Bagging cross-validated bandwidth selection in nonparametric regression estimation with applications to large-sized samples

Daniel Barreiro-Ures

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Ricardo Cao

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Mario Francisco-Fernández †

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Abstract. Cross-validation is a well-known and widely used bandwidth selection method in nonparametric regression estimation. However, this technique has two remarkable drawbacks: (i) the large variability of the selected bandwidths, and (ii) the inability to provide results in a reasonable time for very large sample sizes. To overcome these problems, bagging cross-validation bandwidths are analyzed in this paper. This approach consists in computing the cross-validation bandwidths for a finite number of subsamples and then rescaling the averaged smoothing parameters to the original sample size. Under a random-design regression model, asymptotic expressions up to a second-order for the bias and variance of the leave-one-out cross-validation bandwidth for the Nadaraya–Watson estimator are obtained. Subsequently, the asymptotic bias and variance and the limit distribution for the bagged cross-validation selector are derived. Suitable choices of the number of subsamples and the subsample size lead to an $n^{-1/2}$ rate for the convergence in distribution of the bagging cross-validation selector, outperforming the rate $n^{-3/10}$ of leave-one-out cross-validation. Several simulations and an illustration on a real dataset related to the COVID-19 pandemic show the behavior of our proposal and its better performance, in terms of statistical efficiency and computing time, when compared to leave-one-out cross-validation.

Keywords: bagging, cross-validation, Nadaraya–Watson, regression, subsampling

1. Introduction

The study of a variable of interest depending on other variable(s) is a common problem that appears in many disciplines. To deal with this issue, an appropriate regression model setting up the possible functional relationship between the variables is usually formulated. As part of this analysis, the unknown regression function, describing the general relationship between the variable of interest and the explanatory variable(s), has to be estimated. This task can be carried out using nonparametric methods that do not

†*Address for correspondence:* Mario Francisco-Fernández, University of A Coruña, Faculty of Computer Science, Campus de Elviña, s/n, A Coruña, Spain. E-mail: mariofr@udc.es

assume any parametric form for the regression function, providing flexible procedures and avoiding misspecification problems. Among the available nonparametric approaches, kernel-type regression estimators (Wand and Jones, 1995) are perhaps the most popular. To compute this type of estimators the user has to select a kernel function (typically a density function) and a bandwidth or smoothing parameter that regulates the amount of smoothing to be used, which in turn determines the trade-off between the bias and the variance of the estimator. Although the choice of the kernel function is of secondary importance, the smoothing parameter plays a crucial role. In this regard, numerous contributions have been made over the last decades, providing methods to select the bandwidth. These approaches include, among others, cross-validation methods (Härdle et al., 1988) and plug-in selectors (Ruppert et al., 1995). In Köhler et al. (2014), a complete review and an extensive simulation study of different data-driven bandwidth selectors for kernel regression are presented. Due to their wide applicability and the good performance obtained in this complete comparison, in the present paper, we focus on analyzing cross-validation bandwidth selection techniques.

Cross-validation is a popular method of model selection that precedes an early discussion of the method by Stone (1974). In its simplest form, cross-validation consists of splitting the dataset under study into two parts, using one part to fit one or more models, and then predicting the data in the second part with the models so-built. In this way, by not using the same data to fit and validate the models, it is possible to objectively compare the predictive capacity of different models. The leave-one-out version of cross-validation (of interest in the present paper) is somewhat more involved. It excludes one datum from the dataset, fits a model from the remaining observations, uses this model to predict the datum left out, and then repeats this process for all the data.

The present paper studies the leave-one-out cross-validation bandwidth selection method and the application of bagging (Breiman, 1996) to this procedure. We derive some asymptotic properties of the corresponding selectors when considering a random-design regression model and the Nadaraya–Watson kernel-type estimator is used. The Nadaraya–Watson estimator can be seen as a particular case of a wider class of nonparametric estimators, the so-called local polynomial estimators (Stone, 1977; Cleveland, 1979; Fan, 1992), when performing a local constant fit. Given a random sample of size n, bagging cross-validation consists of selecting N subsamples of size $r < n$, each without replacement, from the n observations. One then computes a cross-validation bandwidth from each of the N subsets, averages them, and then scales the average down appropriately to account for the fact that $r < n$. It is well-known that the use of bagging can lead to substantial reductions in the variability of an estimator that is nonlinear in the observations (see Friedman and Hall, 2007), as occurs in the case of the cross-validation criterion function. The use of bagging in conjunction with cross-validation for bandwidth selection has already been studied in the case of kernel density estimation by several authors (see, for example Barreiro-Ures et al., 2020; Hall and Robinson, 2009). In addition to the potential improvement in statistical precision, even in the case of small sample sizes, the use of bagging (with appropriate elections of r and N) can drastically reduce computation times, especially for very large sample sizes. Note that the complexity of cross-validation is $O(n^2)$, while the complexity of bagging cross-validation is $O(Nr^2)$. Larger reductions in computation time can also be additionally achieved with the application of binning techniques in the bagging procedure.

Apart from the theoretical analysis of the cross-validation bandwidth selection methods, another goal of this study is to apply the techniques studied in the present paper to a dataset related to the current COVID-19 pandemic. In particular, using a moderately large sample, provided by the Spanish Center for Coordinating Sanitary Alerts and Emergencies, consisting of the age and the time in hospital of people infected with COVID-19 in Spain, we are interested in studying the relationship between those two variables by means of the Nadaraya–Watson estimator. Apart from its purely epidemiological interest and due to the considerable size of the sample, this dataset is also useful to put into practice the techniques analyzed in the present paper.

The remainder of the paper is as follows. In Section 2, the regression model considered, the Nadaraya–Watson regression estimator and the important problem of bandwidth selection are presented. In Section 3, the leave-one-out cross-validation bandwidth selection method is described and some asymptotic properties of the corresponding selector are provided when the Nadaraya–Watson estimator is used. Section 4 considers the use of bagging for cross-validation in bandwidth selection for the Nadaraya–Watson estimator. Asymptotic expressions for the bias and the variance of the proposed bandwidth selector, as well as for its limit distribution, are also derived in this section. In Section 5, an algorithm is proposed to automatically choose the subsample size for the bandwidth selector studied in Section 4. The techniques proposed are empirically tested through several simulation studies in Section 6 and applied to the previously mentioned COVID-19 dataset in Section 7. Finally, concluding remarks are given in Section 8. The detailed proofs and some additional plots concerning the simulation study are included in the accompanying supplementary materials.

2. Regression model and Nadaraya–Watson estimator

Let $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an independent and identically distributed (i.i.d.) sample of size n of the two-dimensional random variable (X, Y), drawn from the nonparametric regression model:

$$Y = m(X) + \varepsilon, \tag{1}$$

where $m(x) = \mathrm{E}(Y \mid X = x)$ denotes the regression function, and ε is the error term, satisfying that $\mathrm{E}(\varepsilon \mid X = x) = 0$ and $\mathrm{E}\left(\varepsilon^2 \mid X = x\right) = \sigma^2(x)$.

The Nadaraya–Watson estimator or local constant estimator (Nadaraya, 1964; Watson, 1964) offers a nonparametric way to estimate the unknown regression function, m. It is given by:

$$\hat{m}_h(x) = \frac{\sum\limits_{i=1}^{n} K_h\left(x - X_i\right) Y_i}{\sum\limits_{i=1}^{n} K_h\left(x - X_i\right)}, \tag{2}$$

where $h > 0$ denotes the bandwidth or smoothing parameter and K the kernel function. As pointed out in the introduction, the value of the bandwidth is of great importance since it determines the amount of smoothing performed by the estimator and, therefore, heavily influences its behavior. Thus, in practice, data-driven bandwidth selection methods are needed.

Optimal bandwidths often refer to smoothing parameter values that mimize some error criterion function. These functions are typically expected loss, in some sense. When the aim is predicting the response variable, Y, given the value of the explanatory variable, X, it is natural to consider expectations conditionally on the observed explanatory sample, $(X_1, \ldots, X_n)$. However, the focus of this paper is estimating the regression function on its own. Thus an unconditional expected loss view is adopted. Of course, there exist arguments in favor of both type of criteria. More details on this issue can be found in Köhler et al. (2014).

When adopting an unconditional view, a possible way to select a (global) optimal bandwidth for (2) consists in minimizing, for instance, the mean integrated squared error (MISE), a (global) optimality criterion defined as:

$$M_n(h) = \mathrm{E}\left[\int \left\{\hat{m}_h(x) - m(x)\right\}^2 f(x)\, dx\right], \tag{3}$$

where f denotes the marginal density function of X. The bandwidth that minimizes (3) is called the MISE bandwidth and it will be denoted by h_{n0}, that is,

$$h_{n0} = \arg\min_{h>0} M_n(h). \tag{4}$$

The MISE bandwidth depends on m and f and, since in practice both functions are often unknown, h_{n0} cannot be directly calculated. However, it can be estimated, for example, using the cross-validation method.

In the following section, we present the leave-one-out cross-validation bandwidth selection criterion and provide the asymptotic properties of the corresponding selector when using the estimator (2) and considering the regression model (1).

3. Cross-validation bandwidth

Cross-validation is a method that offers a criterion for optimality which works as an empirical analogue of the MISE and so it allows us to estimate h_{n0}. The cross-validation function is defined as:

$$CV_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left\{\hat{m}_h^{(-i)}(X_i) - Y_i\right\}^2, \tag{5}$$

where $\hat{m}_h^{(-i)}$ denotes the Nadaraya–Watson estimator constructed using $\mathcal{X} \setminus \{(X_i, Y_i)\}$, that is, leaving out the i-th observation,

$$\hat{m}_h^{(-i)}(x) = \frac{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} K_h\left(x - X_j\right) Y_j}{\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} K_h\left(x - X_j\right)}. \tag{6}$$

Hence, the cross-validation bandwidth, $\hat{h}_{CV,n}$, can be defined as

$$\hat{h}_{CV,n} = \arg\min_{h>0} CV_n(h). \tag{7}$$

It is well-known that under suitable regularity conditions, up to first order,

$$M_n(h) = B_1 h^4 + V_1 n^{-1} h^{-1} + O\left(h^6 + n^{-1} h\right),$$

where

$$\begin{aligned} B_1 &= \frac{1}{4}\mu_2(K)^2 \int \left\{ m''(x) + 2\frac{m'(x)f'(x)}{f(x)} \right\}^2 f(x)\,dx, \\ V_1 &= R(K)\int \sigma^2(x)\,dx, \end{aligned}$$

with $R(g) = \int g^2(x)\,dx$ and $\mu_j(g) = \int x^j g(x)\,dx$, $j = 0, 1, \ldots$, provided that these integrals, as well as B_1 and V_1, exist finite. Then, the first-order term of the MISE bandwidth, h_n, has the expression $h_n = C_0 n^{-1/5}$, where

$$C_0 = \left(\frac{V_1}{4B_1}\right)^{1/5}.$$

In order to obtain the asymptotic properties of (7) as an estimator of (4), it is necessary to study certain moments of (5) and its derivatives. However, the fact that the Nadaraya–Watson estimator has a random denominator makes this a very difficult task. To overcome this problem, it will be useful to work with an approximation of $\hat{m}_h(x)$. For this, note that the Nadaraya–Watson estimator can be written as

$$\hat{m}_h(x) = A + B + C + D + E + F, \tag{8}$$

where

$$\begin{aligned} A &= \frac{\hat{a}}{e}, \\ B &= \frac{a(e-\hat{e})}{e^2}, \\ C &= \frac{(\hat{a}-a)(e-\hat{e})}{e^2}, \\ D &= \frac{a}{e}\frac{(e-\hat{e})^2}{e^2}, \\ E &= \frac{\hat{a}-a}{e}\frac{(e-\hat{e})^2}{e^2}, \\ F &= \frac{\hat{a}}{\hat{e}}\frac{(e-\hat{e})^3}{e^3}, \end{aligned}$$

with

$$\begin{aligned}
a &= m(x)f(x),\\
e &= f(x),\\
\hat{a} &= \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right)Y_i,\\
\hat{e} &= \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right).
\end{aligned}$$

Expression (8) splits $\hat{m}_h(x)$ as a sum of five ratios with no random denominator plus an additional term, F, which has a random denominator. However, both E and F are negligible with respect to the other terms. Thus, one may consider the modified version of the Nadaraya–Watson estimator given by $\tilde{m}_h(x) = A + B + C + D$, that is:

$$\tilde{m}_h(x) = m(x) + \frac{1}{n^2 f(x)^2}\sum_{j=1}^{n}\sum_{k=1}^{n} K_h\left(x - X_j\right)\left\{Y_j - m(x)\right\}\left\{2f(x) - K_h\left(x - X_k\right)\right\}, \quad (9)$$

which can be seen as a quadratic approximation of $\hat{m}_h(x)$, where the terms E and F are omitted due to their “cubic negligibility”. In practice, (9) is unobservable and, therefore, it does not define an estimator but a theoretical approximation of (2). This decomposition of $\hat{m}_h(x)$ is in turn inspired by a similar approach proposed in Barbeito (2020). There, a linear approximation of the Nadaraya–Watson estimator was considered and so only the terms A and B were taken into account, leading to the simpler expression

$$\bar{m}_h(x) = m(x) + \frac{1}{nf(x)}\sum_{i=1}^{n} K_h\left(x - X_i\right)\left\{Y_i - m(x)\right\}. \quad (10)$$

Following this approach, (9) could be used to define a theoretical approximation of the MISE function defined in (3), namely

$$\tilde{M}_n(h) = \int \left[\mathrm{E}\left\{\tilde{m}_h(x)\right\} - m(x)\right]^2 f(x)\,dx + \int \operatorname{var}\left\{\tilde{m}_h(x)\right\} f(x)\,dx.$$

The bandwidth that minimizes $\tilde{M}_n(h)$ will be denoted by $\tilde{h}_{n0}$. On the other hand, (9) can also be used to define a modified version of the cross-validation criterion,

$$\widetilde{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{m}_h^{(-i)}(X_i) - Y_i\right\}^2, \quad (11)$$

where $\tilde{m}_h^{(-i)}$ denotes the leave-one-out version of (9) without the i-th observation, that is,

$$\begin{aligned}
\tilde{m}_h^{(-i)}(x) &= m(x) + \frac{1}{(n-1)^2 f(x)^2}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i}}^{n} K_h\left(x - X_j\right)\left\{Y_j - m(x)\right\}\\
&\quad \left\{2f(x) - K_h\left(x - X_k\right)\right\}. \quad (12)
\end{aligned}$$

The bandwidth that minimizes (11) will be denoted by $\tilde{h}_{CV,n}$. Using Taylor expansions, the following approximation can be obtained:

$$\begin{aligned}\tilde{h}_{CV,n} - \tilde{h}_{n0} &\approx -\frac{\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})}{\tilde{M}''_n(\tilde{h}_{n0})} \\ &+ \frac{\left\{\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})\right\}\left\{\widetilde{CV}''_n(\tilde{h}_{n0}) - \tilde{M}''_n(\tilde{h}_{n0})\right\}}{\tilde{M}''_n(\tilde{h}_{n0})^2}, \qquad (13)\end{aligned}$$

where the second term of (13) is negligible with respect to the first one and is assumed not to contribute to the bias and the variance of $\tilde{h}_{CV,n}$. Since the first-order terms of $\mathrm{E}\left\{\widetilde{CV}^{k)}_n(h)\right\}$ and $\tilde{M}^{k)}_n(h)$ coincide for every $k \geq 1$, we need to calculate the second-order terms of both $\mathrm{E}\left\{\widetilde{CV}'_n(\tilde{h}_{n0})\right\}$ and $\tilde{M}'_n(\tilde{h}_{n0})$ in order to analyze the bias of the modified cross-validation bandwidth. As for the variance of the modified cross-validation bandwidth, calculating the first-order term of $\operatorname{var}\left\{\widetilde{CV}'_n(\tilde{h}_{n0})\right\}$ will be enough, and so it will be useful to work with the simpler, linear approximation of $\hat{m}_h(x)$ given by (10).

3.1. Asymptotic results

The asymptotic bias and variance of the cross-validation bandwidth minimizing (11) are derived in this section. For this, some previous lemmas are proved. The detailed proof of these results can be found in the supplementary material. The following assumptions are needed:

A1. K is a symmetric and differentiable kernel function.

A2. For every $j = 0, \ldots, 6$, the integrals $\mu_j(K)$, $\mu_j(K')$ and $\mu_j(K^2)$ exist and are finite.

A3. The functions m and f are eight times differentiable.

A4. The function σ^2 is four times differentiable.

Lemma 3.1 provides expressions for the first and second order terms of both the bias and the variance of (9).

