TEST
https://doi.org/10.1007/s11749-022-00838-7

ORIGINAL PAPER

Bandwidth selection for statistical matching and prediction

Inés Barbeito[1] · Ricardo Cao[1] · Stefan Sperlich[2]

Received: 25 October 2021 / Accepted: 4 November 2022

Abstract

While there exist many bandwidth selectors for estimation, bandwidth selection for statistical matching and prediction has hardly been studied so far. We introduce a computationally attractive selector for nonparametric out-of-sample prediction problems like data matching, impact evaluation, scenario simulations or imputing missings. Even though the method is bootstrap based, we can derive closed expressions for the criterion function which avoids the need of Monte Carlo approximations. We study both, asymptotic and finite sample performance. The derived consistency, convergence rate and extensive simulation studies show the successful operation of the selector. The method is illustrated by applying it to real data for studying the gender wage gap in Spain. Specifically, the salary of Spanish women is predicted nonparametrically by the wage equation estimated for men while conditioned on their own (i.e., women's) characteristics. An important discrepancy between observed and predicted wages is found, exhibiting a serious gender wage gap.

Keywords Bandwidth selection · Statistical matching · Counterfactual analysis · Nonparametric prediction · Smooth bootstrap

Mathematics Subject Classification 62G05 · 62G08 · 62G09

✉ Inés Barbeito
ines.barbeito@udc.es

Ricardo Cao
rcao@udc.es

Stefan Sperlich
stefan.sperlich@unige.ch

[1] Research group MODES, Department of Mathematics, Faculty of Computer Science, 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, 1211 Genève, Switzerland

Published online: 10 December 2022

1 Introduction

While there exists a considerable literature on bandwidth selection for kernel based nonparametric densities and regression, the problem of bandwidth selection for nonparametric prediction has largely been ignored. To our knowledge, such selection methods do not exist despite the relevance and frequency of such prediction problems in practice. Examples are statistical matching or data matching (see Eurostat 2013, and references therein), problems of imputation of missing data (for recent compendia see, Rubin 2004; Su et al. 2010, and van Buuren 2018), and the simulation of scenarios such as those mentioned below. Another prominent example is the prediction of counterfactuals like in treatment effect estimation, see the recent compendium of Frölich and Sperlich (2019). This relates our selection problem also to Vansteelandt et al. (2012) who look more generally on model selection for causal inference. We are not thinking of extrapolation outside of the support of observed covariates (Li and Heckman 2003), a problem that would go beyond the here described one. We neither refer to the bandwidth selection problem for doing forecasting with time series, see the review of Antoniadis et al. (2009) or Tschernig and Yang (2000).

To be more specific, the situations we are thinking of have the following features in common: one has two samples referring to potentially different populations, say the source and the target group. In the source group one has observed the response Y, being the variable of interest, and some auxiliary information X, i.e., some observed covariates. However, for the target group the (potential) response Y is not observed—or the actually observed response is not the wanted counterpart of what we observed in the source group. If one is willing to assume $(Y|X)_{\text{target}} \sim (Y|X)_{\text{source}}$ one can use the source sample to predict Y for the target in order to calculate some parameter of interest like the treatment effect in causal analysis. Consider the following examples: In counterfactual exercises one typically has a response Y also observed for the target sample, but under a different situation, say, under 'treatment,' and needs therefore to predict it under 'no treatment.' The average difference between the observed and imputed Y gives the so-called 'treatment effect for the treated.' In data matching, response Y is not sampled in the target sample but in different sources. Similarly, when facing missing values where one assumes that they are missing at random when conditioning on X, one may consider as target sample the part of your data set that contains Y, and as source sample the rest. Finally, one can imagine scenarios where the explanatory random variable X of the target refers to an artificial, maybe future, population, for which we still cannot observe Y. In most of these examples, and especially for the first and the last one, the distribution of X in the target will be different from the one in the source. For being so-called 'confounders' in causal inference, this is even a requirement. However, one needs either to suppose that $\text{support}(X)_{\text{target}} \subseteq \text{support}(X)_{\text{source}}$, or to use methods made for extrapolating to obtain predictions of Y for the entire $\text{support}(X)_{\text{target}}$. For simplicity, one can simply think of both a common support (for source and target).

In this article, we concentrate on the mean functions, so that the underlying hypothesis is not to have $(Y|X)_{\text{target}} \sim (Y|X)_{\text{source}}$ but the same regression function, say $\mathbb{E}[Y|X = x]$, in both groups. For example, in our application, Y is salary, and we estimate for men (source population) $\mathbb{E}[Y|X]$ containing education, sector and professional experience. This is used then to predict the corresponding expected wages of women (target population) to compare those predictions with the women's realized wages. The difference gives an estimate of the gender wage gap. Evidently, this strategy can be used in many different contexts, also for artificial populations like in scenarios. This describes a quite relevant statistical problem for which we need a regularization parameter like a bandwidth for kernel smoothing. In treatment effect, estimation via matching this type of a bandwidth selection problem has been studied by Frölich (2005), Häggström and Luna (2014) and Galdo et al. (2008). The former two were not interested in the nonparametric estimators but only the total averages. Only the latter considered a problem similar to ours, but proposed weighted cross-validation (CV). Moreover, Frölich (2005) only studied if the use of standard selectors for regression yield acceptable results but did not propose a problem-specific alternative. Also for flexible methods for imputing missings this seems still to be an open problem, see van Buuren (2018). For scenario calculations, we could not even find a profound discussion of this problem in the existing literature.

However, there certainly exists a vast literature on bandwidth selection, see the reviews of Heidenreich et al. (2013) and Köhler et al. (2014). As the prediction problem we discuss is regression based, we should first recall the existing bandwidth selection procedures for regression. As outlined in the above reviews, they can essentially be divided into two groups: cross-validation (CV) and plug-in methods. Likewise one may distinguish between those that try to minimize the integrated or averaged squared error (ISE or ASE) on the one side, and those that try to minimize the expected ISE or ASE, known as MISE and MASE, on the other side. Without pronouncing in favor of one or the other in general, in many regression problems there are good reasons to be more interested in minimizing the ASE (rather than the MASE) as people want to get the best bandwidth for their specific sample fit. Also the implementation seems to be simpler for CV methods; both together explains most of their popularity. In our case, however, the first argument is no longer valid as we may want to get predictions for different samples, and then would prefer the MASE criterion. Also the second argument does no longer hold as adapting CV to our problem is not easy as can be seen from Galdo et al. (2008). Our proposal relies on MASE minimization via smooth bootstrap (Cao and González-Manteiga 1993).

Next, we concentrate on global bandwidths as this is the most popular approach in practice. We call a bandwidth global if it is supposed to be applied over the whole support of the regressor, and is therefore supposed to minimize some global loss function. Local bandwidths are optimal for (just) one point of the support and had therefore to be calculated for all regressor values at which one wants to estimate (or predict). It is therefore hardly ever used. We further limit our presentation to so-called local constant prediction, i.e., based on the classical Nadaraya–Watson estimator. For an extension to local linear methods, see the thesis of Barbeito (2020). While local linear estimators have doubtless their advantages, she showed that for bandwidth selection in our context they render the method much more cumbersome and computationally quite unattractive. Note finally that, although the theory for our method is developed along the problem of a particular bandwidth choice, it also applies more generally to model selection for prediction.

Smooth bootstrap aims to draw bootstrap samples from a nonparametric pilot estimate of the joint distribution of (X, Y). For the source and each bootstrap sample $m(x) := \mathbb{E}[Y|X]$ is estimated. The estimates allow us to approximate the mean squared error of $\hat{m}(x)$. These errors are averaged over the x_i observed in the target sample. It can be shown that there exists a closed analytical form for the resulting MASE bootstrap estimate. This simplifies the procedure importantly making it quite attractive. One may argue that the exactness of this MASE approximation hinges on the pilot estimate. Yet, in order to find the optimal bandwidth or model, it suffices that the MASE approximations take their minimum at the same point as the true, unknown MASE. Our simulations show that this is indeed the case.

In Sect. 2, we introduce the prediction problem and bandwidth selector considered, introduced for the case with one explanatory variable. Assumptions and asymptotic properties are also provided in that section. The simulations in Sect. 3 demonstrate an excellent performance of this method. Section 4 extends our approach to situations with one continuous explanatory variable and several categorical ones and illustrates the method along a study of the gender wage gap in Spain. Section 5 concludes and discusses further extensions like the one to boundary kernels, multivariate continuous covariates and local linear estimation. Asymptotic theory on the pilot choice is deferred to the Appendix, and technical proofs to the Supplementary Material.

2 The bandwidth selection method

2.1 Closed-expression for the criterion functions

Suppose a complete sample $\{(X_i^0, Y_i^0)\}_{i=1}^{n_0}$ from the source population is provided, with $X^0 \sim f^0$, and (re-)define $m(x) := \mathbb{E}\left(Y^0|X^0 = x\right)$. For the target population, we are provided with observations $\{X_i^1\}_{i=1}^{n_1}$ from density f^1 which is potentially different from f^0, and we assume $\mathbb{E}\left[Y^1|X^1 = x\right] = m(x)$. Then, we may consider two different, though related problems: (a) predicting the unobserved $\left\{Y_i^1\right\}_{i=1}^{n_1}$, or (b) estimating $\mathbb{E}\left[Y^1\right] = \mathbb{E}\left[\mathbb{E}\left[Y^1 \middle| X^1\right]\right] = \mathbb{E}\left[m(X^1)\right]$.[1] Recall, if some outcome is observed for the target group, its conditional expectation might nonetheless seriously differ from $m(\cdot)$, like in our study in which we observe women's wages but want to predict their wages as if they were paid like men.

We estimate $m(\cdot)$ by a Nadaraya–Watson estimator $\hat{m}_h$ with bandwidth, h. Let us suppress for a moment the upper index and concentrate on the source sample. The challenge is to find a bandwidth, h, which is optimal for problems (a) and (b). The pointwise MSE, and afterward the MASE are approximated by their bootstrap versions. Similarly as Cao and González-Manteiga (1993), consider two pilot bandwidths g_X, g_Y. They propose to obtain the bootstrap resamples either sampling from:

SB1 $\hat{F}_g(x, y) = n^{-1} \sum\limits_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}} \int\limits_{-\infty}^{x} K_g(t - X_i)\, dt$, or

[1] Our notation differs from the one used in the treatment literature: There, Y^1 refers to outcome under treatment, Y^0 to outcome under control. In contrast, we use the upper index to indicate the group.

SB2 $\tilde{F}_{g_X,g_Y}(x,y) = n^{-1}\sum_{=1}^{n} \mathbb{K}\left(\frac{x-X_i}{g_X}\right)\mathbb{K}\left(\frac{y-Y_i}{g_Y}\right)$.

