Efficient construction of a smooth nonparametric family of empirical distributions and calculation of bootstrap likelihood

Computational Statistics manuscript No.
(will be inserted by the editor)

Efficient construction of a smooth nonparametric family of empirical distributions and calculation of bootstrap likelihood

Bruce J. Worton

Received: date / Accepted: date

Abstract This paper considers the efficient construction of a nonparametric family of distributions indexed by a specified parameter of interest and its application to calculating a bootstrap likelihood for the parameter. An approximate expression is obtained for the variance of log bootstrap likelihood for statistics which are defined by an estimating equation resulting from the method of selecting the first-level bootstrap populations and parameters. The expression is shown to agree well with simulations for artificial data sets based on quantiles of the standard normal distribution, and these results give guidelines for the amount of aggregation of bootstrap samples with similar parameter values required to achieve a given reduction in variance. An application to earthquake data illustrates how the variance expression can be used to construct an efficient Monte Carlo algorithm for defining a smooth nonparametric family of empirical distributions to calculate a bootstrap likelihood by greatly reducing the inherent variability due to first-level resampling.

Keywords Aggregating samples · Bootstrap likelihood · Estimating equations · Exponential tilting · Nonparametric tilting · Smoothing populations

1 Introduction

Davison, Hinkley and Worton (1992) used a nested bootstrap method to generate an analogue of partial likelihood. The basic procedure when applied to an estimator T for a parameter θ and a data set $x_1,\ldots,x_n$, which is assumed to be a random sample from a distribution function F, proceeds as follows. Generate a first-level bootstrap sample $x_1^*,\ldots,x_n^*$ from $\widehat{F}$, the empirical distribution of $x_1,\ldots,x_n$, and calculate the

Bruce J. Worton
School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, U.K.
E-mail: bruce.worton@ed.ac.uk

Algorithm 1 Bootstrap likelihood procedure

1. *First-level resampling:* Generate M first-level bootstrap samples, and for each sample calculate the value of the parameter estimate t^*. Each sample is treated as a population with parameter value t^*.
2. *Second-level resampling:* For each population from step 1, generate second-level bootstrap samples, and for each sample calculate the value of the parameter estimate t^{**}.
3. *Density estimation:* Estimate the bootstrap likelihood associated with parameter value t^* by estimating the conditional density of t^{**} given t^* at the observed value of the parameter estimate for the original data, t.
4. *Curve fitting:* Scatterplot smooth (on the log scale) the M bootstrap likelihood points obtained from step 3 to obtain a curve of bootstrap likelihood.

estimate t^* associated with the sample $x_1^*,\ldots,x_n^*$. We consider such a first-level bootstrap sample as a population $\mathscr{P}^*$ with parameter value t^*. Repeat this step M times to produce populations $\mathscr{P}_1^*,\ldots,\mathscr{P}_M^*$ with parameter values $t_1^*,\ldots,t_M^*$. For each $\mathscr{P}_m^*$ and t_m^*, $m=1,\ldots,M$, use a second-level of bootstrapping to estimate the density of T^{**}, the estimator computed from a second-level sample $x_1^{**},\ldots,x_n^{**}$. In its most general form this is done by using Monte Carlo simulation and kernel density estimation. However, it is usually much more efficient to use a density approximation, e.g. a saddlepoint method (Davison and Hinkley 1988; Kuonen 2005), if it is available to replace the direct simulation at the nested second-level of bootstrap resampling. Evaluating each density at t, the observed value of T for the original data set, gives M likelihood points at $t_1^*,\ldots,t_M^*$. A complete likelihood can be computed by applying a curve-fitting algorithm to these points. Algorithm 1 gives a summary of the steps of the numerical procedure.

An inherent source of variation in the above algorithm is due to the random mechanism used to select the first-level populations and parameter values. However, for each value of t^* in a suitable interval for parameter values, our objective is to determine the expectation of the log bootstrap likelihood with respect to first-level populations with parameters very close to t^*. In Algorithm 1 this is simply achieved by the use of step 4 to remove the substantial variation about the expectation, but it seems highly desirable to reduce this variation at an earlier stage by directly averaging the first-level populations themselves.

To illustrate the variability of simulated populations consider first-level bootstrap samples for the earthquake data set studied in Section 4. The observations are the time intervals in days between the worldwide earthquakes from 1990 to 2010 which resulted in at least 1000 people being killed. In this example there are $n=24$ observations, and the parameter of interest is the mean interval between earthquakes. Figure 1 shows four first-level $\mathscr{P}_m^*$ populations. For each population, some values in the original data set do not appear in the population while other values can appear several times. For example, in population $\mathscr{P}_4^*$ none of the lower values appear but a value just below 500 days appears four times. In fact, the populations shown in Fig. 1 have been simulated to have a parameter value t^*, to be the same as the mean of the original data set, i.e. $t=297.75$, but even so it is evident that the populations differ significantly, and thus have considerable variability, due to the nature of sampling. We note that for this value of the parameter, with $t^*\approx t$, the expected population should

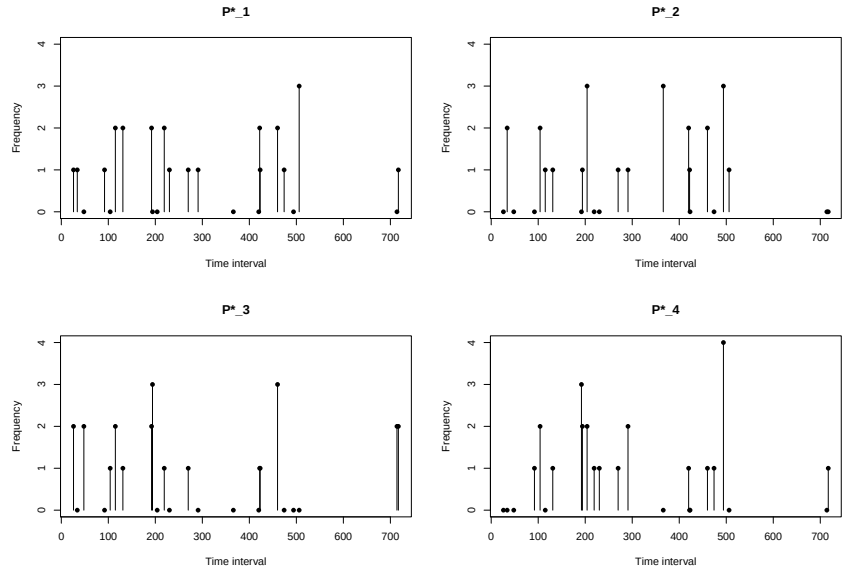

Fig. 1 Frequencies, $np^*_{mi}, i = 1,\ldots,n$, of four first-level $\mathscr{P}^*_m$ populations for the earthquake data set with $n = 24$ discussed in Section 4. Each of these four first-level populations has a parameter value of $t^* = 297.75$ (297.75 is the mean of the original sample)

have n^{-1} at x_i, $i = 1,\ldots,n$, but each of the simulated populations deviates from its expectation fairly markedly.