LEMMA 3.1. *Under assumptions* A1–A4*, the bias and the variance of the modified version of the Nadaraya–Watson estimator defined in* (9) *satisfy:*

$$\begin{aligned}\mathrm{E}\left\{\tilde{m}_h(x)\right\} - m(x) &= \mu_2(K)\left\{\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}h^2 \\ &+ \left[\mu_4(K)\left\{\frac{1}{24}m^{4)}(x) + \frac{1}{6}\frac{m'''(x)f'(x)}{f(x)} + \frac{1}{4}\frac{m''(x)f''(x)}{f(x)}\right.\right. \\ &+ \left.\left.\frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right\} - \mu_2(K)^2\frac{f''(x)}{f(x)}\left\{\frac{1}{4}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}\right]h^4 \\ &+ O\left(h^6 + n^{-1}\right)\end{aligned}$$

and

$$\begin{aligned}
\operatorname{var}\{\tilde{m}_h(x)\} &= R(K)\sigma^2(x)f(x)^{-1}n^{-1}h^{-1}\\
&+ \Big[\mu_2(K^2)f(x)^{-2}\left\{\varphi_3(x)+\frac{1}{2}m(x)^2f''(x)-2\varphi_1(x)m(x)f(x)\right\}\\
&- R(K)\mu_2(K)\sigma^2(x)f(x)^{-2}f''(x)\Big]\,n^{-1}h\\
&+ O(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}).
\end{aligned}$$

It follows from Lemma 3.1 that

$$\tilde{M}_n(h)=B_1h^4+V_1n^{-1}h^{-1}+B_2h^6+V_2n^{-1}h+O\left(h^8+n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}\right),$$

where

$$\begin{aligned}
B_2 &= 2\mu_2(K)\int\left\{\frac{1}{2}m''(x)+\frac{m'(x)f'(x)}{f(x)}\right\}\Big[\mu_4(K)\left\{\frac{1}{24}m^{4)}(x)+\frac{1}{6}\frac{m'''(x)f'(x)}{f(x)}\right.\\
&+ \left.\frac{1}{4}\frac{m''(x)f''(x)}{f(x)}+\frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right\}-\mu_2(K)^2\frac{f''(x)}{f(x)}\left\{\frac{1}{4}m''(x)+\frac{m'(x)f'(x)}{f(x)}\right\}\Big]\,f(x)\,dx,\\
V_2 &= \int\Big[\mu_2(K^2)f(x)^{-2}\left\{\frac{1}{2}f''(x)\sigma^2(x)+m'(x)^2f(x)+\frac{1}{2}\sigma^{2''}(x)f(x)+f'(x)\sigma^{2'}(x)\right\}\\
&- R(K)\mu_2(K)\sigma^2(x)f(x)^{-2}f''(x)\Big]\,f(x)\,dx.
\end{aligned}$$

are assumed to exist finite.

Lemma 3.2 provides expressions for the first and second order terms of both the expectation and variance of $\widetilde{CV}'_n(h)$.

Lemma 3.2. *Let us define*

$$\begin{aligned}
A_1 &= 12\mu_2(K)\mu_4(K)\int f(x)^{-1}\left\{\frac{1}{24}m^{(4)}(x)f(x)+\frac{1}{6}m'''(x)f'(x)+\frac{1}{4}m''(x)f''(x)\right.\\
&+ \left.\frac{1}{6}m'(x)f'''(x)\right\}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right\}\,dx\\
&- 6\mu_2(K)^3\int f''(x)f(x)^{-2}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right\}^2,\\
A_2 &= \mu_2\left(K^2\right)\int f(x)^{-1}\left[\frac{1}{2}f''(x)\sigma^2(x)+f'(x)(\sigma^2)'(x)\right.\\
&+ \left.f(x)\left\{\frac{1}{2}(\sigma^2)''(x)+m'(x)^2\right\}\right]\,dx\\
&- R(K)\mu_2(K)\int\sigma^2(x)f''(x)f(x)^{-1}\,dx,\\
R_1 &= 32R(K)^2\mu_2(K)^2\int\sigma^2(x)f(x)^{-1}\left\{\frac{1}{4}m''(x)^2f(x)^2+m'(x)m''(x)f(x)f'(x)\right.\\
&+ \left.m'(x)^2f'(x)^2\right\}\,dx,\\
R_2 &= 4\mu_2\left\{(K')^2\right\}\int\sigma^2(x)^2\,dx.
\end{aligned}$$

Then, under assumptions A1–A4, *and assuming that* B_1, V_1, A_1, A_2, R_1 *and* R_2 *exist finite:*

$$\begin{aligned} \mathrm{E}\left\{\widetilde{CV}'_n(h)\right\} &= 4B_1h^3 - V_1n^{-1}h^{-2} + A_1h^5 + A_2n^{-1} + O\left(h^7 + n^{-1}h^2\right), \\ \mathrm{var}\left\{\widetilde{CV}'_n(h)\right\} &= R_1n^{-1}h^2 + R_2n^{-2}h^{-3} + O\left(n^{-1}h^4 + n^{-2}h^{-1}\right). \end{aligned}$$

Finally, Theorem 3.1, which can be derived from (13), Lemma 3.1 and Lemma 3.2, provides the asymptotic bias and variance of the cross-validation bandwidth that minimizes (11).

Theorem 3.1. *Under the assumptions of Lemma 3.2 and assuming that* B_2 *and* V_2 *exist finite, the asymptotic bias and variance of the bandwidth that minimizes* (11) *are:*

$$\begin{aligned} \mathrm{E}\left(\tilde{h}_{CV,n}\right) - \tilde{h}_{n0} &= \mathcal{B}n^{-3/5} + o\left(n^{-3/5}\right), \\ \mathrm{var}\left(\tilde{h}_{CV,n}\right) &= \mathcal{V}n^{-3/5} + o\left(n^{-3/5}\right), \end{aligned}$$

where

$$\begin{aligned} \mathcal{B} &= \frac{6B_2C_0^5 + V_2 - A_1C_0^5 - A_2}{12B_1C_0^2 + 2V_1C_0^{-3}}, \\ \mathcal{V} &= \frac{R_1C_0^2 + R_2C_0^{-3}}{\left(12B_1C_0^2 + 2V_1C_0^{-3}\right)^2}. \end{aligned}$$

Corollary 3.1. *Under the assumptions of Theorem 3.1, the asymptotic distribution of the bandwidth that minimizes* (11) *is given by:*

$$n^{3/10}\left(\tilde{h}_{CV,n} - \tilde{h}_{n0}\right) \xrightarrow{d} \mathrm{N}(0, \mathcal{V}),$$

where the constant $\mathcal{V}$ *was defined in Theorem 3.1.*

Remark 3.1. *Although the results presented so far involve only the modified cross-validation bandwidth, defined as the bandwidth that minimizes* (11), *it seems reasonable to think that these asymptotic results also apply to the standard cross-validation bandwidth defined in* (7), *this being the rationale behind the decomposition of the Nadaraya–Watson estimator proposed in* (8). *Under suitable assumptions, it can be proved that, as the sample size increases,*

$$\tilde{h}_{CV,n} - \tilde{h}_{n0} = \hat{h}_{CV,n} - h_{n0} + O_p\left(n^{-2/5}\right). \tag{14}$$

Moreover, since $\tilde{h}_{n0} - h_{n0} = O\left(n^{-4/5}\right)$, *it follows that*

$$\tilde{h}_{CV,n} = \hat{h}_{CV,n} + O_p\left(n^{-2/5}\right).$$

A sketch of the proof of (14) *and some other related results are included in the supplementary material.*

4. Bagged cross-validation bandwidth

While the cross-validation method is very useful to select reliable bandwidths in nonparametric regression, it also has the handicap of requiring a high computing time if the sample size is very large. This problem can be partially circumvented by using bagging (Breiman, 1996), a statistical technique belonging to the family of ensemble methods (Opitz and Maclin, 1999), in the bandwidth selection procedure. In this section, we explain how bagging may be applied in the cross-validation context. Additionally, the asymptotic properties of the corresponding selector are obtained. Apart from the obvious reductions in computing time, the bagging cross-validation selector also presents better theoretical properties than the leave-one-out cross-validation bandwidth. This will be corroborated in the numerical studies presented in Sections 6 and 7.

Let $\mathcal{X}^* = \{(X_1^*, Y_1^*), \ldots, (X_r^*, Y_r^*)\}$ be a random sample of size $r < n$ drawn without replacement from the i.i.d sample $\mathcal{X}$ defined in Section 2. This subsample is used to calculate a cross-validation bandwidth, $\hat{h}_{CV,r}$. A rescaled version of $\hat{h}_{CV,r}$, given by $(r/n)^{1/5}\hat{h}_{CV,r}$, can be viewed as a feasible estimator of the optimal MISE bandwidth, h_{n0}, for $\hat{m}_h$. Bagging consists of repeating this resampling procedure independently N times, leading to N rescaled bandwidths, $(r/n)^{1/5}\hat{h}_{CV,r,1}, \ldots, (r/n)^{1/5}\hat{h}_{CV,r,N}$. The bagging bandwidth is then defined as:

$$\hat{h}(r, N) = \frac{1}{N}\left(\frac{r}{n}\right)^{1/5}\sum_{i=1}^{N}\hat{h}_{CV,r,i}. \tag{15}$$

In the case of kernel density estimation, both the asymptotic properties and the empirical behavior of this type of bandwidth selector have been studied in Hall and Robinson (2009) for $N = \infty$ and generalized in Barreiro-Ures et al. (2020), where the asymptotic properties of the bandwidth selector are derived for the more practical case of a finite N. Furthermore, as discussed there, an alternative approach is to apply bagging to the cross-validation curves, wherein one averages the cross-validation curves from N independent resamples of size r, finds the minimizer of the average curve, and then rescales the minimizer as before. The asymptotic properties of the two approaches are equivalent, but we prefer bagging the bandwidths since doing so does not require as much communication between resamples and allows for parallel computing.

Following the same ideas employed in the previous section, a modified version of (15) can be defined. This modified bagged bandwidth uses modified cross-validation bandwidths $\tilde{h}_{CV,r,i}$ instead of $\hat{h}_{CV,r,i}$, for $i = 1, \ldots, N$, and it is given by

$$\tilde{h}(r, N) = \frac{1}{N}\left(\frac{r}{n}\right)^{1/5}\sum_{i=1}^{N}\tilde{h}_{CV,r,i}. \tag{16}$$

In the next section, the asymptotic bias and variance of the bagging bandwidth (16) when using the Nadaraya–Watson estimator (2) and the regression model (1) are obtained. Moreover, its asymptotic distribution is also derived. From these results and considering Remark 3.1, similar results for (15) could be obtained.

4.1. Asymptotic results

Expressions for the asymptotic bias and the variance of (16) are given in Theorem 4.1. The following additional assumption is needed:

A5. As $r, n \to \infty$, $r = o(n)$ and N tends to a positive constant or ∞.

THEOREM 4.1. *Under assumptions* A1–A5*, the asymptotic bias and the variance of the bagged cross-validation bandwidth $\tilde{h}(r, N)$ are:*

$$\begin{aligned} \mathrm{E}\left\{\tilde{h}(r,N)\right\} - \tilde{h}_{n0} &= (\mathcal{B} + C_1) r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right), \\ \mathrm{var}\left\{\tilde{h}(r,N)\right\} &= V r^{-1/5} n^{-2/5} \left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\} + o\left(\frac{r^{-1/5} n^{-2/5}}{N} + r^{9/5} n^{-12/5}\right), \end{aligned}$$

where the constants $\mathcal{B}$ and V were defined in Theorem 3.1 and the constant C_1 is defined in expression (48) *in the supplementary material.*

COROLLARY 4.1. *Under the assumptions of Theorem 4.1, the asymptotic distribution of the bagged cross-validation bandwidth $\tilde{h}(r, N)$ is:*

$$\frac{r^{1/10} n^{1/5}}{\sqrt{\frac{1}{N} + \left(\frac{r}{n}\right)^2}} \left\{\tilde{h}(r,N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0, V),$$

where the constant V was defined in Theorem 3.1. In particular, assuming that $r = o\left(n/\sqrt{N}\right)$, then,

$$r^{1/10} n^{1/5} \sqrt{N} \left\{\tilde{h}(r,N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0, V).$$

Using (14) in Remark 3.1, it could be proved that similar results to those in Corollary 4.1 hold when considering $\hat{h}(r, N) - h_{n0}$ instead of $\tilde{h}(r, N) - \tilde{h}_{n0}$. It should be noted that, while $\hat{h}_{CV,n} - h_{n0}$ converges in distribution at the rate $n^{-3/10}$, this result can be improved with the use of bagging and letting r and N tend to infinity at adequate rates. For example, if both r and N tended to infinity at the rate $\sqrt{n}$, then $\hat{h}(r, N) - h_{n0}$ would converge in distribution at the rate $n^{-1/2}$, which is indeed a faster rate of convergence than $n^{-3/10}$.

5. Choosing an optimal subsample size

In practice, an important step of our approach is, for fixed values of n and N, choosing the *optimal* subsample size, r_0. A possible optimality criterion, considering the modified bandwidths, could be to select the value of r that minimizes the main term of the variance of $\tilde{h}(r, N)$. In this case, we would get:

$$r_0^{(1)} = \frac{n}{3\sqrt{N}}$$

and the variance of the bagging bandwidth would converge to zero at the rate

$$\mathrm{var}\left\{\tilde{h}\left(r_0^{(1)}, N\right)\right\} \sim n^{-3/5} N^{-9/10},$$

which is a faster rate of convergence than that of the standard (modified) cross-validation bandwidth. In particular,

$$\frac{\operatorname{var}\left\{\tilde{h}\left(r_0^{(1)}, N\right)\right\}}{\operatorname{var}\left(\tilde{h}_{CV,n}\right)} \sim N^{-9/10}.$$

The obvious drawback of this criterion is that it would not allow any improvement in terms of computational efficiency, since the complexity of the algorithm would be the same as in the case of standard cross-validation, $O(n^2)$. This makes this choice of r_0 inappropriate for very large sample sizes. Another possible criterion for selecting r_0 would be to minimize, as a function of r, the asymptotic mean squared error (AMSE) of $\tilde{h}(r, N)$, given by:

$$\operatorname{AMSE}\left\{\tilde{h}(r, N)\right\} = (\mathcal{B} + C_1)^2 r^{-4/5} n^{-2/5} + V r^{-1/5} n^{-2/5} \left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\}. \tag{17}$$

Since $\mathcal{B}$, C_1 and V are unknown, we propose the following method to estimate

$$r_0 = \underset{r>1}{\arg\min}\, \operatorname{AMSE}\left\{\tilde{h}(r, N)\right\}.$$

Step 1. Consider s subsamples of size $p < n$, drawn without replacement from the original sample of size n.

Step 2. For each of these subsamples, obtain an estimate, $\hat{f}$, of the marginal density function of the explanatory variable (using kernel density estimation, for example) and an estimate, $\hat{m}$, of the regression function (for instance, by fitting a polynomial of a certain degree). Do the same for the required derivatives of both f and m.

Step 3. Use the estimates obtained in the previous step to compute the constants $\mathcal{B}^{[i]}$, $C_1^{[i]}$ and $V^{[i]}$ for each subsample, where i $(i = 1, \ldots, s)$ denotes the subsample index.

Step 4. Compute the bagged estimates of the unknown constants, that is,

$$\begin{aligned} \hat{\mathcal{B}} &= \frac{1}{s}\sum_{i=1}^{s} \mathcal{B}^{[i]}, \\ \hat{C}_1 &= \frac{1}{s}\sum_{i=1}^{s} C_1^{[i]}, \\ \hat{V} &= \frac{1}{s}\sum_{i=1}^{s} V^{[i]}, \end{aligned}$$

and obtain $\widehat{\operatorname{AMSE}}\left\{\tilde{h}(r, N)\right\}$ by plugging these bagged estimates into (17).

Step 5. Finally, estimate r_0 by:

$$\hat{r}_0 = \underset{r>1}{\arg\min}\, \widehat{\operatorname{AMSE}}\left\{\tilde{h}(r, N)\right\}.$$

Additionally, assuming that $r = o\left(n/\sqrt{N}\right)$, then

$$r_0^{(2)} = \left\{-\frac{4(\mathcal{B}+C_1)^2}{V}N\right\}^{5/3}$$

and the rate of convergence to zero of the AMSE of the bagging bandwidth would be:

$$\text{AMSE}\left\{\tilde{h}\left(r_0^{(2)}, N\right)\right\} \sim n^{-2/5}N^{-4/3}.$$

Hence,

$$\frac{\text{AMSE}\left\{\tilde{h}\left(r_0^{(2)}, N\right)\right\}}{\text{AMSE}\left(\tilde{h}_{CV,n}\right)} \sim n^{1/5}N^{-4/3},$$

and this ratio would tend to zero if N tended to infinity at a rate faster than $n^{3/20}$. Furthermore, if we let $N = n^{3/20}$ and $r = r_0^{(2)}$, then the computational complexity of the algorithm would be $O\left(n^{13/20}\right)$, much lower than that of standard cross-validation. In fact, by selecting r_0 in this way, the complexity of the algorithm will only equal to that of standard cross-validation when N tends to infinity at the rate $n^{6/13}$.

6. Simulation studies

The behavior of the leave-one-out and bagged cross-validation bandwidths is evaluated by simulation in this section. We considered the following regression models:

M1: $Y = m(X) + \varepsilon$, $m(x) = 2x$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

M2: $Y = m(X) + \varepsilon$, $m(x) = \sin(2\pi x)^2$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

M3: $Y = m(X) + \varepsilon$, $m(x) = x + x^2\sin(8\pi x)^2$, $X \sim \text{Beta}(3,3)$, $\varepsilon \sim \text{N}(0, 0.1^2)$,

whose regression functions are plotted in Figure 1. The Gaussian kernel was used for computing the Nadaraya–Watson estimator throughout this section. Moreover, to reduce computing time in the simulations, we used binning to select the ordinary and the bagged cross-validation bandwidths. The R (R Development Core Team, 2021) package `baggingbwsel` (Barreiro-Ures et al., 2021) was employed to carry out the simulation experiments.

