Comput Stat (2014) 29:1727–1748
DOI 10.1007/s00180-014-0515-0

ORIGINAL PAPER

Targeted smoothing parameter selection for estimating average causal effects

Jenny Häggström · Xavier de Luna

Received: 3 May 2013 / Accepted: 30 June 2014 / Published online: 25 July 2014

Abstract The non-parametric estimation of average causal effects in observational studies often relies on controlling for confounding covariates through smoothing regression methods such as kernel, splines or local polynomial regression. Such regression methods are tuned via smoothing parameters which regulates the amount of degrees of freedom used in the fit. In this paper we propose data-driven methods for selecting smoothing parameters when the targeted parameter is an average causal effect. For this purpose, we propose to estimate the exact expression of the mean squared error of the estimators. Asymptotic approximations indicate that the smoothing parameters minimizing this mean squared error converges to zero faster than the optimal smoothing parameter for the estimation of the regression functions. In a simulation study we show that the proposed data-driven methods for selecting the smoothing parameters yield lower empirical mean squared error than other methods available such as, e.g., cross-validation.

Keywords Causal inference · Double smoothing · Local linear regression

1 Introduction

In observational studies where the interest lies in estimating the average causal effect of a binary treatment z on an outcome of interest y, non-parametric estimators are typically based on controlling for confounding covariates x with smoothing regression methods (kernel, splines, local polynomial regression, series estimators; see, e.g., the reviews by Imbens 2004, and Imbens and Wooldridge 2009). A useful modeling

J. Häggström (✉) · X. de Luna
Department of Statistics, Umeå School of Business and Economics,
Umeå University, 90187 Umeå, Sweden
e-mail: jenny.haggstrom@stat.umu.se

framework in this context was introduced by Neyman (1923) and Rubin (1974), where in particular two potential outcomes are considered for each unit in the study, the outcome that would be observed if the unit is treated, $y(1)$, and the outcome that would be observed if the unit is not treated, $y(0)$. The causal effect at the unit level is defined as $y(1)-y(0)$. Population parameters are targeted by the inference, and we focus here on average causal effects of the type $E(y(1)-y(0))$, where the expectation is taken over a given population of interest. Inference on such expectations is complicated by the fact that the two potential outcomes are not observed for all units in the sample (missing data problem) and assumptions, e.g., on the missingness mechanism must be made in order for the parameter of interest to be identified. In this paper, we consider situations described in Sect. 2, where the causal effect conditional on an observed covariate x (or a score function summarizing a set of observed covariates), $E(y(1) \mid x)-E(y(0) \mid x)$, is identified and can be estimated by fitting two curves, functions of x, $E(y(1) \mid x, z=1)$ and $E(y(0) \mid x, z=0)$ non-parametrically. An estimate of the targeted average causal effect is obtained by averaging the estimated curves over the relevant distribution for x to target $E(y(1)-y(0)) = E(E(y(1) \mid x))-E(E(y(0) \mid x))$, where the missing outcomes are imputed by predictions from the fitted curves. A tuning parameter for each fitted curve is used to regulate the smoothness of the fit. Cheng (1994) showed that when using kernel regression to estimate the average of a curve, say here $E(E(y(1) \mid x))$, with missing $y(1)$ for some units, as described above, then the optimal (in mean squared error, MSE, sense) smoothing parameter for the estimation of the regression curve $E(y(1) \mid x, z=1)$ is not optimal for the estimation of the average $E(E(y(1) \mid x))$. More precisely the optimal rate of convergence towards zero of the smoothing parameter (when the sample size increases) is different in both situations, and one need typically to asymptotically undersmooth $E(y(1) \mid x, z=1)$ when targeting $E(E(y(1) \mid x))$. We show in this paper that a similar result holds when using local linear regression instead of kernel regression, and when two curves (implying the choice of two tunining parameters), are fitted and then averaged to target $E(y(1)-y(0))$.

As a main contribution of the paper, we propose a novel data-driven method geared for selecting the smoothing parameters which minimizes the mean squared error of non-parametric estimators of the average causal effect. Imbens et al. (2005) also proposes a data-driven method based on the estimation of this mean squared error. The two estimators are, however, different. While Imbens et al. (2005) estimates an asymptotic approximation of the population MSE which involves the estimation of the propensity score, the probability of ending up in one of the treatment groups (say $z=1$) given the covariates, our estimator targets the exact population MSE by using a double smoothing technique previously used by Härdle et al. (1992) for estimating regression curves and Häggström (2013) in semi-parametric additive models. Note that Frölich (2005) also derived asymptotic approximation of MSE to obtain smoothing parameter selectors although those were outperformed by cross-validation in finite sample simulations. With simulations we study the finite sample properties of the different data-driven methods. The results suggest that the cross-validation choice, which is known to be optimal in MSE sense to estimate smooth curves (Fan 1992), can indeed be improved by using either Imbens et al. (2005) or our proposal, with the latter often being superior.

In the next section we introduce the potential outcome framework dating back to Neyman (1923) and Rubin (1974), which allows us to define the parameter of

interest, the average causal effect, and commonly used identifying assumptions and estimators. The selection of smoothing parameters is discussed in Sect. 3, where we present asymptotic results based on the use of local linear regression. We also introduce in this section a novel data-driven method. Section 4 presents a simulation study. The paper is concluded in Sect. 5.

2 Model and estimation

2.1 Neyman–Rubin model for causal inference

Suppose we have n units indexed by $i = 1, \ldots, n$. For each unit i a binary treatment z_i is assigned:

$$z_i = \begin{cases} 1 & \text{if unit } i \text{ receives treatment 1,} \\ 0 & \text{if unit } i \text{ receives treatment 0.} \end{cases}$$

Further, each unit i is characterised by two potential outcomes $y_i(1)$ and $y_i(0)$, where $y_i(1)$ is the response that is observed if the unit is given treatment $z_i = 1$ and $y_i(0)$ the response if the unit is given treatment $z_i = 0$. Only one treatment assignment is possible for each unit and, therefore, only one of the two potential outcomes is observed. Denote by $y_i = y_i(0)(1 - z_i) + y_i(1)z_i$ the observed outcome. Finally, let all units have a vector of d background characteristics $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$ (called covariates). We assume in the sequel that the n units corresponds to a random sample from the distribution law of the random variables $(y_i(1), y_i(1), z_i, \mathbf{x}_i)$, and that only $(y_i, z_i, \mathbf{x}_i)$ is actually observed. We use the same notation to denote random variables and their realisations, letting the context make the distinction.

The parameter of interest herein is an average causal effect,

$$\tau = E\big(y_i(1) - y_i(0)\big).$$

If treatment assignment is not randomized, τ is identified if we have available a vector of covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$ not affected by treatment assignment and such that the following assumptions hold,

$$y_i(1), y_i(0) \perp\!\!\!\perp z_i | \mathbf{x}_i,$$

often called unconfoundedness assumption, and

$$0 < \Pr(z_i = 1 | \mathbf{x}_i) < 1,$$

often called overlap assumption. The sign $\perp\!\!\!\perp$ is used here to mean "is independent of" (Dawid 1979). We have unconfoundedness if all covariates affecting both treatment assignment and the potential outcomes are included in $\mathbf{x}_i$. This is a strong assumption which must be based on subject-matter reasoning. A sensitivity analysis to this assumption is often advocated (e.g., de Luna and Lundin 2014). The assumption of overlap states that, for a unit with covariate vector $\mathbf{x}_i$, the probability of receivingeither

treatment should be bounded away from 0. This assumption can be investigated empirically (e.g.,Imbens and Wooldridge 2009). Under these assumptions identifiability of τ is then a consequence of

$$\begin{aligned}\tau &= E\big(y_i(1) - y_i(0)\big)\\ &= E\big(E(y_i(1)|\mathbf{x}_i) - E(y_i(0)|\mathbf{x}_i)\big)\\ &= E\big(E(y_i(1)|z_i = 1, \mathbf{x}_i) - E(y_i(0)|z_i = 0, \mathbf{x}_i)\big)\\ &= E\big(E(y_i|z_i = 1, \mathbf{x}_i) - E(y_i|z_i = 0, \mathbf{x}_i)\big). \qquad (1)\end{aligned}$$

In the sequel we focus on the case $d = 1$ since when $d > 1$, the covariate vector $\mathbf{x}_i$ can be replaced by a scalar, e.g., $p(\mathbf{x}_i) = \Pr(z_i = 1|\mathbf{x}_i)$, the propensity score (Rosenbaum and Rubin 1983, Hansen 2008). Indeed, Rosenbaum and Rubin (1983) showed that it is sufficient to condition on the propensity score, i.e., under the above assumptions we have $y_i(1), y_i(0) \perp\!\!\!\perp z_i|p(\mathbf{x}_i)$, and $0 < \Pr(z_i = 1|p(\mathbf{x}_i)) < 1$. In applications the propensity score need to be modelled and fitted to the data and such situations are considered in the simulation study of Sect. 4. Typically parametric models are used to fit the propensity score, although these do not need to be correctly specified as shown in Waernbaum (2010). Note also that covariate selection procedures may be used to reduce the dimensionality of $\mathbf{x}_i$ (Luna et al. 2011).

