Statistics and Probability Letters 83 (2013) 245–256

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

Density estimation using bootstrap bandwidth selector

Arup Bose [a], Santanu Dutta [b,*]

[a] Stat-Math Unit, Indian Statistical Institute, 203 B.T. Road, Kolktata 700108, India
[b] Mathematical Sciences Department, Tezpur University Napaam, 784028, Tezpur, Assam, India

ARTICLE INFO

Article history:
Received 14 May 2012
Received in revised form 29 August 2012
Accepted 30 August 2012
Available online 17 September 2012

MSC:
62G07
62G09
62G20

Keywords:
Kernel density estimator
Bootstrap
Plug-in
Cross-validation
Automatic bandwidth

ABSTRACT

Smoothing methods for density estimators struggle when the shape of the reference density differs markedly from the actual density. We propose a bootstrap bandwidth selector where no reference distribution is used. It performs reliably in difficult cases and asymptotically outperforms well known automatic bandwidths.

1. Introduction

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with an unknown density $f(\cdot)$. The *kernel density estimator* (KDE) of f, based on the kernel $K(\cdot)$ and bandwidth $h \equiv h_n$, is defined as

$$K_n(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - X_i}{h}\right) \tag{1.1}$$

where $h \to 0$ and $nh \to \infty$ as $n \to \infty$. The *mean integrated squared error (MISE)* of $K_n(\cdot)$ is a global measure of accuracy of $K_n(\cdot)$. It has enjoyed great popularity, especially in the context of optimal bandwidth selection of a KDE. See for instance, Taylor (1989), Faraway and Jhun (1990) and Hall et al. (1992). In this article we consider the problem of *bandwidth selection* with a view to achieve the minimum possible value of the MISE (call it M).

Bandwidth selection procedures with this goal in mind have been widely studied over the past decade and several procedures to choose this bandwidth have been proposed in the literature. In particular, the Sheather and Jones (1991) plug-in bandwidth (say h_{SJPI}) and the smooth bootstrap bandwidth proposed by Cao et al. (1994) (say h_{Cao}), have been suggested as new standard methods. See Cao et al. (1994) and Jones et al. (1996) for a detailed comparison of a number of automatic bandwidths. The latter have suggested that bandwidths such as h_{SJPI} be considered as the benchmark of good performance. However, Loader (1999) observed that h_{SJPI} often over-smooths and misses important features when given difficult problems. As we shall see later this criticism is also relevant for h_{Cao}.

[*] Corresponding author.
E-mail addresses: bosearu@gmail.com (A. Bose), tezpur1976@gmail.com (S. Dutta).

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.08.027

A common feature in these bandwidth selectors is that any unknown functional $T(f)$ is approximated by $T(f_n)$, where f_n is another KDE using the same kernel K and a "pilot bandwidth" λ. Loader (1999) pointed out that these bandwidth selectors are heavily dependent on the specification of λ. For instance in the smooth bootstrap method of Cao et al. (1994), λ is chosen with an aim to estimate $\int [f^{(2)}(x)]^2 dx$ accurately. In Jones et al. (1991), λ is selected with a view to minimize asymptotic (relative) MSE for the selected bandwidth. In all these methods, the best choice of λ depends on some functional of the density or its derivatives. For instance, Cao (1993) and Cao et al. (1994) have proposed the choice $\lambda = \frac{C}{n^{1/7}}$ where C depends on $\int [f^{(3)}(y)]^2 dy$. The unknown constants in λ are usually estimated by approximating the underlying density using a reference distribution. If this reference distribution is far removed from f, the smooth bootstrap bandwidths struggle. For instance, Jones et al. (1991, p. 1925) have observed that for densities which are somewhat far from the Gaussian in terms of shape, the performance of their bootstrap bandwidth selector is not so good.

The plug-in bandwidth selectors, such as h_{SJPI}, also exhibit this demerit. In this method, the optimal choice of h is expressed as a function of $\int [f^{(2)}(x)]^2 dx$ (see Loader, 1999), which is approximated using $\int [f_n^{(2)}(x)]^2 dx$. By varying λ, a wide range of "optimal" values of h can be selected. The plot of h against a broad range of values of λ is referred to as the "actual" relation between λ and h. To choose an appropriate value of λ, a common approach is to "assume" a relation between λ and h. Plug-in methods differ with respect to the choice of this relation (see for example, Sheather and Jones, 1991). The Sheather and Jones method uses a complicated "assumed" relation, based on estimating the density derivatives using a reference normal distribution. As a consequence, if f is substantially different from a normal distribution in shape, h_{SJPI} suffers.

The above mentioned bandwidth selectors use some reference distribution to estimate the unknown constants in λ. When the shape of f and the reference density differ widely, the resulting estimates perform poorly. We propose a new smooth bootstrap method where the choice of λ does not involve any pilot estimate, and no reference distribution is used at any stage. A smooth bootstrap bandwidth $\hat{h}$ equals

$$\hat{h} = \text{minimizer of } M^*(h), \quad h \in I,$$

where I is a compact interval and $M^* \equiv M^*(h)$ is a smooth bootstrap estimator of M. It is defined using (another) KDE K_n^0 with kernel K^0 and bandwidth λ. See (3.2) for the definition of M^*.

From (A.7) in the Appendix it is easy to see that for $n\lambda \to \infty$ and $h \in I$,

$$E|M^*(h)/M(h) - 1| = O\left(\frac{1}{n^{1/(2s+1)}}\sqrt{\frac{1}{n\lambda} + \lambda^{2p}} + \sqrt{\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right).$$

Hence the asymptotic accuracy of M^* depends on the accuracy of $K_n^{0(s)}$ in estimating $f^{(s)}$. Our choice of λ is motivated by the following inequality, established in Lemma 1 in the Appendix. Here p, C_1, C_2 are constants which do not depend on f, but depend on the kernel K^0 and the order s of the original kernel K.

$$\int E[K_n^{0(s)}(y) - f^{(s)}(y)]^2 dy \le \frac{C_1}{n\lambda^{1+2s}} + C_2\lambda^{2p} \int [f^{(s+p)}(y)]^2 dy.$$

The minimizer of the right side of the above inequality equals

$$\lambda = \frac{C_3}{\left[\int [f^{(s+p)}(y)]^2 dy\right]^{1/(2s+2p+1)}} n^{-1/(2s+2p+1)},$$

where C_3 is a constant which depends on K and K^0. The coefficient $C_3/[\int [f^{(s+p)}(y)]^2]^{1/(2s+2p+1)}$ varies widely depending on the choice of f. We observe that within a class of mixed normal densities, this coefficient varies approximately from $\frac{1}{9}$ to 1.3 depending on the choice of f. Through extensive simulations we find that

$$\lambda = \frac{1}{8} n^{-1/(2s+2p+1)}, \quad \text{where } s, p \ge 2,$$

works very well. With this choice of λ, let $\hat{h}^*$ be the bandwidth minimizing M^* in I. This is our recommended bootstrap bandwidth and it works well in capturing important features of a wide variety of densities. In particular, for a second order kernel K, $p = s = 2$.

In Section 2 we report a detailed simulation study and analysis of a real data set. Simulations demonstrate that for a second order kernel, our bootstrap bandwidth can perform much better than h_{SJPI} and h_{Cao} bandwidths in a number of difficult problems – especially when f exhibits a number of peaks and sample size is moderate. In Theorem 1 of Section 3, we obtain the L_1 rate at which $\hat{h}^*$ succeeds in minimizing the M as sample size is increased. Its proof is given in the Appendix.