In a first step, we empirically checked how close the bandwidths that minimize the MISE of (2) and (9) are. For this, we simulated 100 samples of sizes 1000 and 5000 from models M1, M2 and M3 and compute the corresponding MISE curves for the standard Nadaraya–Watson estimator and for its modified version, given in (9). For the sake of brevity, the plot containing these curves is included in the accompanying supplementary materials. That plot shows that the bandwidth that minimizes the MISE of (9) and the MISE of the standard Nadaraya–Watson estimator appear to be quite close for both sample sizes, although the distance between the minima of both curves seems to tend to zero as the sample size increases. Moreover, the standard cross-validation bandwidths

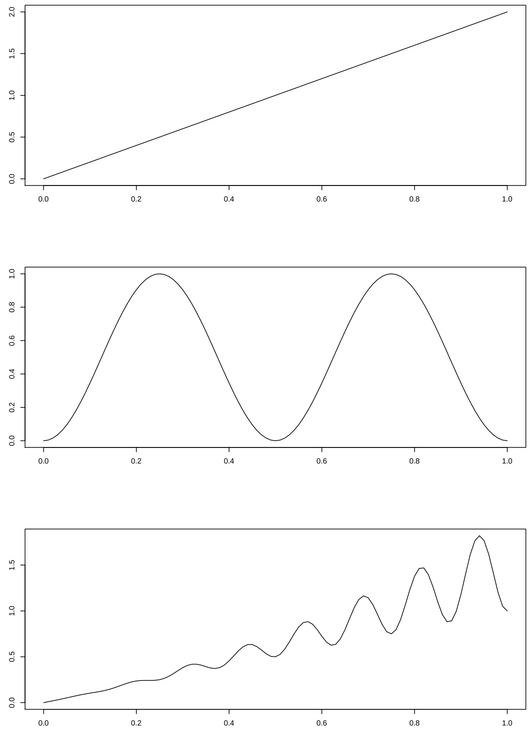

Figure 1. Regression function of models M1 (top), M2 (middle) and M3 (bottom).

and the modified cross-validation selectors (using the standard and the modified version of the Nadaraya–Watson estimator, respectively) are obtained for samples of sizes ranging from 600 to 5000 drawn from model M2. The corresponding figure is also included in the supplementary material. It shows that both bandwidth selectors provide similar results, which in turn get closer as n increases.

In a second step, we checked how fast the statistic $S_n = n^{3/10}\left(\hat{h}_{CV,n} - h_{n0}\right)$ approaches its limit distribution. For this, 1000 samples of size n were simulated from model M2 (with values of n ranging from 50 to 5000) and the corresponding values of S_n were computed. Figure 2 shows the kernel density estimates and boxplots constructed using these samples of S_n. The empirical behavior observed in Figure 2 is in agreement with the result derived from Corollary 3.1 (considering Remark 3.1), since the sampling distribution of S_n seems to tend to a normal distribution with zero mean and constant variance. Similar plots were obtained when considering models M1 and M3. They are not shown here for the sake of brevity.

In the next part of the study, we focused on empirically analyzing the performance of the bagged cross-validation bandwidth $\hat{h}(r, N)$, given in (16), for different values of n, r and N. Figure 3 shows the sampling distribution of $\hat{h}_n/h_{n0}$, where $\hat{h}_n$ denotes either the ordinary or the bagged cross-validation bandwidth. For this, 1000 samples of size

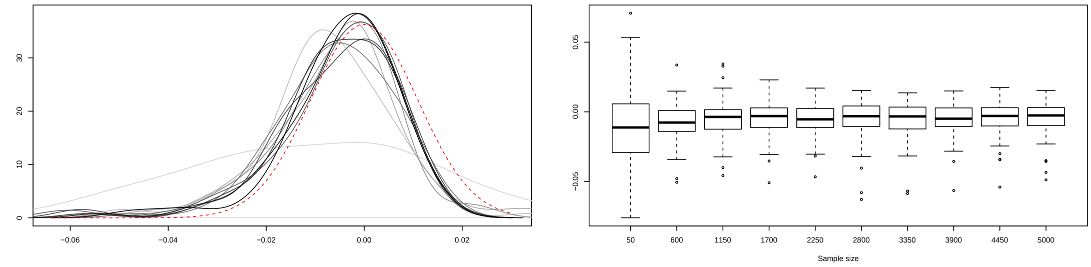

Figure 2. Sampling distribution of $S_n = n^{3/10}(\hat{h}_{CV,n} - h_{n0})$: kernel density estimates (left panel) and boxplots (right panel), for samples drawn from model M2 and considering values of n between 50 and 5000. In the left panel, sharper lines correspond to larger values of n. The limit distribution of S_n is also shown (dashed line).

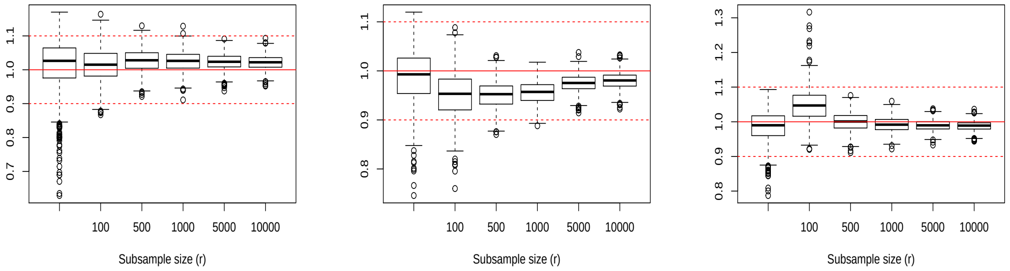

Figure 3. Sampling distribution of $\hat{h}_{CV,n}/h_{n0}$ (first boxplot on each panel) and $\hat{h}(r,N)/h_{n0}$ (second to sixth boxplots on each panel) for models M1 (left panel), M2 (central panel) and M3 (right panel), where the considered subsample sizes are $r \in \{100, 500, 1000, 5000, 10^4\}$ and the number of subsamples is $N = 25$. The original sample size is $n = 10^5$. Dashed lines are plotted at values 0.9 and 1.1 for reference.

$n = 10^5$ from models M1, M2 and M3 were generated, considering in the case of $\hat{h}(r,N)$ the values $r \in \{100, 500, 1000, 5000, 10000\}$ and $N = 25$. For all three models, it is observed how the bias and variance of the bagging bandwidth decrease as the subsample size increases and how its mean squared error seems to stabilize for values of r close to 5000. Moreover, the behavior of the bagging selector turns out to be quite positive even when considering subsample sizes as small as $r = 100$, perhaps excluding the case of model M3 for which the variance of the bagging bandwidth is still relatively high for $r = 100$, although it undergoes a rapid reduction as the subsample size increases slightly.

The effect that r has on the mean squared error of the bagged bandwidth is also illustrated in Table 1, which shows the ratio of the mean squared errors of the bagged bandwidth and the ordinary cross-validation bandwidth, $\mathrm{MSE}\{\hat{h}(r,N)\}/\mathrm{MSE}(\hat{h}_{CV,n})$, for the three models.

Apart from a better statistical precision of the cross-validation bandwidths selected

Table 1. Ratio of the mean squared errors of the bagged and the ordinary cross-validation bandwidths for models M1–M3. Different values of r and $N = 25$ were considered, for a sample size of $n = 10^5$.

	Model		
	M1	M2	M3
Subsample size (r)	MSE ratio		
100	0.47	1.47	2.16
500	0.32	1.06	0.33
$1,000$	0.26	0.80	0.23
$5,000$	0.19	0.30	0.17
$10,000$	0.16	0.22	0.16

using bagging, another potential advantage of employing this approach is the reduction of computing times, especially with large sample sizes. To analyze this issue, Figure 4 shows, as a function of the sample size, n, the CPU elapsed times for computing the standard and the bagged cross-validation bandwidths. Both variables are shown on a logarithmic scale. In the case of the bagging selector, three different subsample size values, r, depending on n were considered: $r = n^{0.7}$, $r = n^{0.8}$ and $r = n^{0.9}$. Calculations were performed in parallel using an Intel Core i5-8600K 3.6GHz CPU. Different sample sizes, $n \in \{5000, 28750, 52500, 76250, 10^5\}$, and a fixed number of subsamples, $N = 25$, were used. In this experiment, binning techniques were employed using a number of bins of $0.1n$ for standard cross-validation and $0.1r$ in the case of bagged cross-validation. The time required to compute the bagged cross-validation bandwidth was measured considering the three possible growth rates for r, mentioned above.

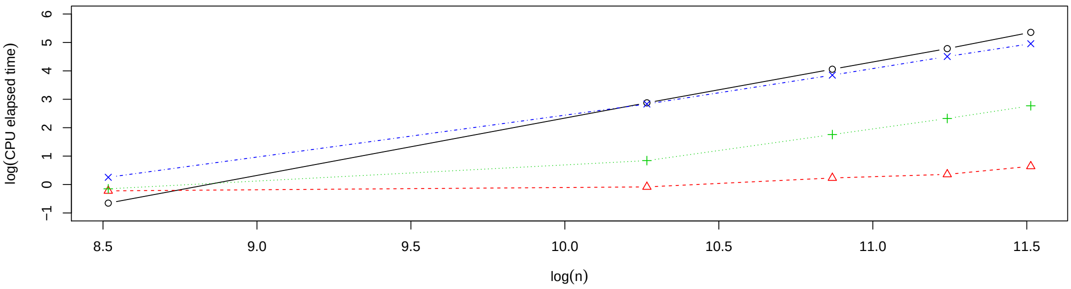

Figure 4. CPU elapsed time (seconds) as a function of the sample size of standard cross-validation (solid line-circles) and bagged cross-validation. Both variables are shown on a logarithmic scale. A fixed number of subsamples was used, $N = 25$. Three growth rates for r were considered, namely, $r = n^{0.7}$ (dashed line-triangles), $r = n^{0.8}$ (dotted line-pluses) and $r = n^{0.9}$ (dashed-dotted line-crosses).

Fitting an appropriate model, these CPU elapsed times could be used to predict the computing times of the different selectors for larger sample sizes. Considering Figure 4,

Table 2. Predicted CPU elapsed time for the standard and the bagging cross-validation method using three different choices for the subsample size.

	Sample size (n)		
	10^6	10^7	10^8
Method	Computing time		
Standard CV	6 hours	24 days	7 years
Bagged CV ($r = n^{0.7}, N = 25$)	40 seconds	25 minutes	16 hours
Bagged CV ($r = n^{0.8}, N = 25$)	16 minutes	17 hours	45 days
Bagged CV ($r = n^{0.9}, N = 25$)	3 hours	11 days	2 years

the following log-linear model was used:

$$T(n) = \alpha n^{\beta}, \tag{18}$$

where $T(n)$ denotes the CPU elapsed time as a function of the original sample size, n. In the case of the bagged cross-validation bandwidths, there is a fixed time corresponding to the one required for the setting up of the parallel socket cluster. This time, which does not depend on n, r or N, but only on the CPU and the number of cores used in the parallelization, was estimated to be 0.79. Using this value, the corrected CPU elapsed times obtained for the bagged bandwidths, $T - 0.79$, were employed to fit the log-linear model (18) estimating $\alpha, \beta > 0$ by least squares and, subsequently, to make predictions. Table 2 shows the predicted CPU elapsed time for ordinary and bagged cross-validation for large sample sizes. Although we should take these predictions with caution, the results in Table 2 serve to illustrate the important reductions in computing time that bagging can provide for certain choices of r and N, especially for very large sample sizes.

Next, the influence of the number of subsamples, N, in the computing times of the bagged badwidths was studied. Similarly to Figure 4, Figure 5 shows the CPU elapsed times for computing the cross-validation bandwidths (standard and bagged). For the bagging method, the number of subsamples, N, was selected depending on the original sample size (n) by $N = \sqrt{n}$. The growth rates used for r are the same as in the case of Figure 4.

It should also be stressed that although the quadratic complexity of the cross-validation algorithm is not so critical in terms of computing time for small sample sizes, even in these cases, the use of bagging can still lead to substantial reductions in mean squared error of the corresponding bandwidth selector with respect to the one selected by ordinary cross-validation. In order to show this, 1000 samples from model M1 of sizes $n \in \{50, 500, 5000\}$ were simulated and the ordinary and bagged cross-validation bandwidths for each of these samples were computed. In the case of the bagged cross-validation bandwidth, both the size of the subsamples and the number of subsamples were selected depending on n, choosing $r = N = 4\sqrt{n}$. Figure 6 shows the sampling distribution of $\hat{h}_n/h_{n0}$, where $\hat{h}_n$ denotes either the ordinary or the bagged cross-validation bandwidth. In the three scenarios, it can be observed the considerable reductions in variance produced by bagging more than offset the slight increases in bias, thus obtaining significant reductions in mean squared error with respect to the ordinary cross-validation bandwidth selector. Specifically, the relative reductions in mean squared error achieved

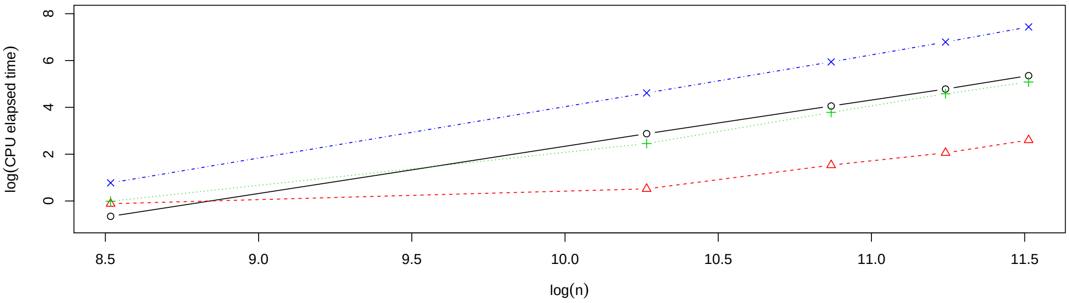

Figure 5. CPU elapsed time (seconds) as a function of the sample size of standard cross-validation (solid line-circles) and bagged cross-validation. Both variables are shown on a logarithmic scale. The number of subsamples grows with n at the rate $N = \sqrt{n}$. Three growth rates for r were considered, namely, $r = n^{0.7}$ (dashed line-triangles), $r = n^{0.8}$ (dotted line-pluses) and $r = n^{0.9}$ (dashed-dotted line-crosses).

by the bagged bandwidth turned out to be 69.3%, 90.1% and 93.8% for $n = 50$, $n = 500$ and $n = 5000$, respectively. This experiment was repeated with models M2 and M3, obtaining similar results.

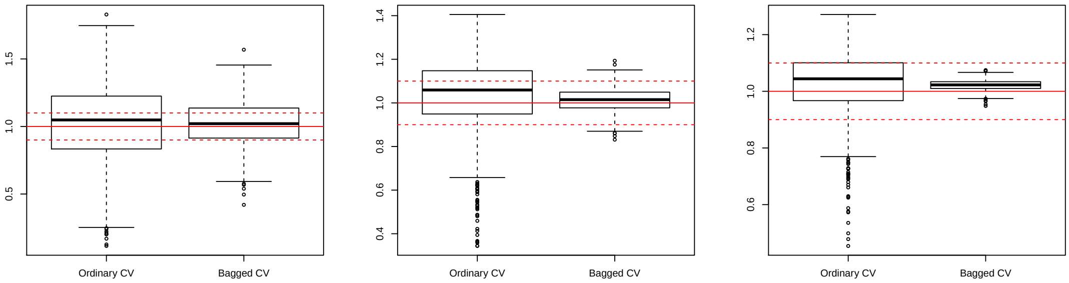

Figure 6. Sampling distribution of $\hat{h}_n/h_{n0}$, where $\hat{h}_n$ denotes either the ordinary or bagged cross-validation bandwidth, for samples of size $n = 50$ (left panel), $n = 500$ (central panel) and $n = 5,000$ (right panel) drawn from model M1. The values of r and N were chosen as $r = N = 4\sqrt{n}$. Dashed lines are plotted at values 0.9 and 1.1 for reference.

7. Application to COVID-19 data

In order to illustrate the performance of the techniques studied in the previous sections, the COVID-19 dataset briefly mentioned in the introduction is considered. It consists of a sample of size $n = 105,235$ which contains the age (the explanatory variable) and the time in hospital (the response variable) of people infected with COVID-19 in Spain from

January 1, 2020 to December 20, 2020. Due to the high number of ties in this dataset and in order to avoid problems when performing cross-validation, we decided to remove the ties by jittering the data. The actual age differs from the observed age, rounded down to years, by an amount that is in the interval $(0, 1)$. Thus, it is reasonable to model this difference between actual and observed age using the uniform distribution in the interval $(0, 1)$. On the other hand, the hospitalization time was calculated as the difference between the day of discharge and the day of admission to the hospital. The specific time of discharge and admission would be obtained by adding uniform variables, with support in the interval $(0, 1)$, to each of the two dates. In particular, three independent random samples of size n, U_1, U_2 and U_3, drawn from a continuous uniform distribution defined on the interval $(0, 1)$, were generated. Then, U_1 was added to the original explanatory variable and $U_2 - U_3$ to the original response variable. Figure 7 shows scatterplots for the complete sample as well as for three randomly chosen subsamples of size $1,000$.