The latter is equivalent to resampling from the bivariate density $\hat{f}_{g_X,g_Y}(x,y) = n^{-1}\sum_{i=1}^{n} K_{g_X}(x-X_i)K_{g_Y}(y-Y_i)$. In the following, only SB2 will be considered because SB1 is just the limit case of SB2 when $g_Y \to 0^+$, for fixed n. Denote $\hat{m}_h(x) = \hat{\Psi}_h(x)/\hat{f}_h(x)$, where $\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)$ and $\hat{\Psi}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x-X_i)Y_i$. For $\Psi(x) := f(x)m(x)$,

$$\hat{m}_h(x) - m(x) = \frac{\hat{\Psi}_h(x)}{\hat{f}_h(x)} - \frac{\Psi(x)}{f(x)} = \left(\frac{\hat{\Psi}_h(x)}{\hat{f}_h(x)} - \frac{\Psi(x)}{f(x)}\right)\left[\frac{\hat{f}_h(x)}{f(x)} + \left(1 - \frac{\hat{f}_h(x)}{f(x)}\right)\right]$$
$$= \frac{\hat{\Psi}_h(x) - m(x)\hat{f}_h(x)}{f(x)} + \frac{\left(\hat{m}_h(x) - m(x)\right)\left(f(x) - \hat{f}_h(x)\right)}{f(x)}, \tag{1}$$

since $\hat{\Psi}_h(x) - m(x)\hat{f}_h(x) = \hat{\Psi}_h(x) - \Psi(x) + m(x)\left(f(x) - \hat{f}_h(x)\right)$. The second term on the right hand side of (1) is negligible. So it suffices to consider a proxy estimator of $m(x)$ that corresponds to considering only the first term, and we get

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

Given that $\mathbb{E}^*\left(Y^* \vert\, X^* = x\right) = \hat{m}_{g_X,g_Y}(x)$, the bootstrap analog of the proxy estimator is

$$\tilde{m}_h^*(x) - \hat{m}_{g_X}(x) = \frac{1}{n\hat{f}_{g_X}(x)}\sum_{i=1}^{n} K_h(x-X_i^*)(Y_i^* - \hat{m}_{g_X}(x)), \tag{3}$$

where X^* has bootstrap marginal density $\hat{f}_{g_X}$. Using convolution $(K_h * q_x)(x) := \int K_h(x-y)q_x(y)\,dy$, the point-wise MSE and its bootstrap analog are in Theorem 1, assuming

(A1) K is a bounded symmetric density function

Theorem 1 *If (A1) holds, x is an interior point of the support of X, $\{X_i\}_{i=1}^{n}$ an i.i.d. sample, and $f(x) \neq 0$, then the point-wise mean squared error of $\tilde{m}_h$ can be expressed as*

$$\mathrm{MSE}_x(h) := \mathbb{E}\left[(\tilde{m}_h(x) - m(x))^2\right] = \frac{n-1}{nf(x)^2}(K_h * q_x)^2(x) + \frac{1}{nf(x)^2}\left[(K_h)^2 * p_x\right](x), \tag{4}$$

where $p_x(z) := \left(\sigma^2(z) + (m(z)-m(x))^2\right)f(z)$, $q_x(z) := (m(z)-m(x))f(z)$ and $\sigma^2(x) := Var\left(Y \vert\, X = x\right)$ stands for the volatility function.

If further $\hat{f}_{g_X}(x) \neq 0$, then the smoothed bootstrap version of $\mathrm{MSE}_x(h)$ *is*

$$\mathrm{MSE}_x^*(h) = \frac{1}{n\hat{f}_{g_X}^2(x)}\left[\frac{g_Y^2\mu_2(K)}{n}\sum_{i=1}^{n}\left[(K_h)^2 * K_{g_X}\right](x-X_i) + \left[(K_h)^2 * \hat{p}_{x,g_X}\right](x)\right]$$

$$+\frac{n-1}{n\hat{f}_{g_X}^2(x)}\left(K_h * \hat{q}_{x,g_X}\right)^2(x), \text{ with} \tag{5}$$

$\hat{p}_{x,g_X}(z) = \left(\hat{\sigma}_{g_X}^2(z) + (\tilde{m}_{g_X}(z) - \tilde{m}_{g_X}(x))^2\right)\hat{f}_{g_X}(z)$, $\hat{q}_{x,g_X}(z) = (\tilde{m}_{g_X}(z) - \tilde{m}_{g_X}(x))\hat{f}_{g_X}(z)$, $\mu_r(K) = \int t^r K(t)\,dt$, $\hat{\sigma}_{g_X}^2(z) = \tilde{m}_{2,g_X}(z) - \tilde{m}_{g_X}^2(z)$, $\tilde{m}_{2,g_X}(z) = \frac{\sum_{i=1}^n K_{g_X}(z - X_i)\,Y_i^2}{\sum_{i=1}^n K_{g_X}(z - X_i)}$.

The proof of Theorem 1 is provided in the Supplementary Material. It is to be emphasized that (5) is exact and not just an approximation. In other words, there is no need to draw bootstrap resamples as (5) can be calculated directly.

The next step is notationally cumbersome because for the MASE, we need to carefully distinguish between source and target sample. Therefore, we use the upper indices again. As said in the introduction, for the sake of presentation suppose f^0 and f^1 have common support. In order to select an optimal bandwidth, our aim is to minimize

$$\mathbb{E}\left[\left(\int \left(\hat{m}_h(x) - m(x)\right)\, dF_1(x)\right)^2\right]. \tag{6}$$

However, thanks to Proposition 1, where an upper bound is obtained, we minimize expression $\mathbb{E}\left[\int \left(\hat{m}_h(x) - m(x)\right)^2\, dF_1(x)\right]$ instead. Let us define now our objective function and then give its upper bound:

Definition 1 Denote $Y_1^1, \ldots, Y_{n_1}^1$ the unobserved values of Y for the target population. Taking $\hat{m}\left(X_i^1\right)$ as a prediction of Y_i^1, $i = 1, \ldots, n_1$, the average prediction error is

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(Y_i^1 - \hat{m}_h\left(X_i^1\right)\right)^2\right].$$

Proposition 1 *If F_1 is the distribution function of the target population, and $\hat{m}_h$ the estimated regression function. Then, an upper bound for expression (6) is given by*

$$\mathbb{E}\left[\left(\int \left(\hat{m}_h(x) - m(x)\right)\, dF_1(x)\right)^2\right] \leq \mathbb{E}\left[\int \left(\hat{m}_h(x) - m(x)\right)^2\, dF_1(x)\right].$$

On the other hand, the average prediction error is given by

$$\mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left(Y_i^1 - \hat{m}_h\left(X_i^1\right)\right)^2\right] = \int \sigma^2(x)\, dF_1(x) + \mathbb{E}\left[\int \left(\hat{m}_h(x) - m(x)\right)^2\, dF_1(x)\right].$$

Then, for finding a globally optimal bandwidth h we minimize

$$\text{MASE}_{\tilde{m}_h, X^1}(h) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left[\mathbb{E}_0\left[\left(\tilde{m}_h(X_j^1) - m(X_j^1)\right)^2\right]\right], \tag{7}$$

where, for any random variable Z, $\mathbb{E}_0\left[Z\right] = \mathbb{E}\left[Z \mid X_j^1, \forall j \in \{1, \ldots, n_1\}\right]$ refers to the expectation in the source population conditioned on the target sample. Similarly to Theorem 1, we can state for the MASE and its bootstrap analog:

Theorem 2 *Assume (A1) and let $\{(X_i^0, Y_i^0)\}_{i=1}^{n_0}$ be a simple random sample coming from the source population, and $\{X_i^1\}_{i=1}^{n_1}$ a simple random sample coming from the target population. Then, the MASE prediction error is*

$$\begin{aligned}\mathrm{MASE}_{\tilde{m}_h, X^1}(h) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{f^0(X_j^1)^2}\Bigg[&\left(1-\frac{1}{n_0}\right)\cdot\left[K_h * q^0_{X_j^1}\right]^2(X_j^1)\\ &+\frac{1}{n_0}\left[(K_h)^2 * p^0_{X_j^1}\right](X_j^1)\Bigg],\end{aligned} \tag{8}$$

where $q_x^0(z) := (m(z) - m(x)) f^0(z)$ and $p_x^0(z) := \left(\sigma_0^2(z) + (m(z) - m(x))^2\right) f^0(z)$. Similarly,

$$\begin{aligned}\mathrm{MASE}^*_{\tilde{m}_h, X^1}(h) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{\hat{f}^0_{g_X}(X_j^1)^2}\Bigg[&\left(1-\frac{1}{n_0}\right)\cdot\left[K_h * \hat{q}^0_{X_j^1, g_X}\right]^2(X_j^1)\\ &+\frac{1}{n_0}\left[(K_h)^2 * \hat{p}^0_{X_j^1, g_X}\right](X_j^1) + \frac{g_Y^2\mu_2(K)}{n_0}\left[(K_h)^2 * \hat{f}^0_{g_X}\right](X_j^1)\Bigg],\end{aligned} \tag{9}$$

with $\hat{p}^0_{x,g_X}(z) = \left(\hat{\sigma}^2_{0,g_X}(z) + (\hat{m}_{g_X}(z) - \hat{m}_{g_X}(x))^2\right)\hat{f}^0_{g_X}(z)$, $\hat{q}^0_{x,g_X}(z) = (\hat{m}_{g_X}(z) - \hat{m}_{g_X}(x))\hat{f}^0_{g_X}(z)$ and $\hat{\sigma}^2_{0,g_X}(z) = \tilde{m}_{2,g_X}(z) - \tilde{m}^2_{g_X}(z)$ with $\tilde{m}_{2,g_X}(z) = \dfrac{\sum_{i=1}^n K_{g_X}\left(z - X_i^0\right)\left(Y_i^0\right)^2}{\sum_{i=1}^n K_{g_X}\left(z - X_i^0\right)}$.

Furthermore, working out the convolutions in expression (9) we get

Corollary 1 *If K is a Gaussian kernel, then expression (9) can be rewritten as follows:*

$$\begin{aligned}\mathrm{MASE}^*_{\tilde{m}_h, X^1}(h) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{\hat{f}^0_g\left(X_j^1\right)^2}\Bigg[&\frac{n_0-1}{n_0^3}\cdot\left[\sum_{i=1}^{n_0} K_h * K_{g_X}\left(X_j^1 - X_i^0\right)\cdot\left(Y_i^0 - \hat{m}_{g_X}\left(X_j^1\right)\right)\right]^2\\ &+\frac{1}{n_0^2}\sum_{i=1}^{n_0}\left[(K_h)^2 * K_{g_X}\right](X_j^1 - X_i^0)\cdot\left[Y_i^0 - \hat{m}_{g_X}\left(X_j^1\right)\right]^2\\ &+\frac{g^2\mu_2(K)}{n_0^2}\sum_{i=1}^{n_0}\left[(K_h)^2 * K_g\right]\left(X_j^1 - X_i^0\right)\Bigg].\end{aligned}$$

In the special case of considering SB1, that is, if $g_Y = 0$, one gets

Corollary 2 *If* $g_Y = 0$, *then expression* (9) *for* $\mathrm{MASE}^*_{\tilde m_h, X^1}(h)$ *becomes*

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{\hat f^0_{g_X}\left(X^1_j\right)^2}\left[\left(1-\frac{1}{n_0}\right)\left[K_h*\hat q^0_{X^1_j,g_X}\right]^2\left(X^1_j\right)+\frac{1}{n_0}\left[(K_h)^2*\hat p^0_{X^1_j,g_X}\right]\left(X^1_j\right)\right]. \tag{10}$$

Proofs of Theorem 2 and Corollary 1 are in the Supplementary Material, whereas the proof of Corollary 2 is immediate. For the sake of simplicity, we consider $g = g_X = g_Y$. A bootstrap bandwidth selector for prediction can now be defined as

$$h_{\mathrm{BOOT}} = h^*_{\mathrm{MASE}_{\tilde m_h, X^1}} = \arg\min_{h>0} \mathrm{MASE}^*_{\tilde m_h, X^1}(h).$$

As we could see, computation of h_{BOOT} does not require the use of Monte Carlo approximation nor the nonparametric estimation of the density f^1 of the target population.