In order to reduce this undesirable feature, an average population with a parameter value near θ_0 can be defined as the aggregate of several first-level bootstrap samples for which $\theta_0 - \varepsilon \leq t^* \leq \theta_0 + \varepsilon$, with ε chosen as a small positive constant. To aggregate M_{θ_0} such populations we can average the relative frequencies over the first-level bootstrap samples. Specifically, if p^*_{mi} denotes the relative frequency at point x_i in population $\mathscr{P}^*_m$, $i = 1,\ldots,M_{\theta_0}$, then the relative frequency at point x_i in the *aggregated* population is given by

$$p^*_i(\theta_0) = \frac{1}{M_{\theta_0}} \sum_{m=1}^{M_{\theta_0}} p^*_{mi}.$$

For example, in the case of the earthquake data set from Section 4, we could combine the populations with a parameter value of about 297.75, such as those shown in Fig. 1, to obtain a smoothed population, and for this particular parameter value we would expect a population that is similar to the original sample. In practice we would take many more than $M_{\theta_0} = 4$ first-level populations to reduce the variance of each element of $(p^*_1(\theta_0),\ldots,p^*_n(\theta_0))^T$ with respect to the first-level sampling, and thus the variance of log bootstrap likelihood at the parameter value θ_0, to an acceptable level.

Empirical evidence (Davison, Hinkley and Worton 1992, 1995; Davison and Hinkley 1997; Ventura 2002) suggests that such population smoothing is an effective variance reduction technique. However, currently there are relatively limited practical

guidelines on how to best select the number of samples to aggregate when constructing smoothed empirically defined populations. One initial consideration is that some care is needed in resampling as with conventional resampling most populations will have a parameter value fairly similar to the value of the statistic for the original data set, as indicated in Section 2. Furthermore, below we demonstrate that we need to generate many more populations in the regions associated with low likelihood, but unfortunately these regions correspond to parameter values with very low probability of generating populations.

In the present paper we study the use of aggregation in the above nested bootstrap algorithm in more detail for estimators which can be written as the solution of a smooth estimating equation, including the case of the mean. In theory we could conduct a large simulation experiment for each data set considered to determine log bootstrap likelihood variability based on either unsmoothed or various smoothed populations with different levels of aggregation. However, this is not really a viable practical solution, and we thus propose and investigate a fast approximate technique. This can then be routinely applied in each problem to construct a tailor-made Monte Carlo simulation algorithm for defining populations for the calculation of bootstrap likelihood. Certainly there would be no practical advantage in using a far more computationally expensive technique to determine variability over the approximate methods which we investigate in this paper.

An alternative but extremely efficient approach to calculate a form of bootstrap likelihood was proposed by Pawitan (2000), and this avoids the need for the second-level of resampling and for the curve fitting. However, in the current paper, we restrict our attention to the original definition of bootstrap likelihood based on nested resampling, and generating nonparametric populations in order to define a bootstrap likelihood. Within this framework we aim to construct an efficient approach to generate the nonparametric populations.

In Section 2 we obtain an approximate expression for the variance of a log bootstrap likelihood conditional on the value of the parameter t^* which results from the first-level bootstrap resampling variability. Section 3 compares the results obtained by using this formula with variances obtained by direct simulation for artificial data sets based on the quantiles of a normal distribution. In Section 4 the approximate variance is used to design an efficient Monte Carlo simulation algorithm in which the t^* values are selected so that when populations are aggregated the variance of log bootstrap likelihood is approximately constant over the parameter values, and at an acceptable level.

2 Variance approximation for log bootstrap likelihood

We now obtain an approximate expression for the conditional variance of log bootstrap likelihood in the general case of an estimator T which is defined by the unique solution to the estimating equation $n^{-1}\sum_{i=1}^{n} u(x_i,T) = 0$, where $u(x,\theta)$ is a monotonic decreasing function of θ, and is smooth at the data points for the values of θ considered. Here we have a parameter θ determined implicitly by $E\{u(X,\theta)\} = \int u(x,\theta)dF(x) = 0$.

First, the second-level bootstrap density estimator is approximated by (Hinkley and Shi 1989)

$$T^{**} \sim \mathrm{N}(t^*, n^{-1}s^{*2}), \quad s^{*2} = n^{-1}\sum_{i=1}^{n} I_i^{*2},$$

where $I_i^* = I(x_i^*, \mathscr{P}^*) = \frac{u(x_i^*, t^*)}{v^*}$, with $v^* = -n^{-1}\sum_{i=1}^{n}\{\partial u(x_i^*, \theta)/\partial\theta\}|_{\theta=t^*}$, is the empirical influence function of T at the ith sample point x_i^* when sampling from population $\mathscr{P}^*$ with parameter value t^*.

By evaluating this approximate density of T^{**} at the observed value for the original data set, t, we obtain an approximate (partial) bootstrap likelihood for the parameter θ at t^*. Taking the natural logarithm of this likelihood, the approximate log bootstrap likelihood at t^* based on $\mathscr{P}^*$ up to an additive constant is given by

$$l(t^*) = -\tfrac{1}{2}\log s^{*2} - \frac{n(t-t^*)^2}{2s^{*2}}. \tag{1}$$

If $\mathbf{p}^* = (p_1^*, \ldots, p_n^*)^T$ denotes the relative frequencies of the values $x_1, \ldots, x_n$ associated with population $\mathscr{P}^*$, that is $p_i^* = \#(x_j^* = x_i)/n$, then we can express s^{*2} in terms of $\mathbf{p}^*$ and t^* as

$$s^{*2} = \frac{\sum_{i=1}^{n} p_i^* u^2(x_i, t^*)}{\{\sum_{i=1}^{n} p_i^* u'(x_i, t^*)\}^2},$$

where $u'(x_i, \theta) = \partial u(x_i, \theta)/\partial\theta$.

