Statistica Sinica 2(1992), 171-188

METHOD OF PAO-ZHUAN YIN-YU: A METHOD OF STOCHASTIC POINT ESTIMATION

James C. Fu and Lung-An Li

University of Manitoba and Academia Sinica

Abstract: A computer intensive resampling technique called the method of Pao-Zhuan Yin-Yu is systematically developed as an alternative to the bootstrap and the jackknife. The method is a sequential parametric resampling scheme which searches for an optimal estimator (in the minimum variance unbiased estimator sense) and provides an estimate for the variance of the optimal estimator. Several numerical examples are given, including inference for a coefficient in an autoregressive model where the observations are dependent. Numerical results show that the method performs extremely well in almost all cases.

Key words and phrases: Pao-Zhuan Yin-Yu, resampling, contourization, empirical conditional expectation, Rao-Blackwell theorem, bootstrap.

1. Introduction

Let us consider the following statistical model:

$$M = \{X, R, f(x|\theta), \theta \in \Theta\},$$

where the X is a random variable defined on the sample space R (real line) having a density function $f(x|\theta)$ and parameter θ is an element in the parameter space Θ. Let $x_1, \ldots, x_n$ be n independent identically distributed (i.i.d.) random observations from the statistical model M. Our aim is to estimate the unknown parameter θ based on the n observations $x_0 = (x_1, \ldots, x_n)$. There are two questions often asked in point estimation;

(a) What estimator should we use for estimating the unknown parameter θ?

(b) Having chosen a particular estimator, say $\widehat{\theta}_n = \widehat{\theta}_n(x_1, \ldots, x_n)$, how accurate is the estimator $\widehat{\theta}_n$?

To choose an optimal estimator is a difficult task. Theoretically, there are many principles and criteria which can be used as guides for selecting an optimal estimator. For instance, minimum variance unbiased estimation (mvue), maximum likelihood estimation (mle), and Bayes estimation are each optimal in a certain sense. Except in a few special cases, the analytic forms of these optimal

estimators are usually hard to obtain, e.g., the Bayes estimator when the prior distribution is not conjugate with respect to the underlying distribution. Typically, the statistician will often select an estimator which is comfortable, familiar, and intuitively sound.

The accuracy of an estimator $\widehat{\theta}_n(x_1,\ldots,x_n)$ is traditionally measured by its variance $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$, where F_θ stands for the underlying cumulative distribution function (cdf) of the random variable X. Having chosen an estimator $\widehat{\theta}_n$, the variance of the estimator $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$ is usually difficult to obtain, especially when the estimator $\widehat{\theta}_n$ and the underlying distribution F_θ are complex. There are several general methods to estimate the variance $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$, for example, the delta-method (see Cramér (1946)), the jackknife-method (see Miller (1974)), and the bootstrap-method (see Efron (1982)). Due to the availability of high speed computers, the jackknife and bootstrap methods have recently become very popular among applied statisticians.

The jackknife-estimator for the variance $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$ is defined by

$$\widehat{\sigma}_J^2 = \frac{n-1}{n}\sum_{j=1}^{n}(\widehat{\theta}_n(j)-\overline{\theta}_J)^2, \tag{1.1}$$

where $\widehat{\theta}_n(j)=\widehat{\theta}_n(x_0\backslash x_j)$, $x_0\backslash x_j=(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)$ =(the sample with the jth observation x_j deleted), and $\overline{\theta}_J$ is the average of $\widehat{\theta}_n(j)$, $j=1,2,\ldots,n$.

The bootstrap-estimator for the variance $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$ is defined by

$$\widehat{\sigma}_B^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\widehat{\theta}_n(x_i^*)-\overline{\theta}_B)^2, \tag{1.2}$$

where $x_i^*=(x_{i1}^*,\ldots,x_{in}^*)$, $i=1,\ldots,k$, are k resamples of size n from the empirical distribution F_n, $\widehat{\theta}_n(x_i^*)=\widehat{\theta}_n(x_{i1}^*,\ldots,x_{in}^*)$ and $\overline{\theta}_B$ is the average of $\widehat{\theta}_n(x_i^*)$, $i=1,\ldots,k$.

The monograph given by Efron (1982) provides an excellent review of both the jackknife and bootstrap methods for estimating the variance $\mathrm{Var}(\widehat{\theta}_n|F_\theta)$.

Bootstrap and Jackknife methods are mainly used for computing the variance of an estimator. They are not designed to search for the optimal estimator. The main purpose of this manuscript is to develop a new computer intensive resampling algorithm, which we call the Method of Pao-Zhuan Yin-Yu, which will be able (a) to find an improved estimator $\tilde{\theta}_n$ from an initial estimator $\widehat{\theta}_n$, in the sense of having smaller variance (or mean square error), and (b) to estimate the variance $\mathrm{Var}(\tilde{\theta}_n|F_\theta)$ for this improved estimator $\tilde{\theta}_n$. Several numerical examples are given, including an example of an autoregressive model where the observations are dependent. The results show that the method of Pao-Zhuan Yin-Yu performs well in almost all cases.

We were unable to find a proper English name for our method so we decided to use a Chinese name for our method. In Chinese, the phrase "Pao-Zhuan Yin-Yu" means "taking a humble initiative in hoping that some excellent results will be generated from the initiation". The meaning of the phrase aptly describes our procedure, that is, to start with a simple unbiased estimator, then generate the minimum variance unbiased estimator gradually. This is the reason we named the method "Pao-Zhuan Yin-Yu".

2. Method of Pao-Zhuan Yin-Yu

In order to introduce and justify our method, we begin with several well-known statistical results in point estimation without giving the proofs.

Let $L(\theta|x) = f(x|\theta)$ be the likelihood function of θ pertaining to the sample $x(x = (x_1, \ldots, x_n))$.

Likelihood Principle: If x and y are observed and their likelihood ratio $\lambda(\theta; x, y) = L(\theta|x)/L(\theta|y)$ is independent of θ, then the statistical inference of θ based on the sample x should be the same as the sample y.

