Subsampling of Biased Statistics with Application to Nonparametric Density and Intensity Estimation

Andreas FUTSCHIK
Univ. of Vienna and UC Berkeley
Universitätsstr. 5/9, A-1010 Vienna, Austria

Abstract

We consider the application of bootstrap and subsampling to cases where biased statistics are of interest. A situation where such statistics frequently occur, is in nonparametric estimation problems. We focus on the case of nonparametric density/intensity estimation. An application is the identification times of maximum arrival, for instance of ships in a harbor.

1 Introduction

A frequent statistical goal is to approximate the distribution of some functional $R_n = R_n(\mathbf{X}_n, \theta_n(P))$ of the data $\mathbf{X}_n = (X_1, \ldots, X_n)$ and some parameter sequence $\theta_n(P)$. This is useful when constructing confidence and prediction intervals, doing hypothesis tests, among others. For instance, a confidence interval for the parameter $\theta_n(P) = \theta(P) = \mathbf{E}_P X_1$, for i.i.d. real valued random variables $X_1, \ldots, X_n$ can be obtained by approximating the distribution of $R_n(\mathbf{X}_n, \theta(P)) := n^{1/2}\frac{\bar{X}_n - \theta(P)}{S_n}$, where $\bar{X}_n$ is the sample mean and $S_n = [\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2]^{1/2}$ the sample standard deviation. Obviously $P(-\gamma \leq R_n \leq \gamma) = 1 - \alpha$ implies that $[\bar{X}_n - \gamma\frac{S_n}{\sqrt{n}}, \bar{X}_n + \gamma\frac{S_n}{\sqrt{n}}]$ is an $1-\alpha$ confidence set. Alternatively, confidence intervals could be based on the α trimmed mean

$$\bar{X}_{n,\alpha} = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor}^{n-\lfloor \alpha n \rfloor} X_{[i]}$$

instead of $\bar{X}_n$. An approximation of the distribution of

$$R_n(\mathbf{X}_n, \theta(P)) := n^{1/2}\left(\bar{X}_{n,\alpha} - \theta(P)\right)$$

leads to confidence interval based on the trimmed mean.

As elaborated for instance in Beran and Ducharme (1991), the bootstrap approximation of the distribution of $R_n(\mathbf{X}_n, \theta_n(P))$ is given by the distribution of

$$R^*_{n,n} = R_n(\mathbf{X}^*_n, \theta_n(P_n)),$$

where P_n is an estimate of P and $\mathbf{X}_n^*$ is a sample of size n from P_n. If P_n is chosen to be the empirical distribution function, the resulting approximation is called nonparametric bootstrap. Alternative estimates of P, like a smoothed version of the empirical cdf or a parametric estimate of P lead to different types of bootstrap. Furthermore, if $\theta_n(P_n)$ is not well defined in a problem, it is often replaced by a suitable estimate of $\theta_n(P)$.

Bootstrap is said to "work", if the asymptotic distribution of $R_n(\mathbf{X}_n, \theta(P))$ is reproduced correctly, as $n \to \infty$. This is a minimal condition for the applicability of bootstrap, stating that the distribution estimated by bootstrap approaches the correct distribution as the sample size increases.

Unfortunately there seems to be no easy to check, generally applicable rule to find out whether bootstrap works in a particular situation. However, several examples have been found where bootstrap fails, and classes of functionals and distributions have been identified where the bootstrap is known to work. (See for instance Bickel et al. (1997).)

An example where the bootstrap fails, is for the family of distributions $\mathcal{P} = \{P : \mathbf{E}_P X < \infty\}$ when the goal is to estimate $|\theta(P)|$ with $\theta(P) = \mathbf{E}X_1$, and P is such that $\mathbf{E}_P(X) = 0$. Consider the functionals $R_n(\mathbf{X}_n, \theta(P)) = \sqrt{n}(|\bar{X}_n| - |\theta(P)|)$. Then, with $Z \sim N(0, \sigma^2[X_1])$, $R_n(\mathbf{X}_n, \theta(P))$ converges in distribution against $|Z|$. The bootstrap approximation $R_n^*(\mathbf{X}_n, \theta(P_n))$, on the other hand, converges weakly against $|Z^* + Z| - |Z|$. (See Beran & Srivastava (1985), Dümbgen (1993).)

Subsampling (also known as m-bootstrap) provides an alternative to bootstrap that is known to work under much weaker conditions than bootstrap does. When subsampling is done, the distribution of $R_n(\mathbf{X}_n, \theta_n(P))$ is approximated by that of

$$R_{m,n}^* := R_m(\mathbf{X}_m^*, \theta_m(P_n)),$$

where $\mathbf{X}_m^*$ is a sample of size m from P_n.

Subsampling can be done with or without replacement. Politis and Romano (1994) show that for a sequence of functionals of type

$$R_n(\mathbf{X}_n, \theta(P_n)) = \tau_n(\theta(P_n) - \theta(P)),$$

weak convergence of R_n against some nondegenerate limiting distribution $\mathcal{L}(P)$ is sufficient for subsampling without replacement to work, provided that $\tau_n \to \infty$, $m \to \infty$ and $m/n \to 0$. Often subsampling also works when done with replacement. A detailed discussion concerning consistency and advantages of bootstrap and subsampling with and without replacement can be found in Bickel et al. (1997).

2 Bootstrap and Subsampling in the context of nonparametric estimation problems

Here we focus on resampling applied to density and/or intensity estimation. Intensity functions occur in the context of inhomogeneous Poisson processes, applications include arrival and departure times of ships etc. Nonparametric estimates have been proposed by several authors, including Diggle (1985), Diggle and Marron (1988), and Leadbetter and Wold (1983). See Figure 1 below for an example involving an estimated intensity function.

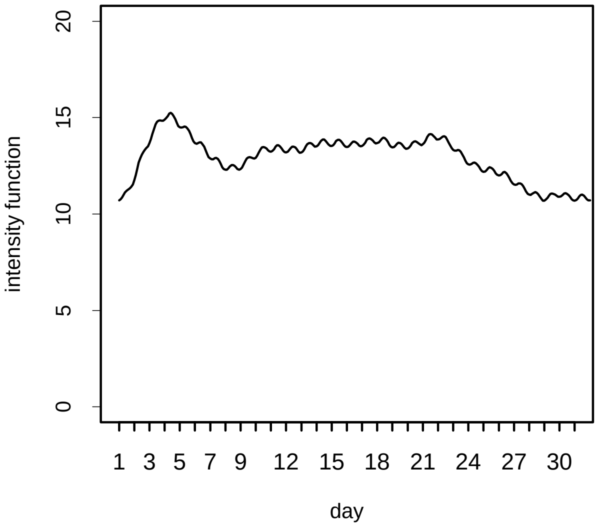

Figure 1: Estimated intensity function for ships arriving at Keelung harbor (Taiwan) in January.

Estimating intensity functions is related to the problem of estimating probability densities. To see this, we will state the problems more formally. When estimating intensities, we assume to have a random number N of observations $X_1, \ldots, X_N$ from an inhomogeneous Poisson process $X(t)$ with intensity function λ on some time interval. To avoid trivialities, assume that λ is positive at least on some interval. Then the intensity density λ^* is related to the intensity function via

$$\lambda^*(x) = \frac{\lambda(x)}{\int_0^t \lambda(s)\,ds}, \quad 0 \leq t \leq 1.$$

Since conditionally on $N = n$, the jump points X_j have the same joint distribution as the order statistics of a sample of n independent observations $X_1, \ldots, X_n$ from density λ^*, the problems of intensity and density estimation are related. Indeed, intensity functions can be estimated in two steps. First estimate λ^* by using a kernel density estimate and then estimate a normalizing constant (the cumulative intensity). An asymptotic analysis of the behavior of intensity estimates is possible in two ways. It might either be assumed that the intensity function is periodic and that the number of observation periods increases. Or, more technically, one might assume that there is a sequence of unknown intensities $\lambda_l = l\mu$ for some density μ and let $l \to \infty$. Both approaches lead to similar results, but we will focus on the first approach involving periodic intensities.

In nonparametric estimation problems, bias is typically non-negligible. Consider for instance a distribution P with twice differentiable density f whose second derivative is bounded. The classical kernel density estimate on $\mathbb{R}^d$ for a sample of $X_1, X_2, \ldots, X_n$ of size n is given by

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Assume that the bandwidth h is chosen of optimal order $h = cn^{-1/(d+4)}$. It is then well known that $\hat{\tau}_n[\hat{f}_n(x) - f(x)]$, with

$$\hat{\tau}_n = \left(\frac{\hat{f}_n(x) \int K^2(u)\,du}{nh^d}\right)^{-1/2}$$

has an asymptotic normal $N(\gamma_f, 1)$ distribution, with $\gamma_f = c^{(d+4)/2} \frac{B_x}{V_x^{1/2}}$,

$$V_x = f(x) \int K^2(u)\,du$$

and

$$B_x = \frac{1}{2} \sum_{|\alpha|=2} D^\alpha f(x) \int u^\alpha K(u)\,du.$$

(See for instance Rosenblatt (1991).) Thus the bias does not vanish asymptotically.