Under simple random sampling of the data $x_1, \ldots, x_n$ to generate the first-level bootstrap samples $x_1^*, \ldots, x_n^*$, $n\mathbf{p}^*$ has a multinomial distribution (Efron and Tibshirani 1993 p. 286), that is

$$\mathbf{p}^* \sim n^{-1}\mathrm{Mult}(n, \mathbf{p}),$$

with $\mathbf{p} = (n^{-1}, \ldots, n^{-1})^T$.

Using large sample properties of the vector $\mathbf{p}^*$, we have that $\mathbf{p}^*$ is approximately multivariate normal with mean vector $\mathbf{p}$ and covariance matrix $\mathbf{V}/n$, where $\mathbf{V} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$ (Mardia, Kent and Bibby 1979 p. 52).

For a particular value of t^*, say t_0^*, in $n^{-1}\sum_{i=1}^{n} u(x_i^*, t_0^*)$, make the transformation from $\mathbf{p}^*$ to $\mathbf{q}^* = (q_1^*, \ldots, q_n^*)^T = \mathbf{A}_{t_0^*}\mathbf{p}^*$, where

$$\mathbf{A}_{t_0^*} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ u(x_1, t_0^*) & u(x_2, t_0^*) & u(x_3, t_0^*) & \cdots & u(x_{n-1}, t_0^*) & u(x_n, t_0^*) \end{pmatrix}.$$

Therefore, using properties of multivariate normal random vectors, the vector $\mathbf{q}^*$ is asymptotically multivariate normal with mean vector $(n^{-1}, \ldots, n^{-1}, n^{-1}\sum_{i=1}^{n} u(x_i, t_0^*))^T$, and covariance matrix $\mathbf{B} = \mathbf{A}_{t_0^*}\mathbf{V}\mathbf{A}_{t_0^*}^T/n$.

It follows that the conditional distribution of $\mathbf{r}^* = (r_1^*, \ldots, r_{n-1}^*)^T = (q_1^*, \ldots, q_{n-1}^*)^T$ given $q_n^* = n^{-1}\sum_{i=1}^n u(x_i^*, t_0^*) = \sum_{i=1}^n p_i^* u(x_i, t_0^*) = 0$ is approximately multivariate normal with mean vector and covariance matrix given by

$$\mathbf{m}_{t_0^*} = E(\mathbf{r}^* | q_n^* = 0) = (n^{-1}, \ldots, n^{-1})^T + \mathbf{B}_{12} B_{22}^{-1} \{-n^{-1} \sum_{i=1}^n u(x_i, t_0^*)\}$$

$$\mathbf{C}_{t_0^*} = \mathrm{var}(\mathbf{r}^* | q_n^* = 0) = \mathbf{B}_{11} - \mathbf{B}_{12} B_{22}^{-1} \mathbf{B}_{21},$$

where $\mathbf{B}_{11}, \mathbf{B}_{12}, \mathbf{B}_{21}$, and B_{22} partition the matrix $\mathbf{B}$ such that

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & B_{22} \end{pmatrix}.$$

We note that in the present context B_{22} is a scalar as the parameter is a scalar.

We can view the mean vector, $\mathbf{m}_{t_0^*}$, as tilted probabilities to ensure the constraint that the population $\mathscr{P}^*$ has specified parameter value t_0^*, rather than the mean of the first-level bootstrap statistics t^* over populations $\mathscr{P}^*$ being t as is the case for the unconstrained and untilted sampling corresponding to simple random sampling (Hinkley and Shi 1989).

Applying the multivariate delta method for large samples (Bishop, Fienberg and Holland 1975 p. 493) to the conditional distribution of $\mathbf{r}^*$ given the particular value t_0^*, yields the variance of $l(t^*)$ at t_0^* as

$$v(t_0^*) = \frac{\partial l(t_0^*)}{\partial \mathbf{r}^*}^T \mathbf{C}_{t_0^*} \frac{\partial l(t_0^*)}{\partial \mathbf{r}^*}, \tag{2}$$

with $\mathbf{r}^*$ evaluated at $\mathbf{m}_{t_0^*}$. The ith element of $\partial l(t_0^*)/\partial \mathbf{r}^*$ in (2), evaluated at $\mathbf{r}^* = \mathbf{m}_{t_0^*}$, is given by

$$\frac{n k_{t,t_0^*}}{2 v_{t_0^*}^2 s_{t_0^*}^4} \left[u(x_i, t_0^*) \{u(x_i, t_0^*) - u(x_n, t_0^*)\} + 2 v_{t_0^*} s_{t_0^*}^2 \left\{ u'(x_i, t_0^*) - \frac{u(x_i, t_0^*) u'(x_n, t_0^*)}{u(x_n, t_0^*)} \right\} \right],$$

where $k_{t,t_0^*} = \{(t - t_0^*)^2 - n^{-1} s_{t_0^*}^2\}$, $s_{t_0^*}^2 = v_{t_0^*}^{-2} \sum_{i=1}^n \mu_i u^2(x_i, t_0^*)$, $v_{t_0^*} = -\sum_{i=1}^n \mu_i u'(x_i, t_0^*)$ and μ_n is determined by $\sum_{i=1}^n \mu_i u(x_i, t_0^*) = 0$ if $(\mu_1, \ldots, \mu_{n-1})^T = \mathbf{m}_{t_0^*}$, provided that $u(x_n, t_0^*) \neq 0$.