Let T be the sub-σ-field generated by the likelihood ratio function $\lambda(\theta; x, y)$. It follows from the definition of sufficiency that T is the minimal sufficient sub-σ-field. Suppose the underlying statistical model is complete and the initial estimator $\widehat{\theta}_n$ is an unbiased estimator; then (by Rao-Blackwell Theorem) the estimator defined by

$$\widehat{\varphi}_n = E(\widehat{\theta}_n|T) \tag{2.1}$$

is the minimum variance unbiased estimator (mvue) for the unknown parameter θ, i.e. $E\widehat{\varphi}_n \equiv \theta$ and $\mathrm{Var}(\widehat{\varphi}_n) \leq \mathrm{Var}(\widehat{\theta}_n)$ for every unbiased estimator $\widehat{\theta}_n$.

For convenience and simplicity, we assume that the initial estimator $\widehat{\theta}_n$ is an unbiased estimator throughout this paper unless it is specified otherwise. If the initial estimator $\widehat{\theta}_n$ is a biased estimator, then the numerical bias correction method can be used to correct the bias. We shall discuss these details of bias correction in Section 4.

The method of Pao-Zhuan Yin-Yu is based on empirical interpretation of the Rao-Blackwell Theorem by sequential resampling. It contains four fundamental parts; (i) resampling from the population $f(x|\widehat{\theta}_n(x_0))$, (ii) partitioning of resamples, (iii) improving the initial estimator $\widehat{\theta}_n$, and (iv) estimating the variance $\mathrm{Var}(\tilde{\theta}_n|F_\theta)$ of the improved estimator $\tilde{\theta}_n$.

(i) Resampling:

Suppose we take k_1 (k_1 large) independent samples of size n from the pop-

ulation $f(x|\widehat{\theta}_n(x_0))$, say

$$\begin{aligned}(x^*_{11},\ldots,x^*_{1n}) &= x^*_1,\\ (x^*_{21},\ldots,x^*_{2n}) &= x^*_2,\\ &\vdots\\ (x^*_{k_1 1},\ldots,x^*_{k_1 n}) &= x^*_{k_1}.\end{aligned} \tag{2.2}$$

Let $x^*_0 = x_0$ and

$$\Omega^*_1 = \{x^*_0, x^*_1, \ldots, x^*_{k_1}\} \tag{2.3}$$

be the empirical sample space which contains $k_1 + 1$ sample points.

(ii) Contourization:

Define

$$\begin{aligned}[x^*_0]_1 &= \text{A contour on } \Omega^*_1 \text{ generated by } x^*_0 \text{ and likelihood ratio statistic}\\ &= \{x^*_j : x^*_j \in \Omega^*_1, \frac{L(\theta|x^*_j)}{L(\theta|x^*_0)} \text{ independent of } \theta\},\\ &\vdots\\ [x^*_{l_1}]_1 &= \{x^*_j : x^*_j \in \Omega^*_1 \text{ and } \frac{L(\theta|x^*_j)}{L(\theta|x^*_{l_1})} \text{ independent of } \theta\},\end{aligned} \tag{2.4}$$

as $l_1 + 1 (0 \le l_1 \le k_1)$ contours on the empirical sample space Ω^*_1. Hence, the contours $\{[x^*_0]_1, \ldots, [x^*_{l_1}]_1\}$ form a partition of the empirical sample space Ω^*_1, i.e.,

$$\Omega^*_1 = \bigcup_{i=0}^{l_1} [x^*_i]_1 \tag{2.5}$$

and

$$[x^*_i]_1 \cap [x^*_j]_1 = \phi, \text{ for all } i \neq j, \text{ where } \phi \text{ is the empty set.} \tag{2.6}$$

The sub-σ-field T_1 generated by the partition $\{[x^*_0]_1, \ldots, [x^*_{l_1}]_1\}$ will be referred to as the empirical minimal sufficient sub-σ-field.

(iii) Empirical conditional expectation:

Suppose the initial estimator $\widehat{\theta}_n$ is an unbiased estimator. In view of the Rao-Blackwell theorem, we define a new estimator $\tilde{\theta}_n$ on the empirical minimum sufficient-σ-field T_1, as follows; for every $x^* \in [x^*_i]_1$,

$$\begin{aligned}\tilde{\theta}_n(x^*) = \tilde{\theta}_n([x^*_i]_1) &= E^*(\widehat{\theta}_n|[x^*_i]_1)\\ &= \frac{1}{\sharp([x^*_i]_1)} \sum_{x^*_j \in [x^*_i]_1} \widehat{\theta}_n(x^*_j), \quad i = 0, 1, \ldots, l_1\end{aligned} \tag{2.7}$$

where $\sharp([x_i^*]_1)$ stands for the number of resamples on the contour $[x_i^*]_1$ and $E^*(\cdot|[x_i^*]_1)$ stands for the empirical conditional expectation given the contour $[x_i^*]_1$. In particular,

$$\tilde{\theta}_n(x_0) = \tilde{\theta}_n([x_0^*]_1) = \frac{1}{\sharp([x_0^*]_1)} \sum_{x_j^* \in [x_0^*]_1} \hat{\theta}_n(x_j^*). \tag{2.8}$$

In plain words, the new estimator $\tilde{\theta}_n$ at x_0 is defined to be the average of the initial estimator $\hat{\theta}_n$ over the contour $[x_0^*]_1$. Since all the resamples on the contour $[x_j^*]_1$ have the same likelihood function, generally speaking the new estimator $\tilde{\theta}_n$ also can be viewed as a direct consequence of the likelihood principle.

Let us go back to step (i) resampling another $k_2 (k_2 > k_1)$ independent samples of size n from the new population $f(x|\tilde{\theta}_n([x_0^*]_1))$. Denote Ω_2^* as the empirical sample space generated by the new resamples. Repeating steps (ii) and (iii) of contourization and taking empirical conditional expectation over the space Ω_2^*, respectively, yields the improved estimator $\tilde{\theta}_n([x_0^*]_2)$. Repeat this procedure again and again and stop when the sequence $\{\tilde{\theta}_n([x_0^*]_m)\}$ of improved estimators becomes stable. For example, the sequence of improved estimators is stable if

$$\sum_{j=m-l}^{m} |\tilde{\theta}_n([x_0^*]_j) - \tilde{\theta}_n([x_0^*]_{j-1})| < \delta, \tag{2.9}$$

where l is a fixed integer, usually $l = 2, 3, 4$, and δ is a predetermined small positive constant, usually $\delta = 0.001, 0.00001$, or 0.000001. If the procedure is stopped at the mth resampling, we define

$$\tilde{\theta}_n(x_0) = \tilde{\theta}_n([x_0^*]_m), \tag{2.10}$$

as the improved estimator at x_0.