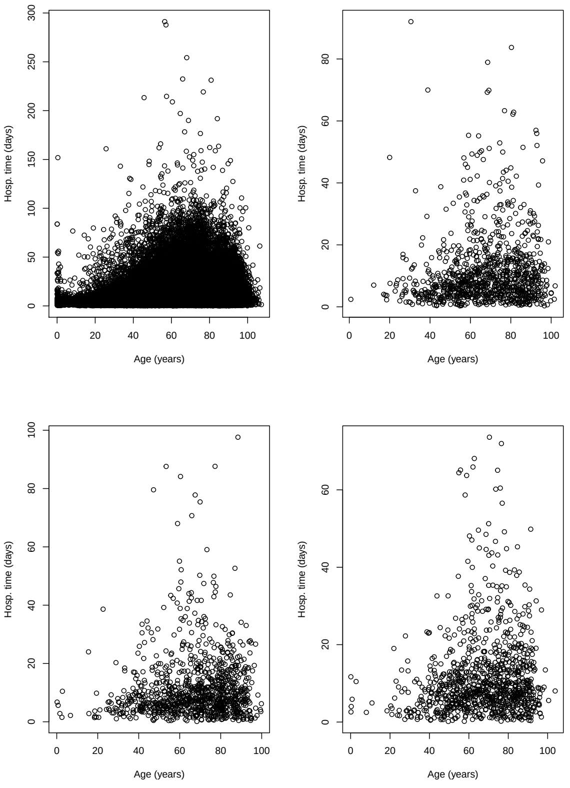

Figure 7. Whole COVID-19 sample (top left panel) as well as three randomly chosen subsamples of size 1000.

To compute the standard cross-validation bandwidth using binning, the number of bins was set to $10,000$, that is, roughly 10% of the sample size. The value of the bandwidth thus obtained was 1.84 and computing it took 72 seconds. For the bagged bandwidth, 10 subsamples of size $30,000$ were considered. Binning was used again for each subsample, fixing the number of bins to $3,000$. The calculations associated with each subsample were performed in parallel using 5 cores. The value of the bagged bandwidth was 1.52 and its computing time was 33 seconds. Figure 8 shows the Nadaraya–Watson estimates with both standard and bagged cross-validation bandwidths. For comparative purposes, the local linear regression estimate with direct plug-in bandwidth (Ruppert et al., 1995) was also computed.

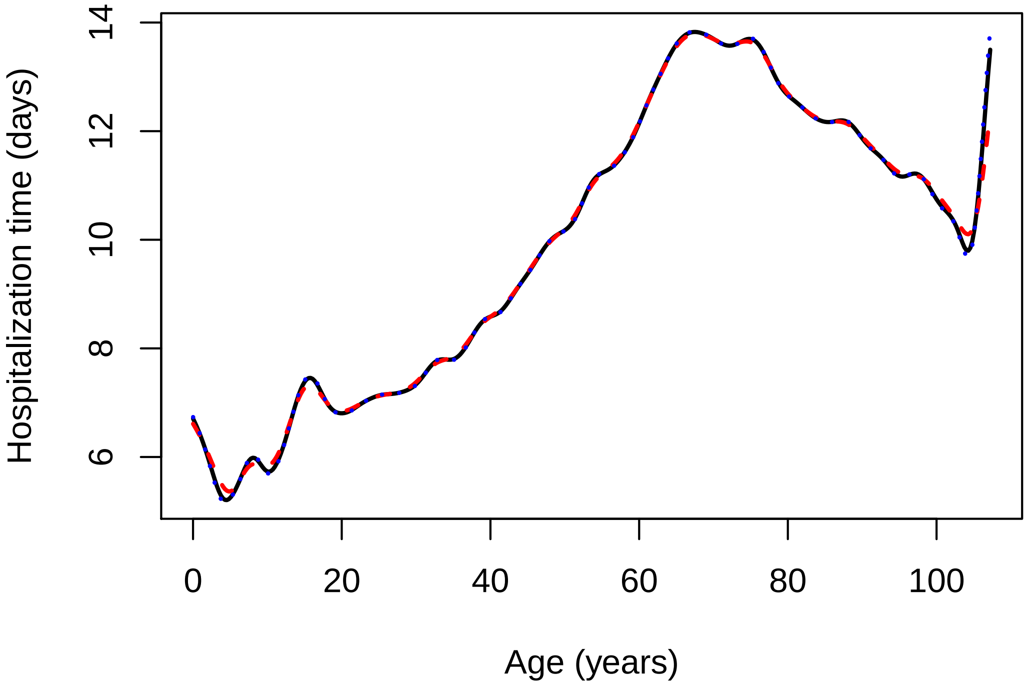

Figure 8. Kernel regression estimates for the COVID-19 data. The Nadaraya–Watson estimator with standard (dashed line) and bagged (solid line) cross-validation bandwidths are shown. Additionally, the local linear estimator with plug-in bandwidth (dotted line) is also presented.

Figure 8 shows that the Nadaraya–Watson estimator with standard cross-validation bandwidth produces a slightly smoother estimate than the one obtained with the bagged bandwidth, the latter being almost indistinguishable from the local linear estimate computed with direct plug-in bandwidth. One can conclude that the expected time that a person infected with COVID-19 will remain in hospital increases non-linearly with age for people under approximately 70 years. This trend is reversed for people aged between 70 and 100 years. This could be due to the fact that patients in this age group are more likely to die and, therefore, end the hospitalization period prematurely. Finally, the expected hospitalization time grows again very rapidly with age for people over 100 years of age, although this could be caused by some boundary effect, since the number of observations for people over 100 years old is very small, specifically 155, which corresponds to roughly 0.15% of the total number of observations. In order to avoid this possible boundary effect, the estimators were also fitted to a modified version of the sample in which the explanatory variable was transformed using its own empirical distribution function. The resulting estimators are shown in Figure 9, where the explanatory variable was returned to its original scale by means of its empirical quantile function.

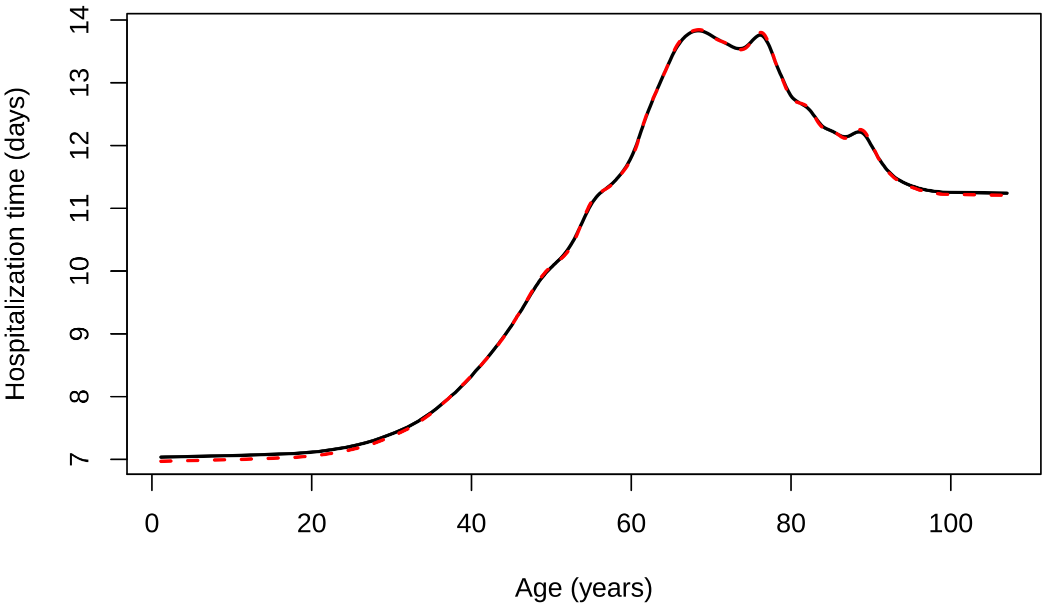

Figure 9. Kernel regression estimators for the COVID-19 data, removing boundary effects. The Nadaraya–Watson estimator with standard (dashed line) and bagged (solid line) cross-validation bandwidths are shown.

Finally, the same procedure was followed to estimate the expected time in hospital but splitting the patients by gender, as shown in Figure 10. This figure shows that the expected time in hospital is generally shorter for women, except for ages less than 30 years or between 65 and 85 years. Anyhow, the difference in mean time in hospital for men and women never seems to exceed one day. In Figure 10, only the Nadaraya–Watson estimates computed with the bagged cross-validation bandwidths ($h = 0.03$ for men and $h = 0.028$ for women) are shown. Both the Nadaraya–Watson estimates with standard cross-validation bandwidths ($h = 0.028$ for men and $h = 0.023$ for women) and the local linear estimates with direct plug-in bandwidths produced very similar and graphically indistinguishable results from those shown in Figure 10.

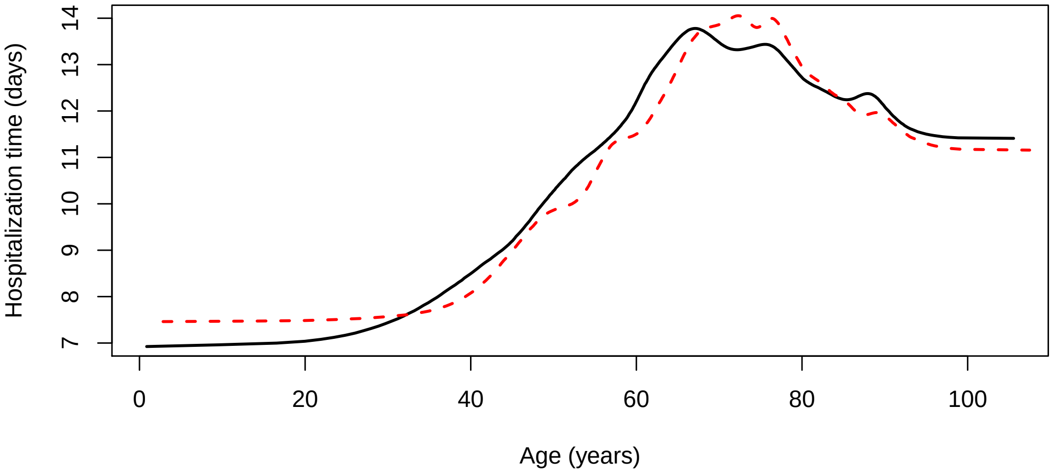

Figure 10. Kernel regression estimators for the COVID-19 data by gender, removing boundary effects. The Nadaraya–Watson estimators with bagged cross-validation bandwidths are shown for male (solid line) and female (dashed line) patients.

8. Discussion

The asymptotic properties of the leave-one-out cross-validation bandwidth for the Nadaraya–Watson estimator considering a regression model with random design have been studied in this paper. Additionally, a bagged cross-validation selector have been also analyzed (theoretically and empirically) as an alternative to standard leave-one-out cross-validation. The advantage of this bandwidth selector is twofold: (i) to gain computational efficiency with respect to standard leave-one-out cross-validation by applying the cross-validation algorithm to several subsamples of size $r < n$ rather than a single sample of size n, and (ii) to reduce the variability of the leave-one-out cross-validation bandwidth. Although the new bandwidth selector studied in the present paper can outperform the behavior of the standard cross-validation selector even for moderate sample sizes, improvements in computation time become truly significant only for large-sized samples.

The methodology presented in this paper can be applied to other bandwidth selection techniques, apart from cross-validation, as mentioned in Barreiro-Ures et al. (2020). Extensions to bootstrap bandwidth selectors is an interesting topic for a future research. The bootstrap resampling plans proposed by Cao and González-Manteiga (1993) can be used to derive a closed form for the bootstrap criterion function in nonparametric regression estimation, along the lines presented by Barbeito et al. (2021) who have dealt with matching and prediction.

Another interesting future research topic is the extension of the results presented in this paper to the case of the local linear estimator, whose behavior is known to be superior to that of the Nadaraya–Watson estimator, especially in the boundary regions.

Acknowledgments

This research has been supported by MINECO Grants MTM2017-82724-R and PID2020-113578RB-I00, and by 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. The authors would like to thank the Spanish Center for Coordinating Sanitary Alerts and Emergencies for kindly providing the COVID-19 hospitalization dataset.

References

Barbeito, I. (2020) *Exact bootstrap methods for nonparametric curve estimation.* Ph.D. thesis, Universidade da Coruña. `https://ruc.udc.es/dspace/handle/2183/26466`.

Barbeito, I., Cao, R. and Sperlich, S. (2021) Bandwidth selection for statistical matching and prediction. *Tech. rep.*, University of A Coruña. Department of Mathematics. `http://dm.udc.es/preprint/Bandwidth_Selection_Matching_Prediction_NOT_BLINDED.pdf` and `http://dm.udc.es/preprint/SuppMaterial_Bandwidth_Selection_Matching_Prediction_NOT_BLINDED.pdf`.

Barreiro-Ures, D., Cao, R., Francisco-Fernández, M. and Hart, J. D. (2020) Bagging cross-validated bandwidths with application to big data. *Biometrika.* `https://doi.org/10.1093/biomet/asaa092`.

Barreiro-Ures, D., Hart, J. D., Cao, R. and Francisco-Fernández, M. (2021) *baggingbwsel: Bagging Bandwidth Selection in Kernel Density and Regression Estimation.* R package version 1.0. `https://cran.r-project.org/package=baggingbwsel`.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Cao, R. and González-Manteiga, W. (1993) Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics*, **2**, 379–388.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Friedman, J. H. and Hall, P. (2007) On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, **137**, 669–683.

Hall, P. and Robinson, A. P. (2009) Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, **96**, 175–186.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83**, 86–95.

Köhler, M., Schindler, A. and Sperlich, S. (2014) A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review / Revue Internationale de Statistique*, **82**, 243–274.

Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability & Its Applications*, **9**, 141–142.

Opitz, D. and Maclin, R. (1999) Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, **11**, 169–198.

R Development Core Team (2021) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Stone, C. J. (1977) Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–620. URL: https://doi.org/10.1214/aos/1176343886.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing.* London: Chapman and Hall.

Watson, G. S. (1964) Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 359–372.

Supplementary material for "Bagging cross-validated bandwidth selection in nonparametric regression estimation with applications to large-sized samples"

Daniel Barreiro-Ures

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Ricardo Cao

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Mario Francisco-Fernández †

Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

Abstract. This supplementary material for "Bagging cross-validated bandwidth selection in nonparametric regression estimation with applications to large-sized samples" contains the proofs of the theoretical results included in the main paper. In addition, some plots completing the simulation study presented in the main paper are also provided. Specifically, a figure showing empirically the closeness between the MISE bandwidths when considering the Nadaraya–Watson estimator and when using its modified version, given in equation (9) of the main paper, in different scenarios, is included. Moreover, a figure presenting the relationship between the standard cross-validation bandwidths and the corresponding modified cross-validation selectors (using the standard and modified version of the Nadaraya–Watson estimator, respectively) is also added.

1. Theoretical results

This section includes the proofs of Lemmas 3.1 and 3.2, Theorems 3.1 and 4.1, and Corollaries 3.1 and 4.1 of the main paper. A sketch of the proof of Remark 3.1 is also included. The following assumptions are needed:

A1. K is a symmetric and differentiable kernel function.

A2. For every $j = 0, \ldots, 6$ the integrals $\mu_j(K)$, $\mu_j(K')$ and $\mu_j(K^2)$ exist and are finite, where $\mu_j(g) = \int x^j g(x)\, dx$.

A3. The functions m and f are eight times differentiable.

A4. The function σ^2 is four times differentiable.

A5. As $r, n \to \infty$, $r = o(n)$ and N tends to a positive constant or ∞.

†*Address for correspondence:* Mario Francisco-Fernández, University of A Coruña, Faculty of Computer Science, Campus de Elviña, s/n, A Coruña, Spain. E-mail: mariofr@udc.es

Lemma 3.1. *Under assumptions* A1–A4, *the bias and the variance of the modified version of the Nadaraya–Watson estimator defined just after equation* (8) *of the main paper satisfy:*

$$
\begin{aligned}
\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x) &= \mu_2(K)\left\{\frac{1}{2}m''(x)+\frac{m'(x)f'(x)}{f(x)}\right\}h^2\\
&+ \left[\mu_4(K)\left\{\frac{1}{24}m^{4)}(x)+\frac{1}{6}\frac{m'''(x)f'(x)}{f(x)}+\frac{1}{4}\frac{m''(x)f''(x)}{f(x)}\right.\right.\\
&+ \left.\left.\frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right\}-\mu_2(K)^2\frac{f''(x)}{f(x)}\left\{\frac{1}{4}m''(x)+\frac{m'(x)f'(x)}{f(x)}\right\}\right]h^4\\
&+ O\left(h^6+n^{-1}\right),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}\left\{\tilde{m}_h(x)\right\} &= R(K)\sigma^2(x)f(x)^{-1}n^{-1}h^{-1}\\
&+ \left[\mu_2(K^2)f(x)^{-2}\left\{\varphi_3(x)+\frac{1}{2}m(x)^2f''(x)-2\varphi_1(x)m(x)f(x)\right\}\right.\\
&- \left.R(K)\mu_2(K)\sigma^2(x)f(x)^{-2}f''(x)\right]n^{-1}h+O(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}),
\end{aligned}
$$

where $R(K)=\int K(x)^2\,\mathrm{d}x$.