2.2 Estimating average causal effects

Let $\beta_0(x_i) = E(y_i|z_i = 0, x_i)$ and $\beta_1(x_i) = E(y_i|z_i = 1, x_i)$ be unknown smooth functions, $Var(y_i|x_i, z_i) = \sigma^2_\varepsilon$, $i = 1, \ldots, n$. Note that the assumption of constant conditional variance could be relaxed without changing in essence the results of this paper. We consider this assumption to alleviate the notational burden. The non-constant variance case is further discussed in the concluding section. From (1), we have that

$$\tau = E\big(\beta_1(x_i)\big) - E\big(\beta_0(x_i)\big).$$

Thus, a natural way to estimate τ is to first estimate the two regression functions $\beta_1(x_i)$ and $\beta_0(x_i)$, based on the treated and the non-treated, respectively, and then take the average over all the observed x_is of the differences between the estimated functions. This estimator of τ is called the imputation estimator in Imbens et al. (2005). They use series estimators for estimating the regression functions but any smoother, e.g., kernel, splines and local polynomial regression (Fan and Gijbels 1996, pp. 14–45), may be used.

Denote $\mathbf{y}^0 = (y^0_1, \ldots, y^0_{n_0})^T$ and $\mathbf{x}^0 = (x^0_1, \ldots, x^0_{n_0})^T$ the observed response and covariate for the n_0 units with treatment $z_i = 0$, and similarly $\mathbf{y}^1 = (y^1_1, \ldots, y^1_{n_1})^T$ and $\mathbf{x}^1 = (x^1_1, \ldots, x^1_{n_1})^T$ for the n_1 units with treatment $z_i = 1$. The smoothers cited above are linear in the sense that the corresponding estimator of $\beta_j(\mathbf{x}) = (\beta_j(x_1), \ldots, \beta_j(x_n))^T$, can be written as

$$\hat{\beta}_j^{h_j}(\mathbf{x}) = S_j^{h_j}[\mathbf{x}]\mathbf{y}^j, \quad j = 0, 1,$$

where $\mathbf{x} = (\mathbf{x}^{0T}, \mathbf{x}^{1T})^T$ and $S_j^{h_j}[\mathbf{x}]$ the smoothing matrix regressing $\mathbf{y}^j$ on $\mathbf{x}^j$, using smoothing parameter h_j. The imputation estimator of τ mentioned above is

$$\hat{\tau}^{imp} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}^{imp}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\beta}_1^{h_1}(x_i) - \hat{\beta}_0^{h_0}(x_i)\right).$$

In this paper we base our results on a specific linear smoother, the local linear regression smoother, although we anticipate that most results should hold for any other linear smoother.

Local linear regression (Cleveland 1979; Fan and Gijbels 1996), consists in fitting a straight line at every x_i, $i = 1, \ldots, n$, using only the part of data that is deemed to be sufficiently close to the target point x_i. Consider estimating the regression function $\beta_j(\cdot)$, $j = 0, 1$. The fit, at x_i, is

$$\hat{\beta}_j^{h_j}(x_i) = \mathbf{e}_1^T(\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{X}_i^j)^{-1}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{y}^j = S_j^{h_j}[x_i]\mathbf{y}^j$$

where $\mathbf{e}_1 = (1, 0)^T$,

$$\mathbf{X}_i^j = \begin{pmatrix} 1 & \left(x_1^j - x_i\right) \\ \vdots & \vdots \\ 1 & \left(x_{n_j}^j - x_i\right) \end{pmatrix}$$

and

$$\mathbf{W}_i^{h_j} = \mathrm{diag}(K\left((x_1^j - x_i)/b_{ji}\right)/b_{ji}, \ldots, K\left((x_{n_j}^j - x_i)/b_{ji}\right)/b_{ji}).$$

$K(\cdot)$ is a kernel function such that $\int K(u)du = 1$ and $\int uK(u)du = 0$. An example is the tricube kernel defined as

$$K(u) = \left\{ \begin{array}{ll} \frac{70}{81}(1 - |u|^3)^3, & \text{if } |u| < 1 \\ 0, & \text{if } |u| \geq 1 \end{array} \right\}.$$

The definition of b_{ji}, $i = 1, \ldots, n$, depends on the type of bandwidth we use. With a constant bandwidth $b_{j1} = \cdots = b_{jn} = h_j$. For a nearest neighbor type bandwidth, assuming no ties, b_{ji} is the Euclidian distance from x_i to the $(h_j n_j)$:th nearest among the x_k^j:s for $x_k^j \neq x_i$, $h_j \in [1/n_j, 1]$, $k = 1, \ldots, n_j$, and the smoothing parameter h_j is the proportion of observations being used to produce the local fit.

3 Selection of smoothing parameters

3.1 Mean squared errors

Many smoothing parameter selection methods are developed with the purpose of estimating the regression function $\beta_j(x_i)$, $j = 0, 1, i = 1, \ldots, n$ and attempts to select the smoothing parameter minimizing the average conditional mean squared error:

$$\begin{aligned}
&\frac{1}{n_j}\sum_{i=1}^{n_j} E\left(y_i^j - \hat{\beta}_j^{h_j}(x_i^j)|\mathbf{x}^j\right)^2 \\
&= \frac{1}{n_j}\sum_{i=1}^{n_j} Var\left(\hat{\beta}_j^{h_j}(x_i^j)|\mathbf{x}^j\right) + \frac{1}{n_j}\sum_{i=1}^{n_j} E\left(\hat{\beta}_j^{h_j}(x_i^j) - \beta_j(x_i^j)|\mathbf{x}^j\right)^2 \\
&= \frac{\sigma_\varepsilon^2}{n_j}\sum_{i=1}^{n_j} S_j^{h_j}[x_i^j]S_j^{h_j}[x_i^j]^T + \frac{1}{n_j}\sum_{i=1}^{n_j}\left(S_j^{h_j}[x_i^j]\beta_j(\mathbf{x}^j) - \beta_j(x_i^j)\right)^2. \qquad (2)
\end{aligned}$$

One frequently used selection procedure that attempts to select the smoothing parameter minimizing (2) is leave-one-out cross-validation. In this setting, cross-validation selects the smoothing parameter h_j minimizing

$$\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_i^j - \hat{\beta}_j^{h_j,-i}\left(x_i^j\right)\right)^2, \qquad (3)$$

where $\hat{\beta}_j^{h_j,-i}(x_i^j)$ is the cross-validated estimate at x_i^j computed without (x_i^j, y_i^j). Asymptotically, for local linear regression, the smoothing parameter minimizing (2) is proportional to $n_j^{-1/5}$ (Fan 1992), and, hence, proportional to $n^{-1/5}$ since $n_j = n\Pr(z = j) + o_p(n)$. However, it is known that for estimating a functional of $\beta_j(x_i)$ such as $E(\beta_j(x_i))$, the smoothing parameter minimizing (2) is not optimal, in the sense that it does not result in $\sqrt{n}$-consistent estimation of the functional (e.g., Cheng 1994).