In order to verify whether this procedure works, we have first to answer two questions: Is the MASE of approximation (2), i.e., when $\tilde m_h$ substitutes $\hat m_h$, a useful approximation for the MASE of $\hat m_h$? And if so, is the MASE bootstrap analog a useful approximation for the former? Figure 1 reveals that the optimal bandwidths for both estimators are very close. It also shows that the first approximation is much more sensitive to the bandwidth than the original one which makes it numerically even easier to find the minimum. Furthermore, for almost all simulations the minimizer in the bootstrap world is very close to the true one. This answers both questions with a clear 'yes.'

2.2 Asymptotic theory

We now derive the asymptotic rate of convergence for the bandwidth selector. To do this we have to look at the asymptotic analogs of $\mathrm{MASE}_{\tilde m_h, X^1}$ and its bootstrap version, say

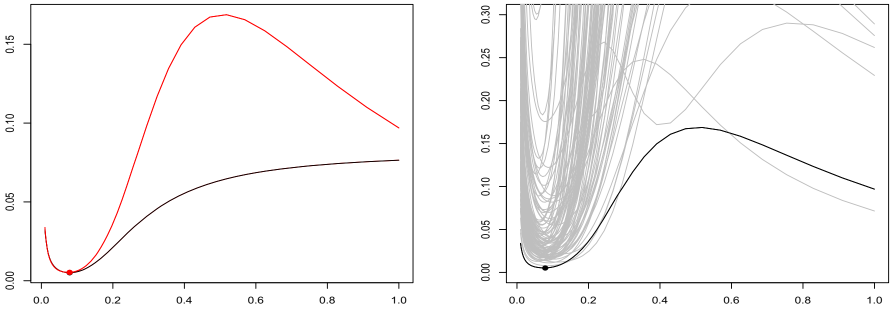

Fig. 1 Left: MASE of the Nadaraya–Watson estimator using Monte Carlo (black line) and its approximated version (9) (red line). Right: MASE of the approximation (7) (black line) and its bootstrap version (9) for 100 different samples (gray lines) $\{X^0, Y^0\}$ of size $n_0 = n_1 = 100$. Specifically, X^0 is $\beta(2, 4)$ distributed, $Y^0 = m(X^0) + 0.4\epsilon$, where $m(x) = 2x^{1/2}$, and ϵ standard normal. X^1 is $\beta(4, 2)$ distributed (color figure online)

at $\mathrm{MISE}^a(h) := \mathbb{E}\left[\int (\tilde{m}_h(x) - m(x))^2 \, \mathrm{d}F_1(x)\right]$ and its bootstrap version. Note that MISE^a is the MISE but with the proxy estimator (2). The proof of theorems, lemmas, propositions and corollaries of this section are in the Supplementary Material.

2.2.1 Asymptotic expression for the criterion function

We need some regularity conditions for the kernel, source density and regression function:

(B1) K is a positive second order kernel.
(B2) f^0 is four times differentiable and $f^0(x) \neq 0, \forall x \in \text{support}\left(X^1\right)$.
(B3) m is four times differentiable.
(B4) σ^2 is two times differentiable.

Using a Taylor expansion and a change of variable, we obtain for the $\mathrm{MISE}^a(h)$:

Lemma 1 *Under the regularity conditions (B1)-(B4), the* MISE^a *can be expressed as*

$$\mathrm{MISE}^a(h) = \frac{R(K)}{n_0\, h} \int \gamma(x) \mathrm{d}x + \frac{h^4 \mu_2(K)^2}{4} \int \beta(x) \, \mathrm{d}x + \mathcal{O}(h^6) + \mathcal{O}\left(\frac{h}{n_0}\right), \quad (11)$$

$$\beta(x) \quad = \quad \left[m''(x)^2 + \frac{4m'(x)m''(x) \left(f^0\right)'(x)}{f^0(x)} + \frac{4m'(x)^2 \left(f^0\right)'(x)^2}{f^0(x)^2} \right] f^1(x) \quad \textit{and}$$

$\gamma(x) = \dfrac{\sigma^2(x) f^1(x)}{f^0(x)}$. *The asymptotic version of expression (11) is given by:*

$$\mathrm{AMISE}^a(h) = \frac{R(K)}{n_0\, h} \int \gamma(x) \mathrm{d}x + \frac{h^4}{4} \mu_2(K)^2 \int \beta(x) \, \mathrm{d}x. \quad (12)$$

Minimizing expression (12), we obtain the AMISE^a bandwidth, namely

$$h_{\mathrm{AMISE}^a} = \left(\frac{R(K) \int \sigma^2(x) f^1(x) \left(f^0(x)\right)^{-1} \mathrm{d}x}{\mu_2(K)^2 \int \beta(x) \, f^1(x) \, \mathrm{d}x} \right)^{1/5} \cdot n_0^{-1/5}. \quad (13)$$

With (11) and (13), the MISE^a of h_{AMISE^a} becomes

$$\mathrm{MISE}^a \left(h_{\mathrm{AMISE}^a}\right) = \frac{R(K)}{n_0^{4/5} c_0} \int \gamma(x) \mathrm{d}x + \frac{c_0^4 \, \mu_2(K)^2}{4\, n_0^{4/5}} \int \beta(x) f^1(x) \, \mathrm{d}x + \mathcal{O}\left(n_0^{-6/5}\right),$$

which tends to zero at the rate $n_0^{-4/5}$ as $n_0 \longrightarrow \infty$. From this we can conclude:

Theorem 3 *Under the regularity conditions (B1)–(B3), the bandwidth which minimizes* MISE^a *has the asymptotic expression:*

$$h_{\mathrm{MISE}^a} = \left(\frac{R(K) \int \sigma^2(x) f^1(x) \left(f^0(x)\right)^{-1} \mathrm{d}x}{\mu_2(K)^2 \int \beta(x) \, f^1(x) \, \mathrm{d}x} \right)^{1/5} n_0^{-1/5} + \mathcal{O}\left(n_0^{-2/5}\right). \quad (14)$$

It remains to study how accurate the approximation MISE^a for the MISE is. As M refers to the 'mean,' we have to study the difference between ISE^a and ISE. For this we need some conditions on the kernel (C1) and density function (C2), cf. Silverman (1978); as well as conditions for the regression function (C3), cf. Mack and Silverman (1982):

(C1) (a) K is uniformly continuous with modulus of continuity w_K and bounded variation $V(K)$. K is absolutely integrable with respect to the Lebesgue measure.
(b) $K(x) \longrightarrow 0$ as $|x| \longrightarrow \infty$.
(c) $\int |x \log |x||^{1/2} \, |\mathrm{d}K(x)| < \infty$.

(C2) (a) f^0 is uniformly continuous.

(C3) (a) $\mathbb{E}(Y)^s < \infty$ and $\sup\limits_x \int |y|^s f_{XY}(x, y) \, \mathrm{d}y < \infty, s \geq 2$, where f_{XY} is the joint density.
(b) The marginal density of X^0, f^0, the joint density function, f_{XY}, and $m(x) f^0(x)$, the theoretical analog of $\hat{\Psi}_h$, are continuous in an open interval containing J, where J is a bounded interval in which f^0 is bounded away from zero.

Proposition 2 *Suppose (C1) to (C3) are fulfilled. Consider a sequence of bandwidths* h_{n_0} *such that* $\sum\limits_{n_0} h_{n_0}^{\lambda} < \infty$ *for some* $\lambda > 0$ *and that* $n_0^{\eta} h_{n_0} \longrightarrow \infty$ *for some* $\eta < 1-s^{-1}$. *Assume* $(n_0 \, h)^{-1/2} \log\left(h^{-1}\right) \longrightarrow 0$, $h \longrightarrow 0$ *and* $n_0 \, h \longrightarrow \infty$ *as* $n_0 \longrightarrow \infty$. *Then,*

$$\text{ISE}(h) = \text{ISE}^a(h) + \mathcal{O}_P\left(h^6\right) + \mathcal{O}_P\left(\frac{h}{n_0} \log \frac{1}{h}\right) + \mathcal{O}_P\left(\frac{h^{7/2}}{n_0^{1/2}}\right) + \mathcal{O}_P\left(\frac{\log \frac{1}{h}}{n_0^{3/2} h^{3/2}}\right), \tag{15}$$

where $\text{ISE}(h) = \int \left(\hat{m}_h(x) - m(x)\right)^2 \mathrm{d}F^1(x)$ *and* $\text{ISE}^a(h) = \int \left(\tilde{m}_h(x) - m(x)\right)^2 \mathrm{d}F^1(x)$.

This proposition tells us that the difference rapidly vanishes.

2.2.2 Asymptotic expressions for the bootstrap criterion functions

Next, the smoothed bootstrap version of MISE^a is to be studied. We start by computing $\text{MISE}^{a*}(h) := \mathbb{E}^*\left[\int \left(\tilde{m}_h^*(x) - \hat{m}_g(x)\right) \mathrm{d}F_1\right]$, i.e., the theoretical analog of $\text{MASE}^*_{\tilde{m}_h, X^1}$. For studying MISE^{a*}, i.e., the bootstrap MISE of $\tilde{m}_h$, given in (2), define $\hat{A}_g = \int \hat{\gamma}_g(x) \, \mathrm{d}x$, $\hat{B}_g = \int \hat{\beta}_g(x) \, \mathrm{d}x$, $A = \int \gamma(x) \, \mathrm{d}x$ and $B = \int \beta(x) \, \mathrm{d}x$, with

$\hat{\gamma}_g(x) = \hat{\sigma}_g^2(x)\hat{f}_g^1(x)/\hat{f}_g^0(x)$ and

$$\hat{\beta}_g(x) = \left[\hat{m}_g''(x)^2 + \frac{4\hat{m}_g'(x)\hat{m}_g''(x)\left(\hat{f}_g^0\right)'(x)}{\hat{f}_g^0(x)} + \frac{4\hat{m}_g'(x)^2\left(\hat{f}_g^0\right)'(x)^2}{\hat{f}_g^0(x)^2}\right]\hat{f}_g^1(x)$$

Lemma 2 *Under the regularity conditions (B1)–(B4), the function* MISE^{a*} *is*

$$\begin{aligned}\text{MISE}^{a*}(h) = {}& \frac{R(K)}{n_0\,h}\hat{A}_g + \frac{h^4}{4}\mu_2(K)^2\hat{B}_g + \mathcal{O}_P\left(h^6\,n_1^{-1}\,g^{-7}\left(g^{-2}+g^{-1}+1\right)\right)\\ &+\mathcal{O}_P(h^8\,n_1^{-1}g^{-9}) + \mathcal{O}_P\left(h^{-1}\,g^2\,n_1^{-1}\right)\\ &+\mathcal{O}_P\left(h\,n_1^{-1}\left(1+g^{-1}+g^{-2}\right)\right).\end{aligned} \tag{16}$$

Thus, the dominant part of expression (16), namely $\text{AMISE}^{a*}(h)$*, is given by:*

$$\text{AMISE}^{a*}(h) = \frac{R(K)}{n_0\,h}\hat{A}_g + \frac{h^4}{4}\mu_2(K)^2\hat{B}_g. \tag{17}$$

By minimizing expression (17), we can state an AMISE^{a*} bandwidth, namely

$$h^*_{\text{AMISE}^a} = \left(\frac{R(K)\hat{A}_g}{\mu_2(K)^2\hat{B}_g}\right)^{1/5} n_0^{-1/5}. \tag{18}$$

Next, similarly as in the non-bootstrap context, it remains to check the accuracy of the theoretical approximation we have considered for MISE^*, i.e., our MISE^{a*}. Again we can show that this approximation is appropriate in terms of ISE^*.

Proposition 3 *Assume conditions (C1) to (C3), and suppose further* $n_0 \longrightarrow \infty$, $h \longrightarrow 0$ *and* $n_0\,h \longrightarrow \infty$. *Consider a sequence of bandwidths* h_{n_0} *such that* $\sum_{n_0} h_{n_0}^{\lambda} < \infty$ *for some* $\lambda > 0$, *and* $n_0^{\eta}h_{n_0} \longrightarrow \infty$ *for some* $\eta < 1 - s^{-1}$. *Assume also* $(n_0\,h)^{-1/2}\log\left(h^{-1}\right) \longrightarrow 0$. *Then,*

$$\begin{aligned}\text{ISE}^*(h) = {}& \text{ISE}^{a*}(h) + \mathcal{O}_{P^*}\left(h^6\right) + \mathcal{O}_{P^*}\left(\frac{h}{n_0}\log\frac{1}{h}\right) + \mathcal{O}_{P^*}\left(\frac{h^{7/2}}{n_0^{1/2}}\right)\\ &+\mathcal{O}_{P^*}\left(\frac{\log\frac{1}{h}}{n_0^{3/2}h^{3/2}}\right),\end{aligned} \tag{19}$$

almost sure with respect to P, *where*

$$\text{ISE}^*(h) = \int\left(\hat{m}_h^{NW*}(x) - \hat{m}_g(x)\right)^2\,\mathrm{d}\hat{F}_g^1(x), \text{ and } \text{ISE}^{a*}(h)$$

$$= \int \left(\tilde{m}_h^{NW*}(x) - \hat{m}_g(x)\right)^2 \mathrm{d}\hat{F}_g^1(x).$$

We can also establish the convergence rate for $\left(h_{\mathrm{MISE}^{a*}} - h_{\mathrm{MISE}^a}\right)/h_{\mathrm{MISE}^a}$, using expressions (16), (18) and (36) in the Appendix.

Theorem 4 *Consider h_{MISE^a}, its bootstrap version $h^*_{\mathrm{MISE}^a}$, and h_{AMISE^a}, which are the minimizers of expressions (11), (16) and (17), respectively. Under the regularity conditions (B1)–(B4) and assuming that g is of order $n_0^{-1/2}$, it holds that*

$$h^*_{\mathrm{MISE}^a} - h_{\mathrm{MISE}^a} = \mathcal{O}_P\left(n_0^{-7/10}\right), \text{and } \frac{h^*_{\mathrm{MISE}^a} - h_{\mathrm{MISE}^a}}{h_{\mathrm{MISE}^a}} = \mathcal{O}_P\left(n_0^{-1/2}\right). \quad (20)$$

It remains to discuss the selection of g. Expression (16) reveals that this pilot bandwidth should be selected to minimize the error produced by estimating expression (12) via (17). So one should minimize the expected squared $\hat{\delta}_g(h) := \mathrm{AMISE}^{a*}(h) - \mathrm{AMISE}^a(h)$, i.e.,

$$g_{\mathrm{OPT}} := \arg\min_{g>0} \mathbb{E}\left[\hat{\delta}_g^2(h)\right]. \quad (21)$$

It is easy to see that

$$\hat{\delta}_g(h) = \frac{R(K)}{n_0 h}\left(\hat{A}_g - A\right) + \frac{h^4}{4}\mu_2(K)^2\left(\hat{B}_g - B\right), \quad (22)$$

implying that g_{OPT} depends in turn on bandwidth h. This suggests to consider

$$g_{\mathrm{OPT}} := \arg\min_{g>0} \mathbb{E}\left[\hat{\delta}_g^2(h_{\mathrm{AMISE}^a})\right], \quad (23)$$

where h_{AMISE^a} was defined in (13). It can be shown that

$$\mathbb{E}\left[\hat{\delta}_g^2(h_{\mathrm{AMISE}^a})\right] = \frac{\mu_2(K)^{4/5}R(K)^{8/5}B^{2/5}}{n_0^{8/5}A^{2/5}}\mathbb{E}\left\{\left[\left(\hat{A}_g - A\right) + \frac{A}{4B}\cdot\left(\hat{B}_g - B\right)\right]^2\right\}, \quad (24)$$

which leads us to

$$g_{\mathrm{OPT}} = \arg\min_{g>0}\left\{\mathbb{E}\left[\alpha_g^2\right] + \frac{A^2}{16B}\mathbb{E}\left[\xi_g^2\right] + \frac{A}{2B}\mathbb{E}\left[\alpha_g \cdot \xi_g\right]\right\}, \quad (25)$$

where $\xi_g = \hat{B}_g - B$ and $\alpha_g = \hat{A}_g - A$. Denote $\hat{\Psi}_{\ell,g}(x) = n_0^{-1}\sum_{i=1}^{n_0} K_g\left(x - X_i^0\right)\left(Y_i^0\right)^\ell$ and $\Psi_\ell(x) = m_\ell(x) f^0(x), \forall \ell \in \{0,1,2\}$, where $m_\ell(x) = \mathbb{E}\left(Y^{0\ell}\big|_{X^0=x}\right)$ if $\ell \in \{1,2\}$, and $m_\ell = 1$ if $\ell = 0$. Unfortunately, the computation of the expectations in (25) is extremely tedious, implying the calculation of more than three hundred U-statistics of order $n_0 \cdot n_1$. An alternative is to find an upper bound for expression (24), say

$$\mathbb{E}\left[\hat{\delta}_g^2(h_{\mathrm{AMISE}^a})\right] \leq 2\mathbb{E}\left[\left(\hat{A}_g - A\right)^2\right] + \frac{A^2}{8\,B^2}\mathbb{E}\left[\left(\hat{B}_g - B\right)^2\right].$$

This requires to work with $\hat{A}_g - A$ and $\hat{B}_g - B$. Some technical results concerning the quantification of the error of both approximations are collected in Lemmas 3, 4 and Corollary 3 in the Appendix. As can be seen in the Supplementary Material, the optimal g for the upper bounds obtained for the terms $\hat{A}_g - A$ and $\hat{B}_g - B$ happens to be $n_0^{-1/2}$.

3 Simulation study of performance

Our method is not only, but particularly interesting for the so-called matching methods and scenarios. Both are applied in many different situations. Although we certainly studied many more simulation designs, we found out that they could all be decomposed and/or classified into the following three different situations.

Scenario 1 In the source population, the distribution of X^0 is a $\beta(2,4)$ and $Y^0 = m(X^0) + 0.3\epsilon$, where $m(x) = \sin(\pi x)$ and ϵ is drawn from a standard normal distribution. The target population, X^1, has distribution $\beta(4,2)$.

Scenario 2 In the source population, X^0 has distribution $\beta(2,5)$ and $Y^0 = m(X^0) + 0.3\epsilon$, where $m(x) = \sin\left(1 + \left(\frac{\pi x}{2}\right)^4\right)$ and ϵ is drawn from a standard normal distribution. The target population, X^1, has distribution $\beta(5,2)$.

Scenario 3 In the source population, X^0 has distribution $\beta(5,2)$ and $Y^0 = m(X^0) + 0.3\epsilon$, where $m(x) = \sin\left(1 + \left(\frac{\pi x}{2}\right)^4\right)$ and ϵ is drawn from a standard normal distribution. The target population X^1 has distribution $\beta(2,5)$.

We therefore decided to limit our presentation to these exemplifying situations which can furthermore be motivated by the following examples.

Imagine an impact evaluation problem with unbalanced treatment vs control group in which one has to estimate the effect of an IT training course for unemployed on their chance to find a job within the next 4 months (after the training) or on their next salary. As these courses are not compulsory, the age of participants may have a right skewed distribution, whereas the one of the nonparticipants may have a left-skewed one. It is well known that the likelihood to find a job and the next salary plotted on age are inverted U-shaped curves. This situation is well reflected in Scenario 1.

For many poverty or inequality as well as for health studies, one needs to perform data matching or imputing missing values as in the main data set the variable of interest is simply not observed or exhibits many missing values but one believes these can be 'explained' (not necessarily in the causal but a stochastic sense) by observed X (see, e.g., Dai et al. (2016) for related problems we consider). For certain tax interventions on luxury goods, one may expect most variation of the regression curve in the areas of higher expenditures. However, in surveys regarding income (X) and expenditures (say for luxury goods, Y), lower- and middle income groups have much less missing values. Here, the source is the subsample with, the target the subsample without Y. This gives scenario 2.

Similarly, for scenarios and ex-ante evaluation one considers the present society versus a potential future one. Imagine the regression function in Scenarios 2 and 3 had an upward bump (what for the simulation outcome does not make any difference). We are not only concerned about the pension system but also look at other aspects like the health system. Most of the variation in doctor visits and health expenditures happens (a) for people above 65, mainly men, and (b) young women. Thinking only of men, for a strongly aging society one has a similar scenario as in 2 but with an upward bump instead of a downward one in the regression function. In contrast, when looking only on men and heavy road accidents, in many countries one has an important upward bump between the age of 18 and 28, followed by a relatively flat curve. Then, one has to flip the source and target group and end up in a (numerically) equivalent situation as in Scenario 3.

Next, think of a counterfactual exercise to study discrimination in wages by gender or race. To do so one may want to estimate the wage gap controlling for (by conditioning on) certain factors like sector, studies and years of experience or age. Such conditioning is not only important for a fair comparison, but also to understand better channels of wage differences and discrimination. Notice that again, the distributions of the three mentioned covariates differ a lot between gender. We do, however, not know if the regression is equally smooth on the main support of both distributions (Scenario 1), or if we are in Scenarios 2 or 3. So it is clear that we should apply therefore our bandwidth selection method. This is the situation we face in the next section, but is covered by these three simulation scenarios.

It is important to notice that Scenario 1 is equivalent to a situation where the distributions f^0, f^1 are quite similar, independently of the functional form of the regression function. In such situation, the density-weighted smoothness of the regression function is the same in both groups as it is in Scenario 1. It is clear that in such situation our bandwidth selector cannot do better than the corresponding counterpart for regression in the source sample. Unless one has done some prior studies about equality of distributions or smoothness of the regression, in practice one does not know if one is in such a special situation. However, our Scenario 1 will show that even then our method does not much worse, and Scenarios 2 and 3 will show that else one does much better with our method.

3.1 Description of the study

For each scenario, 100 random samples of size $n_0 = n_1 = 500$ are drawn. The Gaussian kernel is used to avoid divisions by zero. Recall that the bandwidth selector is the minimizer of an empirical function. As it does not have an explicit expression, numerical methods are used to approximate it. The algorithm is as follows:

Step 1 Consider a grid of 50 values of h in the $[0.01, 0.2]$, equispaced on logarithmic scale.

Step 2 h_{OPT_1} is the value that minimizes the $\mathrm{MASE}^*_{\tilde{m}_h}$, given in expression (10). Selection of g_X is discussed in Sect. 3.2.

Step 3 From the bandwidth grid take the previous and the next one to h_{OPT_1}, and construct an equally spaced grid of five values in logarithmic scale between them.

Step 4 Steps 2 and 3 are repeated three times, retaining the optimal bandwidth value in the last stage. That is taken as our numerical approximate of the optimal h_{BOOT}.

Recall from Fig. 1 that for numerical reasons, the calculated $\mathrm{MASE}^*_{\tilde{m}_h}$ might decrease for very large h. In order to avoid oversmoothing due to this phenomenon, h_{BOOT} is considered as the local minimizer of $\mathrm{MASE}^*_{\tilde{m}_h}$ closest to but larger than zero.

We are interested in the performance of our bandwidth selector in terms of prediction or data matching. To this aim, we compare our h_{BOOT}, made for prediction, (h_1 in the following) with a bandwidth selector made for regression, say h_0. Clearly, there exist many bandwidth selectors for regression. To make it comparable to h_1, notice that a different way to look at the bandwidth selection problem for nonparametric regression is to minimize the MASE in the source population. Then, h_0 is the result of minimizing

$$\begin{aligned}\mathrm{MASE}^*_{\tilde{m}_h,X^{0'}}(h) = \frac{1}{n_0}\sum_{j=1}^{n_0}\frac{1}{\hat{f}^0_{g_X}\left(X_j^{0'}\right)^2}&\left[\left(1-\frac{1}{n_0}\right)\cdot\left(\left[K_h * \hat{q}^0_{X_j^{0'},g_X}\right]\left(X_j^{0'}\right)\right)^2\right.\\ &+\frac{1}{n_0}\left[(K_h)^2 * \hat{p}^0_{X_j^{0'},g_X}\right]\left(X_j^{0'}\right)\\ &\left.+\frac{g_Y^2\mu_2(K)}{n_0^2}\sum_{i=1}^{n_0}\left[(K_h)^2 * K_{g_X}\right]\left(X_j^{0'}-X_i^0\right)\right]. \qquad (26)\end{aligned}$$

As before, estimators in (26) are calculated with the original sample $\left\{\left(X_j^0, Y_j^0\right)\right\}_{j=1}^{n_0}$, but evaluated on a different source sample $\left\{\left(X_j^{0'}, Y_j^{0'}\right)\right\}_{j=1}^{n_0}$. The selector is

$$h_0 = h_{\mathrm{MASE}^*_{\tilde{m}_h,X^{0'}}} = \arg\min_{h>0}\mathrm{MASE}^*_{\tilde{m}_h,X^{0'}}(h). \qquad (27)$$

Certainly, if one wanted to use our method to select a bandwidth for regression in practice, one would set $\left\{\left(X_j^{0'}, Y_j^{0'}\right)\right\}_{j=1}^{n_0} := \left\{\left(X_j^0, Y_j^0\right)\right\}_{j=1}^{n_0}$. This, however, produces an additional bias which disadvantages the resulting selector compared to h_1. One might conclude then in favor of h_1 only because of this bias. Notice that by construction, $h_0 \approx h_1$ if source and target population are very similar. Bandwidths h_0 and h_1 are compared by means of

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\left[m(X_i^1) - \hat{m}_{h_j^{NW}}(X_i^1)\right]^2\ ,\ j = 1, 2 \quad \text{and} \qquad (28)$$

$$\int \left[m(x) - \hat{m}_{h_j^{NW}}(x) \right]^2 \mathrm{d}F_1(x) \ , \ j = 1, 2, \tag{29}$$

where $\hat{m}_{h_j^{NW}}$ stands for the Nadaraya–Watson regression estimator obtained with source sample $\left(X^0, Y^0\right)$. The results of are presented in Figs. 2, 3and 4.

3.2 Selection of *g*

The pilot bandwidth, setting $g = g_X = g_Y$, was $g = h_{\mathrm{SJ}} n_0^{4/45}$, where h_{SJ} is the plug-in bandwidth selector proposed by Sheather and Jones (1991) for kernel density estimation. Accordingly, g has order $n_0^{-1/9}$, being the optimal rate for smoothed bootstrap, see Cao and González-Manteiga (1993). Nonetheless, an additional simulation study was carried out for this pilot bandwidth g. Consider $g = h_C n_0^{1/5-\alpha}$ with $\alpha \in \left\{ \frac{1}{5}, \frac{1}{9} \right\}$, where $h_C = h_{\mathrm{SJ}}, h_{\mathrm{CV}}$, the latter being the cross-validation bandwidth for regression. Our criteria are for $\theta = \mathbb{E}[Y^1]$.

$$\mathrm{MSE}\left[\tilde{\theta}\right] = \mathbb{E}\left[\left(\tilde{\theta} - \theta\right)^2 \right], \ \text{with } \tilde{\theta} = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{m}_{h_j}\left(X_i^1\right), j = 0, 1 \tag{30}$$

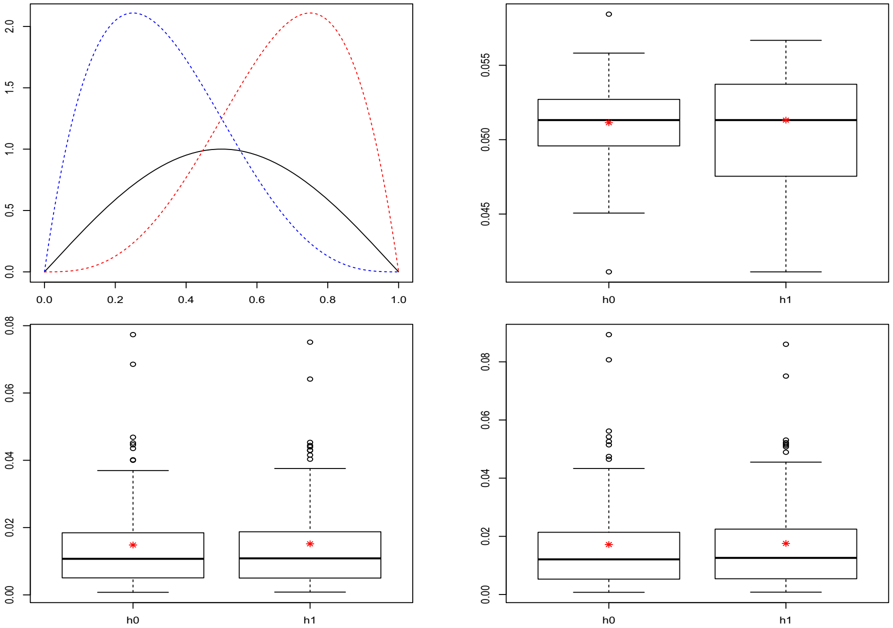

Fig. 2 Boxplots obtained for Scenario 1. Top left: regression function considered in black, density functions of the source (blue) and target (red) population. Top right: bandwidths obtained. Bottom: realized values for ASE and ISE, i.e., expressions (28) (left) and (29) (right). Red points indicate the mean, i.e., MASE and MISE in the lower panel (color figure online)

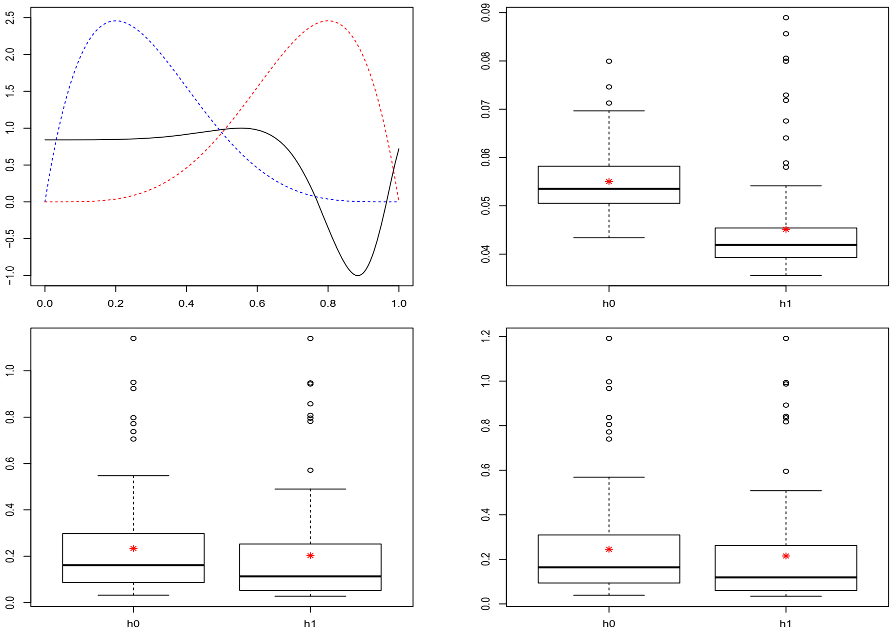

Fig. 3 Boxplots obtained for Scenario 2. Top left: regression function considered in black, density functions of the source (blue) and target (red) population. Top right: bandwidths obtained. Bottom: realized values for ASE and ISE, i.e., expressions (28) (left) and (29) (right). Red points indicate the mean, i.e., MASE and MISE in the lower panel (color figure online)

$$\text{and MeSE}\left[\tilde{\theta}\right] = \text{Median}\left[\left(\tilde{\theta} - \theta\right)^2\right], \tag{31}$$

where $\hat{m}_{h_j}$ is the Nadaraya–Watson estimator computed with the source sample. The results in Table 1suggest that α and h_C do not have a great impact on the performance of h_1. The values for (30) and (31) in Table 1 suggest to choose $\alpha = 1/9$ and $h_C = h_{\text{SJ}}$.

3.3 Discussion of simulation results

Recall that for Scenario 1 is equivalent to a situation where both densities are the same, as densities and regression function are just mirrored; one would therefore expect h_0 to perform at least as good as h_1. For all other cases, we expect h_1 to outperform h_0 (if our method works in practice). And indeed, Figs. 2, 3 and 4 reveal that in Scenario 1, h_0 is at least as good as h_1, but else h_1 clearly beats h_0. Not unexpected we find the following situations:

1. h_0 is much larger than h_1 (Scenario 2, Fig. 3). This is because the regression function considered is almost flat in the main support of the source population, but oscillates in the main support of the target population.

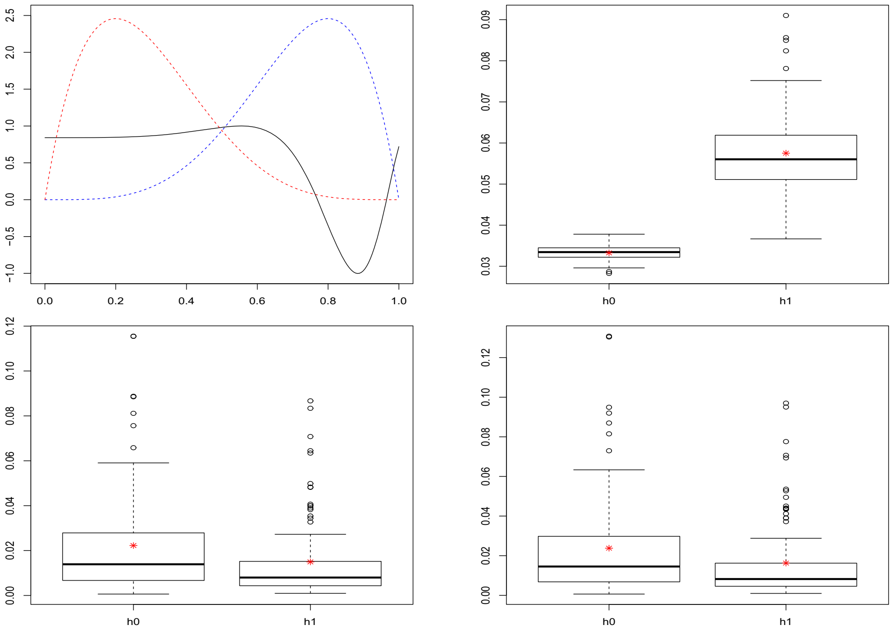

Fig. 4 Boxplots obtained for Scenario 3. Top left: regression function considered in black, density functions of the source (blue) and target (red) population. Top right: bandwidths obtained. Bottom: realized values for ASE and ISE, i.e., expressions (28) (left) and (29) (right). Red points indicate the mean, i.e., MASE and MISE in the lower panel (color figure online)

Table 1 Mean and median of ASE (28), and ISE (29), as well as expressions (30) and (31) for $\alpha = 1/5$ and $\alpha = 1/9$