From the general case of an estimating equation considered above we can deduce an approximate expression for the conditional variance of log bootstrap likelihood in the special case of the sample mean estimator, $T = n^{-1}\sum_{i=1}^n X_i$, of the population mean, $\theta = E(X) = \int x dF(x)$. With this case we have $u(x_i, t) = x_i - t$, $v^* = 1$, $I_i^* = x_i^* - t^*$ and $s^{*2} = \sum_{i=1}^n p_i^* (x_i - t^*)^2$ with $t^* = \overline{x^*}$. As a consequence of the simpler form of s^{*2} than in the more general case, the corresponding expression for the ith element of $\partial l(t_0^*)/\partial \mathbf{r}^*$ in (2) can be shown to have a simpler form, and is given by

$$\frac{n(x_n - x_i)(t_0^* - x_i)\{(t - t_0^*)^2 - n^{-1} s_{t_0^*}^2\}}{2 s_{t_0^*}^4},$$

where $s_{t_0^*}^2 = \sum_{i=1}^n \mu_i (x_i - t_0^*)^2$.

We can exploit the analysis developed in this section in a practical way to not only study the variance of log bootstrap likelihood as a function of the parameter of interest but also to give guidelines on how much aggregation of first-level bootstrap samples should be applied for a particular data set, and whether the amount of aggregation should vary with the value of the parameter. Studying the results used to derive (2), we can deduce that if K samples with similar values of $t^* \approx t_0^*$ are aggregated to produce an 'average' smoothed population with parameter value $t^* \approx t_0^*$ then the denominator of n in the covariance matrix $\mathbf{B}$ is replaced by nK. Consequently both $\mathbf{C}_{t_0^*}$ and $v(t_0^*)$ are reduced by a factor of K. Of course, the precise parameter value for such an aggregated population may be determined exactly. In Section 3 we will demonstrate that variances of log bootstrap likelihood vary considerably for different values of the parameter. This means that different amounts of aggregation will be appropriate to obtain comparable final variability for different values of t_0^*. However, at a particular value of t_0^*, we can select an appropriate K, say $K_{t_0^*}$, to achieve the variability we require. In Section 4 we employ this result directly to construct a method for combining samples over values of the parameter of interest, with the objective of reducing the variation of log bootstrap likelihood to acceptable levels over a range of suitable parameter values.

3 Numerical study: normal reference

In this section we apply the approximate variance expression obtained in Section 2 to artificial data sets of various sample sizes based on the normal distribution for the case of the sample mean. The data sets were taken as

$$x_i = \Phi^{-1}\left(\frac{i}{n+1}\right), \qquad i = 1,\ldots,n,$$

where Φ^{-1} is the inverse of the standard normal distribution function. The results may be used as a normal-type reference for the level of variation we might expect at a given value of t^*, and thus to provide guidelines for the number of samples it is necessary to aggregate at t^* to achieve a specified reduction in variance.

Approximate variance (2), together with particular values determined by direct simulation, for the case $n = 20$ are plotted in Fig. 2 and illustrate the dependence of the variance on the distance between the first-level population parameter value and the observed value of the statistic for the original data set. We can see that the variance is far from constant as a function of $|t^* - t|$ and is extremely high for the distance of about $|t^* - t| \approx 2sn^{-\frac{1}{2}}$.

Table 1 shows that the variances obtained by applying approximation (2) agree well with simulated values obtained under simple random sampling and exponential tilted random sampling, which is discussed in Section 4, of the data. The results for the larger sample sizes of $n = 100$ and $n = 250$ illustrate that the approximation obtained in Section 2, which was based on a large sample argument, improves as sample size increases. For the smaller sample sizes of $n = 15$ and $n = 20$, it can be seen that the approximate variances obtained from (2) are lower than those obtained using the numerically intensive simulation methods for $\delta^* = 1.5$, but are slightly higher when

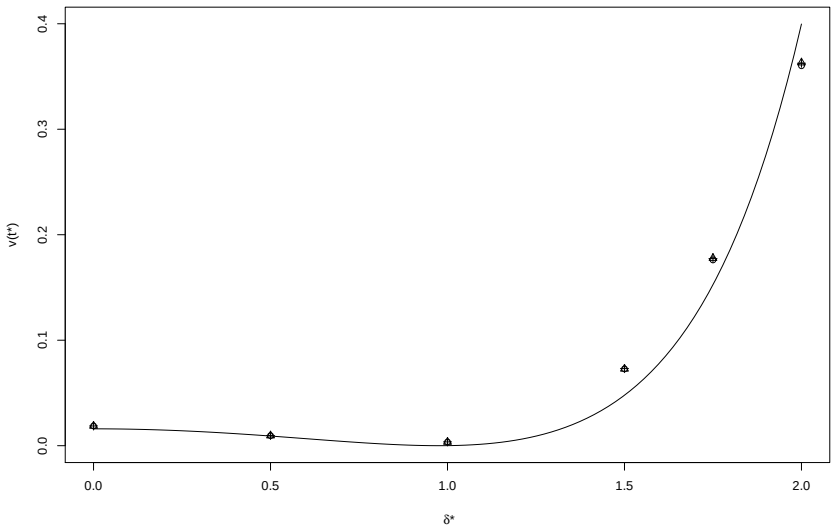

Fig. 2 Approximate conditional variance (2) of log bootstrap likelihood given t^* plotted against $\delta^* = n^{\frac{1}{2}}|t^* - t|/s$, for the artificial data set based on the quantiles of a normal distribution with sample size $n = 20$. Each point shown on the graph was estimated from 50,000 simulated log bootstrap likelihoods with $|T^* - t^*| \leq 0.0005sn^{-\frac{1}{2}}$ calculated under simple random sampling (symbol: $\circ$) and exponential tilted sampling with $\alpha = z/(sn^{\frac{1}{2}})$, $z = 1$ (symbol: $\triangle$), 2 (symbol: $+$)

$\delta^* = 2.0$. However, overall, the approximate variances still give a good indication of the variability relative to the variability at other values of δ^*, and therefore the approximation seems highly suitable for the method which is proposed in Section 4. We also can see that at $\delta^* = 1.0$ the variance is low, a feature which has been observed in empirical studies (e.g., see Fig. 5).