Imbens et al. (2005) suggest that one should select h_0 and h_1 by minimizing the conditional mean squared error of $\frac{1}{n}\sum_{i=1}^n \hat{\beta}_j^{h_j}(x_i)$, for $j = 0, 1$ respectively, i.e.,

$$\begin{aligned}
MSE_{\bar{\hat{\beta}}_j} = &\frac{\sigma_\varepsilon^2}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} S_j^{h_j}[x_i]S_j^{h_j}[x_k]^T \\
&+ \frac{1}{n^2}\left[\sum_{i=1}^{n}\left(S_j^{h_j}[x_i]\beta_j(\mathbf{x}^j) - \beta_j(x_i)\right)\right]^2. \qquad (4)
\end{aligned}$$

We argue that, in order to estimate τ optimally, it may be more suitable to select the combination of (h_0, h_1) minimizing the conditional mean squared error of $\hat{\tau}^{imp}$

$$MSE_{\hat{\tau}} = \frac{\sigma_\varepsilon^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(S_1^{h_1}[x_i]S_1^{h_1}[x_j]^T + S_0^{h_0}[x_i]S_0^{h_0}[x_j]^T\right) + \left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(S_1^{h_1}[x_i]\beta_1(\mathbf{x}^1) - \beta_1(x_i)\right) - \left(S_0^{h_0}[x_i]\beta_0(\mathbf{x}^0) - \beta_0(x_i)\right)\right)\right]^2. \tag{5}$$

Note that

$$MSE_{\hat{\tau}} = MSE_{\bar{\hat{\beta}}_1} + MSE_{\bar{\hat{\beta}}_0} - 2\left(\frac{1}{n}\sum_{i=1}^{n}\left(S_1^{h_1}[x_i]\beta_1(\mathbf{x}^1) - \beta_1(x_i)\right)\right) \times \left(\frac{1}{n}\sum_{i=1}^{n}\left(S_0^{h_0}[x_i]\beta_0(\mathbf{x}^0) - \beta_0(x_i)\right)\right).$$

Hence, criterion (5) differs from (4) when both average bias terms in the latter expression are different from zero.

3.2 Asymptotics

Asymptotic approximations can be used to describe optimal bandwidth choices as the sample size tends to infinity. The results presented here are deduced in Appendix, Sect. 6.2, where regularity conditions also used in Ruppert and Wand (1994) are given. For local linear regression with constant bandwidth such that $h_j \to 0$ and $nh_j \to \infty$ as $n \to \infty$ we have the following approximations for the conditional bias and variance of $\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)$. For $j = 0, 1$,

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i) - \frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i)|\mathbf{x}\right) = B_1(j)h_j^2 + o_p\left(h_j^2\right), \tag{6}$$

and

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)|\mathbf{x}\right) = \frac{V_1(j)}{n} + \frac{V_2(j)}{n^2 h_j} + V_3(j)\frac{h_j^2}{n} + o_p\left(n^{-1} + n^{-2}h_j^{-1} + n^{-1}h_j^2\right), \tag{7}$$

with constants

$$B_1(j) = \frac{1}{2}\int \beta_j^{(2)}(x)f(x)dx \int u^2 K(u)du,$$
$$V_1(j) = \sigma_\varepsilon^2 \int \frac{f(x)}{Pr(z = j|x)}dx,$$

$$V_2(j) = \sigma_\varepsilon^2 \int K(u)^2 du \int \frac{1}{Pr(z=j|x)} dx,$$
$$V_3(j) = -2\sigma_\varepsilon^2 \int u^2 K(u) du \int \frac{f^{(1)}(x)^2}{f(x) Pr(z=j|x)} dx - 2\sigma_\varepsilon^2 \int u^2 K(u) du \int \frac{f^{(1)}(x) P^{(1)}(z=j|x)}{Pr(z=j|x)^2} dx,$$

where $\beta_j^{(m)}(x)$ the m:th derivative of the function $\beta_j(x)$ and $f(x)$ is the density of x. Hence,

$$MSE_{\hat{\bar{\beta}}_j} = \frac{V_1(j)}{n} + \frac{V_2(j)}{n^2 h_j} + V_3(j)\frac{h_j^2}{n} + B_1^2(j)h_j^4 + o_p\left(n^{-1} + n^{-2}h_j^{-1} + n^{-1}h_j^2 + h_j^4\right) \tag{8}$$

and

$$MSE_{\hat{\tau}} = \frac{V_1(1) + V_1(0)}{n} + \frac{V_2(1)}{n^2 h_1} + \frac{V_2(0)}{n^2 h_0} + V_3(1)\frac{h_1^2}{n} + V_3(0)\frac{h_0^2}{n} + B_1^2(1)h_1^4 + B_1^2(0)h_0^4 - 2B_1(1)B_1(0)h_1^2 h_0^2 + o_p\left(n^{-1} + n^{-2}h_1^{-1} + n^{-2}h_0^{-1} + n^{-1}h_1^2 + h_0^2 n^{-1} + h_1^4 + h_0^4 + h_1^2 h_0^2\right). \tag{9}$$

Let us first consider the optimal smoothing parameter for estimating $E(\beta_j(x))$ and assume $nh_j^3 \to 0$ as $n \to \infty$, $j = 0, 1$. An asymptotic approximation to the bandwidth minimizing (8) is

$$h_j^{opt} = \arg\min_{h_j} \frac{V_2(j)}{n^2 h_j} + B_1^2(j)h_j^4 = \left(\frac{V_2(j)}{4B_1^2(j)}\right)^{1/5} n^{-2/5}.$$

Hence, the optimal bandwidths are of order $n^{-2/5}$, so that the optimal bandwidths for the estimation of the average functional τ is smaller than the optimal bandwidths for the estimation of the regression functions $\beta_j(\cdot)$, the latter being of order $n^{-1/5}$. Thus, the regression functions must be undersmooth when the target of the inference is τ. A similar result was shown in Cheng (1994) for kernel regression. Turning to the minimization of (9), this must be done simultaneously in h_0 and h_1. A reasonable assumption, however, is that these two smoothing parameters have the same rate of convergence to zero. Under this assumption we may replace h_1 by ch_0, for c a constant, in (9). Minimizing the latter for h_0 yields as above an optimal bandwidth of order $n^{-2/5}$.

Another related result, deduced from (6) and (7), is that as $n \to \infty$, if $h_j \propto n^r$, for $-1 < r < -1/4$, then (see Appendix, Sect. 6.2)

$$E\left[\sqrt{n}(\bar{\hat{\beta}}_j - E(\beta_j(x_i))) \mid \mathbf{x}\right] = o_p(1), \tag{10}$$

$$E\left[\sqrt{n}(\hat{\tau}^{imp} - \tau) \mid \mathbf{x}\right] = o_p(1), \tag{11}$$

$$Var\left[\sqrt{n}(\bar{\hat{\beta}}_j - E(\beta_j(x_i))) \mid \mathbf{x}\right] = V_1(j) + o_p(1), \tag{12}$$

$$Var\left[\sqrt{n}(\hat{\tau}^{imp} - \tau) \mid \mathbf{x}\right] = V_1(0) + V_1(1) + o_p(1). \tag{13}$$

The results above show that selecting the smoothing parameters minimizing (4) will lead to $\sqrt{n}$-consistent estimation of τ. This is in accordance with previous results (e.g., Speckman 1988) where it was shown that asymptotic undersmoothing of the regression function is needed for the $\sqrt{n}-$consistent estimation of a functional of the regression function.

3.3 Estimating MSEs

Imbens et al. (2005) propose the following estimator of (4), for $j = 0, 1$,

$$\begin{aligned}\widehat{MSE}^{INR}_{\bar{\hat{\beta}}_j} = {} & \frac{\hat{\sigma_\varepsilon}^2}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} S_j^{h_j}[x_i]S_j^{h_j}[x_k]^T \\ & + \frac{1}{n^2}\left[\sum_{i=1}^{n_j}\frac{1}{\hat{p}(x_i^j)}\left(y_i^j - \hat{\beta}_j^{h_j}(x_i^j)\right)\right]^2 \\ & - \frac{\hat{\sigma_\varepsilon}^2}{n^2}\hat{\mathbf{p}}_j^T\left(I_{n_j} - S_j^{h_j}[\mathbf{x}^j]\right)\left(I_{n_j} - S_j^{h_j}[\mathbf{x}^j]\right)^T\hat{\mathbf{p}}_j,\end{aligned} \tag{14}$$

where $\hat{\mathbf{p}}_j = (1/\hat{p}(x_1^j), \ldots, 1/\hat{p}(x_{n_j}^j))^T$ and I_{n_j} is the $n_j \times n_j$ identity matrix. It is worth noting that one need to estimate the propensity score (Waernbaum 2010), in addition to σ_ε^2, in order to use this selection procedure. The error variance σ_ε^2 may be estimated by

$$\hat{\sigma_\varepsilon}^2 = \frac{\left[\mathbf{y} - \left(\hat{\beta}_0^{h_{\varepsilon_0}}(\mathbf{x}^0)^T, \hat{\beta}_1^{h_{\varepsilon_1}}(\mathbf{x}^1)^T\right)^T\right]^T\left[\mathbf{y} - \left(\hat{\beta}_0^{h_{\varepsilon_0}}(\mathbf{x}^0)^T, \hat{\beta}_1^{h_{\varepsilon_1}}(\mathbf{x}^1)^T\right)^T\right]}{n - \text{trace}(2S_0^{h_{\varepsilon_0}}[\mathbf{x}^0] - S_0^{h_{\varepsilon_0}}[\mathbf{x}^0]S_0^{h_{\varepsilon_0}}[\mathbf{x}^0]) - \text{trace}(2S_1^{h_{\varepsilon_1}}[\mathbf{x}^1] - S_1^{h_{\varepsilon_1}}[\mathbf{x}^1]S_1^{h_{\varepsilon_1}}[\mathbf{x}^1])}, \tag{15}$$

where $\mathbf{y} = (\mathbf{y}^{0T}, \mathbf{y}^{1T})^T$ and h_{ε_j}, $j = 0, 1$, could be equal to h_j or selected separately, see, e.g., Opsomer et al. (1995) for further discussion on this issue.