(iv) Estimating the variance of the improved estimator:

If the above procedure is stopped at the mth resampling, the variance $\mathrm{Var}(\tilde{\theta}_n|F_\theta)$ of the improved estimator $\tilde{\theta}_n$ can be estimated by the sample variance,

$$\hat{\sigma}_p^2(m) = \frac{1}{\sharp(\Omega_m^*)} \sum_{j=0}^{l_m} \sharp([x_j^*]_m)(\tilde{\theta}_n([x_j^*]_m) - \bar{\tilde{\theta}}_p)^2, \tag{2.11}$$

where

$$\bar{\tilde{\theta}}_p = \sum_{j=0}^{l_m} \frac{\sharp([x_j^*]_m)}{\sharp(\Omega_m^*)} \tilde{\theta}_n([x_j^*]_m). \tag{2.12}$$

The variance $\mathrm{Var}(\tilde{\theta}_n|F_\theta)$ can be estimated in every stage of the resampling. The sequence of variance estimators $\{\hat{\sigma}_p^2(m)\}$ usually stabilizes faster than the

sequence of estimators $\{\tilde{\theta}([x_0^*]_m)\}$. This can be seen from the numerical results in the next section.

If the underlying distribution is *discrete* and the parameter space Θ contains a finite set of points (or Θ is compact), then we expect the following mathematical results to be true as

$$k_i \to \infty, \ i \to \infty. \tag{2.13}$$

Theorem 1. *If F_θ is complete and condition* (2.13) *is satisfied, then*
(i) *for every $x_0 = (x_1, \ldots, x_n)$,*

$$\tilde{\theta}_n([x_0^*]_i) \to \hat{\varphi}_n(x_0) \tag{2.14}$$

as $i \to \infty$, where $\hat{\varphi}_n$ is minimum variance unbiased estimator defined by (2.1).
(ii) *for every $x_0 = (x_1, \ldots, x_n)$,*

$$\hat{\sigma}_p^2(i) \to \mathrm{Var}(\hat{\varphi}_n | F_{\hat{\varphi}_n(x_0)}) \tag{2.15}$$

as $i \to \infty$.

For continuous distributions, with probability one, no resample will be in the same contour (every contour has probability measure zero). To avoid this problem, we modify our method by the following grouping procedure.

Assuming $f(x|\theta)$ is the density function of a continuous random variable with parameter space Θ the whole line. Let η be a small positive constant and M be a large positive integer. Denote $\{\theta_i\}_{i=1}^m$ as m equally spaced points in the interval with $\theta_1 = -M$ and $\theta_m = M$. For given η, M and m, the contours defined on the observed sample space Ω_1^* are given by

$$\begin{aligned}
[x_0^*]_1 &= \left\{ x_j^* : x_j^* \in \Omega_1^* \text{ and } \left| \max_{1 \le i \le m} \frac{L(\theta_i|x_j^*)}{L(\theta_i|x_0^*)} - \min_{1 \le i \le m} \frac{L(\theta_i|x_j^*)}{L(\theta_i|x_0^*)} \right| \le \eta \right\} \\
&\vdots \\
[x_{l_1}^*]_1 &= \left\{ x_j^* : x_j^* \in \Omega_1^* \text{ and } \left| \max_{1 \le i \le m} \frac{L(\theta_i|x_j^*)}{L(\theta_i|x_{l_1}^*)} - \min_{1 \le i \le m} \frac{L(\theta_i|x_j^*)}{L(\theta_i|x_{l_1}^*)} \right| \le \eta \right\}.
\end{aligned} \tag{2.16}$$

These contours defined here may depend on η, M and m.

Theorem 2. *If F_θ is complete and condition* (2.13) *is satisfied, then as $\eta \to 0$, $M \to \infty$ and $m \to \infty$*
(i) *for every $x_0 = (x_1, \ldots, x_n)$,*

$$\tilde{\theta}_n([x_0^*]_i) \to \hat{\varphi}_n(x_0) \tag{2.17}$$

as $i \to \infty$, where $\hat{\varphi}_n$ is mvue *defined by* (2.1),

(ii) *for every* $x_0 = (x_1, \ldots, x_n)$,

$$\hat{\sigma}_p^2(i) \to \mathrm{Var}(\hat{\varphi}_n | F_{\hat{\varphi}_n(x_0)}) \tag{2.18}$$

as $i \to \infty$.

We omit the mathematical proofs of the theorems in this section. The details of the proofs will be given in the appendix.

Since F_n is the nonparametric maximum likelihood estimate (mle) of the unknown distribution F (see Kiefer and Wolfowitz (1956)), the bootstrap estimate $\hat{\sigma}_B$ is the nonparametric mle of $\sigma(F)$. The Pao-Zhuan Yin-Yu resampling scheme can be viewed as sequential parametric resampling based on the minimum variance unbiased estimate of the value of θ. For n moderate or large, the standard error $\hat{\sigma}_p$ given by (2.11) can be used to obtain approximate confidence intervals for the unknown parameter θ,

$$\tilde{\theta}_n \pm z_{\alpha/2}\hat{\sigma}_p, \tag{2.19}$$

where $z_{\alpha/2}$ is the $\frac{\alpha}{2} \times 100$ percentile point of the standard normal random variable. We expect the above confidence interval (2.19) will perform as well as the bootstrap or jackknife confidence intervals.

Several numerical examples are given in the next section to illustrate the algorithm of the method of Pao-Zhuan Yin-Yu and to demonstrate its efficiency.

3. Numerical Examples

The numerical studies of the method of Pao-Zhuan Yin-Yu provided in this section will be concentrated in two major areas, (i) the rate of convergence according to which the improved estimator $\tilde{\theta}_n$ tends to the minimum variance unbiased estimator $\hat{\varphi}_n$ at the observed sample point x_0, and (ii) the rate of convergence according to which the variance of improved estimator tends to the variance of minimum variance unbiased estimator $\hat{\varphi}_n$ evaluated at $\theta = \hat{\varphi}_n(x_0)$.

For the bootstrap and the jackknife, the assumptions of independence and identical distribution of the observations are vital. It is a great advantage of the method of Pao-Zhuan Yin-Yu that it does not need these strong conditions. It can be easily extended to the observations from a general stationary sequence. To illustrate this, an example is given for estimating the coefficient in an autoregressive model where the observations are dependent.

The method of Pao-Zhuan Yin-Yu includes an algorithm to improve the initial estimator $\hat{\theta}_n$ which the bootstrap and jackknife do not have. Thus, we cannot compare these methods directly on an equal footing as it would be unfair to the bootstrap and jackknife. Therefore, in example (iv) our numerical comparison of

these methods will be carried out only without improving the initial estimator $\widehat{\theta}_n$.

(i) Normal Distribution

Assume we are interested in estimating the parameter θ, the mean of the population, when the variance $\sigma^2 = 1$ is known. Two naive estimators, the sample median $x_{(1/2)}$ and the average of the largest and the smallest observations $(x_{[n]} + x_{[1]})/2$, are considered as initial estimators for a numerical study of the Pao-Zhuan Yin-Yu method. The reasons for starting with this simple example are two-fold. The first reason is that we know the answer that the sample mean $\overline{x}_n$ is the minimum variance unbiased estimator and has a variance $1/n$. Therefore, it is expected that both initial estimators will converge to the $\overline{x}_n$ numerically. The second reason to use this simple example is to make the method of Pao-Zhuan Yin-Yu more transparent.

Suppose an initial sample of forty-nine ($n = 49$) i.i.d. observations is taken from the standard normal population $N(0,1)$ say $x_0 = (x_1, \ldots, x_{49})$, which has mean $\overline{x}_n = 0$, median $x_{(1/2)} = -0.1$ and $(x_{[n]} + x_{[1]})/2 = -0.1$.

(A) Initial naive estimator $= x_{(1/2)}$, $k_i = 700 \times i$, $i = 1, 2, \ldots$, $m_0 =$ the number of resamples on $[x_0^*]_i$ ($\eta = 0.0001$, $M = 100$, $m = 1000$).

Table 1. The rates of convergence of improved estimators and their variances.

i	m_0	$\tilde{\theta}_n([x_0^*]_i)$	$\widehat{\sigma}_p^2(i)$
1	29	0.011	0.0220
2	65	0.001	0.0205
3	96	0.002	0.0203
4	120	0.001	0.0204
5	153	0.000	0.0204

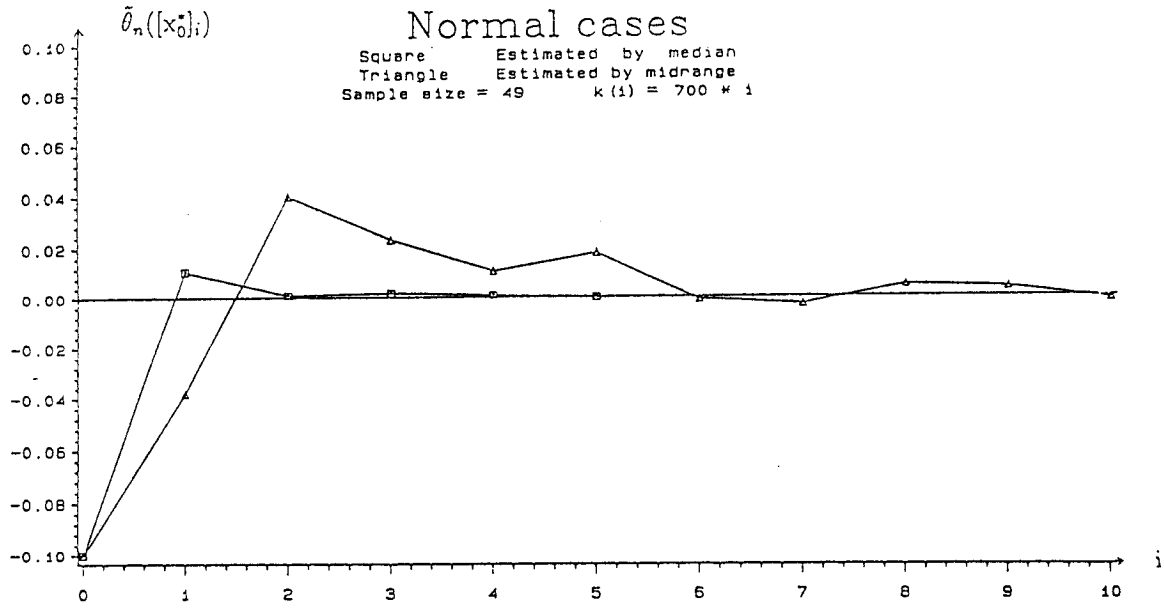
(B) Initial naive estimator $= (x_{[1]} + x_{[n]})/2$, $k_i = 700 \times i$, $i = 1, 2, \ldots$.

Table 2. The rates of convergence of improved estimators and their variances.

$(\eta = 0.0001, M = 100, m = 1000)$

i	m_0	$\tilde{\theta}_n([x_0^*]_i)$	$\widehat{\sigma}_p^2(i)$
1	21	0.038	0.0225
2	55	0.040	0.0221
3	97	0.023	0.0218
4	136	0.011	0.0213
5	189	0.018	0.0208
6	216	−0.001	0.0201
7	246	−0.003	0.0207
8	285	0.005	0.0205
9	313	0.004	0.0204
10	356	−0.001	0.0204

The following Figure 1 illustrates the numerical results in (A) and (B), which shows the rates of convergence of naive estimators toward the mvue at $x_0(\overline{x}_{49} = 0)$.



i: The number of times of resampling

Figure 1. The rates of convergence of improved estimators.

From the above numerical results and Figure 1, we observe several facts. For both initial estimators, the sequences of improved estimators eventually converge to the mvue at x_0, the sample mean $\overline{x}_0 = 0$. To make our comparison of the rates of convergence more transparent and meaningful, we have purposely made $\overline{x}_0 = 0$ and $x_{(1/2)} = (x_{[1]} + x_{[n]})/2 = -0.1$. (This can be done because we do know the sufficient statistic in the case of normal distribution). The rate at which the sequence of improved estimators $\{\tilde{\theta}_n([x_0^*]_i)\}$ with the initial estimator sample median, $x_{(1/2)}$, converges to the mvue, $\overline{x}_0 = 0$, is much faster than the rate of the sequence of improved estimators with the initial estimator $(x_{[1]} + x_{[n]})/2$. It shows very clearly that the procedure of obtaining the optimal estimator (mvue) numerically is independent of selecting the initial unbiased estimator. However, the rate of reaching the optimal estimator does depend on the initial estimator. In general, if the initial estimator is closer to the mvue then it has a better rate of convergence. Similarly, the variances of improved estimators converge to the variance of the sample mean in both cases. Furthermore, the convergence of the sequence $\{\hat{\sigma}_p^2(i)\}$,

$$\lim_{i\to\infty} \hat{\sigma}_p^2(i) = \mathrm{Var}(\hat{\varphi}_n | F_{\hat{\varphi}_n(x_0)}) = 0.0204$$

is independent of the value $\hat{\varphi}_n(x_0)$ (since θ is a location parameter and the variance of $\hat{\varphi}_n$ is independent of the parameter θ) and its rate of convergence is faster than the rate at which the sequence $\{\tilde{\theta}_n([x_0^*]_i)\}$ tends to $\hat{\varphi}_n(x_0)$.