For each sample size, although the variability is low at $|t^* - t| = sn^{-\frac{1}{2}}$, it increases very fast beyond $|t^* - t| = 1.5sn^{-\frac{1}{2}}$, indicating that much greater levels of aggregation are required in regions of low likelihood. Comparing the results for the columns of Table 1 relating to $\delta^* = 0.0$ and $\delta^* = 2.0$ we can see that about 25 times as much first-level bootstrap sample aggregation would be necessary at $\delta^* = 2.0$ as at $\delta^* = 0.0$ to achieve similar levels of variance of log bootstrap likelihood in the case of the sample size of $n = 15$. The corresponding levels of increase are about 20 and 10 times as much aggregation for $n = 20$ and $n = 50$ respectively. Furthermore, if we are using the log bootstrap likelihood to set 95% confidence limits, for example, then variability at $|t^* - t| \approx 2sn^{-\frac{1}{2}}$ is of interest. For these data sets, we should aggregate approximately 70, 40 and 10 first-level samples close to this point for $n = 15$, 20 and 50 respectively to reduce the variance of the log bootstrap likelihood to about 0.1^2, under each of the resampling methods considered. Although to achieve a variance of about 0.05^2 at $|t^* - t| \approx 2sn^{-\frac{1}{2}}$ we would need to aggregate approximately 270, 150

Table 1 Variance of log bootstrap likelihood for artificial data sets with $x_i = \Phi^{-1}\{i/(n+1)\}$, $i = 1,\ldots,n$, at various values of $\delta^* = n^{\frac{1}{2}}|t^* - t|/s$, for sample sizes of $n = 15$, 20, 50, 100 and 250. The approximate conditional variance was calculated using (2), while each variance under simple random sampling (SRS) and exponential tilted sampling with $\alpha = z/(sn^{\frac{1}{2}})$, $z = 1,2$, was estimated from 50,000 simulated log bootstrap likelihoods with $|T^* - t^*| \leq 0.0005sn^{-\frac{1}{2}}$

		δ^*					
n		0.0	0.5	1.0	1.5	1.75	2.0
15	Approximate Eq. (2)	0.0200	0.0115	0.0001	0.0750	0.2559	0.7371
	SRS (untilted)	0.0259	0.0116	0.0065	0.1298	0.2753	0.6679
	Exp. tilted sampling ($z = 1$)	0.0260	0.0117	0.0064	0.1312	0.2751	0.6643
	Exp. tilted sampling ($z = 2$)	0.0254	0.0117	0.0068	0.1331	0.2893	0.6501
20	Approximate Eq. (2)	0.0160	0.0091	0.0000	0.0478	0.1526	0.3997
	SRS (untilted)	0.0187	0.0093	0.0023	0.0702	0.1739	0.3580
	Exp. tilted sampling ($z = 1$)	0.0189	0.0092	0.0023	0.0700	0.1716	0.3548
	Exp. tilted sampling ($z = 2$)	0.0189	0.0093	0.0023	0.0697	0.1742	0.3618
50	Approximate Eq. (2)	0.0075	0.0043	0.0000	0.0151	0.0436	0.0995
	SRS (untilted)	0.0076	0.0041	0.0002	0.0161	0.0414	0.0846
	Exp. tilted sampling ($z = 1$)	0.0077	0.0041	0.0002	0.0159	0.0421	0.0870
	Exp. tilted sampling ($z = 2$)	0.0076	0.0042	0.0002	0.0162	0.0430	0.0874
100	Approximate Eq. (2)	0.0041	0.0023	0.0000	0.0073	0.0203	0.0446
	SRS (untilted)	0.0040	0.0023	0.0000	0.0072	0.0192	0.0404
	Exp. tilted sampling ($z = 1$)	0.0041	0.0023	0.0000	0.0073	0.0196	0.0406
	Exp. tilted sampling ($z = 2$)	0.0041	0.0023	0.0000	0.0073	0.0195	0.0406
250	Approximate Eq. (2)	0.0018	0.0010	0.0000	0.0029	0.0080	0.0172
	SRS (untilted)	0.0018	0.0010	0.0000	0.0029	0.0077	0.0163
	Exp. tilted sampling ($z = 1$)	0.0018	0.0010	0.0000	0.0029	0.0079	0.0167
	Exp. tilted sampling ($z = 2$)	0.0018	0.0010	0.0000	0.0029	0.0078	0.0167

and 40 first-level samples for $n = 15$, 20 and 50 respectively, highlighting the need to generate a sufficient number of samples in regions of low likelihood.

Although the above study gives some indication of the results we might expect in an idealised situation, it does have limitations due to the particular form of the data sets used. Therefore, to investigate how results may vary over simulated data sets, Fig. 3 shows the results of applying the approximation to various data sets of sample size 20 simulated from a normal distribution. For comparison, results obtained by computationally expensive direct simulation are also included, and are similar to the computationally inexpensive approximation. We can see that the variance of log bootstrap likelihood has broadly similar features to the results for the artificial data set of size 20: the variance is extremely high for δ^* near ± 2, but low when δ^* is close to ± 1, and moderate for $\delta^* \approx 0$.

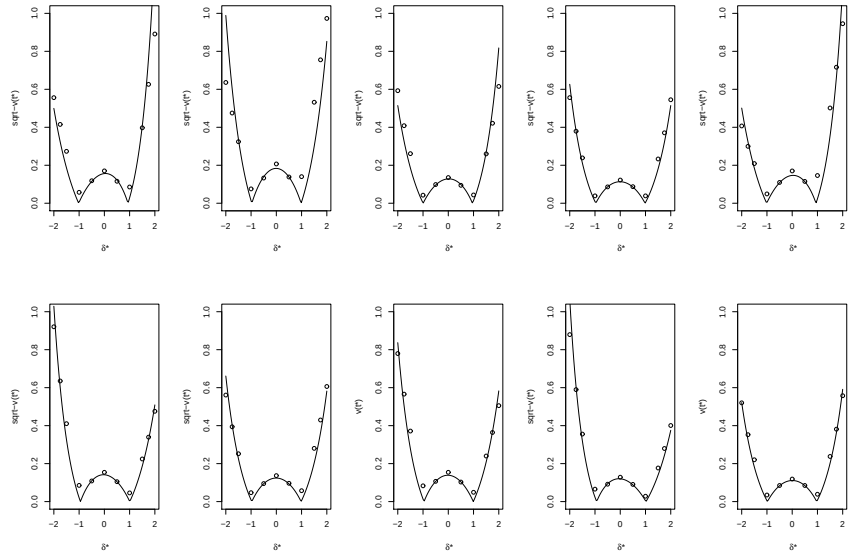

Fig. 3 Approximate conditional variance (2) of log bootstrap likelihood given t^* plotted against $\delta^* = n^{\frac{1}{2}}(t^* - t)/s$, for various simulated normal data sets each with sample size $n = 20$. Each point shown on the graphs was estimated from 50,000 simulated log bootstrap likelihoods with $|T^* - t^*| \leq 0.0005sn^{-\frac{1}{2}}$. (Note that the y-axis is plotted on the square-root scale, as the variances are extremely high for the extreme δ^* values.)