	Scenario	Expression (28)		Expression (29)		Expr. (30)	Expr. (31)
		Mean	Median	Mean	Median	–	–
$\alpha = 1/5$							
$h_{\mathrm{CV}} n_0^{1/5-\alpha}$	1	0.0168	0.0129	0.0194	0.0151	0.0065	0.0045
	2	0.0419	0.0365	0.0434	0.0375	0.0103	0.0064
	3	0.0133	0.0078	0.0144	0.0078	0.0126	0.0119
$h_{\mathrm{SJ}} n_0^{1/5-\alpha}$	1	0.0163	0.0124	0.0189	0.0142	0.0073	0.0046
	2	0.0202	0.0113	0.0215	0.0119	0.0234	0.0229
	3	0.0149	0.0079	0.0162	0.0082	0.0127	0.0120
$\alpha = 1/9$							
$h_{\mathrm{CV}} n_0^{1/5-\alpha}$	1	0.0318	0.0317	0.0361	0.0355	0.0030	0.0010
	2	0.0776	0.0903	0.0809	0.0939	0.0027	0.0016
	3	0.0176	0.0072	0.0161	0.0079	0.0116	0.0111
$h_{\mathrm{SJ}} n_0^{1/5-\alpha}$	1	0.0157	0.0118	0.0181	0.0128	0.0075	0.0044
	2	0.0206	0.0113	0.0219	0.0119	0.0234	0.0023
	3	0.0161	0.0080	0.0174	0.0089	0.0128	0.0121

Table 2 CPU times (in seconds) obtained using Scenario 2, $n_0 = n_1 = 100$ and 500, for the Nadaraya–Watson regression estimator, where h^*_{MSE} is the bandwidth selector that minimizes (5) and h^*_{MASE} the one that minimizes (9)

Trials	$n_0 = n_1 = 100$		$n_0 = n_1 = 500$	
	h^*_{MSE}	h^*_{MASE}	h^*_{MSE}	h^*_{MASE}
1	0.75	0.71	18.01	17.47
100	72.19	79.36	1784.98	1758.63

Table 3 Medians of expressions (28) and (29) obtained from 100 runs, considering h^*_{MASE} minimizing (9) for different sample sizes in Scenario 3, n_0 referring to the source, n_1 to the target sample

n_1	n_0							
	25		50		100		1000	
Expression	(28)	(29)	(28)	(29)	(28)	(29)	(28)	(29)
25	0.0438	0.0337	0.0416	0.0337	0.0401	0.0393	–	–
50	0.0279	0.0215	0.0354	0.0276	0.0287	0.0299	–	–
100	0.0218	0.0173	0.0206	0.0161	0.0177	0.0179	–	–
200	0.0165	0.0136	0.0120	0.0112	0.0101	0.0103	–	–
1000	–	–	–	–	–	–	0.0038	0.0030

2. h_0 is smaller than h_1 (Scenario 3, Fig. 4). This is the opposite situation to the one mentioned before. Now $m(x)$ is quite flat in the main support of the target population, but oscillates a lot in the main support of the source population.
3. h_0 and h_1 are close (Scenario 1, Fig. 2). As discussed, this is expected because $m(\cdot)$ is just mirrored for the main supports of the respective populations.

Results of CPU times are collected in Table 2. This shows the practical behavior of the closed expression obtained for $\text{MASE}^*_{\tilde{m}_h,X^1}$ of the Nadaraya–Watson estimator given in expression (9) in terms of computer time. The order of computational complexity of this method is $\mathcal{O}(n_0 \cdot n_1)$, as shown in Table 2. This is empirically compared with the efficiency in terms of computer time of expression (5), which is the closed expression of MSE^*_x. Packages `parallel` and `Bolstad` of the free software `R` have been used. Notice that selecting our global bandwidth is about as fast as calculating one MSE-based local one. Finally, as said in the introduction and in Sect. 5, when using our selection method for the local linear estimator, see Barbeito (2020), computation time increases by a factor > 60.

In Table 3, we study the performance of h^*_{MASE} for Scenario 3 (calculated from 100 random samples) over different sample sizes, namely $n_0 \in \{25, 50, 100, 200\}$ combined with $n_1 \in \{25, 50, 100\}$, as well as the case of $n_0 = n_1 = 1000$ (for computational reasons without combinations). As expected, the performance of the bandwidth selector clearly improves as n_0 increases, while the target sample size, n_1, does not matter much.

Table 4 Mean and Median of expressions (28) and (29) obtained for 100 runs with $n_0 = n_1 = 100$ in Scenario 3, when using h^*_{MASE} compared to h_{CV}

h^*_{MASE}		h_{CV}		Expression(28)		Expression (29)	
Mean	Median	Mean	Median	Mean	Median	Mean	Median
0.0855	0.0824	–	–	0.0231	0.0177	0.0248	0.0179
–	–	0.0069	0.0068	0.0892	0.0733	0.0838	0.0748

If we consider a classic approach such as the cross-validation bandwidth (as presented in Marron 1985), we have for Scenario 3, 100 runs and $n_0 = n_1 = 100$ a very small median for the values obtained for the bandwidths selected, but large values for expressions (28) and (29), see Table 4, i.e., our method beats the classical CV selector by far.

4 Estimating the gender wage gap in Spain

Consider the problem of estimating the gender wage gap based on a 2014 survey in Spain, i.e., after its recovery from the financial crisis. Wage gaps have been of long-standing political concern as an indicator of discrimination. Yet, the fact that women are paid lower wages than men may well be due to differences in education, experience and/or the sector they work in. The challenge is then to account for factors X that might explain differences in wages. Until today, gender wage gap studies that account for those characteristics are mainly based on fully parametric models. Often they separate the male and female populations for looking at the Blinder–Oaxaca decomposition, see Blau and Kahn (2017) for a recent review. Moral-Arce et al. (2012) introduced a semiparametric version of this decomposition and extended it to quantiles to study the heterogeneity of the gap. They studied the gender wage gap in Spain before the economic crisis, specifically the development around the millennium. For more references consult these two articles.

Our approach is different in several aspects. First, it is fully nonparametric. Having done all nonparametrically, any found difference cannot be explained by the choice of the (semi-)parametric wage model. One may argue that we missed other important factors. However, as gender is externally given, such objection only shifts the discussion to the definition of the gender wage gap. Second, instead of deriving a nonparametric Blinder–Oaxaca decomposition, we explore the heterogeneity of the wage gap over sectors and education. Third, we calculate the counterfactual wages of women as if they were paid like men, to compare these with their realized wages. One may say that in the parametric approach one could do something similar but with the simpler linear model. However, apart from the fact that this can be a poor predictor, its parameters are the result of least square projections inside one population. That means, the coefficients of the linear approximation of men's wages are not made for predicting the counterfactual wages of women. Consequently, they are inappropriate, except if the characteristics of women and their jobs were the same as for men.

The data set used is the EPA of 2014 of the National Institute for Statistics (INE) in Spain, provided at webpage http://www.ine.es. It consists of 64383 observations for women, and 108134 for men, giving the gross annual wage, which is our response variable Y. We only consider full-time employees, being aware that already the distribution of full- and part-time jobs might be discriminatory; that is, our findings are conditioned on the full-time employed population. We are provided with the following covariates:

- Years and months of service, which is a quantitative variable that indicates the professional experience of the individual.
- CNAE, which stands for the National Classification of Economics Activities. It is a qualitative variable that splits the active population into eighteen groups: (B) Extractive industry (anthracite, oil, coal and lignite extraction), (C) Manufacturing industry, (D) Electricity and gas supply, (E) Water supply, (F) Construction, (G) Wholesale and retail trade activities, (H) Transport and storage, (I) Hotel industry, (J) Information and communication, (K) Financial activities and insurances, (L) Real-estate sector, (M) Professional, scientific and technical activities, (N) Administrative activities, (O) Public, defense and social security administration, (P) Education, (Q) Health care system activities, (R) Arts and (S) Other services (computer repair…)
- Studies, a qualitative ordinal variable that divides the population into: (1) Primary education, (2) First stage of secondary education, (3) Second stage of it, (4) Vocational training (FP), (5) Bachelor's degree, (6) Masters degree and/or Ph.D. (7) No studies

Consider the two different populations; men (source) and women (target). We have one (pseudo-)continuous covariate (*Years and months of service*, denoted as X^0 in the source with density f^0, X^1 in the target with density f^1). The other variables are categorical and denoted as Z^0 and Z^1, respectively. We observe the men's Y^0 such that we can estimate their $m(\cdot)$. Under the hypothesis of no wage discrimination, women are paid by the same wage function, such that $m(\cdot)$ can be used to predict their wages Y^1. Therefore, we are not interested in the optimal bandwidth for predicting men's mean wage function, but for predicting women's counterfactual wages, i.e., we need h_1.

In order to account for the categorical variables CNAE and education nonparametrically, and because we want to explore the heterogeneity of the wage gap over these variables, we split the sample into the eighteen times seven subsamples. However, those with no studies were tiny; and it is not always clear what 'no studies' actually means. Therefore, these were not further considered. One may ask if this is necessary or whether educational levels may have similar distributions for men and women once we fixed the sector. Figure 5shows the differences in distribution of *Studies* between men and women for the examples of sectors F and R. Indeed, in some sectors like R the distributions looks somewhat more similar; but in most sectors they clearly do not. Therefore, we keep all information and split along sector and education. Notice that such split is equivalent to a nonparametric estimation with the full sample, and setting the smoothing parameter for these two variables to zero.

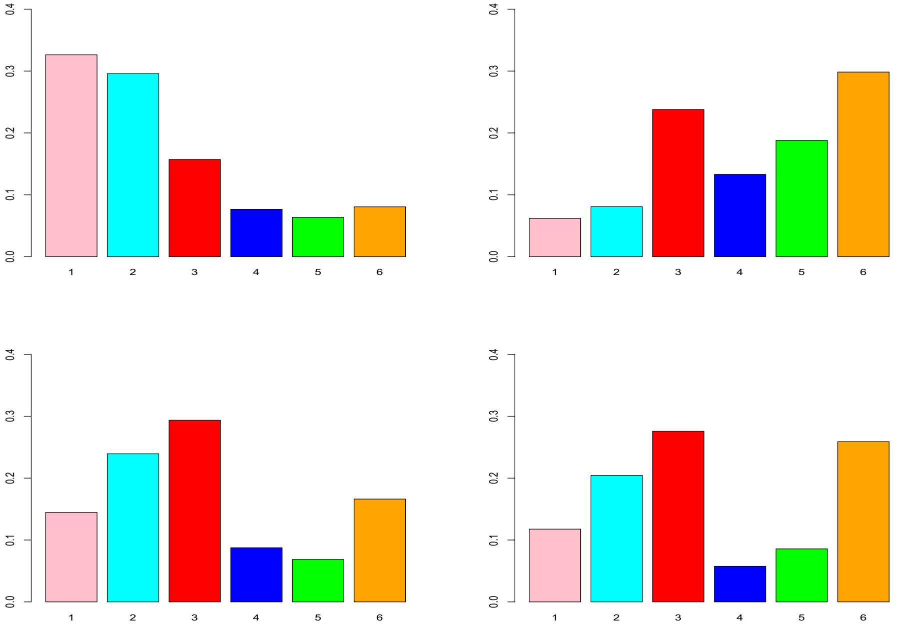

Fig. 5 Distribution of Studies in sector F (upper) and R (lower) for men (left) and women (right), respectively