We propose below novel double smoothing estimators of (4) and (5), respectively:

$$\widehat{MSE}^{DS}_{\tilde{\beta}_j} = \frac{\hat{\sigma_\varepsilon}^2}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} S_j^{h_j}[x_i]S_j^{h_j}[x_k]^T + \frac{1}{n^2}\left[\sum_{i=1}^{n}\left(S_j^{h_j}[x_i]\hat{\beta}_j^{g_j}(\mathbf{x}_j) - \hat{\beta}_j^{g_j}(x_i)\right)\right]^2, \tag{16}$$

and

$$\widehat{MSE}^{DS}_{\hat{\tau}} = \frac{\hat{\sigma_\varepsilon}^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(S_1^{h_1}[x_i]S_1^{h_1}[x_j]^T + S_0^{h_0}[x_i]S_0^{h_0}[x_j]^T\right) + \left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(S_1^{h_1}[x_i]\hat{\beta}_1^{g_1}(\mathbf{x}^1) - \hat{\beta}_1^{g_1}(x_i)\right) - \left(S_0^{h_0}[x_i]\hat{\beta}_0^{g_0}(\mathbf{x}^0) - \hat{\beta}_0^{g_0}(x_i)\right)\right)\right]^2, \tag{17}$$

where g_0, g_1 are pilot smoothing parameters. Because the purpose of these pilots parameters is to estimate β_0 and β_1 respectively, we suggest using leave-one-out cross-validation; see (3). In specific situations one may want to check whether the results are sensitive to changes in the choice of the pilot parameters. The double smoothing (DS) estimation concept was utilized by Härdle et al. (1992), although for the estimation of the entire regression function $\beta_j(\cdot)$. One could, as mentioned by Härdle et al. (1992), specify the pilot bandwidths as $g_j = n_j^{-c}$, for an appropriate constant c which would result in good asymptotic performance. This would also reduce the computational burden of the method, although a relevant choice of the arbitrary constant c is problematic. Finally, note that a difference between $\widehat{MSE}^{INR}_{\tilde{\beta}_j}$ and $\widehat{MSE}^{DS}_{\tilde{\beta}_j}$ is that the former is based on an asymptotic approximation of (4) while the double smoothing estimator targets (4) directly.

4 Simulation study

In this section, we study the finite sample properties of different methods for the selection of constant and nearest neighbour type bandwidths, and in particular the resulting empirical MSE when estimating the average causal effect τ.

4.1 Design of the study

Data were generated according to the model

$$y_i = \beta_0(x_i) + \tau(x_i)z_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{18}$$

with $x_i \sim \text{Uniform}(0, 2\pi)$, $z_i|x_i \sim \text{Bernoulli}(p(x_i))$, $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$, $\tau(x_i) = \beta_1(x_i) - \beta_0(x_i)$, $\sigma_\varepsilon^2 \approx Var\left(\beta_0(x_i) + \tau(x_i)z_i\right)$, $n = 100, 200, 500, 1{,}000$. Since z_i is

Table 1 Specification of the six designs used to generate data according to model (18)

Design	$\beta_1(x_i)$	$\beta_0(x_i)$
1	$4\pi + 5 - 2\pi x_i + x_i^2 + 5\sin(2x_i)$ $-4\cos(x_i)$	$\sin(2x_i) - 4\cos(x_i) + 5$
2	$4\big(x_i + \sin(x_i) + \sin(2x_i)\big) + 3$	$2\big(x_i + \sin(x_i) + \sin(2x_i)\big) + 3$
3	$4\pi - \pi x_i + \frac{x_i^2}{2}$	$\pi x_i - \frac{x_i^2}{2}$
4	$4\pi - \pi x_i + \frac{x_i^2}{2}$	$\pi x_i - \frac{x_i^2}{2}$
5	$4\pi + 5 - 2\pi x_i + x_i^2 + 5\sin(2x_i)$ $-4\cos(x_i)$	$\sin(2x_i) - 4\cos(x_i) + 5$
6	$10 + x_i(2\pi - x_i)$ $\times \sin(2\pi(2\pi + 0.05)/(x_i + 0.05))$	$8 + 1.5\sin(2x_i - 4) + 6exp(-16(2x_i - 2.5)^2)$
Design	$\tau(x_i)$	$p(x_i)$
1	$4\pi - 2\pi x_i + x_i^2 + 4\sin(2x_i)$	$[e^{-3.5+x_i}]/[1 + e^{-3.5+x_i}]$
2	$2x_i + 2\sin(x_i) + 2\sin(2x_i)$	$[e^{-3.5+x_i}]/[1 + e^{-3.5+x_i}]$
3	$4\pi - 2\pi x_i + x_i^2$	$[e^{-3.5+x_i}]/[1 + e^{-3.5+x_i}]$
4	$4\pi - 2\pi x_i + x_i^2$	$(5\sin 2x_i - 4\cos x_i + 4\pi - 2\pi x_i + x_i^2)/11.3$
5	$4\pi - 2\pi x_i + x_i^2 + 4\sin(2x_i)$	$(5\sin 2x_i - 4\cos x_i + 4\pi - 2\pi x_i + x_i^2)/11.3$
6	$2 + x_i(2\pi - x_i)\sin(\frac{2\pi(2\pi+0.05)}{x_i+0.05})$ $-1.5\sin(2x_i - 4) + 6exp(-16(2x_i - 2.5)^2)$	$(5\sin 2x_i - 4\cos x_i + 4\pi - 2\pi x_i + x_i^2)/11.3$

a Bernoulli draw dependent on x_i generated from a uniform distribution, n_1 and n_0 are stochastic. Table 1 and Fig. 1 display the six designs generated. Bandwidths h_0 and h_1 considered are, for the constant bandwidth setting, 40 equally spaced values within the interval $[h_{min}, 2\pi]$, where h_{min} is the smallest bandwidth value such that at least 10 observations are used for the local fits. For the nearest neighbour bandwidth setting, we consider 40 equally spaced values within the intervals [0.1, 1] for $n = 100, 200$ and [0.02, 1] for $n = 500, 1{,}000$, and, e.g., $h = 0.1$ implies using 10 % of the data for the local fits. The propensity score, $p(x)$, in (14) is estimated by logistic regression with correctly specified model for Design 1–3 (i.e., `glm(z~x, family=binomial)` in R) and misspecified model for Design 4–6 (i.e., `glm(z ~I(sin(2*x))+I(cos(x))+x+I(x^2), family=binomial)` in R). The variance estimator (15) is used in (14), (16) and (17) with h_{ε_j}, $j = 0, 1$, selected by leave-one-out cross-validation (3). These cross-validation bandwidths are also used as pilot bandwidths in the DS estimators in (16) and (17) .