4 Application to earthquake data set

An application of the proposed methodology to calculation of a bootstrap likelihood for earthquake data is considered in this section. Table 2 gives the dates and locations of earthquakes from 1990 to 2010 in which at least 1000 people died. The last column lists the number of days between the earthquakes for the $n = 24$ intervals over the total of 7146 days.

The parameter of interest θ is the mean time interval in days between the earthquakes. The mean interval for the earthquake data set is $t = 297.75$. Figure 4 shows the approximate variance curve $v(t^*)$ for this problem which we will employ to design an efficient method for constructing populations that can be used to calculate bootstrap likelihood. Also shown on the plot are some estimates obtained by direct simulation. As with the artificial data sets of Section 3, for the earthquake data set the variability is extremely high for values of the parameter t^* far from the sample mean of the data, and thus first-level populations associated with these regions require more aggregation than populations associated with the central parameter region near $t = 297.75$. Considering the results in terms of the number of aggregated first-level bootstrap samples required to reduce the variance to, for example, 0.05^2 as presented in the right axis of Fig. 4 shows the extremely high levels of aggregation needed for the higher and lower parameter regions.

Table 2 Earthquake data set (U.S. Geological Survey, 2010). Worldwide earthquakes from June 1990 to January 2010 with at least 1000 deaths

Date (year/month/day)	Location	Interval between earthquakes (days)
1990 06 20	Western Iran	
1990 07 16	Luzon, Philippine Islands	26
1991 10 19	Northern India	460
1992 12 12	Flores Region, Indonesia	420
1993 09 29	Latur-Killari, India	291
1995 01 16	Kobe, Japan	474
1995 05 27	Sakhalin Island	131
1997 05 10	Northern Iran	714
1998 02 04	Hindu Kush Region, Afghanistan	270
1998 05 30	Afghanistan-Tajikistan Border Region	115
1998 07 17	Papua New Guinea	48
1999 01 25	Colombia	192
1999 08 17	Turkey	204
1999 09 20	Taiwan	34
2001 01 26	Gujarat, India	494
2002 03 25	Hindu Kush Region, Afghanistan	423
2003 05 21	Northern Algeria	422
2003 12 26	Southeastern Iran	219
2004 12 26	Sumatra	366
2005 03 28	Northern Sumatra, Indonesia	92
2005 10 08	Pakistan	194
2006 05 26	Indonesia	230
2008 05 12	Eastern Sichuan, China	717
2009 09 30	Southern Sumatra, Indonesia	506
2010 01 12	Haiti Region	104

We first give some general practical considerations before describing an efficient method for implementing calculation of bootstrap likelihood. The results of Section 2 indicate that to achieve approximate homogeneity of variance at M' (usually $M' \ll M$) selected nominal points $\theta_1, \ldots, \theta_{M'}$, with populations defined as the aggregate of first-level bootstrap samples having parameters within bins $\theta_m - \varepsilon \leq t^* \leq \theta_m + \varepsilon$, where $0 < \varepsilon \ll sn^{-\frac{1}{2}}$ for s defined below, we require the number of samples in the mth bin to be proportional to $v(\theta_m)$. Ordinary (untilted) bootstrap sampling at the first-level is clearly a very poor method for generating the populations $\mathscr{P}^*$ as T^* is approximately $\mathrm{N}(t, n^{-1}s^2)$, with $s^2 = n^{-1}\sum_{i=1}^n I_i^2$, where $I_i = x_i - t$ is the empirical influence function of T at x_i when sampling from $\widehat{F}$, and this density is far from being proportional to the variance shown in Fig. 4. We could incorporate rejection sampling to combine K_m samples in the mth population such that $v(\theta_m)/K_m$ is equal to a constant, for example 0.05^2, but this is extremely inefficient for values of θ_m which are far from the mean of the original sample $t = 297.75$.

In practical applications it is much more efficient to generate populations with t^* values that are (in effect) continuous, with reference to $v(t^*)$ as a guide on where t^* values are required, rather than use rejection sampling with a limited set of parameter values. This can be done by use of importance sampling. Exponential tilted

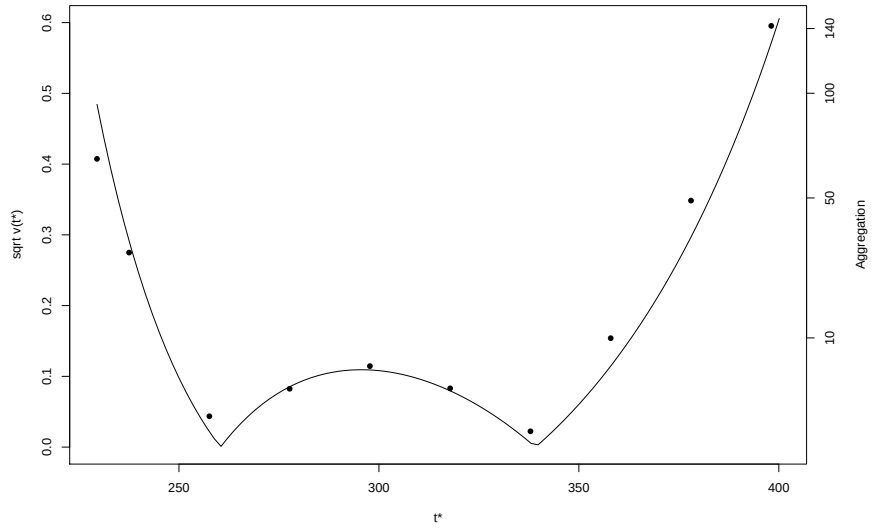

Fig. 4 Approximate conditional variance (2) of log bootstrap likelihood given t^* for the earthquake data with sample size $n = 24$. Each point shown on the graph was estimated from 50,000 simulated log bootstrap likelihoods with $|T^* - t^*| \leq 0.0005sn^{-\frac{1}{2}}$. The right axis gives the number of first-level bootstrap samples it is necessary to aggregate to reduce the variance to 0.05^2. (Note that the y-axis is plotted on the square-root scale, as the variances are extremely high for the lower and higher t^* values.)

