proceedings

Extended Abstract

Computationally Efficient Bootstrap Expressions for Bandwidth Selection in Nonparametric Curve Estimation [†]

Inés Barbeito * and Ricardo Cao

Research group MODES, Department of Mathematics, CITIC, Universidade da Coruña, 15071 A Coruña, Spain; rcao@udc.es

* Correspondence: ines.barbeito@udc.es; Tel.: +34-881-011-301

† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 17 September 2018

Abstract: Bootstrap methods are used for bandwidth selection in: (1) nonparametric kernel density estimation with dependent data (smoothed stationary bootstrap and smoothed moving blocks bootstrap), and (2) nonparametric kernel hazard rate estimation (smoothed bootstrap). In these contexts, four new bandwidth parameter selectors are proposed based on closed bootstrap expressions of the MISE of the kernel density estimator (case 1) and two approximations of the kernel hazard rate estimation (case 2). These expressions turn out to be very useful since Monte Carlo approximation is no longer needed. Finally, these smoothing parameter selectors are empirically compared with the already existing ones via a simulation study.

Keywords: hazard rate; Kernel Method; Mean integrated squared error; moving blocks bootstrap; Smooth Bootstrap; smoothing parameter; stationary bootstrap; Stationary Processes

1. Introduction

This work deals with the well known problem of data-driven choice of smoothing parameters in nonparametric density and hazard rate estimation (see [1–4]). Our aim is also to propose new bootstrap procedures for nonparametric density estimation considering dependent data. On the other hand, hazard rate estimation is considered and two bootstrap bandwidth selectors based on some approximation of the kernel hazard rate estimator are proposed.

2. Nonparametric Density Estimation

Let us consider a random sample, $(X_1, \ldots, X_n)$, coming from a population with density f and the kernel density estimator (see [5,6]), which strongly depends on a bandwidth selector, h. In fact, its choice is really important since it regulates the degree of smoothing applied to the data.

In this context, the smoothed stationary bootstrap (SSB) resampling plan has been proposed (see the Appendix for a detailed description of the algorithm and [7]), as well as a bandwidth selector, namely h^*_{SSB}. It is the result of minimizing the SSB version of the MISE. A closed expression for the bootstrap MISE is also obtained by [7]. On the other hand, smoothed moving blocks bootstrap (SMBB) has been proposed (see the Appendix for a complete description of the method), as well as a bandwidth selector, h^*_{SMBB}, which is the minimizer in h of the closed expression for the $MISE^*_{SMBB}$ (see [8] for a deeper insight on the topic). It is worth mentioning that the exact expressions for the $MISE^*_{SSB}(h)$ and $MISE^*_{SMBB}(h)$ are really useful since Monte Carlo approximation is no longer necessary.

3. Nonparametric Hazard Rate Estimation

Let us consider $(X_1, X_2, \ldots, X_n)$, a simple random sample coming from a population with continuous density f and cumulative distribution function F. Consider, additionally, the nonparametric hazard rate estimator (see [3,4]), the kernel density estimator $\hat{f}_h$ and the kernel distribution estimator $\hat{F}_h$. In order to establish a bootstrap bandwidth selector for the hazard rate estimator, two approximations of the hazard rate estimator are considered. The two hazard rate approximated versions are given by:

$$\begin{aligned}\tilde{r}_{h,1}(x) &= \frac{\hat{f}_h(x)}{1-F(x)}.\\ \tilde{r}_{h,2}(x) &= \frac{1}{1-F(x)}\hat{f}_h(x) + \frac{f(x)}{(1-F(x))^2}\hat{F}_h(x) - \frac{f(x)}{(1-F(x))^2} + r(x).\end{aligned}$$

Closed-form expressions of the MISE of $\tilde{r}_{h,1}$ and $\tilde{r}_{h,2}$, as well as their bootstrap versions can be found in [9]. Moreover, two bootstrap bandwidth selectors, namely h_{BOOT1} and h_{BOOT2}, are defined as the minimizers of $MISE^*_{\tilde{r}_{h,1},w}(h)$ and $MISE^*_{\tilde{r}_{h,2},w}(h)$, respectively (see [9] for a deeper insight on the approach). It is worth mentioning that Monte Carlo approximation is not required.

4. Simulation Results

A simulation study is now carried out in order to check the good empirical behaviour of the new smoothing parameter selectors in both contexts. These are the models considered:

1. **Density estimation:** An AR(1) model given by $X_t = -0.6X_{t-1} + 0.8a_t$, where $a_t \overset{d}{=} N(0,1)$.
2. **Hazard rate estimation:** A Gumbel model such that $f(x) = e^{-x}e^{-e^{-x}}, \forall x \geq 0$.

5. Discussion

Figure 1 shows that h^*_{SSB} and h^*_{SMBB} display a similar performance, actually the best one. According to Table 1, h_{BOOT1} and h_{BOOT2} display the overall best performance.

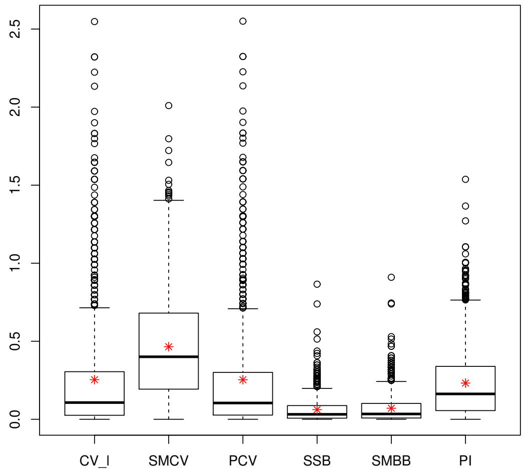

Figure 1. Boxplot of $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$, $n = 100$, where $\hat{h} = h_{CV_l}$ (first box), h_{SMCV} (second box), h_{PCV} (third box), h^*_{SSB} (fourth box), h^*_{SMBB} (fifth box) and h_{PI} (sixth box).

Table 1. Mean and median of $ISE(\hat{h})$, $n = 100$, where $\hat{h} = h_{CV}$ (third column), h_{DO} (fourth column), h_{BOOT1} (fifth column), h_{BOOT2} (sixth column) and h^*_{GCM} (seventh column).

		CV	**DO**	**BOOT1**	**BOOT2**	**GCM**
Gumbel model	Mean	0.1656	0.01651	0.02914	0.02882	0.03595
	Median	0.15527	0.01037	0.012844	0.01282	0.01739

Funding: The authors acknowledge partial support by MINECO grants MTM2014-52876-R and MTM2017-82724-R (EU ERDF support included). Additionally, financial support from the Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva ED431C2016-015) and the European Union (European Regional Development Fund - ERDF), is gratefully acknowledged. The first author aknowledges financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF), the reference of which is ED481A-2017/215. Additionally, the work of the first author has been partially carried out during a visit at the University of California, San Diego, financed by INDITEX, with reference INDITEX-UDC 2017.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MISE	Mean integrated squared error
ISE	Integrated squared error
SSB	Smoothed stationary bootstrap
SMBB	Smoothed moving blocks bootstrap
iid	Independent and identically distributed
h_{DO}	DO-validation bandwidth selector for hazard rate estimation (see [10])
h^*_{GCM}	González-Manteiga, Cao, Marron bandwidth selector for hazard rate estimation (see [11])
h_{PI}	Plug-in bandwidth selector for bandwidth selection with dependent data (see [12])
h_{CV_l}	Leave-$(2l+1)$-out cross-validation for density estimation (see [13])
h_{SMCV}	Modified cross validation for density estimation with dependent data (see [8])
h_{PCV}	Penalized cross validation for density estimation with dependent data (see [8])
h_{CV}	Cross validation bandwidth selector for hazard rate estimation (see [14])
h_{MISE}	Bandwidth selector which minimizes the theoretical MISE(h)

