Computational Statistics and Data Analysis 104 (2016) 130–147

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data

Inés Barbeito *, Ricardo Cao

Research group MODES, Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 1 April 2016
Received in revised form 12 May 2016
Accepted 26 June 2016
Available online 1 July 2016

Keywords:
Kernel method
Mean integrated squared error
Smoothing parameter
Stationary processes
Stationary bootstrap

ABSTRACT

A smoothed version of the stationary bootstrap is established for the purpose of bandwidth selection in density estimation for dependent data. An exact expression for the bootstrap version of the mean integrated squared error under dependence is obtained in this context. This is very useful since implementation of the bootstrap selector does not require Monte Carlo approximation. A simulation study is carried out to show the good practical performance of the new bootstrap bandwidth selector with respect to other existing competitors. The method is illustrated by applying it to two real data sets.

1. Introduction

Nonparametric density estimation is an important research area in Statistics. It is about estimation of the underlying probability density function of a continuous population when only smoothness conditions are assumed. In the i.i.d. case there are classical books such as Silverman (1986) and Devroye (1987), among others, that deal with this fundamental statistical problem.

Let us consider a random sample $(X_1, X_2, \ldots, X_n)$ coming from a population with density f. We focus on the problem of estimating f in a nonparametric way. The Parzen–Rosenblatt (see Parzen, 1962; Rosenblatt, 1956) kernel density estimator is considered:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad (1)$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$, K is a kernel function (typically a probability density function) and $h = h_n > 0$ is the sequence of smoothing parameters or bandwidths. While the kernel is responsible of the regularity of the resulting estimate (continuity, differentiability), the bandwidth is very important to control the degree of smoothing applied to the data. Since this degree of smoothing is a crucial aspect for the quality of the kernel density estimator, a spate of papers have been published, over the past three decades, to propose automatic methods for selecting the smoothing parameter. But the vast majority of them are confined to the i.i.d. case.

After a first bloom of cross-validation methods (see Rudemo, 1982; Chow et al., 1983; Bowman, 1984; Stone, 1984; Marron, 1985, 1987; Hall, 1983; Hall and Marron, 1987a,b; Scott and Terrell, 1987; Stute, 1992; Feluch and Koronacki,

* Corresponding author.
E-mail addresses: ines.barbeito@udc.es (I. Barbeito), rcao@udc.es (R. Cao).

http://dx.doi.org/10.1016/j.csda.2016.06.015

1992), more elaborated plug-in bandwidth selectors (see Park and Marron, 1990; Hall and Marron, 1991; Sheather and Jones, 1991; Jones et al., 1991) and bootstrap bandwidth selectors (see Cao, 1993; Marron, 1992) outperformed the cross-validation approach for i.i.d. data. Extensive comparative simulation studies have been carried out in this context. We only refer to Park and Marron (1990), Cao et al. (1994) and Jones et al. (1996) for the sake of brevity.

If the data are generated by a stochastic process, observed in time, they will no longer be i.i.d. and the dependence structure plays an important role. Under stationarity, $(X_1, X_2, \ldots, X_n)$ is now assumed to be part of a random trajectory of this stochastic process and we focus on the problem of estimating the common marginal density, f, assumed to exist. Of course the estimator in (1) can still be used for dependent data. However its asymptotic properties suffer important changes under dependence. The choice of the smoothing parameter is also a very important issue for dependent data, but very few papers have dealt with data-driven bandwidth selectors under stationary dependence for kernel density estimation. The cross-validation method was modified by Hart and Vieu (1990) in order to produce a more stable procedure under dependence. This method was also theoretically studied by Cox and Kim (1997), who investigated its convergence rates. Hall et al. (1995) obtained an asymptotic expression for the mean integrated squared error and proposed a plug-in bandwidth selector for any data generating process that is an unknown function of a Gaussian process. Some of these methods have been extended for kernel estimation of other related curves under dependence. This is the case of the plug-in bandwidth selector proposed by Quintela del Río (2007) for nonparametric hazard rate estimation. Cao et al. (1993) carried out an extensive simulation study to compare well-known bandwidth selectors (most of them proposed for the i.i.d. case) in a context of serial dependence.