resampling (Johns 1988; Hinkley and Shi 1989) was used as an efficient method to suitably and substantially increase the generation of values of t^* in regions of high log bootstrap likelihood variability. With this approach, the resampling probabilities of the data are taken as

$$\Pr(X^* = x_i) \propto \exp(\alpha I_i), \tag{3}$$

for a specified tilting constant α, to give $T^* \sim \mathrm{N}(t + \alpha s^2, n^{-1}s^2)$. By being able to select the tilting constant α appropriately, based on the form of $v(\cdot)$, we have control over placement of the t^* parameter values in regions where they are most needed. To be specific, the following steps of an inversion-type algorithm were used to determine appropriate first-level populations:

1. Calculate a function $G(u) = \int^u v(u)du$ with integration implemented using a simple numerical procedure.
2. Determine a suitable set of tilting constants, $\alpha_m = (\tau_m - t)/s^2$, with τ_m specified with reference to the variance function $v(\cdot)$, such that $\tau_m = G^{-1}\{m/(M+1)\}$, evaluated using interpolation, $m = 1, \ldots, M$.
3. Generate M first-level bootstrap samples with importance sampling using (3) and the tilting constants obtained in step 2.

We have an additional important requirement that M is sufficiently large to ensure that the variability of log bootstrap likelihood, based on aggregated first-level samples, is expected to be below a specified level over values of the parameter. The right

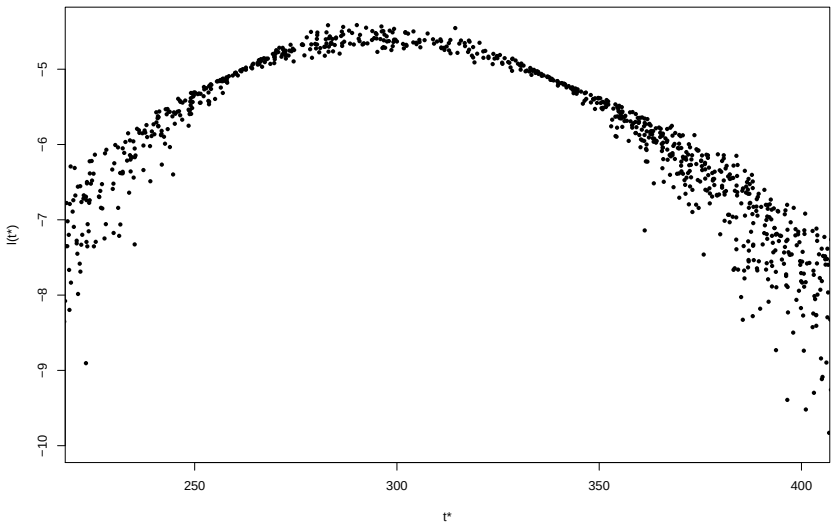

Fig. 5 Log bootstrap likelihood points for the earthquake data using unaggregated first-level populations

axis of Fig. 4 and similar plots for other variance levels provide us with guidance on the selection of an appropriate value for M. In addition, the selection could be based on the local density of the τ_m values, but it is more natural to assess whether sufficient samples have been generated locally over each value in the parameter space by studying the actual t^* values within intervals, as we do in the application to the earthquake data below.

The log bootstrap likelihood points $(t^*, l(t^*))$ from step 3 of Algorithm 1 using the above approach to generate populations $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$ with parameter values $t_1^*, \ldots, t_M^*$ are plotted in Fig. 5 for the earthquake data set. In this application, the value of $M = 1000$ was chosen sufficiently large so that $v(\theta)/K < 0.025^2$, where K is the number of first-level samples with parameter within an interval $\theta - 10 \leq t^* \leq \theta + 10$, for any $240 < \theta < 380$. The likelihood points were obtained by using a density estimator of T^{**} in the second-level bootstrap approximated by $\mathrm{N}(t^*, n^{-1}s^{*2})$, but a saddlepoint approximation density estimator would produce very similar results, with high variability due to the multinomial sampling discussed in Section 2. Note the high levels of variability of log bootstrap likelihood in Fig. 5 for the lower and upper values of t^*, as predicted by Fig. 4.

We now use these populations $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$, which have been generated to suitably increase t^* values in regions of high log bootstrap likelihood variability, to define a smooth family of empirical distributions indexed by the parameter of interest. A convenient way to smoothly aggregate first-level bootstrap samples to obtain a smoothed population with a target value of the parameter, θ^0 say, is to locally average the populations $\mathscr{P}_1^*, \ldots, \mathscr{P}_M^*$ with parameter values $t_1^*, \ldots, t_M^*$, with a kernel-type

smoother (Davison, Hinkley and Worton 1995)

$$p_i^*(\theta^0,\varepsilon) \propto \sum_{m=1}^{M} w\left(\frac{\theta^0 - t_m^*}{\varepsilon}\right) p_{mi}^*, \qquad i = 1,\ldots,n,$$

where $p_{m1}^*,\ldots,p_{mn}^*$ denote the relative frequencies of $x_1,\ldots,x_n$ for population $\mathscr{P}_m^*$, $m = 1,\ldots,M$, with a chosen bandwidth $\varepsilon > 0$ and kernel function $w(.)$. Here $w(.)$ was taken as the standard normal density function. The probabilities $p_i^*(\theta^0,\varepsilon)$ associated with points x_i, $i = 1,\ldots,n$, are now considered as defining a smoothed population $\mathscr{P}^*(\theta^0,\varepsilon)$ for which the precise value of the parameter $\theta(\theta^0,\varepsilon) \approx \theta^0$ can be determined. Although, as noted by Canty, Davison, Hinkley and Ventura (2006) in the context of bootstrap diagnostics, θ^0 and $\theta(\theta^0,\varepsilon)$ are very similar for small or moderate values of the smoothing parameter, e.g. $\varepsilon = 0.2sn^{-\frac{1}{2}}$ to $1.0sn^{-\frac{1}{2}}$. A grid of θ^0 values was used to generate $M' = 100$ populations over an interval of parameter values. Of course, the value of M' is not crucial, but needs to be large enough to give an accurate representation of the curve, as it is the local weighted averaging of the underlying M populations that gives a reduced variance. These smoothed populations were then used in steps 2 and 3 of Algorithm 1 in place of the original unsmoothed samples to compute the bootstrap likelihood.