The criteria (2), (3), (4), (5), (14), (16) and (17) are computed for every bandwidth, 40 values, in the interval. For the minimizing bandwidths $\hat{\tau}^{imp}$ is computed. Due to computer time constraint, we use 200 replicates. On the other hand, we reduce noise in the simulation results by making use of the control variate method (see, e.g., (Wilson 1984) with $\hat{\tau}^{ols}$, the mean of the fitted values resulting from estimating $\tau(x)$ by ordinary least squares with correctly specified model, as control variate. If $\hat{\tau}^{ols}$ is positively correlated with $\hat{\tau}^{imp}$ then $\hat{\tau}^{c} = \hat{\tau}^{imp} - (\hat{\tau}^{ols} - \tau)$ has the same mean as

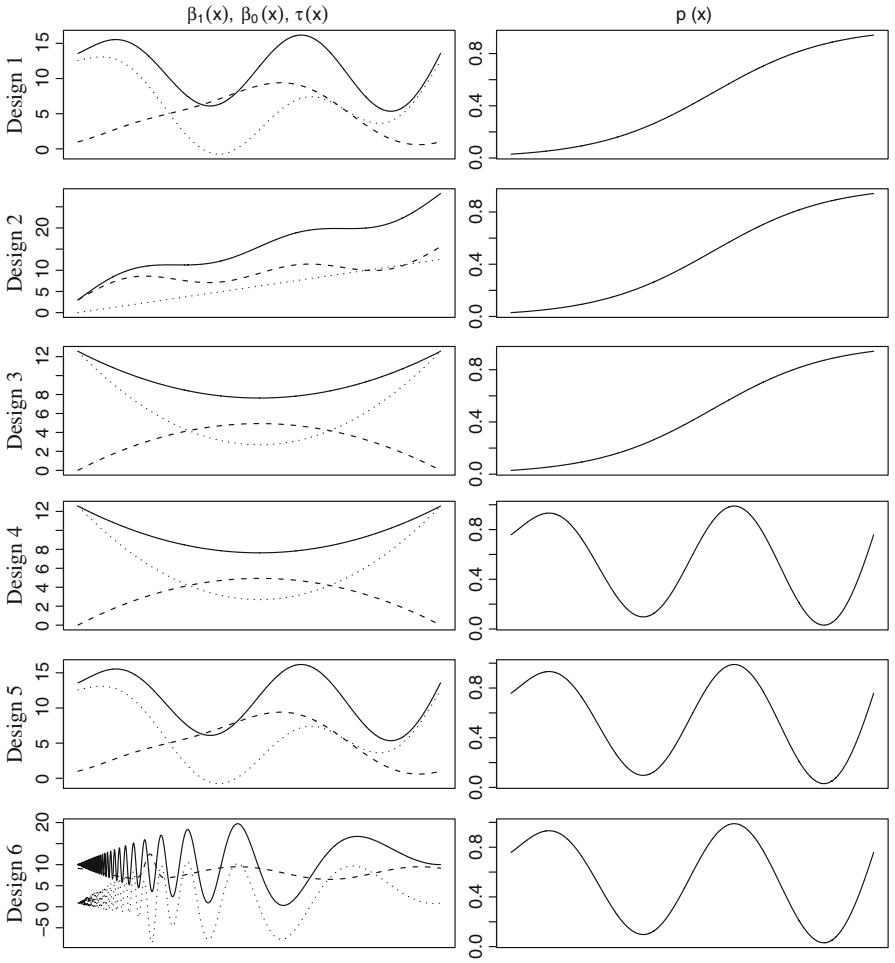

Fig. 1 Design 1–6 (from *top* to *bottom*) used to generate data as specified in Table 1. The first column displays $\beta_1(x_i)$ (*solid line*), $\beta_0(x_i)$ (*dashed*) and $\tau(x_i)$ (*dotted*), and the second column displays $p(x_i)$

$\hat{\tau}^{imp}$ but lower variance. For instance, for $n = 1{,}000$ such correlations varied between 0.39 and 0.96 (Median = 0.82, IQR = 0.18). Results based on the raw replicates are similar to the results reported here utilizing the control variate method, except for an increase in noise. All computations are made in R (Core Team 2014). Studying bandwidth selection by simulation is computationally demanding and this study was made possible by the use of the High Performance Computing Center North (HPC2N) at Umeå University.

4.2 Results

Results for $n = 500$ and 1,000 are displayed in Tables 2 and 3, for both constant and nearest neighbour bandwidths, and in Figs. 2, 3, 4 and 5 (Appendix, Sect. 6.1) for

Table 2 MSE comparison: the table displays the method yielding lowest MSE (in the estimation of τ) among M_β, M_τ and M_y, when either constant or nearest neighbour bandwidth are used

Design	Minimum MSE obtained by			
	n			
	100	200	500	1,000
Constant bandwidth				
1	M_y^{**}	M_τ^{**}	M_β^{**}	M_τ^{**}
2	M_β	M_β	M_β	M_β
3	M_τ	M_τ	M_τ^{**}	M_β^{**}
4	M_τ	M_τ	M_y	M_y^{*}
5	M_τ	M_β^{**}	M_y	M_β^{**}
6	M_y	M_β	M_τ	M_τ
Nearest neighbour bandwidth				
1	M_τ^{**}	M_τ^{**}	M_τ^{**}	M_τ^{**}
2	M_τ	M_τ	M_τ	M_τ
3	M_τ^{**}	M_τ	M_τ^{**}	M_τ^{**}
4	M_β	M_τ^{*}	M_τ	M_τ
5	M_β^{**}	M_τ^{**}	M_τ^{**}	M_τ^{**}
6	M_β	M_τ	M_τ	M_τ

Stars indicate that the method has significantly lower MSE than the next best method, with "*" for a 5 % level test and "**" for a 1 % level test

nearest neighbour bandwidths. Due to the similarity of bandwidth selection patterns, and to save space, analogous figures with results for constant bandwidths are not included. These figures and more detailed results (also for $n = 100, 200$), also left out to save space, can be obtained from the authors. Note first that we can compute the smoothing parameter values minimizing (2), (4) and (5), labeled M_y, M_β and M_τ, respectively, because we know the data generating mechanisms.

We see in Figs. 2, 3, 4 and 5 that the double smoothing methods introduced, (16) and (17), labeled DS_β and DS_τ respectively, mimic quite well their target in terms of selected smoothing parameters. This is not the case for (14), labeled INR, whose selected smoothing parameters are not in accordance with the target M_β. Table 2 summarizes empirical MSE results for the theoretical criteria M_β, M_τ and M_y, by indicating which criterion yielded lowest MSE for the estimation of τ. For constant bandwidhts, the smallest MSE is most often obtained by M_β or M_τ and the largest MSE is most often obtained by M_y. However, only in 17 and 25 % of the cases, respectively, do M_τ and M_β result in significantly lower MSE than M_y. For nearest neighbour bandwidths, we see that M_τ always results in smallest MSE for $n = 200, 500, 1{,}000$, which is, in half of the cases, significantly smaller than the second smallest MSE (achieved by M_β). Both M_τ and M_β result in significantly smaller MSE than M_y in a majority of cases (71 and 67 %, respectively). Table 3 gives information on empirical MSE (similar to Table 2), where comparisons are made between the data-driven criteria DS_β, DS_τ, INR and CV. For both the constant and nearest neighbour bandwidth setting, we see that double smoothing does not always yields lowest empirical MSE, although CV is most often outperformed by the methods targeting the estimation of functional averages (DS and INR − for design 2 when INR performed best, CV was also outperformed by DS_τ but not by DS_β).

Table 3 MSE comparison: the table displays the method yielding lowest MSE (in the estimation of τ) among DS_β, DS_τ, INR and CV, when either constant or nearest neighbour bandwidth are used

Design	Minimum MSE obtained by			
	n			
	100	200	500	1,000
Constant bandwidth				
1	DS_β^{**}	INR**	DS_τ^{**}	DS_β^{**}
2	CV	INR	INR	DS_τ
3	INR**	INR	DS_τ	DS_τ^{**}
4	DS_τ	DS_τ	DS_β	DS_β
5	DS_β	DS_β	DS_β	DS_β
6	CV	DS_β	DS_τ	DS_τ^{**}
Nearest neighbour bandwidth				
1	DS_β	DS_τ	DS_τ	DS_β
2	INR	INR*	INR*	INR**
3	CV	CV	CV**	CV**
4	INR	DS_τ	DS_τ	DS_τ^{**}
5	INR	DS_τ	DS_τ	DS_τ^{*}
6	INR	INR	DS_τ	CV*

Stars indicate that the method has significantly lower MSE than the next best method, with "*" for a 5 % level test and "**" for a 1 % level test

Finally, note that the propensity scores used in the designs of this study are rather extreme in the sense that they may yield probabilities near zero and one. We have also run these experiments by damping these propensity scores to let them vary only between 0.2 and 0.8. The results where similar qualitatively with double smoothing often performing better.