Since the introduction of the bootstrap method (see Efron, 1979) this technique has been extensively used in Statistics to approximate the sampling distribution of statistics of interest. An important problem addressed by the bootstrap is bias and variance estimation. For a comprehensive study about the bootstrap method and its applications, the reader is referred to classical books on the subject such as Efron and Tibishirani (1993) and Davison and Hinkley (1997). As mentioned above, the bootstrap method has been used to produce bandwidth selectors for independent data (see Cao, 1993). The idea is to use the smoothed bootstrap proposed by Silverman and Young (1987) to approximate the mean integrated squared error (MISE) of the kernel density estimator. The bandwidth minimizing the bootstrap version of MISE turns out to be a reasonable empirical analogue for the theoretical smoothing parameter that minimizes MISE.

The bootstrap method has once been used to propose a bandwidth selector for nonparametric kernel density estimation with dependent data. This is the paper by Saavedra and Cao (2001), where a smoothed bootstrap bandwidth selector has been proposed when the data come from a moving average process and the kernel estimator is of convolution-type, adapted to the moving average structure. In this case, the bootstrap mechanism makes an extensive use of the parametric dependence structure as well. The core of that bootstrap resampling plan is just a classical i.i.d. smoothed bootstrap for the residuals of the moving average model.

The bootstrap method has been extensively used in the dependent data setup. The interested reader may consult the review papers by Cao (1999) and Kreiss and Paparoditis (2011) for a general background on the subject. When the dependence structure is general (for instance strong mixing, uniformly mixing, or even general stationary processes) there exist bootstrap methods to deal with relevant inference problems. Just to mention the most popular approaches we consider the Moving Block Bootstrap (MBB) (studied by Künsch, 1989 and Liu and Singh, 1992), the Stationary Bootstrap (SB) (proposed by Politis and Romano, 1994b) and the Subsampling method (proposed by Politis and Romano, 1994a). The stationary bootstrap (see Hwang and Shin, 2012) has been recently used in the context of nonparametric density estimation, but in its original version, i.e. in an unsmooth way. To the best of our knowledge none of the existing bootstrap methods for a general stationary dependence setting has been extended to a smooth bootstrap method. As a consequence none of them have been used or modified to produce a bootstrap bandwidth selector for nonparametric density estimation under general dependence. This is precisely the gap intended to be partially filled by this paper.

The rest of the paper proceeds as follows. Section 2 presents a new resampling plan in this context: the Smoothed Stationary Bootstrap (SSB). An explicit expression for the mean integrated squared error of the kernel density under stationary dependence is included in Section 3. A closed expression for the bootstrap version of MISE is presented in that section and a bootstrap bandwidth selector is proposed. The performance of this bootstrap selector is analyzed via a simulation study in Section 4. Two other bandwidth selectors for dependent data are also considered: the plug-in bandwidth by Hall et al. (1995) and the modified cross-validation bandwidth proposed by Hart and Vieu (1990). Section 5 illustrates the practical performance of the three bandwidth selectors by applying them to the well known lynx and sunspot data sets. Finally, an Appendix A contains the proof of the results stated in Section 3.

2. Smoothed stationary bootstrap (SSB)

In a density estimation context it makes much sense to build a smoothed version of the stationary bootstrap by Politis and Romano (1994b). The key idea is to preserve stationarity of the resampling plan, but producing absolutely continuous resamples, as the regular smoothed bootstrap plan does in the independent data case (see Silverman and Young, 1987).

Let us consider the observed sample, $(X_1, X_2, \ldots, X_n)$, and fix some pilot bandwidth, g. The smoothed stationary bootstrap (SSB) can be presented in two equivalent forms (see Cao, 1999 for the unsmoothed case, SB), proceeding as follows: