proceedings

MDPI

Extended Abstract

Bandwidth Selection in Nonparametric Regression with Large Sample Size †

Daniel Barreiro-Ures ‡, Ricardo Cao ‡ and Mario Francisco-Fernández ‡

Department of Mathematics, Faculty of Computer Science, University of A Coruña, A Coruña 15008, Spain; ricardo.cao@udc.es (R.C.); mario.francisco@udc.es (M.F.-F.)

* Correspondence: daniel.barreiro.ures@udc.es

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

‡ These authors contributed equally to this work.

Published: 17 September 2018

Abstract: In the context of nonparametric regression estimation, the behaviour of kernel methods such as the Nadaraya-Watson or local linear estimators is heavily influenced by the value of the bandwidth parameter, which determines the trade-off between bias and variance. This clearly implies that the selection of an optimal bandwidth, in the sense of minimizing some risk function (MSE, MISE, etc.), is a crucial issue. However, the task of estimating an optimal bandwidth using the whole sample can be very expensive in terms of computing time in the context of Big Data, due to the computational complexity of some of the most used algorithms for bandwidth selection (leave-one-out cross validation, for example, has $\mathcal{O}(n^2)$ complexity). To overcome this problem, we propose two methods that estimate the optimal bandwidth for several subsamples of our large dataset and then extrapolate the result to the original sample size making use of the asymptotic expression of the MISE bandwidth. Preliminary simulation studies show that the proposed methods lead to a drastic reduction in computing time, while the statistical precision is only slightly decreased.

Keywords: nonparametric; regression; bandwidth; Big Data; cross-validation; subsampling

1. Scenario

Let us consider a sample of size n, $\{(x_i, y_i)\}_{i=1,\ldots,n}$, drawn from a nonparametric regression model $y_i = m(x_i) + \varepsilon_i$. We assume random design, $\mathbb{E}[\varepsilon \mid x] = 0$ and $\mathbb{E}[\varepsilon^2 \mid x] = \sigma^2(x) < \infty$. In this context, we deal with the Nadaraya-Watson estimator [1] for the regression function, m, which is characterized by the kernel function K and the bandwidth or smoothing parameter $h > 0$. Under suitable conditions, the asymptotically optimal (in the sense of minimum AMISE) bandwidth satisfies

$$h_{AMISE,n} = c_0 n^{-\frac{1}{5}}. \tag{1}$$

Since we are assuming that the sample size, n, is very large, the task of computing a bandwidth selector using the whole sample would be too computationally expensive. For example, the leave-one-out cross-validation (LOO CV) bandwidth selector has complexity $\mathcal{O}(n^2)$.

2. Bandwidth Selection

The idea behind our proposal is to find the LOO CV bandwidth for several subsamples and then extrapolate the result to the original sample size using the asymptotic expression of the MISE bandwidth (1).

2.1. One Subsample Size (OSS)

The idea behind this method is to draw several subsamples of size r, much smaller than n, then compute the LOO CV selector and finally use Equation (1) to extrapolate the CV bandwidth for the original sample size (this idea was already proposed in [2] in the context of kernel density estimation to reduce the variance of the CV bandwidth selector).

1. Obtain s subsamples of size $r \ll n$ subsampling without replacement from our original dataset.
2. For each subsample, find the LOO CV bandwidth.
3. Let $\hat{h}_r$ denote the average of these bandwidths.
4. We estimate the unknown constant c_0 by $\hat{c}_0 = \hat{h}_r r^{\frac{1}{5}}$.
5. Therefore, our estimate of the AMISE bandwidth would be $\hat{h}_{AMISE,n} = \hat{c}_0 n^{-\frac{1}{5}} = \hat{h}_r \left(\frac{r}{n}\right)^{\frac{1}{5}}$.

2.2. Several Subsample Sizes (SSS)

We now propose a method that considers several subsamples of different sizes.

1. Consider a grid of subsample sizes, $r_1, \ldots, r_s$, with $r_j \ll n$.
2. For each r_j, compute the LOO CV bandwidth, $\hat{h}_j$ (several subsamples of each size could be considered).
3. Solve the ordinary least squares problem (or a robust analogue) given by $(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min} \sum_{i=1}^{s} (\log(\hat{h}_i) - \beta_0 - \beta_1 \log(m_i))^2$, in which case $\hat{c} = e^{\hat{\beta}_0}$ and $\hat{p} = \hat{\beta}_1$ is our estimate of the order of convergence of the AMISE bandwidth.
4. Our estimate of the AMISE bandwidth for the original sample size, n, would be $\hat{h}_{AMISE,n} = \hat{c} n^{\hat{p}}$.

3. Simulation Study

Let us consider samples of size $n = 10^6$ drawn from the model $Y = m(X) + \varepsilon$, where $X \sim Beta(2,2)$, $\varepsilon \sim N(0, 0.2^2)$ and $m(x) = 1 + x \sin(5.5\pi x)^2$. Furthermore, we have considered a Gaussian kernel and, as a weight function, $w(x) = 1_{\{F_X^{-1}(0.05) \leq x \leq F_X^{-1}(0.95)\}}$, where F_X^{-1} denotes the marginal quantile function of X.

It is clear from Figure 1 that the OSS selector outperforms the SSS selector in terms of statistical precision. Moreover, in many cases bandwidths that are quite distant from the optimum do not have an associated large error (in terms of AMISE). On the other hand, as we can observe in Tables 1 and 2, the OSS selector is substantially faster than the SSS selector due to the fact that the former works with a single subsample size which, in turn, is even smaller than most of those considered for the SSS selector). It should be noted that the source code for both selectors was written in C++ and run in parallel on an Intel Core i5-8600K 3.6 GHz.

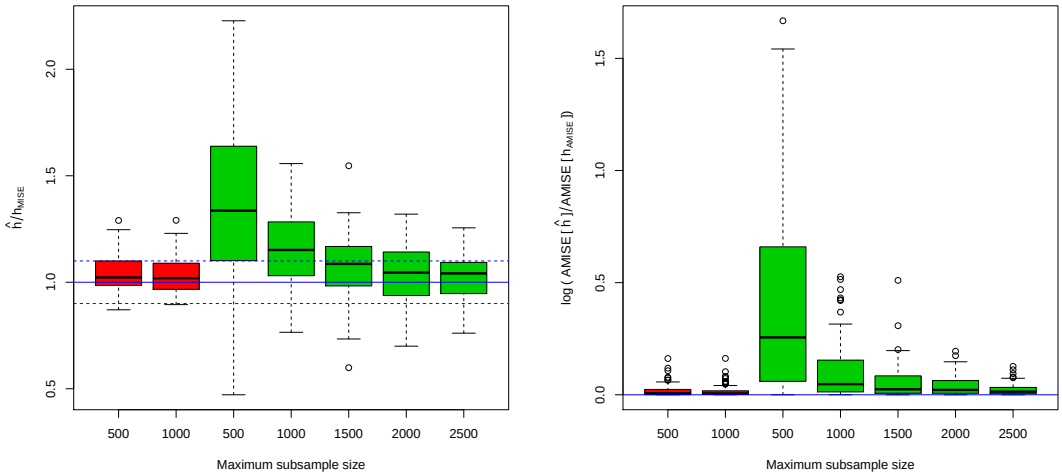

Figure 1. Sampling distributions of $\frac{\hat{h}}{h_{MISE,n}}$ (**left figure**) and $\log\left(\frac{AMISE(\hat{h})}{AMISE(h_{AMISE,n})}\right)$ (**right figure**) for the OSS (red) and SSS (green) bandwidth selectors.

Table 1. CPU elapsed times for the OSS selector with $n = 10^6$. 10 subsamples of the corresponding size were considered.

Subsample Size	CPU Elapsed Time (s)
500	1.62
1000	2.82

Table 2. CPU elapsed times for the SSS selector with $n = 10^6$ considering uniform grids (of 20 elements) of subsample sizes ranging from 100 to the corresponding maximum size. 10 subsamples of each of the corresponding sizes were considered.

Maximum Subsample Size	CPU Elapsed Time (s)
500	9.21
1000	17.9
1500	32.1
2000	51.0
2500	75.1

Author Contributions: Conceptualization, D.B., R.C. and M.F.; Methodology, D.B., R.C. and M.F.; Software, D.B., R.C. and M.F.; Validation, D.B., R.C. and M.F.; Formal Analysis, D.B., R.C. and M.F.; Investigation, D.B., R.C. and M.F.; Resources, D.B., R.C. and M.F.; Data Curation, D.B., R.C. and M.F.; Writing—Original Draft Preparation, D.B., R.C. and M.F.; Writing—Review & Editing, D.B., R.C. and M.F.; Visualization, D.B., R.C. and M.F.; Supervision, D.B., R.C. and M.F.; Project Administration, D.B., R.C. and M.F.; Funding Acquisition, D.B., R.C. and M.F.

Funding: This research received no external funding.

Acknowledgments: This research has been supported by MINECO grant MTM-2014-52876-R and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

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.

References

1. Nadaraya, E.A. On estimating regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142.
2. Wang, Q.; Lindsey, B.G. Improving cross-validated bandwidth selection using subsampling-extrapolation techniques. *Comput. Stat. Data Anal.* **2015**, *89*, 51–71.