5 Conclusion

In this paper we have proposed double smoothing methods for selecting smoothing parameters that target the estimation of functional averages where the latter are average causal effects of interest. In our numerical experiments cross-validation is often outperformed by double smoothing as we expected since the latter criterion is optimized for the estimation of functions underlying the average causal effect, and not the average itself. The methods proposed and studied here have large applicability, and are, for instance, straightforward to adapt to non-parametric estimators based on instruments as those introduced in Frölich (2007). Finally, note that similar results as the one obtained should hold under a non-constant variance assumption (Andrews 1991; Ruppert and Wand 1994). In such cases the estimation of σ_ε^2 need to be replaced by estimators of $Var(y_i|x_i, z_i = 0)$ and $Var(y_i|x_i, z_i = 1)$, e.g. using linear smoothers when regressing y_i^2 on x_i for the units with $z_i = 0$ and $z_i = 1$ respectively.

Acknowledgments We are grateful to Yanyuan Ma and Sara Sjöstedt-de Luna for comments that have helped us to improve the paper. We acknowledge the financial support of the Swedish Research Council through the Swedish Initiative for Research on Microdata in the Social and Medical Sciences (SIMSAM), the Ageing and Living Condition Program and Grant 70246501.

6 Appendix

6.1 Figures with results

See Figs. 2, 3, 4 and 5.

6.2 Asymptotics

In order to derive the results of Sect. 3.2 we focus on local linear regression with constant bandwidth. We use further the following assumptions.

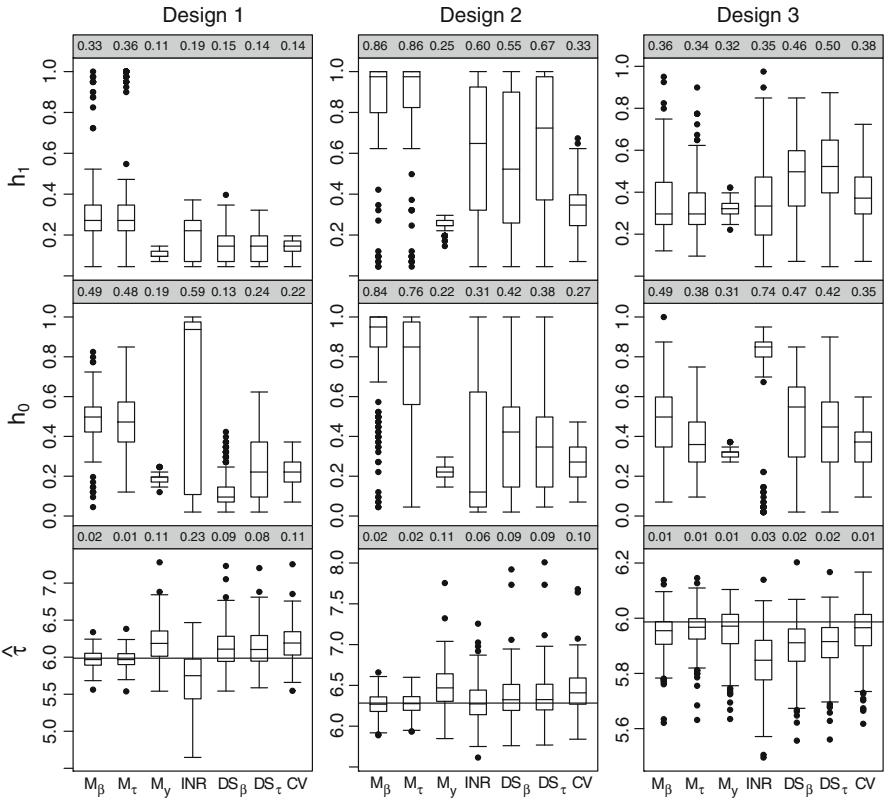

Fig. 2 Design 1–3 columnwise, sample size $n = 500$. Selected bandwidths and resulting $\hat{\tau}$ when using (4), labeled M_β, (5), labeled M_τ, (2), labeled M_y, (14), labeled INR, (16), labeled DS_β, (17), labeled DS_τ, and (3), labeled CV. Average h values are given on *top* of the figures in the *two first rows*, while in the *last row* resulting empirical MSEs of the estimates of τ are displayed on *top* of the boxplots

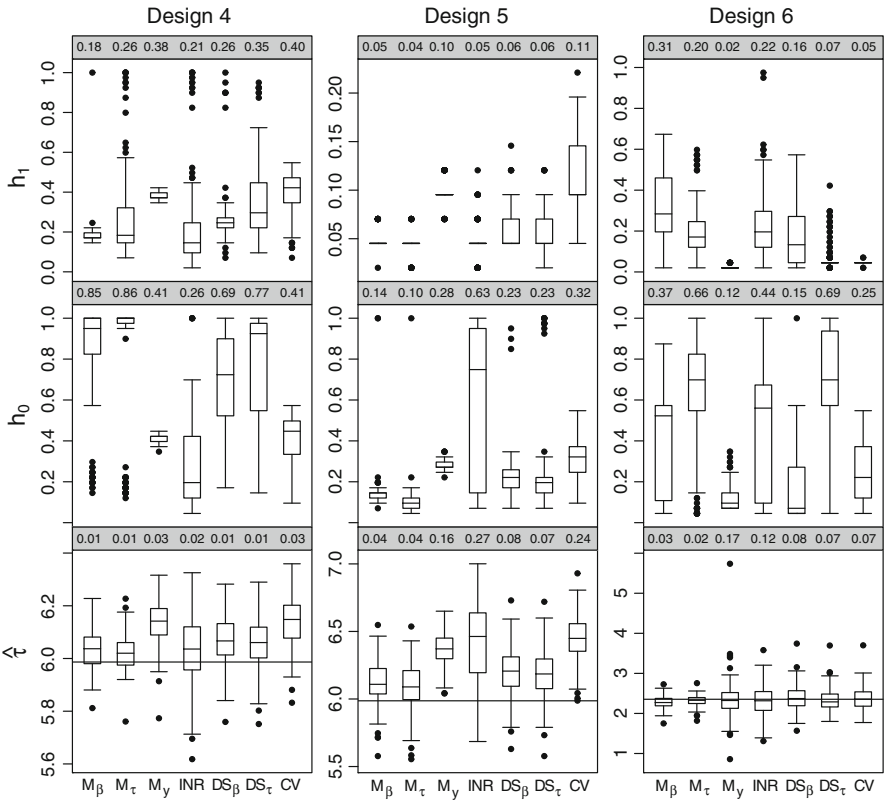

Fig. 3 Design 4–6 columnwise, sample size $n = 500$. Selected bandwidths and resulting $\hat{\tau}$ when using (4), labeled M_β, (5), labeled M_τ, (2), labeled M_y, (14), labeled INR, (16), labeled DS_β, (17), labeled DS_τ, and (3), labeled CV. Average h values are given on *top* of the figures in the *two first rows*, while in the *last row* resulting empirical MSEs of the estimates of τ are displayed on *top* of the boxplots

(A1) The kernel K is a compactly supported, bounded kernel such that $\int u^2 K(u)du \neq 0$. In addition, all odd-order moments of K vanish, that is $\int u^l K(u)du = 0$ for all nonnegative odd integers l.

(A2) The covariate x has density f. The point $\tilde{x}$ is in the interior of $\mathrm{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\}$. At $\tilde{x}$, f is continuously differentiable and all second-order derivatives of β_j, $j = 0, 1$, are continuous.

(A3) For $j = 0, 1$, $h_j \to 0$ and $nh_j \to \infty$ as $n \to \infty$.