Proof. Let us start by defining

$$
\begin{aligned}
\varphi_1(x) &= f(x)^{-1}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)+\frac{1}{2}m(x)f''(x)\right\},\\
\varphi_2(x) &= f(x)^{-1}\left\{\frac{1}{24}m^{4)}(x)f(x)+\frac{1}{6}m'''(x)f'(x)+\frac{1}{4}m''(x)f''(x)+\frac{1}{6}m'(x)f'''(x)\right.\\
&+ \left.\frac{1}{24}m(x)f^{4)}(x)\right\},\\
\varphi_3(x) &= \frac{1}{2}f''(x)\left\{m(x)^2+\sigma^2(x)\right\}+f(x)\left\{m(x)m''(x)+m'(x)^2+\frac{1}{2}\sigma^{2''}(x)\right\}\\
&+ f'(x)\left\{2m(x)m'(x)+\sigma^{2'}(x)\right\}.
\end{aligned}
$$

Let us first study the bias of $\tilde{m}_h$. Recall that $\tilde{m}_h(x)=A+B+C+D$, where A, B, C and D are defined just after equation (8) of the main paper. Then, $\mathrm{E}\left\{\tilde{m}_h(x)\right\}=\mathrm{E}(A)+\mathrm{E}(B)+\mathrm{E}(C)+\mathrm{E}(D)$.

To compute $\mathrm{E}(A)$, we start by expanding the following expectation:

$$\begin{aligned}
\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)Y_i\right\} &= \mathrm{E}\left\{K_h(x-X_1)Y_1\right\} = \mathrm{E}\left\{K_h(x-X_1)\mathrm{E}(Y_1 \mid X_1)\right\} \\
&= E\left\{K_h(x-X_1)m(X_1)\right\} = \int K_h(x-x_1)m(x_1)f(x_1)\,dx_1 \\
&= \int K(u)m(x-hu)f(x-hu)\,du \\
&= m(x)f(x) + h^2\mu_2(K)f(x)\varphi_1(x) \\
&+ h^4\mu_4(K)f(x)\varphi_2(x) + O(h^6).
\end{aligned}$$

Therefore,

$$\mathrm{E}(A) = m(x) + h^2\mu_2(K)\varphi_1(x) + h^4\mu_4(K)\varphi_2(x) + O(h^6). \tag{1}$$

Similarly, to compute $\mathrm{E}(B)$, we compute the following expectation:

$$\begin{aligned}
\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\right\} &= \mathrm{E}\{K_h(x-X_1)\} = \int K_h(x-x_1)f(x_1)\,dx_1 \\
&= \int K(u)f(x-hu)\,du = f(x) + \frac{1}{2}h^2\mu_2(K)f''(x) \\
&+ \frac{1}{24}h^4\mu_4(K)f^{4)}(x) + O(h^6)
\end{aligned}$$

and, hence,

$$\mathrm{E}(B) = -\frac{m(x)}{f(x)}\left\{\frac{1}{2}h^2\mu_2(K)f''(x) + \frac{1}{24}h^4\mu_4(K)f^{4)}(x)\right\} + O(h^6). \tag{2}$$

To compute $\mathrm{E}(C)$, we start by expanding the following expectation:

$$\begin{aligned}
\mathrm{E}\left\{Y_1K_h(x-X_1)^2\right\} &= \mathrm{E}\left\{m(X_1)K_h(x-X_1)^2\right\} = \int K_h(x-x_1)^2m(x_1)f(x_1)\,dx_1 \\
&= h^{-1}\int K(u)^2m(x-hu)f(x-hu)\,du = R(K)m(x)f(x)h^{-1} \\
&+ h\mu_2(K^2)f(x)\varphi_1(x) + h^3\mu_4(K^2)f(x)\varphi_2(x) + O(h^5).
\end{aligned}$$

We now obtain some asymptotic expressions for some of the terms in $\mathrm{E}(C)$:

$$\begin{aligned}
\mathrm{E}(\hat{a}\hat{e}) &= \mathrm{E}\left\{n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_iK_h(x-X_1)K_h(x-X_j)\right\} = n^{-2}\left[n\mathrm{E}\left\{Y_1K_h(x-X_1)^2\right\}\right. \\
&+ \left. n(n-1)\mathrm{E}\left\{Y_1K_h(x-X_1)K_h(x-X_2)\right\}\right] = n^{-1}\mathrm{E}\left\{Y_1K_h(x-X_1)^2\right\} \\
&+ \mathrm{E}\left\{Y_1K_h(x-X_1)\right\}\mathrm{E}\left\{K_h(x-X_1)\right\} = R(K)m(x)f(x)n^{-1}h^{-1} + m(x)f(x)^2 \\
&+ h^2\mu_2(K)\left\{\frac{1}{2}f''(x)m(x)f(x) + f(x)^2\varphi_1(x)\right\} + h^4\left\{\frac{1}{24}\mu_4(K)f^{4)}(x)m(x)f(x)\right. \\
&+ \left.\mu_4(K)f(x)^2\varphi_2(x) + \frac{1}{2}\mu_2(K)^2f''(x)f(x)\varphi_1(x)\right\} + O(h^6 + n^{-1}).
\end{aligned}$$

Therefore,

$$\mathrm{E}(C) = -R(K)m(x)f(x)^{-1}n^{-1}h^{-1} - \frac{1}{2}h^4\mu_2(K)^2 f''(x)f(x)^{-1}\varphi_1(x) + O(h^6 + n^{-1}). \quad (3)$$

To deal with $\mathrm{E}(D)$, we proceed in a similar way:

$$\begin{aligned}\mathrm{E}\left\{K_h(x-X_1)^2\right\} &= \int K_h(x-x_1)^2 f(x_1)\,dx_1 = h^{-1}\int K(u)^2 f(x-hu)\,du \\ &= R(K)f(x)h^{-1} + \frac{1}{2}h\mu_2(K^2)f''(x) + \frac{1}{24}h^3\mu_4(K^2)f^{4)}(x) + O(h^5),\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(\hat{e}^2\right) &= \mathrm{E}\left\{n^{-2}\sum_{i=1}^n\sum_{j=1}^n K_h(x-X_i)K_h(x-X_j)\right\} = n^{-2}\left[n\mathrm{E}\left\{K_h(x-X_1)^2\right\}\right. \\ &+ \left. n(n-1)\mathrm{E}\left\{K_h(x-X_1)K_h(x-X_2)\right\}\right] \\ &= n^{-1}\mathrm{E}\left\{K_h(x-X_1)^2\right\} + \frac{n-1}{n}\mathrm{E}\left\{K_h(x-X_1)\right\}^2 \\ &= R(K)f(x)n^{-1}h^{-1} + f(x)^2 + h^2\mu_2(K)f''(x)f(x) \\ &+ h^4\left\{\frac{1}{12}\mu_4(K)f^{4)}(x)f(x) + \frac{1}{4}\mu_2(K)^2 f''(x)^2\right\} + O(h^6 + n^{-1}),\end{aligned}$$

and, hence,

$$\mathrm{E}(D) = R(K)m(x)f(x)^{-1}n^{-1}h^{-1} + \frac{1}{4}h^4\mu_2(K)^2 f''(x)^2 m(x)f(x)^{-2} + O(h^6 + n^{-1}). \quad (4)$$

Adding (1), (2), (3) and (4), we get

$$\begin{aligned}\mathrm{E}\left\{\tilde{m}_h(x)\right\} - m(x) &= \mu_2(K)\left\{\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}h^2 \\ &+ \left[\mu_4(K)\left\{\frac{1}{24}m^{4)}(x) + \frac{1}{6}\frac{m'''(x)f'(x)}{f(x)} + \frac{1}{4}\frac{m''(x)f''(x)}{f(x)}\right.\right. \\ &+ \left.\left.\frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right\} - \mu_2(K)^2\frac{f''(x)}{f(x)}\left\{\frac{1}{4}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right\}\right]h^4 \\ &+ O\left(h^6 + n^{-1}\right).\end{aligned}$$

Regarding the variance of $\tilde{m}_h(x) = A + B + C + D$, we have that

$$\begin{aligned}\mathrm{var}\left\{\tilde{m}_h(x)\right\} &= \mathrm{var}(A) + \mathrm{var}(B) + \mathrm{var}(C) + \mathrm{var}(D) + 2\left\{\mathrm{cov}(A,B) + \mathrm{cov}(A,C)\right. \\ &+ \left.\mathrm{cov}(B,C) + \mathrm{cov}(A,D) + \mathrm{cov}(B,D) + \mathrm{cov}(C,D)\right\}. \quad (5)\end{aligned}$$

The following second order moment will be needed to handle some of the variance and covariance terms:

$$\begin{aligned}
\mathrm{E}\left\{Y_1^2 K_h(x-X_1)^2\right\} &= \mathrm{E}\left\{K_h(x-X_1)^2 \mathrm{E}\left(Y_1^2 \mid X_1\right)\right\} \\
&= \mathrm{E}\left[K_h(x-X_1)^2 \mathrm{E}\left\{(m(X_1)+\varepsilon_1)^2 \mid X_1\right\}\right] \\
&= \mathrm{E}\left[\left\{m(X_1)^2+\sigma^2(X_1)\right\} K_h(x-X_1)^2\right] \\
&= \int K_h(x-x_1)^2 \left\{m(x_1)^2+\sigma^2(x_1)\right\} f(x_1)\, dx_1 \\
&= h^{-1} \int K(u)^2 \left\{m(x-hu)^2+\sigma^2(x-hu)\right\} f(x-hu)\, du \\
&= R(K)\left\{m(x)^2+\sigma^2(x)\right\} f(x)h^{-1} + \mu_2(K^2)\varphi_3(x)h + O(h^3).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{var}(\hat{a}) &= n^{-2}\mathrm{var}\left\{\sum_{i=1}^n Y_i K_h(x-X_i)\right\} = n^{-1}\mathrm{var}\left\{Y_1 K_h(x-X_1)\right\} \\
&= n^{-1}\left[\mathrm{E}\left\{Y_1^2 K_h(x-X_1)^2\right\} - \mathrm{E}\left\{Y_1 K_h(x-X_1)\right\}^2\right] \\
&= R(K)\left\{m(x)^2+\sigma^2(x)\right\} f(x)n^{-1}h^{-1} - m(x)^2 f(x)^2 n^{-1} \\
&+ \mu_2(K^2)\varphi_3(x)n^{-1}h + O(n^{-1}h^2)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(A) &= R(K)\left\{m(x)^2+\sigma^2(x)\right\} f(x)^{-1}n^{-1}h^{-1} - m(x)^2 n^{-1} \\
&+ \mu_2(K^2)\varphi_3(x)f(x)^{-2}n^{-1}h + O(n^{-1}h^2). \qquad (6)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathrm{var}(\hat{e}) &= n^{-2}\mathrm{var}\left\{\sum_{i=1}^n K_h(x-X_i)\right\} = n^{-1}\mathrm{var}\left\{K_h(x-X_1)\right\} \\
&= n^{-1}\left[\mathrm{E}\left\{K_h(x-X_1)^2\right\} - \mathrm{E}\left\{K_h(x-X_1)\right\}^2\right] \\
&= R(K)f(x)n^{-1}h^{-1} - f(x)^2 n^{-1} + \frac{1}{2}\mu_2(K^2)f''(x)n^{-1}h + O(n^{-1}h^2)
\end{aligned}$$

and, so,

$$\begin{aligned}
\mathrm{var}(B) &= R(K)m(x)^2 f(x)^{-1}n^{-1}h^{-1} - m(x)^2 n^{-1} \\
&+ \frac{1}{2}\mu_2(K^2)m(x)^2 f(x)^{-2}f''(x)n^{-1}h + O(n^{-1}h^2). \qquad (7)
\end{aligned}$$

Straightforward calculations lead to

$$\begin{aligned}
\operatorname{cov}(\hat{a},\hat{e}) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}\left\{Y_iK_h(x-X_i),K_h(x-X_j)\right\}\\
&= n^{-1}\operatorname{cov}\left\{Y_1K_h(x-X_1),K_h(x-X_1)\right\}\\
&= n^{-1}\left[\mathrm{E}\left\{Y_1K_h(x-X_1)^2\right\}-\mathrm{E}\left\{Y_1K_h(x-X_1)\right\}\mathrm{E}\left\{K_h(x-X_1)\right\}\right]\\
&= R(K)m(x)f(x)n^{-1}h^{-1}-m(x)f(x)^2n^{-1}\\
&+ \mu_2(K^2)\varphi_1(x)f(x)n^{-1}h+O(n^{-1}h^2)
\end{aligned}$$

and, hence,

$$\begin{aligned}
\operatorname{cov}(A,B) &= -R(K)m(x)^2f(x)^{-1}n^{-1}h^{-1}+m(x)^2n^{-1}\\
&- \mu_2(K^2)\varphi_1(x)m(x)f(x)^{-1}n^{-1}h+O(n^{-1}h^2).
\end{aligned}\tag{8}$$

Now,

$$\begin{aligned}
&\operatorname{cov}\left\{Y_1K_h(x-X_1),Y_1K_h(x-X_1)K_h(x-X_2)\right\}\\
=\;& \operatorname{var}\left\{Y_1K_h(x-X_1)\right\}\mathrm{E}\left\{K_h(x-X_1)\right\}\\
=\;& R(K)\left\{m(x)^2+\sigma^2(x)\right\}f(x)^2h^{-1}-m(x)^2f(x)^3\\
+\;& \Big[\mu_2(K^2)\varphi_3(x)f(x)\\
+\;& \frac{1}{2}R(K)\mu_2(K)\left\{m(x)^2+\sigma^2(x)\right\}f(x)f''(x)\Big]h+O(h^2)
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{cov}\left\{Y_1K_h(x-X_1),Y_2K_h(x-X_2)K_h(x-X_1)\right\}\\
=\;& \operatorname{cov}\left\{Y_1K_h(x-X_1),K_h(x-X_1)\right\}\mathrm{E}\left\{Y_1K_h(x-X_1)\right\}\\
=\;& R(K)m(x)^2f(x)^2h^{-1}-m(x)^2f(x)^3\\
+\;& \left\{\mu_2(K^2)\varphi_1(x)m(x)f(x)^2\right.\\
+\;& \left.R(K)\mu_2(K)\varphi_1(x)m(x)f(x)^2\right\}+O\left(h^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{cov}(\hat{a},\hat{a}\hat{e}) &= n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\operatorname{cov}\left\{Y_iK_h(x-X_i),Y_jK_h(x-X_j)K_h(x-X_k)\right\}\\
&= n^{-1}\left[\operatorname{cov}\left\{Y_1K_h(x-X_1),Y_1K_h(x-X_1)K_h(x-X_2)\right\}\right.\\
&+ \left.\operatorname{cov}\left\{Y_1K_h(x-X_1),Y_2K_h(x-X_2)K_h(x-X_1)\right\}\right]+O(n^{-1}h^2)\\
&= R(K)\left\{2m(x)^2+\sigma^2(x)\right\}f(x)^2n^{-1}h^{-1}-2m(x)^2f(x)^3n^{-1}\\
&+ \Big[\mu_2(K^2)\varphi_3(x)f(x)+\frac{1}{2}R(K)\mu_2(K)\left\{m(x)^2+\sigma^2(x)\right\}f(x)f''(x)\\
&+ \mu_2(K^2)\varphi_1(x)m(x)f(x)^2+R(K)\mu_2(K)\varphi_1(x)m(x)f(x)^2\Big]n^{-1}h\\
&+ O(n^{-1}h^2)
\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(A,C) &= -R(K)\mu_2(K)\left[\frac{1}{2}\left\{m(x)^2+\sigma^2(x)\right\}f(x)^{-2}f''(x)\right.\\ &+ \left.\varphi_1(x)m(x)f(x)^{-1}\right]n^{-1}h+O(n^{-1}h^2). \qquad (9)\end{aligned}$$

Some auxiliary covariances are needed:

$$\begin{aligned}&\operatorname{cov}\left\{K_h(x-X_1),Y_1K_h(x-X_1)K_h(x-X_2)\right\}\\ =\ &\operatorname{cov}\left\{K_h(x-X_1),Y_1K_h(x-X_1)\right\}\mathrm{E}\left\{K_h(x-X_1)\right\}\\ =\ &R(K)m(x)f(x)^2h^{-1}-m(x)f(x)^3+\left\{\mu_2(K^2)\varphi_1(x)f(x)^2\right.\\ +\ &\left.\frac{1}{2}R(K)\mu_2(K)m(x)f(x)f''(x)\right\}h+O(h^2),\end{aligned}$$

$$\begin{aligned}&\operatorname{cov}\left\{K_h(x-X_1),Y_2K_h(x-X_2)K_h(x-X_1)\right\}\\ =\ &\operatorname{var}\left\{K_h(x-X_1)\right\}\mathrm{E}\left\{Y_1K_h(x-X_1)\right\}\\ =\ &R(K)m(x)f(x)^2h^{-1}-m(x)f(x)^3\\ +\ &\left\{\frac{1}{2}\mu_2(K^2)m(x)f(x)f''(x)+R(K)\mu_2(K)\varphi_1(x)f(x)^2\right\}h\\ +\ &O(h^2).\end{aligned}$$