We are left now in each subsample with wage and professional experience. For the overall average wage gap, one can still integrate over Z. Expressing all this in formulas looks as follows: Say Z is the categorical variable with r modes, $\left(X^0, Z^0, Y^0\right)$ is observed in the source population, $\left(X^1, Z^1\right)$ in the target population. We want $\mathbb{E}[Y^1|x,z] = m(x,z)$ for all observed (X_i^1, Z_i^1) with $m(x,z) := \mathbb{E}[Y^0|x,z]$ being estimated from the source population. Then, one predicts the average counterfactual wage of women by

$$\hat{\theta} = \widehat{\mathbb{E}\left[Y^1\right]} = \frac{1}{n_1}\sum_{j=1}^{n_1} \tilde{m}_h\left(X_j^1, Z_j^1\right).$$

Denote $m_z(x) = \mathbb{E}\left[Y^0 \middle| X^0 = x, Z^0 = z\right]$. In other words, m_z is obtained from the regression of $Y^0\big| Z^0 = z$ on $X^0\big| Z^0 = z$. Define for indicator function $\mathbb{I}_{\{\cdot\}}$

$$n_{0,z} = \sum_{i=1}^{n_0} \mathbb{I}_{\{Z_i^0 = z\}} = \#\left\{i \in \{1,\ldots,n_0\} \,|Z_i^0 = z\right\},$$
$$n_{1,z} = \sum_{j=1}^{n_1} \mathbb{I}_{\left\{Z_j^1 = z\right\}} = \#\left\{j \in \{1,\ldots,n_1\} \,|Z_j^1 = z\right\},$$

$$\tilde{m}_h(x,z) = \hat{m}_{z,h}(x) = \frac{\frac{1}{n_{0,z}}\sum_{i=1}^{n_0} K_h\left(x - X_i^0\right) Y_i^0 \mathbb{I}_{\{Z_i^0=z\}}}{\frac{1}{n_{0,z}}\sum_{i=1}^{n_0} K_h\left(x - X_i^0\right) \mathbb{I}_{\{Z_i^0=z\}}}, \text{ and}$$

$$\hat{\theta}_z = \frac{1}{n_{1,z}}\sum_{j=1}^{n_1} \tilde{m}_h(X_j^1, Z_j^1)\mathbb{I}\left\{Z_j^1 = z\right\} = \frac{1}{n_{1,z}}\sum_{j=1}^{n_1} \frac{\sum_{i=1}^{n_0} K_h\left(X_j^1 - X_i^0\right) Y_i^0 \mathbb{I}_{\left\{Z_i^0=Z_j^1\right\}}}{\sum_{i=1}^{n_0} K_h\left(X_j^1 - X_i^0\right) \mathbb{I}_{\left\{Z_i^0=Z_j^1\right\}}} \mathbb{I}_{\left\{Z_j^1=z\right\}}$$

being the average counterfactual wage of women given z, i.e., for a specific educational level in a given sector. The estimator of $\hat{\theta}$ is a weighted average of the $\hat{\theta}_z$

$$\begin{aligned}\hat{\theta} &= \frac{1}{n_1}\sum_{j=1}^{n_1} \tilde{m}_h\left(X_j^1, Z_j^1\right) = \frac{1}{n_1}\sum_{z\in I}\sum_{j=1}^{n_1} \tilde{m}_h\left(X_j^1, Z_j^1\right)\mathbb{I}_{\left\{Z_j^1=z\right\}} \\ &= \frac{1}{n_1}\sum_{z\in I} n_{1,z}\cdot\frac{1}{n_{1,z}}\sum_{j=1}^{n_1} \tilde{m}_h\left(X_j^1, Z_j^1\right)\mathbb{I}_{\left\{Z_j^1=z\right\}} = \sum_{z\in I}\frac{n_{1,z}}{n_1}\hat{\theta}_z. \qquad (32)\end{aligned}$$

Certainly, it is up to the practitioner which of the r dimensions of Z (s)he wants to integrate out and which to keep. In our case study, we looked at all combinations but present here only the results of all $\hat{\theta}_z$ to explore the heterogeneity of the wage gap over educational level and sectors, see Figs. 6 and 7 . We plotted the absolute and relative wage gap, calculated as the (absolute and relative) difference between

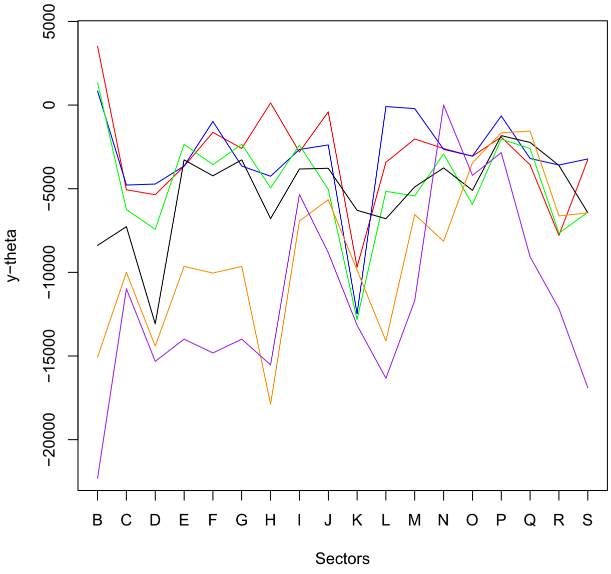

Fig. 6 Plot of $\left(\bar{y}_z^1\right) - \hat{\theta}_z$ (with $\bar{y}_z^1$ being the average of observed wages of women for z) for the 18 sectors and 6 levels of studies (1: red, 2: blue, 3: green, 4: black, 5: orange, 6: purple). Results obtained from our method (color figure online)

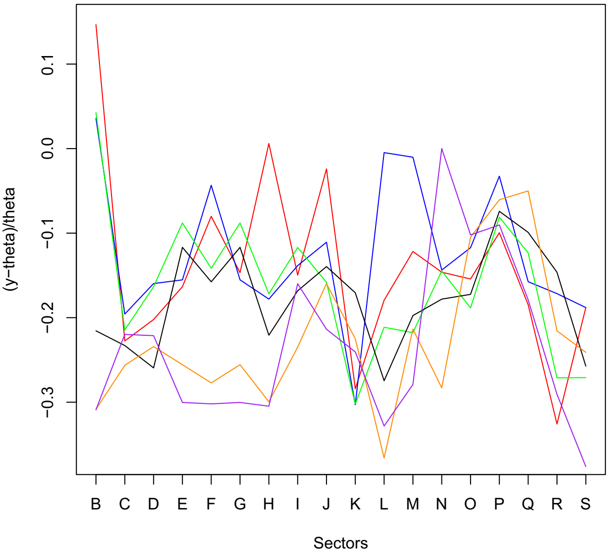

Fig. 7 Plot of $\left\{\left(\bar{y}_z^1\right) - \hat{\theta}_z\right\} / \hat{\theta}_z$ (with $\bar{y}_z^1$ being the average of observed wages of women for z) for the 18 sectors and 6 levels of studies (1: red, 2: blue, 3: green, 4: black, 5: orange, 6: purple). Results obtained from our method (color figure online)

the average realized minus predicted wages per sector and educational level. In the Supplementary Material and in Barbeito (2020) are given many more details like all numerical numbers together with $n_{0,z}$, $n_{1,z}$ and MSE estimates. Note that for each mode z (i.e., all combinations of sectors and education) one may want to calculate the optimal $h_{BOOT,z}$. So we did, and the values are given in Table 6 in the Supplementary Material. We used the Gaussian kernel and $g_z = h_{\text{SJ}} n_{0,z}^{4/45}$. We have compared these results to those provided by h_0, given in expression (27), and a plug-in bandwidth minimizing the MASE (namely, h_{PI}). The bandwidths obtained were quite different. For instance, considering sector F (Construction) and level of studies 3, $h_0 = 2.01$, $h_1 = 0.17$, $h_{\text{PI}} = 5.26$. Considering sector P (Education) and level of studies 5, $h_0 = 2.96$, $h_1 = 8.05$, $h_{\text{PI}} = 3.62$. Not surprisingly, they suggest for those sectors accordingly different wage gaps compared to our method. Recalling the findings of the simulation study, we then trust more in the results provided by our method.

From Figs. 6 and 7, we see that for most of the educational levels and sectors, the observed salaries of women are between 10 and 30% lower than the wage equation obtained from men's salaries would predict. This calculus does not account for potential discrimination by gender regarding equal opportunities, for instance, whether it is harder for women to get into better paying sectors. The same holds for equal opportunities in education and professional experience, e.g., due to maternal leaves. That is, we only study the monetary discrimination in a given sector, once a certain educational level and professional experience was achieved. Discrimination in opportunities comes on top, but is harder to measure and often not translated into monetary terms. This limitation applies in general, and is not particular to our method.

Finally, a reviewer suggested to compare our results with those one obtains using the `gam` command from package `mgcv` in `R` calling `Y=Z1+Z2+Z1Z2+s(x,by=Z1Z2)`, i.e., using continuous variable X and two factors (Z_1,Z_2). This was done using the estimated weights $\hat{f}^1\left(X_i^0\right)/\hat{f}^0\left(X_i^0\right)$ with both densities being estimated by kernel density estimation with their optimal bandwidths along Sheather and Jones (1991). This weighting tries to optimize the prediction similar to the CV modification in Galdo et al. (2008). As can be seen in Fig. 8, this gives by far a larger variability over sectors and education with little credible results like 'positive wage discrimination' of up to 40% and many cases of negative wage discrimination between 40 to almost 60%. Strangely, when we rerun this exercise without weights, these extremes remained but some figures changed. Note also that for these estimates, especially in the case of using estimated weights, we do not know how to obtain confidence intervals or standard errors. We guess that some bootstrap methods could be developed. A check of the manual reveals that the simulated confidence bands provided in `R` are not helpful here.

5 Conclusions and discussion

This paper contributes to the existing, large literature on bandwidth selection, by providing a selector designed for prediction and data-matching problems when the distribution of the covariates, X, in the target population is potentially different from the one in the source population. This is a standard problem in counterfactual exercises

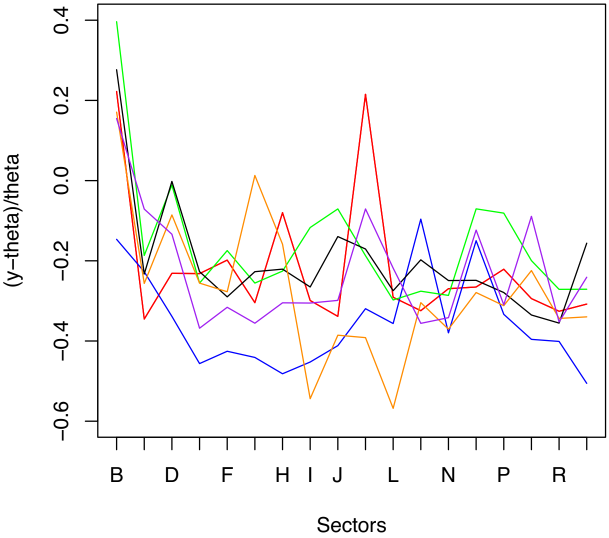

Fig. 8 Plot of $\left\{\left(\bar{y}_z^1\right)-\hat{\theta}_z\right\}/\hat{\theta}_z$ (with $\bar{y}_z^1$ being the average of observed wages of women for z) for the 18 sectors and 6 levels of studies (1: red, 2: blue, 3: green, 4: black, 5: orange, 6: purple). Results obtained by using `gam` with estimated weights $\hat{f}^1\left(X_i^0\right)/\hat{f}^0\left(X_i^0\right)$ (color figure online)

like causal analysis and studies on discrimination, but also quite frequent for the calculus of scenarios. It offers thereby an interesting alternative to Häggström and Luna (2014) and Galdo et al. (2008) which seem to be the only existing contributions in this direction so far. Our method has a closed form, is computationally attractive, asymptotically well understood and shows a very satisfying behavior in simulations, even for moderate and small sample sizes.