We have

$$MSE_{\hat{\tilde{\beta}}_j} = \frac{1}{n^2}\sum_{i=1}^{n} Var(\hat{\beta}_j^{h_j}(x_i)|\mathbf{x}) + \frac{1}{n^2}\sum_{\substack{i=1\\ i\neq l}}^{n}\sum_{l=1}^{n} Cov(\hat{\beta}_j^{h_j}(x_i), \hat{\beta}_j^{h_j}(x_l)|\mathbf{x}) + \left[\frac{1}{n}\sum_{i=1}^{n} E(\hat{\beta}_j^{h_j}(x_i) - \beta_j(x_i)|\mathbf{x})\right]^2.$$

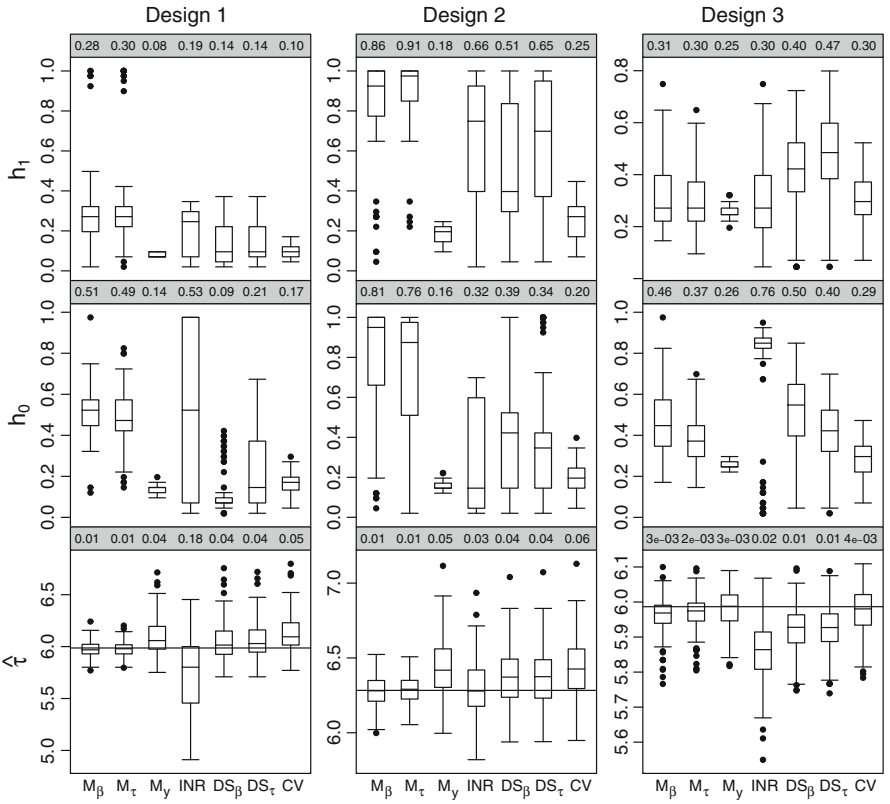

Fig. 4 Design 1–3 columnwise, sample size $n = 1{,}000$. Selected bandwidths and resulting $\hat{\tau}$ when using (4), labeled M_β, (5), labeled M_τ, (2), labeled M_y, (14), labeled INR, (16), labeled DS_β, (17), labeled DS_τ, and (3), labeled CV. Average h values are given on *top* of the figures in the *two first rows*, while in the *last row* resulting empirical MSEs of the estimates of τ are displayed on *top* of the boxplots

Under (A1)–(A2) for $\tilde{x} = x_i$, (Ruppert and Wand (1994), Thm 2.1) states that

$$Var(\hat{\beta}_j^{h_j}(x_i)|\mathbf{x}) = \frac{\sigma_\varepsilon^2}{n_j h_j} f_j(x_i)^{-1} \int K(u)^2 du\{1 + o_p(1)\} \tag{19}$$

and

$$E(\hat{\beta}_j^{h_j}(x_i) - \beta_j(x_i)|\mathbf{x}) = \frac{h_j^2}{2}\beta_j^{(2)}(x_i) \int u^2 K(u) du\{1 + o_p(1)\}, \tag{20}$$

where $f_j(x_i) = f(x_i|z_i = j) = \frac{f(x_i)\Pr(z_i=j|x_i)}{\Pr(z_i=j)}$. It follows from (19) and the fact that $n_j = (-1)^{j+1}\sum_{i=1}^n z_i + n(1-j)$ that

$$\frac{1}{n^2}\sum_{i=1}^n Var(\hat{\beta}_j^{h_j}(x_i)|\mathbf{x}) = \frac{\sigma_\varepsilon^2}{n^2 h_j} \int K(u)^2 du \int \frac{1}{\Pr(z_i = j|x)} dx + o_p(n^{-2}h_j^{-1}). \tag{21}$$

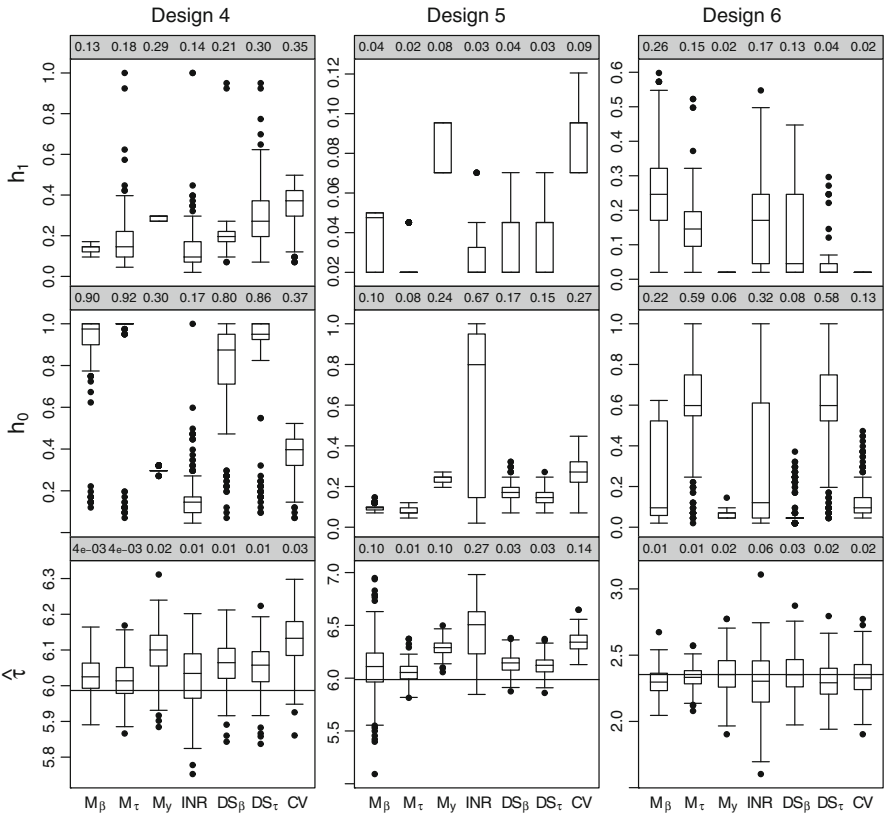

Fig. 5 Design 4–6 columnwise, sample size $n = 1{,}000$. Selected bandwidths and resulting $\hat{\tau}$ when using (4), labeled M_β, (5), labeled M_τ, (2), labeled M_y, (14), labeled INR, (16), labeled DS_β, (17), labeled DS_τ, and (3), labeled CV. Average h values are given on *top* of the figures in the *two first rows*, while in the *last row* resulting empirical MSEs of the estimates of τ are displayed on *top* of the boxplots

Using (20) we have

$$\frac{1}{n}\sum_{i=1}^{n} E(\hat{\beta}_j^{h_j}(x_i) - \beta_j(x_i)|\mathbf{x}) = \frac{h_j^2}{2}\int \beta_j^{(2)}(x) f(x)dx \int u^2 K(u)du + o_p(h_j^2). \tag{22}$$

Now,

$$\begin{aligned}
&\frac{1}{n^2}\sum_{\substack{i=1\\ i\neq l}}^{n}\sum_{l=1}^{n} Cov(\hat{\beta}_j^{h_j}(x_i), \hat{\beta}_j^{h_j}(x_l)|\mathbf{x}) \\
&= \frac{1}{n^2}\sum_{\substack{i=1\\ i\neq l}}^{n}\sum_{l=1}^{n} \sigma_\varepsilon^2 \mathbf{e}_1^T (n_j^{-1}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{X}_i^j)^{-1} n_j^{-2} \\
&\quad \times \mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{W}_l^{h_j}\mathbf{X}_l^j (n_j^{-1}\mathbf{X}_l^{jT}\mathbf{W}_l^{h_j}\mathbf{X}_l^j)^{-1}\mathbf{e}_1.
\end{aligned} \tag{23}$$

According to (Ruppert and Wand (1994), eq. (2.11))

$$(n_j^{-1}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{X}_i^j)^{-1} = \begin{pmatrix} f_j(x_i)^{-1} + o_p(1) & -\frac{f_j^{(1)}(x_i)}{f_j(x_i)^2} + o_p(1) \\ -\frac{f_j^{(1)}(x_i)}{f_j(x_i)^2} + o_p(1) & \{\int u^2 K(u)du f_j(x_i)h_j^2\}^{-1} + o_p(h_j^{-2}) \end{pmatrix}.$$