Hence,

$$\begin{aligned}\operatorname{cov}(\hat{e},\hat{a}\hat{e}) &= n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\operatorname{cov}\left\{K_h(x-X_i),Y_jK_h(x-X_j)K_h(x-X_k)\right\}\\ &= n^{-1}\left[\operatorname{cov}\left\{K_h(x-X_1),Y_1K_h(x-X_1)K_h(x-X_2)\right\}\right.\\ &+ \left.\operatorname{cov}\left\{K_h(x-X_1),Y_2K_h(x-X_2)K_h(x-X_1)\right\}\right]\\ &+ n^{-2}\operatorname{cov}\left\{K_h(x-X_1),K_h(x-X_1)^2Y_1\right\}\\ &= 2R(K)m(x)f(x)^2n^{-1}h^{-1}-2m(x)f(x)^3n^{-1}\\ &+ \left\{\mu_2(K^2)\varphi_1(x)f(x)^2+\frac{1}{2}\mu_2(K^2)m(x)f(x)f''(x)\right.\\ &+ \left.\frac{1}{2}R(K)\mu_2(K)m(x)f(x)f''(x)+R(K)\mu_2(K)\varphi_1(x)f(x)^2\right\}n^{-1}h\\ &+ O(n^{-1}h^2+n^{-2}h^{-2})\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(B,C) &= R(K)\mu_2(K)\left\{\frac{1}{2}m(x)^2f(x)^{-2}f''(x)+\varphi_1(x)m(x)f(x)^{-1}\right\}n^{-1}h\\ &+ O(n^{-1}h^2+n^{-2}h^{-2}). \qquad (10)\end{aligned}$$

Another covariance term is needed:

$$\begin{aligned} & \operatorname{cov}\left\{Y_1K_h(x-X_1), K_h(x-X_1)K_h(x-X_2)\right\} \\ = \; & \operatorname{cov}\left\{Y_1K_h(x-X_1), K_h(x-X_1)\right\} \mathrm{E}\left\{K_h(x-X_1)\right\} \\ = \; & R(K)m(x)f(x)^2h^{-1} - m(x)f(x)^3 + \Big\{\mu_2(K^2)\varphi_1(x)f(x)^2 \\ + \; & \frac{1}{2}R(K)\mu_2(K)m(x)f(x)f''(x)\Big\}\, h + O(h^2). \end{aligned}$$

Therefore,

$$\begin{aligned} \operatorname{cov}(\hat{a}, \hat{e}^2) &= n^{-3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \operatorname{cov}\left\{Y_iK_h(x-X_i), K_h(x-X_j)K_h(x-X_k)\right\} \\ &= 2n^{-1}\operatorname{cov}\left\{Y_1K_h(x-X_1), K_h(X-X_1)K_h(x-X_2)\right\} \\ &+ O(n^{-1}h^2 + n^{-2}h^{-2}) \\ &= 2R(K)m(x)f(x)^2n^{-1}h^{-1} - 2m(x)f(x)^3n^{-1} + \big\{2\mu_2(K^2)\varphi_1(x)f(x)^2 \\ &+ R(K)\mu_2(K)m(x)f(x)f''(x)\big\}\, n^{-1}h + O(n^{-1}h^2 + n^{-2}h^{-2}) \end{aligned}$$

and

$$\operatorname{cov}(A, D) = R(K)\mu_2(K)m(x)^2f(x)^{-2}f''(x)n^{-1}h + O(n^{-1}h^2 + n^{-2}h^{-2}). \tag{11}$$

The following covariance is also needed:

$$\begin{aligned} \operatorname{cov}\left\{K_h(x-X_1), K_h(x-X_1)K_h(x-X_2)\right\} &= \operatorname{var}\left\{K_h(x-X_1)\right] \mathrm{E}\left[K_h(x-X_1)\right\} \\ &= R(K)f(x)^2h^{-1} - f(x)^3 \\ &+ \left\{\frac{1}{2}\mu_2(K^2)f(x)f''(x)\right. \\ &+ \left.\frac{1}{2}R(K)\mu_2(K)f(x)f''(x)\right\} h + O(h^2). \end{aligned}$$

Similarly,

$$\begin{aligned} \operatorname{cov}(\hat{e}, \hat{e}^2) &= n^{-3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \operatorname{cov}\left\{K_h(x-X_i), K_h(x-X_j)K_h(x-X_k)\right\} \\ &= 2n^{-1}\operatorname{cov}\left\{K_h(x-X_1), K_h(x-X_1)K_h(x-X_2)\right\} \\ &+ O(n^{-1}h^2 + n^{-2}h^{-2}) \\ &= 2R(K)f(x)^2n^{-1}h^{-1} - 2f(x)^3n^{-1} \\ &+ \left\{\mu_2(K^2)f(x)f''(x) + R(K)\mu_2(K)f(x)f''(x)\right\} n^{-1}h \\ &+ O(n^{-1}h^2 + n^{-2}h^{-2}) \end{aligned}$$

and

$$\operatorname{cov}(B, D) = -R(K)\mu_2(K)m(x)^2f(x)^{-2}f''(x)n^{-1}h + O(n^{-1}h^2 + n^{-2}h^{-2}). \tag{12}$$

Hence,

$$\begin{aligned}\mathrm{var}(\hat{a}\hat{e}) &= R(K)\left\{4m(x)^2+\sigma^2(x)\right\}f(x)^3n^{-1}h^{-1}-4m(x)^2f(x)^4n^{-1}\\ &+ \left[\mu_2(K^2)\varphi_3(x)f(x)^2+R(K)\mu_2(K)\left\{2m(x)^2+\sigma^2(x)\right\}f(x)^2f''(x)\right.\\ &+ 4R(K)\mu_2(K)\varphi_1(x)m(x)f(x)^3+2\mu_2(K^2)\varphi_1(x)m(x)f(x)^3\\ &+ \left.\frac{1}{2}\mu_2(K^2)m(x)^2f(x)^2f''(x)\right]n^{-1}h+O\left(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}\right)\end{aligned}$$

and

$$\mathrm{var}(C)=O\left(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}\right). \tag{13}$$

The remaining variances and covariances were not explicitly calculated because they are clearly negligible with respect to $n^{-1}h$. In particular, $\mathrm{var}(D)$ and $\mathrm{cov}(C,D)$ are both $O(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3})$.

Therefore, plugging (6)–(13) into (5) yields

$$\begin{aligned}\mathrm{var}\left\{\tilde{m}_h(x)\right\} &= R(K)\sigma^2(x)f(x)^{-1}n^{-1}h^{-1}\\ &+ \left[\mu_2(K^2)f(x)^{-2}\left\{\varphi_3(x)+\frac{1}{2}m(x)^2f''(x)-2\varphi_1(x)m(x)f(x)\right\}\right.\\ &- \left.R(K)\mu_2(K)\sigma^2(x)f(x)^{-2}f''(x)\right]n^{-1}h\\ &+ O(n^{-1}h^2+n^{-2}h^{-2}+n^{-3}h^{-3}).\end{aligned}$$

Lemma 3.2 provides expressions for the first and second order terms of both the expectation and variance of $\widetilde{CV}_n'(h)$, where $\widetilde{CV}_n(h)$ is given in equation (11) of the main paper.

Lemma 3.2. *Let us define*

$$\begin{aligned}A_1 &= 12\mu_2(K)\mu_4(K)\int f(x)^{-1}\left\{\frac{1}{24}m^{(4)}(x)f(x)+\frac{1}{6}m'''(x)f'(x)+\frac{1}{4}m''(x)f''(x)\right.\\ &+ \left.\frac{1}{6}m'(x)f'''(x)\right\}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right\}dx\\ &- 6\mu_2(K)^3\int f''(x)f(x)^{-2}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right\}^2,\\ A_2 &= \mu_2\left(K^2\right)\int f(x)^{-1}\left[\frac{1}{2}f''(x)\sigma^2(x)+f'(x)(\sigma^2)'(x)\right.\\ &+ \left.f(x)\left\{\frac{1}{2}(\sigma^2)''(x)+m'(x)^2\right\}\right]dx\\ &- R(K)\mu_2(K)\int\sigma^2(x)f''(x)f(x)^{-1}\,dx,\\ R_1 &= 32R(K)^2\mu_2(K)^2\int\sigma^2(x)f(x)^{-1}\left\{\frac{1}{4}m''(x)^2f(x)^2+m'(x)m''(x)f(x)f'(x)\right.\\ &+ \left.m'(x)^2f'(x)^2\right\}dx,\\ R_2 &= 4\mu_2\left\{(K')^2\right\}\int\sigma^2(x)^2\,dx.\end{aligned}$$

Then, under assumptions A1–A4, *and assuming that* B_1, V_1, A_1, A_2, R_1 *and* R_2 *exist finite:*

$$\begin{aligned} \mathrm{E}\left\{\widetilde{CV}'_n(h)\right\} &= 4B_1h^3 - V_1n^{-1}h^{-2} + A_1h^5 + A_2n^{-1} + O\left(h^7 + n^{-1}h^2\right), && (14)\\ \mathrm{var}\left\{\widetilde{CV}'_n(h)\right\} &= R_1n^{-1}h^2 + R_2n^{-2}h^{-3} + O\left(n^{-1}h^4 + n^{-2}h^{-1}\right). && (15) \end{aligned}$$

where B_1 *and* V_1 *are the main terms of the bias and the variance of the MISE of the Nadaraya–Watson estimator, given by:*

$$\begin{aligned} B_1 &= \frac{1}{4}\mu_2(K)^2\int\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\}^2 f(x)\,dx,\\ V_1 &= R(K)\int\sigma^2(x)\,dx. \end{aligned}$$

Proof. For the sake of simplicity, we will denote by "$Z(h,n) \overset{2}{=}$" the second order terms of a function $Z(h,n)$. For example, if $Z(h,n) = a_0 + a_1h + a_2h^3 + o\left(h^3\right)$, for some constants a_0, a_1 and a_2, then we would denote $Z(h,n) \overset{2}{=} a_1h$.

If we define

$$\begin{aligned} \alpha_1(u) &= K(u) + uK'(u),\\ \alpha_{1h}(u) &= h^{-1}\alpha_1\left(\frac{u}{h}\right),\\ \Gamma_1(u,v) &= 2K(u)K(v) + K(u)K'(v)v + K(v)K'(u)u,\\ \Gamma_{1h}(u,v) &= h^{-1}\Gamma_1\left(\frac{u}{h},\frac{v}{h}\right)\\ \beta_1(u,v) &= K(u)K(v) + K(u)K'(v)v,\\ \beta_{1h}(u,v) &= h^{-1}\beta\left(\frac{u}{h},\frac{v}{h}\right), \end{aligned}$$

then $\widetilde{CV}'_n(h)$ can be expressed as follows:

$$\begin{aligned} \widetilde{CV}'_n(h) = \frac{2}{n}\sum_{i=1}^{n}\Bigg[& m(X_i) - Y_i + \frac{1}{(n-1)^4hf(X_i)^4}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq i}}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\sum_{\substack{s=1\\s\neq i}}^{n} K_h\left(X_i - X_j\right)\\ & \left\{Y_j - m(X_i)\right\}\left\{2f(X_i) - K_h\left(X_i - X_k\right)\right\}\left\{Y_l - m(X_i)\right\}\\ & \left\{-2f(X_i)\alpha_{1h}\left(X_i - X_l\right) + h^{-1}\Gamma_{1h}\left(X_i - X_l, X_i - X_s\right)\right\}\Bigg] \qquad (16) \end{aligned}$$

and so

$$\mathrm{E}\left\{\widetilde{CV}'_n(h)\right\} = \frac{2}{(n-1)^4h}\mathrm{E}\left\{\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\left(\Lambda_{11}^{j} + \Lambda_{12}^{jk}\right)\left(\Lambda_{21}^{l} + \Lambda_{22}^{ls}\right)\right\}, \qquad (17)$$

where

$$\begin{aligned}
\Lambda_{11}^{j} &= 2f(X_1)^{-3}K_h\left(X_1-X_j\right)\left\{Y_j-m(X_1)\right\},\\
\Lambda_{12}^{jk} &= -f(X_1)^{-4}K_h\left(X_1-X_j\right)K_h\left(X_1-X_k\right)\left\{Y_j-m(X_1)\right\},\\
\Lambda_{21}^{l} &= -2f(X_1)\alpha_{1h}\left(X_1-X_l\right)\left\{Y_l-m(X_1)\right\},\\
\Lambda_{22}^{ls} &= h^{-1}\Gamma_{1h}\left(X_1-X_l,X_1-X_s\right)\left\{Y_l-m(X_1)\right\}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{21}^{l}\right) &= (n-1)^2\left\{(n-1)\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{2}\right)\right.\\
&+ \left.(n-1)(n-2)\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{3}\right)\right\}. \qquad (18)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{2}\right) &= -4\mathrm{E}\left[f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_2\right)\left\{Y_2-m(X_1)\right\}^2\right]\\
&= -4\mathrm{E}\left(f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_2\right)\right.\\
&\quad \left.\left[\sigma^2(X_2)+\left\{m(X_2)-m(X_1)\right\}^2\right]\right)\\
&= -4h^{-1}\iint f(x_1)^{-1}K(u)\alpha_1(u)\left[\sigma^2(x_1-hu)+\left\{m(x_1-hu)-m(x_1)\right\}^2\right]\\
&\quad f(x_1-hu)\,dx_1du\\
&\overset{2}{=} -4h^{-1}\iint f(x_1)^{-1}K(u)\alpha_1(u)h^2u^2f(x_1)\varphi_4(x_1)\,dx_1du\\
&= 2\mu_2\left(K^2\right)h\int\varphi_4, \qquad (19)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{3}\right) &= -4\mathrm{E}\left[f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_3\right)\left\{Y_2-m(X_1)\right\}\left\{Y_3-m(X_1)\right\}\right]\\
&= -4\mathrm{E}\left[f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_3\right)\left\{m(X_2)-m(X_1)\right\}\right.\\
&\quad \left.\left\{m(X_3)-m(X_1)\right\}\right]\\
&= -4\iiint f(x_1)^{-1}K(u)\alpha_1(v)\left\{m(x_1-hu)-m(x_1)\right\}\left\{m(x_1-hv)-m(x_1)\right\}\\
&\quad f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} -4\iiint f(x_1)^{-1}K(u)\alpha_1(v)h^6\left(u^2v^4+u^4v^2\right)f(x_1)^2\varphi_6(x_1)\varphi_7(x_1)\,dx_1dudv\\
&= 24\mu_2(K)\mu_4(K)h^6\int\varphi_6\varphi_7f, \qquad (20)
\end{aligned}$$

where

$$\begin{aligned}\varphi_4(x) &= f(x)^{-1}\left[\frac{1}{2}f''(x)\sigma^2(x)+f'(x)\sigma^{2\prime}(x)+f(x)\left\{\frac{1}{2}\sigma^{2\prime\prime}(x)+m'(x)^2\right\}\right],\\ \varphi_6(x) &= f(x)^{-1}\left\{\frac{1}{24}m^{4)}(x)f(x)+\frac{1}{6}m'''(x)f'(x)+\frac{1}{4}m''(x)f''(x)+\frac{1}{6}m'(x)f'''(x)\right\},\\ \varphi_7(x) &= f(x)^{-1}\left\{\frac{1}{2}m''(x)f(x)+m'(x)f'(x)\right\},\end{aligned}$$

and we have used the fact that

$$\begin{aligned}\int K(u)\alpha_1(u)u^i\,du &= \frac{1-i}{2}\mu_i\left(K^2\right),\\ \iint K(u)\alpha_1(v)u^iv^j\,dudv &= -j\mu_i(K)\mu_j(K).\end{aligned}$$

Then, plugging (19) and (20) into (18) we get:

$$\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{21}^{l}\right) \overset{2}{=} 2\mu_2\left(K^2\right)n^3h\int\varphi_4+24n^4h^6\int\varphi_6\varphi_7 f. \tag{21}$$

We have

$$\begin{aligned}\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{21}^{l}\right) &= (n-1)\left[(n-1)(n-2)(n-3)\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{4}\right)\right.\\ &+ (n-1)(n-2)\left\{\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{21}^{3}\right)+\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{2}\right)\right.\\ &+ \left.\left.\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{3}\right)\right\}\right]+o\left(n^3h+n^4h^6\right).\end{aligned} \tag{22}$$

