x_1, \ldots, x_n)$ is a random sample of fixed size $n$ from a nonparametric family. The profile maximum likelihood estimate can be a natural choice of the hybrid resampling distribution $\widehat{F}_\theta$. Let $p_i = F(\{X_i\})$ and $\widetilde{F}(x) = \sum p_i I(x_i \le x)$. For fixed $\theta$, the profile MLE of $F$ is to maximize

$$\prod_{i=1}^{n} p_i,$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \text{and } \theta(\widetilde{F}) = \theta.$$

When the equation $\theta(F) = \theta$ can be expressed as an integral $\int w(x, \theta)dF(x)$ for a specific function $w$, the profile MLE for $p_i$ is

$$\widehat{p}_i = \frac{1}{n\{1 + \lambda w(x_i, \theta)]\}}, \ i = 1, \ldots, n,$$

where $\lambda$ is the Lagrange multiplier and solves the equation

$$\sum_{i=1}^{n} \frac{w(x_i, \theta)}{1 + \lambda w(x_i, \theta)} = 0.$$

If in addition the test statistic $R(X, \theta)$ is chosen to be the log-likelihood ratio, as given by

$$R(X, \theta) = 2\sum_{i=1}^{n} \log\{1 + \lambda w(x_i, \theta)\},$$

the hybrid resampling method is exactly the empirical likelihood method (Owen (1990)).

One striking similarity between the hybrid method and the empirical likelihood method is that both estimate $F$ for each $\theta$ case by case. The empirical likelihood method restricts itself by only considering $\widehat{F}_\theta$ which satisfy the conditions $\theta(F) = \theta$ for fixed $\theta$, and maximize the empirical likelihood, while the hybrid method offers much more flexibility than the empirical likelihood method. One of the appealing features of the hybrid method is that it does not require an explicit expression of the likelihood function, at least formally. Indeed, this is the most important feature of the hybrid method and distinguishes it from the empirical likelihood method. The example of group sequential trial discussed in the paper demonstrates this clearly, a setting where the empirical likelihood method does not work.

As a nonparametric method, the empirical likelihood method is best known for its advantage of conveniently accommodating auxiliary information. See, for example, Qin and Lawless (1994, 1995), Chen and Qin (1993), Chen and Sitter (1999) and Chen and Chen (1999). The hybrid ressampling method enjoys a similar advantage. One may use auxiliary information to further narrow the range of the "resampling family". The inference utilizing auxiliary information is usually found to be more efficient than not using the information. We conjecture that this is also true in the context of the hybrid resampling method, as the choice of the resampling family may have notable impact on the average size of the confidence intervals and coverage probabilities.

Department of Statistics and Actuarial Science, University of Waterloo, Ontario N2L 3G1, Canada.
E-mail: jhchen@math.uwaterloo.ca

Bowling Green State University
E-mail: hchen@math.bgsu.edu

# COMMENT

Stephen M. S. Lee and G. Alastair Young

*The University of Hong Kong and University of Cambridge*

## 1. Introduction

This article by Chuang and Lai provides a very nice summary of hybrid resampling methods and their properties. We believe that it contributes significantly to the establishment of an effective and reliable resampling methodology for the construction of accurate confidence intervals. While congratulating the authors on the clarity of their discussion, which in particular provides a useful presentation of conventional bootstrap methods as a special case of hybrid resampling, we should like to remark on some specific aspects of the methodology.

## 2. Bootstrap Inconsistency

Of major focus in recent times has been the establishment of resampling methods of inference which are valid, in the sense of consistency, even when the conventional bootstrap fails, and especially for circumstances where it fails for particular values of the model parameter, as in the first-order autoregressive example of Section 5 of the paper. A key tool for this purpose has been the "$m$ out of $n$" bootstrap, as examined by Bickel, Götze and van Zwet (1997). Of interest would be a detailed comparison of the properties of hybrid resampling methods with those of the $m$ out of $n$ bootstrap. A potential disadvantage of the $m$ out of $n$ bootstrap is that, while it may provide a consistent estimate, the accompanying efficiency losses noted by Bickel, Götze and van Zwet (1997) might, in examples such as those considered by Chuang and Lai, produce an order of coverage error inferior to that given by hybrid resampling. Whether hybrid resampling is to be generally preferred, in terms of efficiency loss or its remedy, to the $m$ out of $n$ bootstrap remains an open question.