Appendix A

Smoothed stationary bootstrap

1. Draw $X_1^{*(SB)}$ from F_n, the empirical distribution function of the sample.
2. Define $X_1^* = X_1^{*(SB)} + gU_1^*$, where U_1^* has been drawn with density K and independently from $X_1^{*(SB)}$.
3. Assume we have already drawn $X_1^*, \ldots, X_i^*$ (and, consequently, $X_1^{*(SB)}, \ldots, X_i^{*(SB)}$) and consider the index j, for which $X_i^{*(SB)} = X_j$. We define a binary auxiliary random variable I_{i+1}^*, such that $P^*\left(I_{i+1}^* = 1\right) = 1 - p$ and $P^*\left(I_{i+1}^* = 0\right) = p$. We assign $X_{i+1}^{*(SB)} = X_{(j \mod n)+1}$ whenever $I_{i+1}^* = 1$ and we use the empirical distribution function for $X_{i+1}^{*(SB)}|_{I_{i+1}^*=0}$, where $\mod$ stands for the modulus operator.
4. Once drawn $X_{i+1}^{*(SB)}$, we define $X_{i+1}^* = X_{i+1}^{*(SB)} + gU_{i+1}^*$, where, again, U_{i+1}^* has been drawn from the density K and independently from $X_{i+1}^{*(SB)}$.

Smoothed moving blocks bootstrap

1. Fix the block length, $b \in \mathbb{N}$, and define $k = \min_{\ell \in \mathbb{N}} \ell \geq \frac{n}{b}$

2. Define:

$$B_{i,b} = (X_i, X_{i+1}, \ldots, X_{i+b-1})$$

3. Draw $\xi_1, \xi_2, \ldots, \xi_k$ with uniform discrete distribution on $\{B_1, B_2, \ldots, B_q\}$, with $q = n - b + 1$
4. Define $X_1^{*(MBB)}, \ldots, X_n^{*(MBB)}$ as the first n components of

$$(\xi_{1,1}, \xi_{1,2}, \ldots, \xi_{1,b}, \xi_{2,1}, \xi_{2,2} \ldots, \xi_{2,b}, \ldots, \xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,b})$$

5. Define $X_i^* = X_i^{*(MBB)} + gU_i^*$, where U_i^* has been drawn with density K and independently from $X_i^{*(MBB)}$, for all $i = 1, 2, \ldots, n$

References

1. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall: London, UK, 1986.
2. Devroye, L. *A Course in Density Estimation*; Birkhauser: Boston, MA, USA, 1987.
3. Watson, G.S.; Leadbetter, M.R. Hazard analysis I. *Biometrika* **1964a**, *51*, 175–184.
4. Watson, G.S.; Leadbetter, M.R. Hazard analysis II. *Sankhyā Ser. A* **1964b**, *26*, 101–116.
5. Parzen, E. Estimation of a probability density-function and mode. *Ann. Stat.* **1962**, *33*, 1065–1076.
6. Rosenblatt, M. Estimation of a probability density-function and mode. *Ann. Stat.* **1956**, *27*, 832–837.
7. Barbeito, I.; Cao, R. Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data. *Comput. Stat. Data Anal.* **2016**, *104*, 130–147.
8. Barbeito, I.; Cao, R. A review and some new proposals for bandwidth selection in nonparametric density estimation for dependent data. In *From Statistics to Mathematical Finance: Festschrift in Honour of Winfried Stute*; Ferger, D., González Manteiga, W., Schmidt, T., Wang, J.L., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 173–208, ISBN 978-3-319-50986-0.
9. Barbeito, I.; Cao, R. Smoothed bootstrap bandwidth selection for nonparametric hazard rate estimation. *Preprint* **2018**.
10. Gámiz, M.L.; Mammen, E.; Martínez-Miranda, M.D.; Nielsen, J.P. Double one-sided cross-validation of local linear hazards. *J. R. Stat. Soc. Ser. B Stat.* **2016**, *78*, 775–779.
11. González-Manteiga, W.; Cao, R.; Marron, J.S. Bootstrap Selection of the Smoothing Parameter in Nonparametric Hazard Rate Estimation. *J. Am. Stat. Assoc.* **1996**, *91*, 1130–1140.
12. Hall, P.; Lahiri, S.N.; Truong, Y.K. On bandwidth choice for density estimation with dependent data. *Ann. Stat.* **1995**, *23*, 2241–2263.
13. Hart, J.D.; Vieu, P. Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Stat.* **1990**, *18*, 873–890.
14. Patil, P.N. On the Least Squares Cross-Validation Bandwidth in Hazard Rate Estimation. *Ann. Stat.* **1993**, *21*, 1792–1810.