Now,

$$\begin{aligned}\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{4}\right) &= 2\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\alpha_{1h}\left(X_1-X_4\right)\right.\\ &\quad\left.\left\{Y_2-m(X_1)\right\}\left\{Y_4-m(X_1)\right\}\right]\\ &= 2\iiiint f(x_1)^{-2}K(u)K(v)\alpha_1(w)\left\{m(x_1-hv)-m(x_1)\right\}\\ &\quad\left\{m(x_1-hw)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)f(x_1-hw)\\ &\quad dx_1dudvdw\\ &\overset{2}{=} 2\iiiint f(x_1)^{-2}K(u)K(v)\alpha_1(w)h^6\left\{(w^4v^2+w^2v^4)f(x_1)^3\right.\\ &\quad\left.\varphi_6(x_1)\varphi_7(x_1)+w^2u^2v^2\frac{1}{2}f''(x_1)f(x_1)^2\varphi_7(x_1)^2\right\}f(x_1-hu)\\ &\quad f(x_1-hv)f(x_1-hw)\,dx_1dudvdw\\ &= -2h^6\left\{6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7f+\mu_2(K)^3\int\varphi_7^2f''\right\},\end{aligned} \tag{23}$$

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{21}^{3}\right) &= 2\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)^2\alpha_{1h}\left(X_1-X_3\right)\left\{Y_2-m(X_1)\right\}\right.\\
&\quad \left\{Y_3-m(X_1)\right\}]\\
&= 2h^{-1}\iiint f(x_1)^{-2}K(u)^2\alpha_1(v)\left\{m(x_1-hu)-m(x_1)\right\}\\
&\quad \left\{m(x_1-hv)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} O\left(h^3\right), \qquad (24)
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{2}\right) &= 2\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\right.\\
&\quad \left.\left\{Y_2-m(X_1)\right\}^2\right]\\
&= 2h^{-1}\iint f(x_1)^{-2}K(u)\alpha_1(u)K(v)\left[\sigma^2(x_1-hu)\right.\\
&+ \left.\left\{m(x_1-hu)-m(x_1)\right\}^2\right]f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} 2h^{-1}\iiint f(x_1)^{-2}K(u)\alpha_1(u)K(v)h^2\left\{u^2f(x_1)^2\varphi_4(x_1)\right.\\
&+ \left.v^2\frac{1}{2}\sigma^2(x_1)f(x_1)f''(x_1)\right\}dx_1dudv\\
&= h\left\{\frac{1}{2}R(K)\mu_2(K)\int\sigma^2f''f^{-1}-\mu_2\left(K^2\right)\int\varphi_4\right\} \qquad (25)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{3}\right) &= 2\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\alpha_{1h}\left(X_1-X_3\right)\right.\\
&\quad \left\{Y_2-m(X_1)\right\}\left\{Y_3-m(X_1)\right\}]\\
&= 2h^{-1}\iiint f(x_1)^{-2}K(u)K(v)\alpha_1(v)\left\{m(x_1-hu)-m(x_1)\right\}\\
&\quad \left\{m(x_1-hv)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} O\left(h^3\right), \qquad (26)
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
\iiint K(u)K(v)\alpha_1(w)u^iv^jw^k\,dudvdw &= -k\mu_i(K)\mu_j(K)\mu_k(K),\\
\iint K(u)^2\alpha_1(v)u^iv^j\,dudv &= -j\mu_i\left(K^2\right)\mu_j(K),\\
\iint K(u)\alpha_1(u)K(v)u^iv^j\,dudv &= \frac{1-i}{2}\mu_i\left(K^2\right)\mu_j(K).
\end{aligned}$$

Then, plugging (23), (24), (25) and (26) into (22) we get

$$
\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{21}^{l}\right) &\overset{2}{=} -2n^4h^6\left\{6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+\mu_2(K)^3\int\varphi_7^2 f''\right\} \\
&+ n^3h\left\{\frac{1}{2}R(K)\mu_2(K)\int\sigma^2 f''f^{-1}-\mu_2\left(K^2\right)\int\varphi_4\right\} \quad (27)
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{22}^{ls}\right) &= (n-1)^2(n-2)(n-3)\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{34}\right) \\
&+ (n-1)^2(n-2)\left\{\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{23}\right)+\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{32}\right)\right. \\
&+ \left.\mathrm{E}\left(\Lambda_{11}^{3}\Lambda_{22}^{22}\right)\right\}+o\left(n^3h+n^4h^6\right). \quad (28)
\end{aligned}
$$

Now,

$$
\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{34}\right) &= 2h^{-1}\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_3,X_1-X_4\right)\right. \\
&\quad \left.\{Y_2-m(X_1)\}\{Y_3-m(X_1)\}\right] \\
&= 2\iiiint f(x_1)^{-2}K(u)\Gamma_1(v,w)\{m(x_1-hu)-m(x_1)\} \\
&\quad \{m(x_1-hv)-m(x_1)\}f(x_1-hu)f(x_1-hv)f(x_1-hw) \\
&\quad dx_1dudvdw \\
&\overset{2}{=} 2\iiiint f(x_1)^{-2}K(u)\Gamma_1(v,w)h^6\left\{(u^2v^4+u^4v^2)f(x_1)^3\varphi_6(x_1)\varphi_7(x_1)\right. \\
&+ \left. w^2u^2v^2\frac{1}{2}f''(x_1)f(x_1)^2\varphi_7(x_1)^2\right\}dx_1dudvdw \\
&= -h^6\left\{12\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+4\mu_2(K)^3\int\varphi_7^2 f''\right\}, \quad (29)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{23}\right) &= 2h^{-1}\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_2,X_1-X_4\right)\right. \\
&\quad \left.\{Y_2-m(X_1)\}^2\right] \\
&= 2h^{-1}\iiint f(x_1)^{-2}K(u)\Gamma_1(u,v)\left[\sigma^2(x_1-hu)\right. \\
&+ \left.\{m(x_1-hu)-m(x_1)\}^2\right]f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} 2h^{-1}\iiint f(x_1)^{-2}K(u)\Gamma_1(u,v)h^2\left\{u^2f(x_1)^2\varphi_4(x_1)\right. \\
&+ \left. v^2\frac{1}{2}\sigma^2(x_1)f''(x_1)f(x_1)\right\}dx_1dudv \\
&= -h\left\{\mu_2\left(K^2\right)\int\varphi_4+\frac{3}{2}R(K)\mu_2(K)\int\sigma^2f''f^{-1}\right\}, \quad (30)
\end{aligned}
$$

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{22}^{32}\right) &= 2h^{-1}\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_3,X_1-X_2\right)\right.\\
&\quad\left.\left\{Y_2-m(X_1)\right\}\left\{Y_3-m(X_1)\right\}\right]\\
&= 2h^{-1}f(x_1)^{-2}K(u)\Gamma_1(v,u)\left\{m(x_1-hu)-m(x_1)\right\}\\
&\quad\left\{m(x_1-hv)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} O\left(h^3\right) \qquad (31)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{3}\Lambda_{22}^{22}\right) &= 2h^{-1}\mathrm{E}\left[f(X_1)^{-3}K_h\left(X_1-X_3\right)\Gamma_{1h}\left(X_1-X_2,X_1-X_2\right)\right.\\
&\quad\left.\left\{Y_2-m(X_1)\right\}\left\{Y_3-m(X_1)\right\}\right]\\
&= 2h^{-1}f(x_1)^{-2}K(u)\Gamma_1(v,v)\left\{m(x_1-hu)-m(x_1)\right\}\\
&\quad\left\{m(x_1-hv)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv\\
&\overset{2}{=} O\left(h^3\right), \qquad (32)
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
\iiint K(u)\Gamma_1(v,w)u^iv^jw^k\,dudvdw &= (-k-j)\mu_i(K)\mu_j(K)\mu_k(K),\\
\iint K(u)\Gamma_1(u,v)u^iv^j\,dudv &= \frac{1-i-2j}{2}\mu_i\left(K^2\right)\mu_j(K),\\
\Gamma_1(u,v) &= \Gamma_1(v,u),\\
\iint K(u)\Gamma_1(v,v)u^iv^j\,dudv &= (1-j)\mu_i(K)\mu_j\left(K^2\right).
\end{aligned}$$

Then, plugging (29), (30), (31) and (32) into (28), we get

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{22}^{ls}\right) &\overset{2}{=} -n^4h^6\left\{12\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7f+4\mu_2(K)^3\int\varphi_7^2f''\right\}\\
&- \;n^3h\left\{\mu_2\left(K^2\right)\int\varphi_4+\frac{3}{2}R(K)\mu_2(K)\int\sigma^2f''f^{-1}\right\}.(33)
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{22}^{ls}\right) &= (n-1)(n-2)(n-3)(n-4)\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{45}\right)\\
&+ (n-1)(n-2)(n-3)\left\{\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{22}^{34}\right)+\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{24}\right)\right.\\
&+ \mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{42}\right)+\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{24}\right)+\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{42}\right)\\
&+ \left.\mathrm{E}\left(\Lambda_{12}^{34}\Lambda_{22}^{22}\right)\right\}+o\left(n^3h+n^4h^6\right). \qquad (34)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{45}\right) &= -h^{-1}\mathrm{E}\left[f(X_1)^{-4}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\Gamma_{1h}\left(X_1-X_4,X_1-X_5\right)\right.\\
&\quad \left.\left\{Y_2-m(X_1)\right\}\left\{Y_4-m(X_1)\right\}\right]\\
&= -\int\ldots\int f(x_1)^{-3}K(u)K(v)\Gamma_1(w,z)\left\{m(x_1-hu)-m(x_1)\right\}\\
&\quad \left\{m(x_1-hw)-m(x_1)\right\}f(x_1-hu)f(x_1-hv)f(x_1-hw)f(x_1-hz)\\
&\quad dx_1dudvdwdz\\
&\overset{2}{=} -\int\ldots\int f(x_1)^{-3}K(u)K(v)\Gamma_1(w,z)h^6\left\{(u^4w^2+u^2w^4)f(x_1)^4\right.\\
&\quad \left.\varphi_6(x_1)\varphi_7(x_1)+(u^2w^2v^2+u^2w^2z^2)\frac{1}{2}f(x_1)^3f''(x_1)\varphi_7(x_1)^2\right\}\\
&\quad dx_1dudvdwdz\\
&= h^6\left\{6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7f+3\mu_2(K)^3\int\varphi_7^2f''\right\} \qquad (35)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{24}\right) &= -h^{-1}\mathrm{E}\left[f(X_1)^{-4}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\Gamma_{1h}\left(X_1-X_2,X_1-X_4\right)\right.\\
&\quad \left.\left\{Y_2-m(X_1)\right\}^2\right]\\
&= -\iiiint f(x_1)^{-3}K(u)K(v)\Gamma_1(u,w)\left[\sigma^2(x_1-hu)\right.\\
&+ \left.\left\{m(x_1-hu)-m(x_1)\right\}^2\right]\\
&\quad f(x_1-hu)f(x_1-hv)f(x_1-hw)f(x_1-hz)\,dx_1dudvdw\\
&\overset{2}{=} -h^{-1}\iiiint f(x_1)^{-3}K(u)K(v)\Gamma_1(u,w)h^2\left\{(v^2+w^2)\frac{1}{2}f(x_1)^2\sigma^2(x_1)\right.\\
&\quad \left.f''(x_1)+u^2f(x_1)^3\varphi_4(x_1)\right\}\,dx_1dudvdw\\
&= \frac{1}{2}h\left\{R(K)\mu_2(K)\int\sigma^2f''f^{-1}+\mu_2\left(K^2\right)\int\varphi_4\right\}, \qquad (36)
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
\iiiint K(u)K(v)\Gamma_1(w,z)u^iv^jw^kz^l\,dudvdwdz &= (-k-l)\mu_i(K)\mu_j(K)\mu_k(K)\mu_l(K),\\
\iiint K(u)K(v)\Gamma_1(u,w)u^iv^jw^k\,dudvdw &= \frac{1-i-2k}{2}\mu_i\left(K^2\right)\mu_j(K)\mu_k(K).
\end{aligned}$$

Also, it is straightforward to see that the second order terms of $\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{22}^{34}\right)$, $\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{42}\right)$, $\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{24}\right)$, $\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{42}\right)$, $\mathrm{E}\left(\Lambda_{12}^{34}\Lambda_{22}^{22}\right)$ are $O\left(h^3\right)$.

Then, plugging (35) and (36) into (34), we get

$$\begin{aligned}\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{22}^{ls}\right) &\overset{2}{=} n^4h^6\left\{6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+3\mu_2(K)^3\int\varphi_7^2 f''\right\}\\ &+ \frac{1}{2}n^3h\left\{R(K)\mu_2(K)\int\sigma^2 f''f^{-1}+\mu_2\left(K^2\right)\int\varphi_4\right\}(37)\end{aligned}$$

Finally, plugging (21), (27), (33), and (37) into (17) yields:

$$\begin{aligned}\mathrm{E}\left\{\widetilde{CV}_n'(h)\right\} &= h^5\left\{12\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f-6\mu_2(K)^3\int\varphi_7^2 f''\right\}\\ &+ n^{-1}\left\{\mu_2\left(K^2\right)\int\varphi_4-R(K)\mu_2(K)\int\sigma^2 f''f^{-1}\right\},\end{aligned}$$

which, considering the definitions of φ_4, φ_6 and φ_7 given above, matches the second order terms of $\mathrm{E}\left\{\widetilde{CV}_n'(h)\right\}$ given (14) in Lemma 3.2. Regarding the first order terms of $\mathrm{E}\left\{\widetilde{CV}_n'(h)\right\}$ and as already mentioned, it is well known that these coincide with the main term of $\tilde{M}_n'(h)$.

As for the variance of $\widetilde{CV}_n'(h)$, recall that we are only interested in obtaining its first-order terms. Thus, instead of working with the quadratic approximation of $\hat{m}_h$, namely $\tilde{m}_h$, given in equation (9) of the main paper, we can employ a simpler, linear approximation of $\hat{m}_h$, denoted by $\bar{m}_h$. This linear approximation of the Nadaraya-Watson estimator was already proposed in Barbeito (2020) and it is given in equation (10) of the main paper. Its expression is:

$$\bar{m}_h(x)=m(x)+\frac{1}{nf(x)}\sum_{i=1}^{n}K_h\left(x-X_i\right)\left\{Y_i-m(x)\right\}.$$

Let us now define

$$\begin{aligned}\overline{CV}_n(h) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\bar{m}_h^{(-i)}(X_i)-Y_i\right\}^2,\\ P_{ij} &= \frac{Y_i-m(X_i)}{f(X_i)}\{Y_j-m(X_i)\}\alpha_{1h}(X_i-X_j),\\ Q_{ijk} &= f(X_i)^{-2}\{Y_j-m(X_i)\}\{Y_k-m(X_i)\}\beta_{1h}(X_i-X_j,X_i-X_k).\end{aligned}$$

Then,

$$\operatorname{var}\left\{\overline{CV}_n'(h)\right\}=\frac{4}{n^2(n-1)^4h^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i}}^{n}\sum_{l=1}^{n}\sum_{\substack{r=1\\ r\neq l}}^{n}\sum_{\substack{s=1\\ s\neq l}}^{n}C_{ijklrs},$$

where

$$C_{ijklrs}=\operatorname{cov}\left(P_{ij},P_{lr}\right)-h^{-1}\operatorname{cov}\left(P_{ij},Q_{lrs}\right)-h^{-1}\operatorname{cov}\left(P_{lr},Q_{ijk}\right)+h^{-2}\operatorname{cov}\left(Q_{ijk},Q_{lrs}\right).$$

By counting the possible cases and using $C_{122345} = C_{123455} = 0$, we get

$$\begin{aligned}
\mathrm{var}\left\{\widetilde{CV}'_n(h)\right\} &= \frac{4}{n^2(n-1)^4h^2}\left\{n(n-1)(n-2)(n-3)(n-4)(n-5)C_{123456}\right. \\
&+ n(n-1)(n-2)(n-3)(n-4)\left(C_{123145} + 2C_{123415} + 2C_{123451}\right. \\
&+ \left. 2C_{123455} + C_{123425} + 2C_{123452} + C_{123453}\right) + n(n-1)(n-2)(n-3) \\
&\quad \left(2C_{122134} + C_{123124} + 2C_{123142} + C_{123143} + 2C_{122314} + C_{123214}\right. \\
&+ 2C_{123412} + 2C_{123314} + 2C_{123413} + 2C_{122341} + 2C_{123421} \\
&+ C_{123341} + 2C_{123431} + C_{122344} + C_{123423} + C_{123432} + 2C_{123411} \\
&+ \left. 2C_{122324} + 2C_{122342}\right) + n(n-1)(n-2)\left(C_{122322} + C_{122133}\right. \\
&+ C_{123123} + C_{123132} + C_{123213} + 2C_{123312} + C_{123321} + 2C_{122311} \\
&+ \left. 2C_{123211} + 2C_{123311} + 2C_{123122} + 2C_{123322} + 2C_{122132} + 2C_{122312}\right) \\
&+ \left. n(n-1)\left(C_{122122} + C_{122211}\right)\right\}. \qquad (38)
\end{aligned}$$