Note that, for the case of a normal distribution, the log-likelihood ratio is

$$\log \lambda(\theta; x_i^*, x_j^*) = -n\theta(\bar{x}_i^* - \bar{x}_j^*) + \text{constant}.$$

It follows that $\log \lambda(\theta; x_i^*, x_j^*)$ is independent of θ if, and only if $\bar{x}_i^* = \bar{x}_j^*$. For given η, M and m, the partition $\{[x_0^*]_i, [x_1^*]_i, \ldots, [x_{l_1}^*]_i\}$ defined by (2.16) on Ω_i^* can be obtained easily. Generally speaking, it is easy to establish the empirical partition on the space Ω_i^* in the case of the exponential family of distributions, but it is a tedious task for many other distributions. The procedure of contourization of Ω_i^* numerically becomes essential in the general case.

(ii) Logistic Distribution

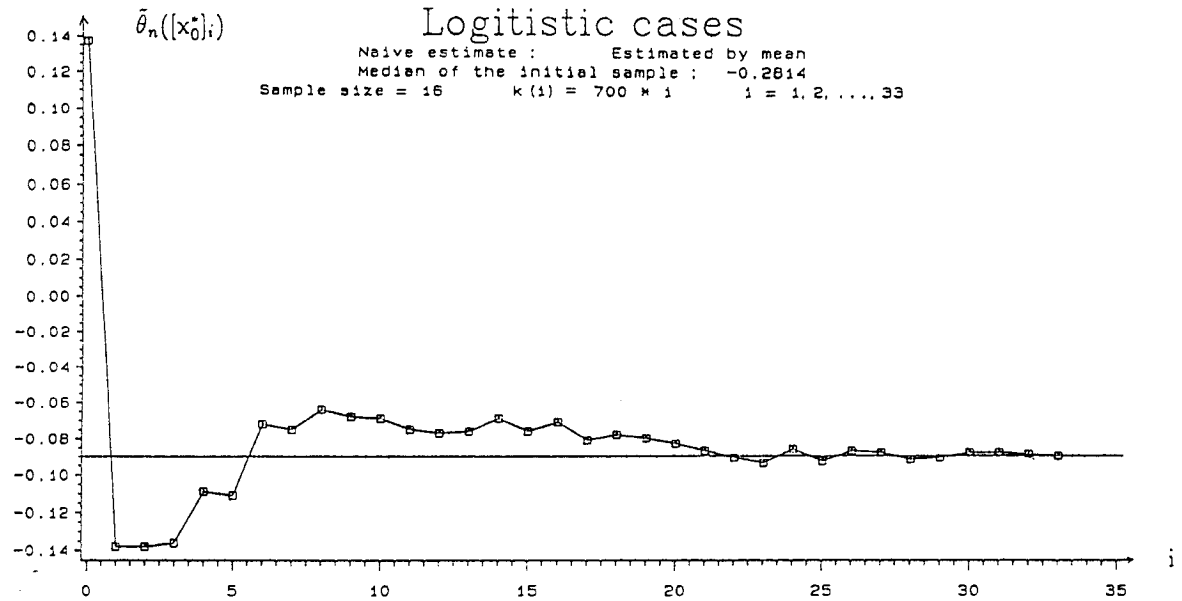
$$f_X(x|\theta) = \frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2}, \quad x \in (-\infty, \infty), \quad \theta \in R.$$

Our numerical study uses $\theta = 0$, $n = 16$, and $k_i = 700 \times i$, $i = 1, 2, \ldots$. The initial sample of sixteen observations $x_0 = (x_1, \ldots, x_{16})$ gives a mean $\bar{x}_n = 0.1373$ and median $x_{(1/2)} = -0.2814$. Suppose the initial naive estimator $\hat{\theta}_n = \bar{x}$ is used. The method of Pao-Zhuan Yin-Yu yields the following numerical results:

Table 3. The rates of convergence of improved estimators.

i	m_0	$\tilde{\theta}_n([x_0^*]_i)$	i	m_0	$\tilde{\theta}_n([x_0^*]_i)$	i	m_0	$\tilde{\theta}_n([x_0^*]_i)$
1	8	−0.1382	12	51	−0.0777	23	103	−0.0944
2	10	−0.1383	13	56	−0.0758	24	106	−0.0868
3	12	−0.1330	14	59	−0.0698	25	115	−0.0931
4	17	−0.1090	15	63	−0.0756	26	121	−0.0944
5	20	−0.1108	16	65	−0.0771	27	124	−0.0876
6	26	−0.0724	17	73	−0.0816	28	133	−0.0924
7	27	−0.0754	18	76	−0.0787	29	140	−0.0913
8	36	−0.0635	19	78	−0.0801	30	144	−0.0863
9	43	−0.0678	20	81	−0.0831	31	147	−0.0882
10	44	−0.0694	21	87	−0.0873	32	152	−0.0897
11	49	−0.0754	22	95	−0.0915	33	157	−0.0906

In this example, we do not know the analytical form of the minimum variance unbiased estimator. It is neither the sample mean $\bar{x}_n$ nor the sample median $x_{(1/2)}$. It can be seen from Figure 2 below that the sequence of improved estimates $\{\tilde{\theta}_n([x_0^*]_i)\}$ converges to -0.09 at a very fast rate. It seems to us that $\hat{\varphi}_n(x_0)$, the minimum variance unbiased estimator at x_0, is approximately -0.09.



i: The number of times of resampling

Figure 2. The rates of convergence of improved estimators.

(iii) Dependent Observations

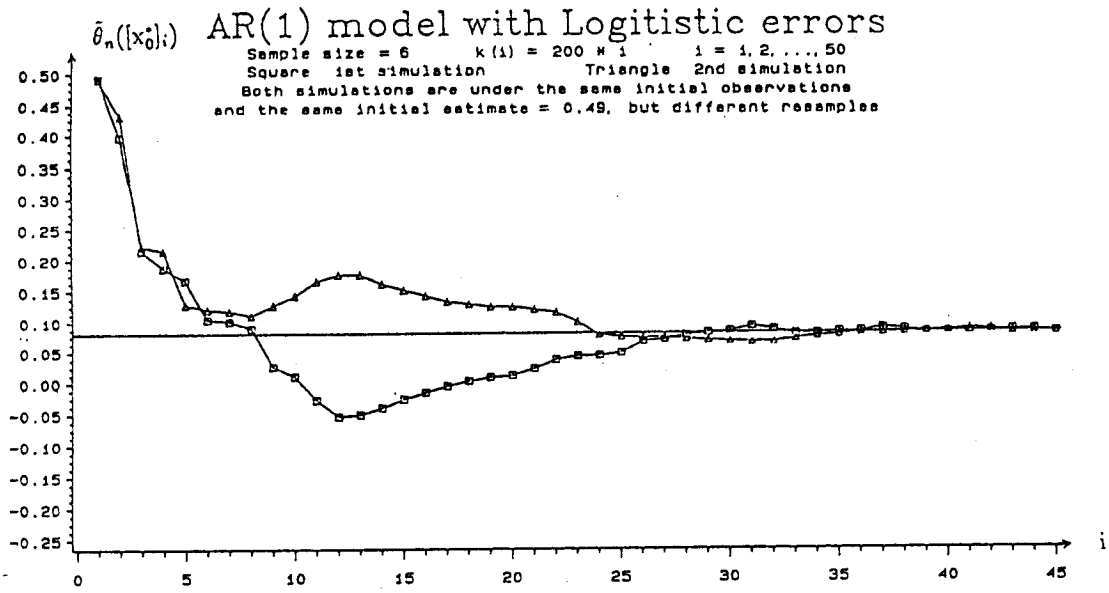
For the bootstrap and the jackknife, the assumptions of independence and identical distribution of the observations are crucial. Künsch (1989) has extended the bootstrap and the jackknife method of estimating standard errors to the case where the observations form a general stationary sequence. His extension is rather complex, artificial and inefficient. The method of Pao-Zhuan Yin-Yu can be easily extended to the observations from a general stationary sequence. Let us consider a simple AR(1) model under the frame work of Künsch (1989),

$$X_t = \beta X_{t-1} + \varepsilon_t, \quad t = 1, 2, \ldots \tag{3.1}$$

where the errors ε_t, $t = 1, \ldots$, are i.i.d. from a logistic distribution with mean zero (without assuming normality to avoid the trivial case). Suppose six observations $x_1 = -0.32$, $x_2 = -0.78$, $x_3 = 1.08$, $x_4 = -0.78$, $x_5 = -3.31$, $x_6 = -3.83$ have been observed from the above AR(1) model. The least squares estimator

$$\widehat{\beta}_0 = \frac{\sum_{t=1}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2} \tag{3.2}$$

is an unbiased estimator for β, but is not a mvue. Assuming $X_0 = 0$ with probability one, it follows from (3.2) that the initial estimate $\widehat{\beta}_0(x_1, \ldots, x_6) = 0.49$. Applying the method of Pao-Zhuan Yin-Yu directly to this model with the least squares estimator as initial estimator and size of resamples $k_i = 700 \times i$, $i = 1, 2, \ldots$, yields the following numerical results (Figure 3).



i: The number of times of resampling

Figure 3. The rates of convergence of improved estimators based on two sets of resamples.

It can be seen that the method of Pao-Zhuan Yin-Yu also performs very well in this case. There is an interesting phenomenon in Figure 3 that, for a fixed initial estimator and size of resamples k_i, the rate of convergence of the sequence of improved estimators $\{\widehat{\beta}([x_0^*]_i)\}$ has very little to do with the different resamples.

(iv) Efron and Tibshirani's Example

How well does the method of Pao-Zhuan Yin-Yu work compared with the bootstrap and the jackknife? To answer this question we present the following example given by Efron and Tibshirani (1985, p.7).

Let R, the real line, be the sample space, $n = 15$, and the estimator $\widehat{\theta}_n$ of interest be the 25% trimmed mean. If the true sampling distribution F_0 is $N(0,1)$, then the true standard error $\sigma(\widehat{\theta}_n|F_0) = 0.286$. The bootstrap estimate $\widehat{\sigma}_B$ is nearly unbiased, averaging 0.287 in a large sampling experiment. The standard error of the bootstrap estimate $\widehat{\sigma}_B$ is itself 0.071 in this case, with a coefficient of variation $0.071/0.287 = 0.25$.

The true standard error of the sample mean is $\sigma(\overline{X}|F_\theta) = 0.258$. Taking the 25% trimmed mean as initial estimate, the Pao-Zhuan Yin-Yu estimate $\widehat{\sigma}_p(\tilde{\theta})$ yields 0.258 in a large sampling experiment. It is independent of the values of improved estimate $\tilde{\theta}_n$. The standard error of the Pao-Zhuan Yin-Yu estimate $\widehat{\sigma}_p$ is 0.012 with a coefficient of variation $0.012/0.258 \doteq 0.047$. Note that the improved estimate $\tilde{\theta}_n$ is the sample mean which has a much smaller standard

error than the 25% trimmed mean. One may feel that the comparison, in this case, is somewhat unfair to the bootstrap since the method of Pao-Zhuan Yin-Yu is based on the optimal estimator (mvue).

In order to have a meaningful comparison, suppose we only apply the Pao-Zhuan Yin-Yu to find the standard error of the trimmed mean. Our numerical results and the Efron and Tibshirani's (1985, p.8) results together yield the following table.

Table 4. Numerical comparison of Pao-Zhuan Yin-Yu, bootstrap, and jackknife.

	Average	Sd. Error	Coeff. Variation
Pao-Zhuan Yin-Yu	0.286	0.052	0.18
Bootstrap	0.287	0.071	0.25
jackknife	0.280	0.084	0.30
Exact	0.286		

One can see from the above results that the method of Pao-Zhuan Yin-Yu performs extremely well compared to both the bootstrap and jackknife methods. However, this comes as no surprise because the Pao-Zhuan Yin-Yu uses a sequential parametric resampling scheme under the correct model and is expected to be more accurate than non-parametric methods such as the bootstrap and jackknife.

Recently, two new sophisticated bootstrap procedures, the percentile-t method and the nested method, have been developed by Efron (1987, 1990). Unlike the general bootstrap procedure (non-parametric resampling from the empirical distribution), both methods take some consideration of the tail probability of resampling population. In many cases, these two new bootstrap procedures perform well, especially for constructing confidence intervals. The two methods mentioned above, and our new method, have one thing in common. They are very computationally demanding. At this moment, we have no strong numerical evidence to say that our method is superior. A large numerical study in this direction is definitely needed.

4. Discussion

In view of all the numerical results given in the previous sections, it is clear that the method of Pao-Zhuan Yin-Yu performs extremely well in almost all cases. It is not only a simple algorithm for searching for the muve, but it also performs well for estimating the variance of the mvue. The rate at which the improved estimator converges to the mvue is highly dependent on the selection of initial estimator. The closer the initial estimator $\hat{\theta}_n$ is to the mvue, the faster the improved estimator $\tilde{\theta}_n$ converges to the muve. The rate also depends on many

other factors, such as, the underlying distribution, the number of resamples k_i, and the structure of the parameter space.

All our numerical results were done on the Digital Vax 8350 computer under the VMS operating system. The SAS program was used for all the computing and the data was generated by the random generator of the SAS program. Given the initial estimator $\widehat{\theta}_n$, the CPU time required to obtain the mvue varies greatly from case to case. For instance, it takes only a few minutes of CPU time for Example 1, but it takes approximately one hour (57 minutes and 14.5 seconds) of CPU time for the logistic distribution in Example 2.

If the parameter space, Θ, contains a finite set of points, say $\theta_1, \ldots, \theta_m$, then it is easy to find the partition $\{[x_0^*]_1, \ldots, [x_{l_1}^*]_1\}$ on the empirical sample space Ω_1^* since, for given x_i^* and x_j^*, we only need check whether the likelihood ratio statistic $L(\theta|x_j^*)/L(\theta|x_i^*)$ is a constant for all m points $\theta_1, \ldots, \theta_m$. If the parameter space, Θ, is an interval of the whole real line and the underlying distribution is continuous, then to find whether two resamples x_1^* and x_j^* belong to the same contour by using a computer is no simple task, and also very time consuming. In general, if η is very small and M and m are very large, the computations can be extensive. Our choices of the quantities η, m and M in above examples are somewhat arbitrary. In order to make the computation manageble, it is our suggestion (based on experience) that η should be the rounding decimal point of the observation but it should never be less than 0.00001 in order to avoid the computer's error entering the computation, M should be around 10 times the largest absolute value of the observations, and m should be less then 10,000.

The conditional distribution of x_j^* given the contour $[x_0^*]_i$ is independent (or nearly independent) of θ and index i. Therefore, we could pool all the resamples on the contours $[x_0^*]_j$, $j = 1, 2, \ldots$, for computing the improved estimator. In other words, regardless of whatever value θ is used to generate the resamples, the value of the improved estimator defined on $[x_0^*]$ is independent of the value θ. On the contrary, the estimator of the variance of improved estimator may depend on the value of θ where the resamples are taken. Hence, one should not pool the previous resamples together to estimate the variance. For i large, all the values of the improved estimator defined on the partition $\{[x_0^*]_i, \ldots, [x_{l_i}^*]_i\}$ and number of resamples on the contours yield an empirical distribution of the minimum variance unbiased estimator $\widehat{\varphi}_n$. This empirical distribution can be used for constructing approximate confidence intervals or α-critical regions of hypothesis testing, etc.

The improved estimator $\tilde{\theta}_n$ given by (2.7) can also be interpreted as a least squares estimator with respect to the initial estimator $\widehat{\theta}_n$ over the contours $\{[x_1^*],$

$\ldots, [x_l^*]\}$ in the following sense: $\tilde{\theta}_n$ minimizes the sum of squares

$$\sum_{x_j^* \in [x_i^*]} (\widehat{\theta}_n(x_j^*) - \theta)^2$$

over all the contours.

If the initial estimator $\widehat{\theta}_n$ is a biased estimator then a bias-reduction procedure can be used to correct the bias. There are many bias-reduction procedures, for example, the procedures based on $E_\theta(\widehat{\theta}_n)$ (see Cox and Hinkley (1974), section 8.4), jackknifing and bootstrapping (see Efron (1982)). Recently, Doss and Sethuraman (1989, p.440) proved that if there does not exist an unbiased estimator for θ then these procedures cannot eliminate the bias completely without make the variance tend to infinity. It is a widely held view that bias reduction is by itself not a desirable property but becomes desirable only if it can be demonstrated that it is also accompanied by a reduction in mean squared error. In view of the above facts, if the minimum variance unbiased estimator is not the prime target (or the mvue does not exist) we can obtain a new and improved estimator by applying bias correction and the method of Pao-Zhuan Yin-Yu to the initial estimator $\widehat{\theta}_n$ simultaneously and continuously. We stop the procedure whenever the bias correction increases the mean square error.

For the Pao-Zhuan Yin-Yu method to work well, one needs to have a parametric model. This is not required by the bootstrap and the jackknife methods. For the non-parametric estimation problems, the bootstrap and jackknife methods remain valuable and indispensable.

With some simple modification of the stopping rule (2.9), the method of Pao-Zhuan Yin-Yu can also be extended directly to the case when both the random variable X and the parameter θ are vectors.

The basic concept of the method of Pao-Zhuan Yin-Yu can be summarized in one sentence: the minimum variance estimator is obtained numerically by interpreting the Rao-Blackwell theorem empirically via sequential resampling. Our interpretation makes the Rao-Blackwell theorem more powerful in its application. With some modification, this fundamental approach can also be extended to obtain other optimal estimators.

5. Appendix

In order to prove our main results, we need the following well-known results. The proofs are omitted.

Let B be the σ-field of the sample space and φ be B-measurable functions.

Lemma 1. *If $E\varphi(x)$ exists and T is an arbitrary sub-σ-field of B then there exists unique equivalent class integrable random variables $E(\varphi|T)$ belonging to T*

and

$$E\varphi = EE(\varphi|T) \tag{5.1}$$

This result follows directly from the Radon-Nikodym theorem (see Halmos (1950) and Chung (1968, pp.276-7)).