As pointed out by Hall (1992), the bootstrap fails in this context, unless the asymptotic bias is zero. To see why, consider the functional

$$R_n = R_n(\mathbf{X}_n, \theta_n(P)) = \hat{\tau}_n[\hat{f}_n(x) - f(x)]$$

whose distribution is of interest when constructing confidence sets. As shown above, R_n converges in distribution to a normal $N(\gamma_f, 1)$ distribution. The bootstrap version is

$$R_n^* = \hat{\tau}_n[\hat{f}_n^*(x) - \hat{f}_n(x)],$$

with $\hat{f}_n^*(x) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{x - X_i^*}{h_n}\right)$ calculated from a resample $X_1^*, X_2^*, \ldots, X_n^*$ with replacement of size n. Since

$$R_n^* = \hat{\tau}_n[\left(\hat{f}_n^*(x) - \mathbf{E}^*\hat{f}_n^*(x)\right) + \left(\mathbf{E}^*\hat{f}_n^*(x) - \hat{f}_n(x)\right)] = \hat{\tau}_n[\hat{f}_n^*(x) - \mathbf{E}^*\hat{f}_n^*(x)],$$

we see that $R_n^* \to N(0,1)$ in distribution.

Subsampling on the other hand gets the bias right, irrespectively whether it is done with or without replacement. To see why, notice that the subsampling version of R_n is given by

$$R^*_{m,n} = \hat{\tau}_n [\hat{f}^*_{m,h_m}(x) - \hat{f}_n(x)],$$

where $\hat{f}^*_{m,h_m}(x) = \frac{1}{mh_m^d} \sum_{i=1}^m K\left(\frac{x-X_i^*}{h_m}\right)$ is based on a resample $X_1^*, X_2^*, \ldots, X_m^*$ of size m. Now

$$\begin{array}{rcl} R^*_{m,n} & = & \hat{\tau}^{-1}_{m,h_m} [\hat{f}^*_{m,h_m}(x) - \mathbf{E}^* \hat{f}^*_{m,h_m}(x)] + \hat{\tau}^{-1}_{m,h_m} [\mathbf{E}^* \hat{f}^*_{m,h_m}(x) - \hat{f}_n(x)] \\ & =: & U_{m,n} + \tilde{U}_{m,n}. \end{array}$$

Since $U_{m,n} \to N(0,1)$ in distribution, the bias is approximated correctly, if $\tilde{U}_{m,n} \to \gamma_f$ in probability. But this follows since

$$\begin{array}{rcl} \tilde{U}_{m,n} & = & \hat{\tau}_m \left[(\mathbf{E} f_m(x) - f(x)) + (\mathbf{E}^* f^*_m(x) - \mathbf{E} f_m(x)) + [f(x) - f_n(x)]\right] \\ & = & \dfrac{B_x}{V_x^{-1/2}} (m h_m^{d+4})^{1/2} + o_P(1). \end{array}$$

I.e. subsampling works, if the rate-optimal bandwidth is adapted together with the sample size.

Notice that whereas the classical nonparametric bootstrap does not work in this situation, smoothed bootstrap provides another alternative that works under suitable assumptions.

In the context of intensity estimation bootstrap faces the same bias problems as described above. See Cowling, Hall and Phillips (1996) for details. Their resampling method 3 is equivalent to the method described here. Subsampling is able to approximate the distribution of kernel based intensity density estimates correctly. If our goal is to estimate the intensity function itself and assume that it has period 1 and observations are collected up to time t, correct results are obtained by multiplying the kernel density estimate by N/t, where N denotes the total number of observations.

References

[1] Beran R., and Ducharme, G.R. Asymptotic theory for bootstrap methods in statistics. Les Publications CRM, 1991.

[2] Beran, R., Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. Annals of Statistics, 13, 95–115.

[3] Bickel, P. J., Götze, F. and Van Zwet, W. R. (1997) Resampling fewer than n observations: gains, losses and remedies for losses. Statistica Sinica, 7, 1–31.

[4] Cowling, A., Hall, P., Phillips, M.J. (1996). Bootstrap confidence regions for the intensity of a Poisson point process. JASA, 91, 1516-1524.

[5] Diggle, P. (1985). A kernel method for smoothing point process data. Applied Statistics, 34, 138–147.

[6] Diggle, P. and Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. JASA, 83, 793–800.

[7] Leadbetter, M. R., Wold, D. (1983) On estimation of point process intensities. In: Contributions to Statistics: Essays in Honor of Norman L. Johnson, 299–312, ed. P.K. Sen. Amsterdam: North Holland.

[8] Hall, P. (1992) Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann. Statist. 20 675-694.

[9] Politis, D.N., Romano, J.P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist. 22, 2031-2050.

[10] Rosenblatt, M. Stochastic curve estimation. NSF-CBMS Regional conference series, Hayward, 1991.