Noting that

$$\begin{aligned} &\{n_j^{-2}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{W}_l^{h_j}\mathbf{X}_l^j\}_{11} \\ &\quad = \frac{1}{n_j^2 h_j^2}\sum_{k=1}^{n_j} K\left(\frac{x_k - x_i}{h_j}\right)K\left(\frac{x_k - x_l}{h_j}\right), \\ &\{n_j^{-2}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{W}_l^{h_j}\mathbf{X}_l^j\}_{12} \\ &\quad = \frac{1}{n_j^2 h_j^2}\sum_{k=1}^{n_j} K\left(\frac{x_k - x_i}{h_j}\right)K\left(\frac{x_k - x_l}{h_j}\right)(x_k - x_l), \\ &\{n_j^{-2}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{W}_l^{h_j}\mathbf{X}_l^j\}_{21} \\ &\quad = \frac{1}{n_j^2 h_j^2}\sum_{k=1}^{n_j} K\left(\frac{x_k - x_i}{h_j}\right)K\left(\frac{x_k - x_l}{h_j}\right)(x_k - x_i), \end{aligned}$$

and

$$\begin{aligned} &\{n_j^{-2}\mathbf{X}_i^{jT}\mathbf{W}_i^{h_j}\mathbf{W}_l^{h_j}\mathbf{X}_l^j\}_{22} \\ &\quad = \frac{1}{n_j^2 h_j^2}\sum_{k=1}^{n_j} K\left(\frac{x_k - x_i}{h_j}\right)K\left(\frac{x_k - x_l}{h_j}\right)(x_k - x_i)(x_k - x_l). \end{aligned}$$

Starting with (23) we arrive at the following result after some calculus (details can be obtained from the authors)

$$\begin{aligned} &\frac{1}{n^2}\sum_{\substack{i=1 \\ i\neq l}}^{n}\sum_{l=1}^{n} Cov(\hat{\beta}_j^{h_j}(x_i), \hat{\beta}_j^{h_j}(x_l)|\mathbf{x}) \\ &\quad = \frac{\sigma_\varepsilon^2}{n}\left[\int \frac{f(x_k)}{\Pr(z_k = j|x_k)}dx_k\right] + o_p(n^{-1}) \\ &\qquad -\frac{2\sigma_\varepsilon^2 h_j^2}{n}\int u^2 K(u)du \int \left(\frac{f^{(1)}(x)^2}{f(x)\Pr(z = j|x)}\right. \\ &\qquad \left. + \frac{f^{(1)}(x)P^{(1)}(z = j|x)}{\Pr(z = j|x)^2}\right)dx + o_p(n^{-1}h_j^2). \end{aligned} \tag{24}$$

It follows from (21) and (24) that

$$
\begin{aligned}
&Var\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)\Big|\mathbf{x}\right)\\
&\quad=\frac{\sigma_\varepsilon^2}{n^2h_j}\int K(u)^2du\int\frac{1}{\Pr(z=j|x)}dx\\
&\qquad+\frac{\sigma_\varepsilon^2}{n}\left[\int\frac{f(x)}{\Pr(z=j|x)}dx\right]\\
&\qquad-\frac{2\sigma_\varepsilon^2h_j^2}{n}\int u^2K(u)du\int\left(\frac{f^{(1)}(x)^2}{f(x)\Pr(z=j|x)}\right.\\
&\qquad\left.+\frac{f^{(1)}(x)P^{(1)}(z=j|x)}{\Pr(z=j|x)^2}\right)dx+o_p(n^{-2}h_j^{-1}+n^{-1}+n^{-1}h_j^2).
\end{aligned}
\tag{25}
$$

Hence, from (25) and (22),

$$
\begin{aligned}
&MSE\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)\Big|\mathbf{x}\right)\\
&\quad=\frac{\sigma_\varepsilon^2}{n^2h_j}\int K(u)^2du\int\frac{1}{\Pr(z=j|x)}dx+\frac{\sigma_\varepsilon^2}{n}\left[\int\frac{f(x)}{\Pr(z=j|x)}dx\right]\\
&\qquad-\frac{2\sigma_\varepsilon^2h_j^2}{n}\int u^2K(u)du\\
&\qquad\times\int\left(\frac{f^{(1)}(x)^2}{f(x)\Pr(z=j|x)}+\frac{f^{(1)}(x)P^{(1)}(z=j|x)}{\Pr(z=j|x)^2}\right)dx\\
&\qquad+\frac{h_j^4}{4}\left[\int\beta_j^{(2)}(x)f(x)dx\right]^2\left[\int u^2K(u)du\right]^2\\
&\qquad+o_p(n^{-2}h_j^{-1}+n^{-1}+n^{-1}h_j^2+h_j^4).
\end{aligned}
$$

Finally, (11)–(13) follows from (6) and (7). By the weak law of large numbers we can write

$$
\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i)-E(\beta_j(x_i))=o_p(1).
$$

Combined with (6) we thus have

$$
E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)\Big|\mathbf{x}\right)-E(\beta_j(x_i))=B_1(j)h_j^2+o_p(h_j^2).
$$

For $h_j \propto n^r$ we have thus

$$\begin{aligned}\sqrt{n}\ E\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)\Big|\mathbf{x}\right) - \sqrt{n}E(\beta_j(x_i)) \\ = n^{1/2}B_1(1)n^{2r} + o_p(n^{1/2}n^{2r}) \\ = O(n^{1/2+2r}) + o_p(n^{1/2}n^{2r}).\end{aligned}$$

Furthermore from (7) and for $h_j \propto n^r$ we can write

$$\begin{aligned}&nVar\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j^{h_j}(x_i)\Big|\mathbf{x}\right) \\ &= V_1(j) + \frac{1}{n^{1+r}}V_2(j) + n^{2r}V_3(j) + o_p(1+n^{-1-r}+n^2) \\ &= V_1(j) + O(n^{-1-r}) + O(n^{2r}) + o_p(1+n^{-1-r}+n^2).\end{aligned}$$

References

Andrews DWK (1991) Asymptotic optimality of generalized cl, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. J Econom 47:359–377

Cheng PE (1994) Nonparametric estimation of mean functionals with data missing at random. J Am Stat Assoc 89:81–87

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74:829–836

Dawid AP (1979) Conditional independence in statistical theory. J R Stat Soc Ser B Stat Methodol 41:1–31

de Luna X, Lundin M (2014) Sensitivity analysis of the unconfoundedness assumption with an application to an evaluation of college choice effects on earnings. J App Stat 41:1–18

de Luna X, Waernbaum I, Richardson TS (2011) Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika 98:861–875

Fan J (1992) Design-adaptive nonparametric regression. J Am Stat Assoc 87:998–1004

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London

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

Frölich M (2007) Nonparametric IV estimation of local average treatment effects with covariates. J Econom 139:35–75

Häggström J (2013) Bandwidth selection for backfitting estimation of semiparametric additive models: a simulation study. Comput Stat Data Anal 62:136–148

Hansen B (2008) The prognostic analogue of the propensity score. Biometrika 95:481–488

Härdle W, Hall P, Marron J (1992) Regression smoothing parameters that are not far from their optimum. J Am Stat Assoc 87:227–233

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29

Imbens GW, Newey W, Ridder G (2005) Mean-squared-error calculations for average treatment effects. IEPR Working Papers 05.34, Institute of Economic Policy Research (IEPR). http://dornsife.usc.edu/IEPR/Working%20Papers/IEPR_05.34_%5bImbens.Newey.Ridder%5d.pdf

Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J. Econ. Lit. 47:5–86

Neyman J (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (1990), translated (with discussion). Stat Sci 5:465–480

Opsomer JD, Sheather S, Wand M (1995) An effective bandwidth selector for local least squares regression. J Am Stat Assoc 90:1257–1270

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Ruppert D, Wand M (1994) Multivariate locally weighted least squares regression. Ann Stat 22:1346–1370
Speckman P (1988) Kernel smoothing in partial linear models. J R Stat Soc Ser B Stat Methodol 50:413–436
Waernbaum I (2010) Propensity score model specification for estimation of average treatment effects. J Stat Plan Inference 140:1948–1956
Wilson JR (1984) Variance reduction techniques for digital simulation. Am J Math Manag Sci 4:277–312