## 3. Choice of Root

Historically, much focus within the bootstrap literature has involved the issues and benefits of studentization and/or prepivoting, the latter taken to include ideas of bootstrap calibration and "double bootstrapping". The paper of Chuang and Lai presents an interesting idea on the choice of the root $R(X, \theta)$ used in construction of the confidence interval, which we believe is practically important, and worthy of further development. They suggest that stability of the hybrid resampling approach in small to moderate sample sizes can be enhanced by use of

a hybrid pivot $R(X, \theta)$, which depends on the value of $\theta$, as exemplified by (3.8) and (3.15) of the paper. As the discussion following (3.15) of the paper makes clear, use of a modified form of pivot can be made to automatically incorporate both studentization and prepivoting ideas. Implementation depends, however, on interpretation, for the problem at hand, of what constitutes '$\widehat{\theta}$ is not too far from $\theta$'. In the examples given in the paper the authors give no specific guidance on how this question should be met. Some adaptive procedure, based on empirical assessment of the stability of the hybrid root $R(X, \theta)$ seems natural.

## 4. Non-parametric Inference

We were particularly interested to read the authors' recommendations, in Section 6 of the paper, on the choice of resampling family recommended for the hybrid resampling methodology in nonparametric problems. Their discussion advocates a particular one-parameter tilting family of distributions, as given by (6.1) of the paper. We have argued in Lee and Young (1999a) the advantages of such a tilting family in the construction of nonparametric likelihood ratio confidence intervals. The simplicity of the tilting family allows us to propose and analyze various asymptotic and bootstrap correction techniques as a means of producing, via the nonparametric likelihood, confidence intervals of low coverage error, comparable to those obtained by more computationally-intensive methods such as the iterated bootstrap. Direct comparison of these methods with hybrid resampling methods would also be of practical interest.

## 5. Iterated Hybrid Resampling

Chuang and Lai discuss the possibility of applying the hybrid resampling method to both non-pivotal and approximately pivotal $R(X, \theta)$, to achieve both second order accuracy and correctness of the hybrid confidence region; the correlation coefficient example is used to illustrate the latter. In fact, it is possible to prove rather stronger results about the effect of using hybrid resampling, rather than conventional bootstrap resampling. We provide here a brief description of these results; full details will be given elsewhere.

Suppose we assume the smooth function model, where $\widehat{\theta}$ is a smooth function of sample means, and consider construction of a (one-sided) nonparametric confidence set for $\theta$ from the non-pivotal root $R(X, \theta) = \sqrt{n}(\widehat{\theta} - \theta)$. Then the conventional bootstrap, which resamples from the empirical distribution function of the observed sample, yields a coverage error of order $O(n^{-1/2})$. Hybrid resampling improves this error to one of order $O(n^{-1})$. On the other hand, if we proceed from the approximately pivotal root $R(X, \theta) = (\widehat{\theta} - \theta)/\widehat{\sigma}$, the conventional bootstrap yields coverage error of order $O(n^{-1})$. Hybrid resampling

improves this to $O(n^{-3/2})$, which is what is achieved by a conventional bootstrap calibration or double bootstrap method; see Martin (1990). The latter method provides, in our view, a satisfactory pragmatic solution to the problem of producing nonparametric confidence intervals of low coverage error, but with appropriate stability, which may not be enjoyed by other more sophisticated bootstrap procedures. It will be interesting to undertake a more extensive empirical analysis of how hybrid resampling intervals and the double bootstrap compare in practice. Evidence presented by Chuang and Lai for the correlation coefficient example suggests that hybrid resampling ought to be capable of challenging the double bootstrap gold standard.

In terms of computational expense, hybrid resampling is clearly preferable to the double bootstrap, as it only requires one level of resampling. But if we are willing to undertake a second level of resampling, might it not be advantageous to iterate the hybrid resampling, rather than use the conventional double bootstrap?

We sketch here the theoretical effects of iterated hybrid resampling. For simplicity of presentation, consider a nonpivotal root $R(X, \theta)$, and denote by $G(\cdot, \theta)$ its sampling distribution, as estimated by the hybrid resampling scheme using the tilting family (6.1) of Chuang and Lai's paper. As we have noted, the confidence limit based on the appropriate quantile of $G(\cdot\,, \theta)$ typically has coverage error of order $O(n^{-1})$. The concept of iteration is that an improved confidence limit can be obtained from the sampling distribution of the root $R_1(X, \theta) = G(R(X, \theta), \theta)$. There are two natural ways of estimating this sampling distribution: (a) by conventional bootstrapping, or (b) by hybrid resampling again.

It turns out that (a) yields a confidence limit with coverage error of order $O(n^{-3/2})$, an improvement in order terms over the $O(n^{-1})$ coverage error obtained by the conventional double bootstrap, and of the same order as the coverage error obtained if the sampling distribution $G(\cdot\,, \theta)$ is estimated by the conventional bootstrap, but hybrid sampling used to estimate the sampling distribution of $R_1(X, \theta)$. However, the benefits of hybrid resampling over conventional bootstrapping ensure that possibility (b) yields a confidence limit with coverage error of order $O(n^{-2})$. This means that a two-level resampling analysis which uses hybrid resampling at both levels, rather than conventional uniform resampling, produces an interval whose error is reduced by two orders of magnitude. Stated simply, single level hybrid resampling has the theoretically beneficial effect of conventional double bootstrapping, while a double level hybrid resampling has an effect similar to a conventional "quadruple" bootstrap.

Of course, as Chuang and Lai discuss, hybrid resampling requires rather more sophisticated computation than ordinary bootstrapping. In their notation, the sampling distribution of the root $R(X, \theta)$ must be simulated under $\widehat{F}_\theta$, for a set of different $\theta$ values, which amounts to weighted bootstrapping if the tilting

family (6.1) is employed. Ordinary boostrapping just requires simulation of the sampling distribution of $R(X^*, \widehat{\theta})$ under the empirical distribution $\widehat{F}$. Iterated hybrid resampling would presumably involve weighted bootstrapping at a number of selected values of $\theta$ at each resampling level, but will still be substantially less expensive computationally than the quadruple bootstrap.

## 6. Monte Carlo Implementation

We should finally like to make some brief remarks on the conventional approach, as adopted in this paper, to the need for Monte Carlo simulation in the implementation of resampling methods of inference. Traditionally, the prevailing attitude has automatically been to seek an implementation which uses the maximum number of Monte Carlo samples possible, within the limitations imposed by the need to control the overall computational burden. We have recently challenged this attitude in showing that there may sometimes be advantage, in terms of coverage accuracy, in more careful control of the Monte Carlo simulation. In Lee and Young (1999b) we provide an analysis of the coverage accuracy of the calibrated percentile method confidence set, which takes into account both the inherent bootstrap and Monte Carlo errors. We demonstrate that by suitable control of the size of the Monte Carlo simulation we may actually reduce the order of coverage error below that of the 'infinite simulation' interval. In times of readily available computational power, it seems appropriate to think more deeply about implementation, not just in terms of overall computational expense.

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.
E-mail:

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, U.K.
E-mail: g.a.young@statslab.cam.ac.uk

# REJOINDER

Chin-Shan Chuang and Tze Leung Lai

Although it has been two decades since Efron's seminal paper on bootstrap methods, there are still many unresolved problems in resampling methodology. As noted by Bickel, Götze and van Zwet (1997), "Practical anecdotal experience seems to support theory in the sense that the bootstrap generally gives reasonable answers but can bomb". Indeed, our motivation for developing hybrid resampling

methods comes from problems in inference following group sequential tests, where the natural analogues of bootstrap methods do not work well. We wish to thank the discussants for their insightful and informative contributions. Bickel classifies the examples discussed in the present paper into three cases, which we shall use in our rejoinder to compare hybrid resampling methods with other existing methods.

**Standard Case**

The standard case, which corresponds to Bickel's Case II, has been well studied in the bootstrap literature. As we have pointed out in Section 3.2, although bootstrapping the root $t(\theta)$ leads to second-order correct and accurate confidence limits, $t(\theta)$ may not behave well in small to moderate samples, especially when the functional of interest is highly nonlinear, as is the case for the correlation coefficient, for which bootstrap calibration of the percentile limits has been observed to give good results in practice. One contribution of the paper is the proposal of a new root (see (3.15) and (3.18)) which depends on a tuning parameter $M$ and gives results similar to bootstrap calibration of the percentile limits, but with less computational burden. The computational burden directly relates to $M$, with $M = 0$ corresponding to calibration and $M = \infty$ to using $t(\theta)$ as the root. Thus we view bootstrap calibration as a complementary rather than as an alternative method. As pointed out by Lee and Young, and also in Section 3.3 of our paper, the new root is a hybrid of studentization and calibration (prepivoting) ideas, and attempts to achieve the accuracy and correctness of the resultant confidence limits of the latter and the computational simplicity of the former.

To obtain second-order correctness and accuracy for these problems, it is not necessary to incorporate test inversion with a resampling family $\{\widehat{F}_\theta\}$. On the other hand, there may be certain advantages in considering a hybrid confidence region of the type (2.3). As Section 6 points out, DiCiccio and Romano (1990) show that test inversion based on the tilting family (6.1) gives second-order correct and accurate confidence limits even when the root is not studentized. Lee and Young have indicated in their discussion that "it is possible to prove stronger results about the effect of using hybrid resampling, rather than conventional bootstrap resampling". We look forward to reading their work in this direction, particularly concerning iterated hybrid resampling.

For small to moderate samples, the discussion by Hall and Section 6 of our paper both suggest that test inversion based on (6.1) for $t(\theta)$ is unlikely to work well. On the other hand, test inversion yields good results when both $t(\theta)$ and the tilting family are modified, as in (3.15) and (6.4). The resultant confidence limits retain the same order of coverage error as those obtained by calibration of the bootstrap-$t$ limits but require substantially less computation.

Hall and the references cited in his discussion suggest other alternatives to the $t$-root and empirical likelihood, while Chen and Chen suggest further modifications of the resampling family to incorporate auxiliary information (constraints) on the unknown parameters.

**Non-standard Cases**

Case I(a) in Bickel's discussion includes the branching process and autoregressive examples, where the root considered is not an asymptotic pivot and its limiting distribution may change drastically even when the unknown parameter changes little. Bickel, Götze and van Zwet (1997) propose the $m$ out of $n$ bootstrap for problems similar to these. Although Bickel has pointed out that the $m$ out of $n$ bootstrap works for these examples, hybrid resampling is considerably more efficient and is also easier to implement since it does not require the user to come up with a suitable choice of $m$, particularly when $n$ is not large, as in Tables 1 and 5.

Bickel's case I(b) consists of inference problems following group sequential tests. Whereas the studentized roots from fixed sample problems are usually asymptotically standard normal, their analogues for group sequential problems are not asymptotic pivots because of the effect of the stopping rule. Although Bickel suggests looking into the possibility of extending the $m$ out of $n$ bootstrap to this situation, such subsampling ideas would not work in sequential settings, particularly in view of the "overshoot" for the last observation (group). Somehow the form of the stopping rule has to be incorporated into the resampling scheme. Woodroofe and Weng, building on previous work cited in their references, discuss the use of asymptotic expansions for these problems. These asymptotic expansions, however, are in the "very weak" sense of Woodroofe (1986). Specifically, they provide confidence intervals $I$ whose integrated coverage errors $\int P_\theta(\theta \notin I)\xi(\theta)d\theta$ differ from the nominal value $2\alpha$ by $o(a^{-1})$ for a large class of smooth probability densities $\xi$, where $a(\to \infty)$ denotes the boundary (or some component thereof) of the stopping rule. This average coverage accuracy differs from the usual sense of second-order accuracy in the bootstrap literature, such as in Theorem 1 of our paper, or in Theorems 1 and 2 of Chuang and Lai (1998) on confidence intervals obtained by hybrid resampling methods following group sequential tests, or in Hall (1992) and Efron and Tibshirani (1993).

The derivation of the second-order accuracy theory in Chuang and Lai (1998) uses an Edgeworth expansion involving a $k$-variate normal distribution, where $k$ is the number of interim analyses ("groups") and is assumed to be fixed. When $k \to \infty$ as $n \to \infty$, as in the case of fully sequential tests, this argument breaks down. However, it is still possible to prove second-order accuracy by using the Edgeworth expansions for smooth functions of randomly stopped sums in Lai

and Wang (1994), who generalized and refined the seminal work of Woodroofe and Keener (1987) in this direction. Unlike the usual (fixed sample size) Edgeworth expansion which only involves moments of the population, the Edgeworth expansions for smooth functions of randomly stopped sums also involve quantities which are related to the fluctuation theory of random walks and which can be expressed in terms of the population characteristic function. The details are given in Chuang and Lai (1999), where it is also shown that test inversion with a resampling family is usually not needed (so direct bootstrapping can be used) for fully sequential (instead of group sequential) tests. In contrast, Woodroofe and his collaborators use Stein's identity to carry out Bayesian calculations, generating expansions for the posterior expectations (for which the stopping rule does not cause difficulties) and then integrating them. Their approach is applicable to both fully sequential and group sequential settings, but only yields asymptotic expansions in the very weak sense instead of the Edgeworth-type expansions needed in the second-order accuracy theory.