Figure 6 shows eight independent log bootstrap likelihood curves for the earthquake data set, obtained by repeat applications of Algorithm 1 but with the smoothed populations replacing the unsmoothed populations. Each curve was calculated by using the above aggregation method, with $\varepsilon = 0.3sn^{-\frac{1}{2}}$, $M = 1000$ and $M' = 100$. Evidently, there is a dramatic reduction in the variability when compared with Fig. 5, both within and between the different curves, and this method provides a much more effective use of first-level bootstrap samples than the basic method, especially if the very inefficient (untilted) simple random resampling were to be used to generate first-level bootstrap samples. Also, and perhaps of equal importance, it seems more desirable for likelihood to vary smoothly over an underlying family rather than erratically jump about as the parameter value varies, and this variability to have to be removed by scatterplot smoothing. Note that for the smoothed nonparametric populations step 4 of Algorithm 1 is not necessary as the population smoothing is sufficient to generate a smooth curve of log bootstrap likelihood.

5 Conclusions

In this paper we have shown that by using the properties of first-level bootstrap samples it is possible to obtain an explicit approximate expression for the variance of log bootstrap likelihood. We have applied this expression to suitably generate first-level bootstrap samples in order to define a smooth family of nonparametric distributions indexed by a parameter of interest. From a computational point of view, we have shown that in its implementation we can employ approaches based on tilted sampling to obtain an efficient method for constructing the smoothed populations which are used to compute a curve of log bootstrap likelihood.

One key feature of the unsmoothed bootstrap populations is that they are inherently variable and in particular do not vary smoothly with the parameter of interest,

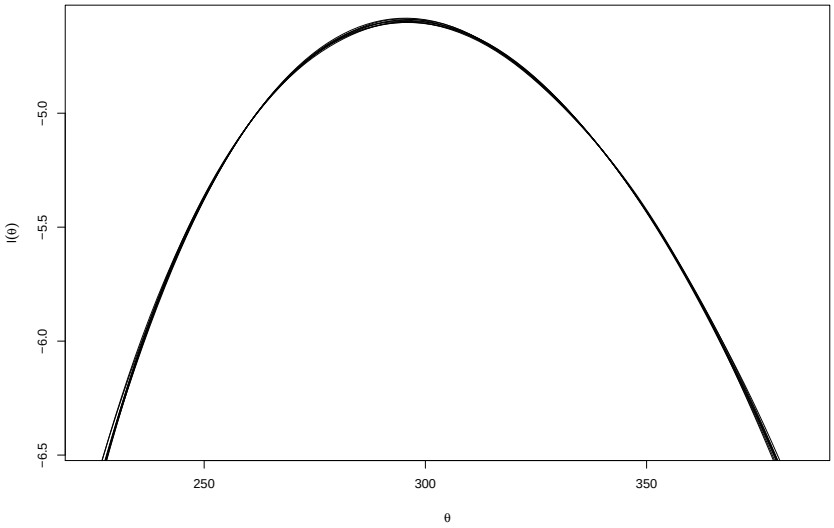

Fig. 6 Eight independent log bootstrap likelihood curves for the earthquake data using smoothed first-level populations. (No scatterplot smoothing, step 4 of Algorithm 1, has been used to generate the curves.)

whichever method of density smoothing is applied to the second-level bootstrap samples. We should note that if kernel density estimation is used, then increasing the level of smoothing leads to grossly biased estimation and is not appropriate in the present context. At the other extreme, it is possible to define a parametric family of empirical distributions which does vary smoothly with the parameter of interest, e.g. an empirical exponential family model. However, by doing this we are not fully exploiting the nonparametric nature of the problem. Therefore, using the smoothed populations seems to be an attractive compromise between using the unsmoothed bootstrap populations which have high variability and using a parametric family of empirical distributions which may place possibly unreasonable constraints on the populations.

Acknowledgements The author would like to thank two anonymous referees for their extremely valuable comments and suggestions on an earlier version of this paper, which substantially improved the content and clarity of the article. The author would also like to thank Professor D.V. Hinkley and Professor A.C. Davison for helpful discussions. The earthquake data set was compiled from various sources and made available by U.S. Geological Survey.

References

1. Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, Massachusetts.
2. Canty AJ, Davison AC, Hinkley DV, Ventura V (2006) Bootstrap diagnostics and remedies. Can J Stat 34:5–27.
3. Davison AC, Hinkley DV (1988) Saddlepoint approximations in resampling methods. Biometrika 75:417–431.

4. Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge.
5. Davison AC, Hinkley DV, Worton BJ (1992) Bootstrap likelihoods. Biometrika 79:113–130.
6. Davison AC, Hinkley DV, Worton BJ (1995) Accurate and efficient construction of bootstrap likelihoods. Stat Comput 5:257–264.
7. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London.
8. Hinkley DV, Shi S (1989) Importance sampling and the nested bootstrap. Biometrika 76:435–446.
9. Johns MV (1988) Importance sampling for bootstrap confidence intervals. J Am Stat Assoc 83:709–714.
10. Kuonen D (2005) Saddlepoint approximations to studentized bootstrap distributions based on M-estimates. Comput Stat 20:231–244.
11. Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London.
12. Pawitan Y (2000) Computing empirical likelihood from the bootstrap. Stat Probab Lett 47:337–345.
13. US Geological Survey (2010) Earthquake Hazards Program. `http://earthquake.usgs.gov`.
14. Ventura V (2002) Non-parametric bootstrap recycling. Stat Comput 12:261–273.