proceedings

MDPI

Proceedings

Bandwidth Selection for Prediction in Regression †

Inés Barbeito [1,*], Ricardo Cao [1] and Stefan Sperlich [2]

[1] Research Group MODES, Department of Mathematics, CITIC, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain

[2] Geneva School of Economics and Management, Université de Genève, Bd du Pont d'Arve 40, CH-1211 Genève, Switzerland

* Correspondence: ines.barbeito@udc.es; Tel.: +34-881011301

† Presented at the 2nd XoveTIC Conference, A Coruña, Spain, 5–6 September 2019.

Published: 5 August 2019

Abstract: There exist many different methods to choose the bandwidth in kernel regression. If, however, the target is regression based prediction for samples or populations with potentially different distributions, then the existing methods can easily be suboptimal. This situation occurs for example in impact evaluation, data matching, or scenario simulations. We propose a bootstrap method to select a global bandwidth for nonparametric out-of-sample prediction. The asymptotic theory is developed, and simulation studies show the successful operation of our method. The method is used to predict nonparametrically the salary of Spanish women if they were paid along the same wage equation as men, given their own characteristics.

Keywords: bandwidth selection; nonparametric prediction; smooth bootstrap

1. Introduction

While there exist a considerable literature on bandwidth selection for kernel based nonparametric density and regression estimation, the problem of nonparametric prediction has largely been ignored. To our knowledge, such selection method does not exist albeit the relevance and frequency of such prediction problems in practise. They include for example any situation for which you want to predict counterfactuals like in impact evaluation (also known as treatment effect estimation). Other examples are statistical matching or data matching (see [1–3], and references therein), the imputation of missings (see e.g., [4–6], and references therein), or the simulation of scenarios. Note that we are not thinking of extrapolation far outside of the support of the observed covariates, a problem that would go beyond the here described ones, see [7]. We do not refer to bandwidth selection in stationary time series. In this context, various bandwidth and other model selection methods have been developed, see e.g. the review of Antoniadis, [8] or [9].

In all these situations have the following three features in common: you can think of a regression model with Y being the left-hand, and X the observed right-hand variables. You have one sample, denoted as 'source', in which both are given such that you can conduct a nonparametric regression. At the same time you have or simulate another sample or population, denoted as 'target', for which the same (as for 'source') potential response Y is not obtained. The basic assumption is that the dependence structure between, or in our case the conditional expectation of Y given X, $m(x) := E[Y|X = x]$ is the same in both populations. In data matching, and similarly when imputing missings, the Y were not sampled for the target sample; in scenarios the X of the target refer to an artificial, maybe future population, for which we just cannot observe any Y; in counterfactual exercises you typically have Y observed for the target sample, but under a different situation, called 'treatment'. Then you use the source sample to impute the potential Y of the target group for the situation 'without treatment'. The difference between the observed Y (under treatment) and the imputed (without treatment) gives the so-called 'treatment effect for the treated'.

Our proposal relies on the so-called smooth bootstrap approach, see [10]. That is, you aim to draw bootstrap samples from a nonparametric pre-estimate of the joint distribution of (X, Y). For the original source sample, and for each bootstrap sample you estimate $m(x)$. These allow us to approximate the mean squared error of $\hat{m}(x)$ for any x inside the support of X. Finally you average these over the x_i observed in the target sample. We said 'you aim' because it can be shown that there exists a closed analytical form for the resulting MASE estimate. This simplifies the procedure drastically making it quite attractive in practise. One may argue that the exactness of this MASE approximation hinges a lot upon the pre-estimate. Yet, for finding the optimal bandwidth (or model) it suffices that our MASE approximations take their minimum at the same bandwidth as the true but unknown MASE. Our simulation studies show that this is actually the case. This work is collected in [11].

2. The Bandwidth Selection Method

Suppose we are provided with a complete sample $\{(x_i^0, y_i^0)\}_{i=1}^{n^0}$ from the source population with $X^0 \sim f^0$ and $m(x) := E_0[Y^0|X^0 = x]$. For the target population we only are provided with observations $\{x_i^1\}_{i=1}^{n_1}$ from density f^1 which is potentially different from f^0. We are interested in predicting the expected $\{y_i^1\}_{i=1}^{n_1}$ assuming that $m(x) = E_1[Y^1|X^1 = x]$, or to estimate $\mathbb{E}\left[Y^1\right] = \mathbb{E}\left[m(X^1)\right]$. Moreover, if some 'outcomes' y_i are observed for the target population, their conditional expectation is supposed to differ from $m(\cdot)$; recall our example of outcome under treatment vs without, or see our application were $m(x)$ is the expected wage given x if you were a man.

For the prediction we have to estimate $m(\cdot)$ by a Nadaraya-Watson estimator $\hat{m}_h^{NW}$ with bandwidth h. Let us suppress for a moment the hyper-indices thinking for now always in the source sample with Y observed. The challenge is to find a bandwidth h which is MASE optimal for our predicting problem. The point-wise MSE, and afterwards the MASE are approximated by their bootstrap versions obtained as follows: Imagine $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ are bootstrap samples drawn from the kernel density $\hat{f}_g(x, y) = n^{-1} \sum_{i=1}^{n} K_g(x - X_i) K_g(y - Y_i)$ with bandwidth g. Then, for $\tilde{m}_h^{NW*}(x)\ f(x) = \hat{f}_h(x)\ \hat{m}_h^{NW}(x)$ we get

$$\tilde{m}_h^{NW*}(x) - \hat{m}_g^{NW}(x) \quad = \quad \frac{1}{n\hat{f}_g(x)} \sum_{i=1}^{n} K_h(x - X_i^*)(Y_i^* - \hat{m}_g^{NW}(x)), \tag{1}$$

where X^* has bootstrap marginal density $\hat{f}_g$, and $\mathbb{E}\left[Y^* \middle| X^* = x\right] = \hat{m}_g^{NW}(x)$. Clearly, this is the bootstrap analogue to

$$\tilde{m}_h^{NW}(x) - m(x) \quad = \quad \frac{1}{nf(x)} \sum_{i=1}^{n} K_h(x - X_i)(Y_i - m(x)). \tag{2}$$

In order to compute the MASE we need to carefully distinguish between source and target sample, and have therefore to use the hyper-indices again. For finding a globally optimal bandwidth h we would like to minimise

$$MASE_{\tilde{m}_h^{NW}, X^1}(h) \quad = \quad \frac{1}{n_1} \sum_{j=1}^{n_1} \left[\mathbb{E}_0 \left[\left(\tilde{m}_h^{NW}(X_j^1) - m(X_j^1) \right)^2 \right] \right], \tag{3}$$

where $\mathbb{E}_0$ refers to the expectation in the source population. We have in the bootstrap world

$$\begin{aligned} MASE^*_{\tilde{m}_h^{NW}, X^1}(h) \quad = \quad & \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{1}{\hat{f}_g^0(X_j^1)^2} \left[\left(1 - \frac{1}{n_0}\right) \cdot \left(\left[K_h * \hat{q}^0_{X_j^1, g} \right] (X_j^1) \right)^2 \right. \\ & \left. + \frac{1}{n_0} \left[(K_h)^2 * \hat{p}^0_{X_j^1, g} \right] (X_j^1) + \frac{g^2 d_K}{n_0^2} \sum_{i=1}^{n_0} \left[(K_h)^2 * K_g \right] (X_j^1 - X_i^0) \right]. \end{aligned} \tag{4}$$

A bootstrap bandwidth selector for prediction is defined as

$$h_{BOOT}^{NW} = h_{MASE^*_{\hat{m}_h^{NW},X^1}} = \arg\min_{h>0} MASE^*_{\hat{m}_h^{NW},X^1}(h).$$

Note that the computation of h_{BOOT}^{NW} does not require the use of Monte Carlo approximation nor the nonparametric estimation of the density f^1 of the target population.

References

1. De Waal, T.; Pannekoek, J.; Scholtus, S. *Handbook of Statistical Data Editing and Imputation*; John Wiley: New York, NY, USA, 2011.
2. Rässler, S. Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Aust. J. Stat.* **2004**, *33*, 153–171.
3. Eurostat. *Statistical Matching: A Model Based Approach for Data Integration*; Methodologies and Working Papers; Eurostat: Luxembourg, 2013.
4. Horton, N.J.; Lipsitz, S.R. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. *Am. Stat.* **2001**, *55*, 244–254.
5. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley: New York, NY, USA, 2004.
6. Su, Y.-S.; Gelman, A.; Hill, J.; Yajima, M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.* **2010**, *45*, 1–31.
7. Li, X.; Heckman, N.E. Local linear extrapolation. *J. Nonparametr. Stat.* **2003**, *15*, 565–578.
8. Antoniadis, A.; Paparoditis, E.; Sapatinas, T. Bandwidth selection for functional time series prediction. *Stat. Probab. Lett.* **2009**, *79*, 733–740.
9. Tschernig, R.; Yang, L. Nonparametric lag selection for time series. *J. Time Ser. Anal.* **2000**, *21*, 457–487.
10. Cao-Abad, R.; González-Manteiga, W. Bootstrap methods in regression smoothing. *J. Nonparametr. Stat.* **1993**, *2*, 379–388.
11. Barbeito, I.; Cao, R.; Sperlich, S. Bandwidth Selection for Nonparametric Kernel Prediction. Unpublished work, 2019.