The very weak expansions are computationally much more attractive than the Edgeworth-type expansions (which involve difficult fluctuation-theoretic quantities) in Woodroofe and Keener (1987) and Lai and Wang (1994). However, as pointed out by Lai and Wang (1994) and Chuang and Lai (1999), these Edgeworth-type expansions can be implemented indirectly by simulation via the bootstrap which has the additional advantage of not requiring parametric assumptions on the underlying distribution. We are currently working towards extending hybrid resampling methods to construct confidence intervals following clinical trials with failure-time endpoints and interim analyses, as discussed in Gu and Lai (1991, 1999). In these problems, analytic corrections are prohibitively difficult and Monte Carlo simulations are needed for power and coverage probability calculations, in view of the complexity due to staggered patient entry, time-dependent rates of loss to follow-up and noncompliance, and complicated baseline and treatment survival patterns in practice.

## Additional References

Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85**, 535-547.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457-468.

Bickel, P. J. (1992). On the theoretical comparison of bootstrap confidence intervals. *Exploring the Limits of the Bootstrap* (Edited by R. LePage and L. Billard). John Wiley, New York.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than $n$ observations: gains, losses, and remedies for losses. *Statist. Sinica* **7**, 1-31.

Bickel, P. J. and Sakov, A. (1999). On the choice of $m$ in the $m$ out of $n$ bootstrap. Submitted to *Ann. Statist.*

Chen, H. and Chen, J. (1999). Bahadur representations of the empirical likelihood quantile processes. *J. Nonparametr. Statist.* To appear.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107-116.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica* **9**, 385-406.

Chuang, C. and Lai, T. L. (1999). Bootstrap and hybrid resampling methods for confidence intervals in sequential analysis. Technical Report, Department of Statistics, Stanford University.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85**, 967-972.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Hall, P. and Presnell, B. (1999a). Intentionally biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B* **61**, 661-680.

Gu, M. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.* **19**, 1403-1433.

Gu, M. and Lai, T. L. (1999). Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. *Controlled Clin. Trials* **20**, 423-438.

Lai, T. L. and Wang, J. Q. (1994). Asymptotic expansions of stopped random walks and first passage times. *Ann. Probab.* **22**, 1957-1992.

Lee, S. M. S. and Young, G. A. (1999a). Nonparametric likelihood ratio confidence intervals. *Biometrika* **86**, 107-118.

Lee, S. M. S. and Young, G. A. (1999b). The effect of Monte Carlo approximation on coverage error of double bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser. B* **61**, 353-366.

Martin, M. A. (1990) On bootstrap iteration for coverage correction in confidence intervals. *J. Amer. Statist. Assoc.* **85**, 1105-1118.

Putter, H. and van Zwet, W. R. (1998). Empirical Edgeworth expansions for symmetric statistics. *Ann. Statist.* **26**, 1540-1569.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-325.

Qin, J. and Lawless, J. (1995). Estimating equations, empirical likelihood and constraints on parameters. *Canad. J. Statist.* **23**, 145-159.

Siegmund, D. (1985). *Sequential Analysis.* Springer, New York.

Weng, R. C. (1999). *Very Weak Expansions for Sequentially Designed Experimaents.* Ph.D. Thesis, Statistics Department, Uinversity of Michigan.

Weng, R. C. and Woodroofe, M. (2000). Integrable expansions for posterior distributions for multiparameter exponential families with applications to sequential confidence levels. *Statist. Sinica.* **10**, ?-?.

Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika* 79, 347-353.

Woodroofe, M. and Coad, D. S. (1997). Coreected confidence sets for sequentially designed experiments. *Statist. Sinica* **7**, 53-74.

Woodroofe, M. and Keener, R. (1987). Asymptotic expansions in boundary crossing problems. *Ann. Probab.* **15**, 102-114.