Let $\{k_i\}_{i=1}^{\infty}$ be a sequence of positive integer satisfies

$$\lim_{i\to\infty} k_i = \infty, \text{ as } i \to \infty. \tag{5.2}$$

Lemma 2. *If* $\{X_i\}$ *is a sequence of* i.i.d. *random variables with mean* $\mu = EX < \infty$, $\{k_i\}$ *satisfies* (5.2), *and* $\{N_i\}$ *is a sequence of positive integer random variables which satisfies*

$$P(\lim_{i\to\infty} N_i = \infty) = 1, \tag{5.3}$$

then

$$\frac{1}{N_i}\sum_{j=1}^{N_i} X_j \to \mu \tag{5.4}$$

in probability as $i \to \infty$.

This lemma is a simple extension of i.i.d. case of the weak law of large numbers.

Let $\{T_M\}$ be a sequence of sub-σ-fields such that T_M converges to a sub-σ-field $T = \lim_{M\to\infty} T_M$.

Lemma 3. *If* φ *is a* T_M*-measurable and integrable*

$$E(\varphi|T_M) < \infty \tag{5.5}$$

for every M *then*

$$\lim_{M\to\infty} E(\varphi|T_M) = E(\varphi|T). \tag{5.6}$$

The detail proof of this lemma can be found from Chung (1968, Ch.9).

Throughout the following Proof of Theorem 1, we assume that the underlying distribution of the random variable is discrete and the parameter space Θ is compact.

Proof of Theorem 1. It follows from the Rao-Blackwell theorem (see Bickel and Doksum (1977), p.121), that $\widehat{\varphi}_n$ defined by (2.1) is the unique mvue for the parameter θ. For given x_0, it follows

$$\widehat{\varphi}_n(x_0) = E(\widehat{\theta}_n|[x_0^*]). \tag{5.7}$$

Note that the sub-σ-field generated by the likelihood ratio statistics is the minimum sufficient sub-σ-field. Hence, the conditional distribution of x_j^* given x_0^* is independent of θ. Furthermore, the resamples x_j^* on the given contour $[x_0^*]_i$ are independent and identically distributed with the same conditional distribution regardless of the value $\tilde{\theta}_n([x_0^*]_{i-1})$. Let $m_0(i) = \sharp([x_0^*]_i)$ be the number of resamples on the contour $[x_0^*]_i$. For discrete distribution any contour has positive probability, hence, the number of resamples $m_0(i)$ on the contour $[x_0^*]_i$ goes to infinite with probability one as $i \to \infty$. Furthermore, for given i, the sequence of estimates $\{\widehat{\theta}_n(x_j^*) : x_j^* \in [x_0^*]_i, j = 1, 2, \ldots, m_0(i)\}$ are independent and identically distributed. It follows immediately from Lemma 1 and Lemma 2 that

$$\tilde{\theta}_n([x_0^*]_i) = \frac{1}{m_0(i)} \sum_{x_j^* \in [x_0^*]_i} \widehat{\theta}_n(x_j^*) \to \widehat{\varphi}_n(x_0) \tag{5.8}$$

in probability as $i \to \infty (k_i \to \infty)$.

For *given* and *fixed* x_0, it follows from (5.8) that

$$F_{\tilde{\theta}_n([x_0^*]_i)} \to F_{\widehat{\varphi}_n(x_0)} \tag{5.9}$$

as $i \to \infty$. Again, given x_0, it follows from (5.8) and (5.9) that the Pao-Zhuan Yin-Yu variance estimator converges to the variance of mvue: i.e.,

$$\widehat{\sigma}_p^2(i) \to \mathrm{Var}(\widehat{\varphi}_n | F_{\widehat{\varphi}_n(x_0)}) \tag{5.10}$$

as $i \to \infty$. This completes the proof of Theorem 1.

For a continuous random variable, theoretically speaking, with probability zero, resamples could fall into the contour $[x_0^*]_i$ defined by (2.4). The procedure of grouping suggested by (2.16) could void this difficulty. The contour $[x_0^*]_i$ defined by (2.16) may depend on η, M and m. To simplify the proof, we could arrange $\eta = 1/M$ and $m = 2M2^M$ so that the contour $[x_0^*]_i$ depends only on M.

Proof of Theorem 2. It follows from the definition that the sub-σ-field T_M generated by the contours given by (2.16) converges to the minimum sufficient sub-σ-field T as $M \to \infty$, i.e.

$$T_M \to T, \quad \text{as } M \to \infty. \tag{5.11}$$

Denote, for every M,

$$\widehat{\varphi}_M = E(\widehat{\theta} | T_M). \tag{5.12}$$

For each given M, by the same method of proving Theorem 1, we have

$$\widehat{\theta}([x_0^*]_i) \to \widehat{\varphi}_M(x_0) \tag{5.13}$$

as $i \to \infty$. It follows from Lemma 3 that

$$\widehat{\varphi}_M(x_0) \to \widehat{\varphi}(x_0) \tag{5.14}$$

as $M \to \infty$. This proves the first part of the Theorem.

By the same token, the second part of the Theorem also holds. This completes the proof.

Acknowledgments

The authors are grateful to Professor Jerry Klotz and Dr. M. T. Chao for their many useful comments and assistance, and also to a referee for his invaluable suggestions. This work is supported in part by the Natural Science and Engineering Research Council of Canada under grant NSERC A-9216.

References

Bickel, P. J. and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, Inc., San Francisco.

Chung, K. L. (1968). *A Course in Probability Theory.* Harcourt, Brace & World, Inc., New York.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, New Jersey.

Doss, H. and Sethuraman, J. (1989). The price of bias reduction when there is no unbiased estimate. *Ann. Statist.* **17**, 440-442.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. and Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. Technical Report No. 19, Department of Statistics, Stanford University.

Efron, B. (1987). Better bootstrap confidence intervals and bootstrap approximation. *J. Amer. Statist. Assoc.* **82**, 171-185.

Efron, B. (1990). More efficient bootstrap computations. *J. Amer. Statist. Assoc.* **85**, 79-89.

Halmos, P. R. (1950). *Measure Theory.* Litton Educational Publishing.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Künsch, H. (1989). The jackknife and bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217-1241.

Miller, R. G. (1974). The Jackknife–A review. *Biometrika* **61**, 1-15.

Department of Statistics, University of Manitoba, Winnipeg, Canada R3T 2N2.
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

(Received March 1990; accepted June 1991)