Along the quite exemplifying problem of studying the gender wage gap in Spain after the last big economic crises, we do not only motivate our method but also illustrate its usefulness and application for such a highly relevant issue. Moreover, we show a first extension toward its use in a multivariate context. We succeed to carve out the heterogeneity and depth of this wage gap over different sectors and education levels. This application study is completed with some comparisons using alternative selectors or estimation methods.

There are certainly several interesting extensions thinkable, most of them having already been mentioned at some point in this paper. First note that the proposed selection procedure could be interesting for more general model selection problems in the context of prediction and data matching, not only for finding the bandwidth. Recall that modified cross-validation is frequently used for all kind of features selection, in classical kernel regression as well as in various recent machine learning procedures.

A second extension would be to explicitly account for potential boundary effects. Certainly, if the support of f^0 covers the support of f^1 such that all observations x_i made in the target sample are interior points of f^0 or if the conditional expectation is relatively flat at the boundaries, such correction is not needed. Also, unless many x_i of the target sample suffer strongly from boundary effects, the selector itself should hardly be affected as it just searches the minimum of the MASE which in turn is calculated from a comparison of two estimates that would both suffer from the same boundary effects. However, especially for the final estimate, an obvious remedy would be to use boundary kernels like the local linear (or 'equivalent') one of Jones (1993). For our selector, we had then to replace assumption (B1) by (B1') *K is a second order kernel, a density for interior points, else a left, respectively, right boundary kernel.* A careful check of the proof suggests that this would not change our results but make our presentation and implementation somewhat more cumbersome. For our application, we realized that it had hardly any impact for the above mentioned reasons (f^0 spans typically over f^1, etc.).

A third extension has already been studied in the thesis of Barbeito (2020), namely the one toward local linear estimators (Fan and Gijbels 1992). While the empirical results turned out to be alike to the simulation results for the Nadaraya–Watson, the expressions for the bootstrap version of the MASE become extremely complex, leading to a quite important loss of efficiency in the computation of the bandwidth selector.

A fourth extension would be allow for various continuous covariate simultaneously. In practice one could follow then the suggestion of Köhler et al. (2014) and use a multivariate kernel with a bandwidth matrix proportional to $\Sigma_X^{-1/2}$ with Σ_X being the variance–covariance matrix of X in the source group, times a constant to be chosen along our criterion. This approach, however, needs some deeper examinations.

A fifth extension would be to use our selector if one wants to reveal the distribution of the unobserved Y for the target group similar to Dai et al. (2016). This is not only interesting for analyzing poverty and vulnerability; we noticed that also in our discrimination study, in some sectors the residual variance was still pretty large. For those sectors, one would like to compare real with counterfactual distributions, not just the mean.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-022-00838-7.

Acknowledgements This research has been supported by the MICINN Grant PID2020-113578RB-I00. The first two authors have been supported by the MINECO Grant MTM2017-82724-R, and the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. Furthermore, the first author acknowledges financial support from the Xunta de Galicia and the European Union (European Social Fund—ESF), ED481A-2017/215.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Appendix

A Asymptotics for the selection of prior bandwidth *g*

Because of the tedious computations required to find an analytical expression for the expectations involved in expression (25), especially for $\mathbb{E}\left(\xi^2\right)$, we present some asymptotic upper bounds for $\hat{A}_g - A$ and $\hat{B}_g - B$. We analyze the closeness of $\hat{A}_g$ to A, and $\hat{B}_g$ to B. Proofs of Lemmas 3 and 4 are deferred to the Supplementary Material.

Definition 2 In order to analyze the MSE of $\hat{A}_g$ and $\hat{B}_g$, consider the following expression:

$$C^{[s]}_{\nu,\ell,r} := \int \nu(x) \left(\hat{\Psi}^{(r)}_{s,\ell}(x) - \Psi^{(r)}_{s,\ell}(x)\right) dx, \tag{33}$$

where $\nu(x)$ is a function $\nu : \mathbb{R} \to \mathbb{R}$, $\ell \in \mathbb{Z}^+$, $r \in \mathbb{Z}^+$, $s \in \{0,1\}$, $\hat{\Psi}^{(r)}_{s,\ell}(x) = \dfrac{1}{n_s\, g^{r+1}} \sum_{i=1}^{n_s} K^{(r)}\left(\dfrac{x - X_i^s}{g}\right) Y_i^{s\ell}$ and $\Psi^{(r)}_{s,\ell}(x) := \dfrac{\partial^r \left(m_\ell(x) f^s(x)\right)}{\partial x^r}$.

We can state a lemma concerning an approximation for $\hat{A}_g - A$.

Lemma 3 *Consider the expressions for $\hat{A}_g$ and A. Then,*

$$\hat{A}_g - A = \sum_{i=1}^{k_0} a_i C^{[s_i]}_{\nu_i,\ell_i,r_i} + A_1, \tag{34}$$

where $k_0 = 6$, $a_1 = 1, a_2 = -1, a_3 = 1, a_4 = -1, a_5 = -2, a_6 = -2$, $\nu_1(x) = \frac{\sigma^2(x)}{f^0(x)}$, $\nu_2(x) = \frac{\sigma^2(x) f^1(x)}{f^0(x)^2}$, $\nu_3(x) = \frac{f^1(x)}{f^0(x)^2}$, $\nu_4(x) = \frac{f^1(x)\Psi_2(x)}{f^0(x)^3}$, $\nu_5(x) = \frac{f^1(x)\Psi_1(x)}{f^0(x)^3}$, $\nu_6(x) = \frac{f^1(x)\Psi_1^2(x)}{f^0(x)^4}$, $\ell_1 = 0$, $\ell_2 = 0$, $\ell_3 = 2$, $\ell_4 = 0$, $\ell_5 = 1$, $\ell_6 = 0$, $r_1 = 0$, $r_2 = 0$, $r_3 = 0$, $r_4 = 0$, $r_5 = 0$, $r_6 = 0$, $[s_1] = 1$, $[s_2] = 0$, $[s_3] = 0$, $[s_4] = 0$, $[s_5] = 0$, $[s_6] = 0$ *and* $A_1 = \mathcal{O}\left(r_{0,n_0}\right)$, *with*

$$\begin{aligned} r_{0,n_0} = & \int \left(\hat{f}^0_g(x) - f^0(x)\right)^2 dx + \int \left(\hat{f}^0_g(x) - f^0(x)\right) \cdot \left(\hat{\Psi}_{2,g}(x) - \Psi_2(x)\right) dx \\ & + \int \left(\hat{f}^0_g(x) - f^0(x)\right) \cdot \left(\hat{\sigma}^2_g(x)\hat{f}^1_g(x) - \sigma^2(x) f^1(x)\right) dx \\ & + \int \left(\hat{f}^0_g(x)^2 - f^0(x)^2\right) \cdot \left(\hat{\Psi}^2_{1,g}(x) - \Psi^2_1(x)\right) dx + \int \left(\hat{\Psi}_{1,g}(x) - \Psi_1(x)\right)^2 dx \\ & + \int \left(\hat{\sigma}^2_g(x) - \sigma^2(x)\right) \cdot \left(\hat{f}^1_g(x) - f^1(x)\right) dx + \int \left(\hat{f}^0_g(x)^2 - f^0(x)^2\right)^2 dx. \end{aligned}$$

Similarly, we can state a result for the difference $\hat{B}_g - B$.

Lemma 4 *Given the expressions for $\hat{B}_g$ and B, then $\hat{B}_g - B$ consists of a sum of 60 terms similar to those in expression (34). Specifically,*

$$\hat{B}_g - B = \sum_{i=7}^{k_1} a_i C^{[s_i]}_{\nu_i,\ell_i,r_i} + B_1, \tag{35}$$

where $k_1 = 66$. The functions $\nu(x)$ and the values of r, ℓ, a and $[s]$ are collected in Tables 1–4 in the Supplementary Material. Additionally, term B_1 is of order $\mathcal{O}\left(r_{1,n_0}\right)$ with r_{1,n_0} being also given in the Supplementary Material.

As an immediate consequence of the Tchebycheff inequality, we can conclude:

Corollary 3 *An upper bound for expressions $\hat{A}_g - A$ and $\hat{B}_g - B$, under regularity conditions (B1)–(B4) and considering a g of order $n_0^{-1/2}$, is given by:*

$$\hat{A}_g - A = \mathcal{O}_P\left(n_0^{-1/2}\right), \text{and } \hat{B}_g - B = \mathcal{O}_P\left(n_0^{-1/2}\right). \tag{36}$$

References

Antoniadis A, Paparoditis E, Sapatinas T (2009) Bandwidth selection for functional time series prediction. Stat Probab Lett 79:733–740

Barbeito I (2020) Exact bootstrap methods for nonparametric curve estimation (Doctoral dissertation). Universidade da Coruña

Blau F, Kahn L (2017) The gender wage gap: extent, trends, and explanations. J Econ Lit 55:789–865

Cao R, González-Manteiga W (1993) Bootstrap methods in regression smoothing. J Nonparametr Stat 2:379–388

Dai J, Sperlich S, Zucchini W (2016) A simple method for predicting distributions by means of covariates with examples from welfare and health economics. Swiss J Econ Stat 152:49–80

Eurostat (2013) Statistical matching: a model based approach for data integration. In: Methodologies and working papers, Eurostat, Luxembourg

Fan J, Gijbels I (1992) Variable bandwidth and local linear regression smoothers. Ann Stat 20:2008–2036

Frölich M (2005) Matching estimators and optimal bandwidth choice. Stat Comput 15:197–215

Frölich M, Sperlich S (2019) Impact evaluation: treatment effects and causal analysis. Cambridge University Press

Galdo JC, Smith J, Black D (2008) Bandwidth selection and the estimation of treatment effects with unbalanced data. Annales d'Économie et de Statistique 91–92:189–216

Häggström J, Luna X (2014) Targeted smoothing parameter selection for estimating average causal effects. Comput Stat 29:1727–1748

Heidenreich NB, Schindler A, Sperlich S (2013) Bandwidth Selection Methods for Kernel density estimation: a review of fully automatic selectors. AStA Adv Stat Anal 97:403–433

Jones MC (1993) Simple boundary correction in Kernel density estimation. Stat Comput 3:135–146

Köhler M, Schindler A, Sperlich S (2014) A review and comparison of bandwidth selection methods for Kernel regression. Int Stat Rev 82:243–274

Li X, Heckman NE (2003) Local linear extrapolation. J Nonparametr Stat 15:565–578

Mack YP, Silverman BW (1982) Weak and strong uniform consistency of kernel regression estimates. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 61:405–415

Marron J (1985) An asymptotically efficient solution to the bandwidth problem of kernel density-estimation. Ann Stat 13:1011–1023

Moral-Arce I, Sperlich S, Fernández-Saínz A, Roca M (2012) Trends in the Gender Pay Gap in Spain: A Semiparametric Analysis. J Labor Res 33:173–195

Rubin DB (2004) Multiple imputation for nonresponse in surveys. John Wiley, New York

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 53:683–690

Silverman BW (1978) Weak and strong uniform consistency of the kernel esitmate of a density and its derivatives. Ann Stat 6:177–184

Su YS, Gelman A, Hill J, Yajima M (2010) Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Stat Softw 45:1–31

Tschernig R, Yang L (2000) Nonparametric lag selection for time series. J Time Ser Anal 21:457–487

van Buuren S (2018) Flexible imputation of missing data, 2nd edn. Chapman and Hall/CRC

Vansteelandt S, Bekaert M, Claeskens G (2012) On model selection and model misspecification in causal inference. Stat Methods Med Res 21(1):7–30