Among the previous covariances, it can be argued that the only ones that contribute to the dominant term of $\mathrm{var}\left\{\widetilde{CV}'_n(h)\right\}$ are C_{123245}, C_{123425}, C_{123124} and C_{123145}. Before we continue and with the intention of facilitating the calculations of the four C_{ijklrs} that we need, let us obtain general expressions for each of the summands that make up C_{ijklrs}. Since

$$\mathrm{E}\left(P_{ij} \mid X_i, X_j, X_l, X_r, Y_j, Y_r\right) = 0$$

and

$$\mathrm{cov}\left\{Y_i - m(X_i), Y_l - m(X_l) \mid X_i, X_l\right\} = \delta_{il}\sigma^2(X_i)$$

then

$$\begin{aligned}
\mathrm{cov}\left(P_{ij}, P_{lr}\right) &= \mathrm{E}\left\{\mathrm{cov}\left(P_{ij}, P_{lr} \mid X_i, X_j, X_l, X_r, Y_j, Y_r\right)\right\} \\
&= \mathrm{E}\left[f(X_i)^{-1}f(X_l)^{-1}\alpha_{1h}(X_i - X_j)\alpha_{1h}(X_l - X_r)\{Y_j - m(X_i)\}\right. \\
&\quad \left.\{Y_r - m(X_l)\}\mathrm{cov}\left(Y_i - m(X_i), Y_l - m(X_l) \mid X_i, X_l\right)\right] \\
&= \delta_{il}\mathrm{E}\left[f(X_i)^{-2}\alpha_{1h}(X_i - X_j)\alpha_{1h}(X_i - X_r)\{Y_j - m(X_i)\}\right. \\
&\quad \left.\{Y_r - m(X_i)\}\sigma^2(X_i)\right].
\end{aligned}$$

Let us now consider the covariance

$$\begin{aligned}
\mathrm{cov}\left(P_{ij}, Q_{lrs}\right) &= \mathrm{E}\left[f(X_i)^{-1}f(X_l)^{-2}\alpha_{1h}(X_i - X_j)\beta_{1h}(X_l - X_r, X_l - X_s)\right. \\
&\quad \left.\{Y_i - m(X_i)\}\{Y_j - m(X_i)\}\{Y_r - m(X_l)\}\{Y_s - m(X_l)\}\right].
\end{aligned}$$

If $r, s \neq i$ it is clear that $\mathrm{cov}\left(P_{ij}, Q_{lrs}\right) = 0$. Now, for the cases $r = i$ and $s = i$ (both cases imply $i \neq l$), let us define

$$t = \begin{cases} s, \text{ if } r = i \\ r, \text{ if } s = i \end{cases}$$

and note that

$$\begin{aligned}
\operatorname{cov}\{Y_i - m(X_i), Y_i - m(X_l) \mid X_i, X_l\} &= \operatorname{cov}\{\varepsilon_i, \varepsilon_i + m(X_i) - m(X_l) \mid X_i, X_l\} \\
&= \operatorname{var}(\varepsilon_i \mid X_i) + \operatorname{cov}\{\varepsilon_i, m(X_i) - m(X_l) \mid X_i, X_l\} \\
&= \sigma^2(X_i).
\end{aligned}$$

Then, using the law of total covariance:

$$\begin{aligned}
\operatorname{cov}(P_{ij}, Q_{lrs}) &= \mathrm{E}\left[f(X_i)^{-1} f(X_l)^{-2} \alpha_{1h}(X_i - X_j) \beta_{1h}(X_l - X_r, X_l - X_s)\right. \\
&\quad \{Y_j - m(X_i)\}\{Y_t - m(X_l)\}\operatorname{cov}\{Y_i - m(X_i), Y_i - m(X_l) \mid X_i, X_l\}] \\
&= \mathrm{E}\left[f(X_i)^{-1} f(X_l)^{-2} \alpha_{1h}(X_i - X_j) \beta_{1h}(X_l - X_r, X_l - X_s)\right. \\
&\quad \left.\{Y_j - m(X_i)\}\{Y_t - m(X_l)\}\sigma^2(X_i)\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
&\operatorname{cov}(Q_{ijk}, Q_{lrs}) \\
= \; &\mathrm{E}\{\operatorname{cov}(Q_{ijk}, Q_{lrs} \mid X_i, X_j, X_k, X_l, X_r, X_s)\} \\
+ \; &\operatorname{cov}\{\mathrm{E}(Q_{ijk} \mid X_i, X_j, X_k, X_l, X_r, X_s), \mathrm{E}(Q_{lrs} \mid X_i, X_j, X_k, X_l, X_r, X_s)\} \\
= \; &\mathrm{E}\left(f(X_i)^{-2} f(X_l)^{-2} \beta_{1h}(X_i - X_j, X_i - X_k) \beta_{1h}(X_l - X_r, X_l - X_s)\right. \\
&\left.\operatorname{cov}[\{Y_j - m(X_i)\}\{Y_k - m(X_i)\}, \{Y_r - m(X_l)\}\{Y_s - m(X_l)\} \mid X_i, X_j, X_k, X_l, X_r, X_s]\right) \\
+ \; &\operatorname{cov}\{\mathrm{E}(Q_{ijk} \mid X_i, X_j, X_k, X_l, X_r, X_s), \mathrm{E}(Q_{lrs} \mid X_i, X_j, X_k, X_l, X_r, X_s)\}.
\end{aligned}$$

Note that, if $\{j, k\} \cap \{r, s\} = \emptyset$, then

$$\operatorname{cov}[\{Y_j - m(X_i)\}\{Y_k - m(X_i)\}, \{Y_r - m(X_l)\}\{Y_s - m(X_l)\} \mid X_i, X_j, X_k, X_l, X_r, X_s] = 0.$$

Now, regarding the term C_{123245}, since $1 \neq 2$ and $4, 5 \neq 1$, we have

$$\operatorname{cov}(P_{12}, P_{24}) = \operatorname{cov}(P_{12}, Q_{245}) = 0$$

and

$$\begin{aligned}
\operatorname{cov}(P_{24}, Q_{123}) &= \mathrm{E}\left[f(X_2)^{-1} f(X_1)^{-2} \alpha_{1h}(X_2 - X_4) \beta_{1h}(X_1 - X_2, X_1 - X_3)\right. \\
&\quad \left.\{Y_4 - m(X_2)\}\{Y_3 - m(X_1)\}\sigma^2(X_2)\right] \\
&= \mathrm{E}\left[f(X_2)^{-1} f(X_1)^{-2} \alpha_{1h}(X_2 - X_4) \beta_{1h}(X_1 - X_2, X_1 - X_3)\right. \\
&\quad \left.\{m(X_4) - m(X_2)\}\{m(X_3) - m(X_1)\}\sigma^2(X_2)\right] \\
&= \iiiint f(x_2)^{-1} f(x_1)^{-2} \alpha_{1h}(x_2 - x_4) \beta_{1h}(x_1 - x_2, x_1 - x_3) \\
&\quad \{m(x_4) - m(x_2)\}\{m(x_3) - m(x_1)\} f(x_1) f(x_2) f(x_3) f(x_4) \, dx_1 dx_2 dx_3 dx_4
\end{aligned}$$

Making the following changes of variable,

$$\begin{cases} x_4 = x_2 - hu_4 \\ x_3 = x_1 - hu_3 \\ x_2 = x_1 - hu_2 \end{cases}$$

and using the fact that

$$\begin{aligned}&\iiint \alpha_1(u_4)\beta_1(u_2,u_3)u_4^iu_2^ju_3^k\,du_4du_2du_3 = ik\mu_i(K)\mu_j(K)\mu_k(K)=0\\ \iff\ & i=0 \text{ or } k=0 \text{ or } (i,\ j \text{ or } k \text{ is an odd number}),\end{aligned}$$

we obtain that

$$\begin{aligned}\operatorname{cov}(P_{24},Q_{123}) &= h\iiiint f(x_1)^{-1}\alpha_1(u_4)\beta_1(u_2,u_3)\{m(x_1-hu_2-hu_4)-m(x_1-hu_2)\}\\ &\quad \{m(x_1-hu_3)-m(x_1)\}\sigma^2(x_1-hu_2)f(x_1-hu_3)\\ &\quad f(x_1-hu_2-hu_4)\,dx_1du_2du_3du_4\\ &= h\iiiint f(x_1)^{-1}\alpha_1(u_4)\beta_1(u_2,u_3)u_4^2u_3^2h^4\left\{\frac{1}{4}m''(x_1)^2\sigma^2(x_1)f(x_1)^2\right.\\ &+ \left. m'(x_1)^2\sigma^2(x_1)f'(x_1)^2+m'(x_1)m''(x_1)\sigma^2(x_1)f(x_1)f'(x_1)\right\}\\ &\quad dx_1du_2du_3du_4+O\left(h^7\right)\\ &= 4\mu_2(K)^2h^5\int f(x)^{-1}\sigma^2(x)\left\{\frac{1}{4}m''(x)^2f(x)^2+m'(x)^2f'(x)^2\right.\\ &+ \left. m'(x)m''(x)f(x)f'(x)\right\}dx+O\left(h^7\right).\end{aligned}$$

Since $\{2,3\}\cap\{4,5\}=\emptyset$,

$$\begin{aligned}\operatorname{cov}(Q_{123},Q_{245}) &= \operatorname{cov}\left[f(X_1)^{-2}\beta_{1h}(X_1-X_2,X_1-X_3)\{m(X_2)-m(X_1)\}\right.\\ &\quad \{m(X_3)-m(X_1)\},f(X_2)^{-2}\beta_{1h}(X_2-X_4,X_2-X_5)\\ &\quad \left.\{m(X_4)-m(X_2)\}\{m(X_5)-m(X_2)\}\right]\\ &= \mathrm{E}\left[f(X_1)^{-2}f(X_2)^{-2}\beta_{1h}(X_1-X_2,X_1-X_3)\beta_{1h}(X_2-X_4,X_2-X_5)\right.\\ &\quad \{m(X_2)-m(X_1)\}\{m(X_3)-m(X_1)\}\{m(X_4)-m(X_2)\}\\ &\quad \left.\{m(X_5)-m(X_2)\}\right]\\ &- \mathrm{E}\left[f(X_1)^{-2}\beta_{1h}(X_1-X_2,X_1-X_3)\{m(X_2)-m(X_1)\}\right.\\ &\quad \left.\{m(X_3)-m(X_1)\}\right]^2\\ &= O\left(h^{10}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}C_{123245} &= -4\mu_2(K)^2h^4\int f(x)^{-1}\sigma^2(x)\left\{\frac{1}{4}m''(x)^2f(x)^2+m'(x)^2f'(x)^2\right.\\ &+ \left. m'(x)m''(x)f(x)f'(x)\right\}\,dx+O\left(h^6\right).\end{aligned}\tag{39}$$

Regarding the term C_{123425}, since $1\neq 4$, $2,3\neq 4$ and $2,5\neq 1$, then

$$\operatorname{cov}(P_{12},P_{42})=\operatorname{cov}(P_{12},Q_{425})=\operatorname{cov}(P_{42},Q_{123})=0.$$

We have that

$$\begin{aligned}
&h^2 \int \dots \int f(x_1)^{-2} f(x_4)^{-2} \beta_{1h}(x_1 - x_2, x_1 - x_3) \beta_{1h}(x_4 - x_2, x_4 - x_5) \\
&\{m(x_3) - m(x_1)\}\{m(x_5) - m(x_4)\} \left[\sigma^2(x_2) + \{m(x_2) - m(x_1)\}\{m(x_2) - m(x_4)\}\right] \\
&f(x_1) f(x_2) f(x_3) f(x_4) f(x_5) dx_1 dx_2 dx_3 dx_4 dx_5 \\
&= h^2 \int \dots \int f(x_1)^{-1} f(x_1 - hu_2 + hu_4)^{-1} \beta_1(u_2, u_3) \beta_1(u_4, u_5) \left\{m(x_1 - hu_3)\right. \\
- \;& \left. m(x_1)\right\} \left\{m(x_1 - hu_2 + hu_4 - hu_5) - m(x_1 - hu_2 + hu_4)\right\} \left[\sigma^2(x_1 - hu_2)\right. \\
+ \;& \left.\{m(x_1 - hu_2) - m(x_1)\}\{m(x_1 - hu_2) - m(x_1 - hu_2 + hu_4)\}\right] f(x_1 - hu_2) \\
&f(x_1 - hu_3) f(x_1 - hu_2 + hu_4 - hu_5) dx_1 du_2 du_3 du_4 du_5 \\
= \;& 4R(K)^2 \mu_2(K)^2 h^6 \int \sigma^2 f(x)^{-1} \left\{\frac{1}{4}(m'')^2 f^2 + m'm''ff' + (m')^2 (f')^2\right\} + O\left(h^8\right),
\end{aligned}$$

where we have made the following change of variables,

$$\begin{cases} x_2 = x_1 - hu_2 \\ x_3 = x_1 - hu_3 \\ x_4 = x_2 + hu_4 \\ x_5 = x_2 - hu_5 \end{cases}$$

and used the fact that

$$\begin{aligned}
&\iiiint \beta_1(u_2, u_3) \beta_1(u_4, u_5) u_2^i u_3^j u_4^k u_5^l \, du_2 du_3 du_4 du_5 = jl\mu_i(K)\mu_j(K)\mu_k(K)\mu_l(K) = 0 \\
&\iff j = 0 \text{ or } l = 0 \text{ or } (i,\ j,\ k \text{ or } l \text{ is an odd number}).
\end{aligned}$$

Therefore,

$$C_{123425} = 8R(K)^2 \mu_2(K)^2 h^4 \int \sigma^2 f(x)^{-1} \left\{\frac{1}{4}(m'')^2 f^2 + m'm''ff' + (m')^2 (f')^2\right\} + O\left(h^6\right). \quad (40)$$

As for the term C_{123124}, since $2, 4 \neq 1$ and $2, 3 \neq 1$, then

$$\operatorname{cov}(P_{12}, Q_{124}) = \operatorname{cov}(P_{12}, Q_{123}) = 0.$$

We have that

$$\begin{aligned}
\operatorname{cov}(P_{12}, P_{12}) &= \mathrm{E}\left[f(X_1)^{-2} \alpha_{1h}(X_1 - X_2)^2 \sigma^2(X_1) \left\{(m(X_2) - m(X_1))^2 + \sigma^2(X_2)\right\}\right] \\
&= h^{-1} \iint f(x_1)^{-1} \alpha_1(u)^2 \sigma^2(x_1) \left[\sigma^2(x_1 - hu) + \{m(x_1 - hu) - m(x_1)\}^2\right] \\
&\quad f(x_1 - hu)\, dx_1 du = \mu_2\left\{(K')^2\right\} h^{-1} \int (\sigma^2)^2 + O(h),
\end{aligned}$$

where we have used the fact that

$$\int \alpha_1(u)^2 u^i \, du = -i\mu_i(K^2) + \mu_{i+2}\left\{(K')^2\right\} = 0 \iff i \text{ is odd}.$$

On the other hand,

$$\begin{aligned}\operatorname{cov}(Q_{123},Q_{124}) &= \mathrm{E}\big(f(X_1)^{-4}\beta_{1h}(X_1-X_2,X_1-X_3)\beta_{1h}(X_1-X_2,X_1-X_4)\\ &\quad \{m(X_3)-m(X_1)\}\{m(X_4)-m(X_1)\}\left[\sigma^2(X_2)+\{m(X_2)-m(X_1)\}^2\right]\big)\\ &- \mathrm{E}\left[f(X_1)^{-2}\beta_{1h}(X_1-X_2,X_1-X_3)\{m(X_2)-m(X_1)\}\right.\\ &\quad \left.\{m(X_3)-m(X_1)\}\right]^2 = O\left(h^5\right).\end{aligned}$$

Therefore,

$$C_{123124} = \mu_2\left\{(K')^2\right\}h^{-1}\int(\sigma^2)^2 + O\left(h\right). \tag{41}$$

Using similar arguments and calculations we get that

$$C_{123145} = 4\mu_2(K)^2h^4\int f^{-1}\sigma^2\left\{\frac{1}{4}(m'')^2f^2+(m')^2(f')^2+m'm''ff'\right\}+O\left(h^6\right). \tag{42}$$

Finally, considering (38), (39), (40), (41) and (42), we obtain (15).

Now, from the following decomposition (equation (13) of the main paper):

$$\begin{aligned}\tilde{h}_{CV,n}-\tilde{h}_{n0} &\approx -\frac{\widetilde{CV}'_n(\tilde{h}_{n0})-\tilde{M}'_n(\tilde{h}_{n0})}{\tilde{M}''_n(\tilde{h}_{n0})}\\ &+ \frac{\left\{\widetilde{CV}'_n(\tilde{h}_{n0})-\tilde{M}'_n(\tilde{h}_{n0})\right\}\left\{\widetilde{CV}''_n(\tilde{h}_{n0})-\tilde{M}''_n(\tilde{h}_{n0})\right\}}{\tilde{M}''_n(\tilde{h}_{n0})^2},\end{aligned} \tag{43}$$

and using Lemmas 3.1 and 3.2, the asymptotic bias and variance of the cross-validation bandwidth that minimizes the modified version of the cross-validation criterion given in equation (11) of the main paper:

$$\widetilde{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{m}_h^{(-i)}(X_i)-Y_i\right\}^2, \tag{44}$$

can be obtained. Theorem 3.1 contains this result.

Theorem 3.1. *Under the assumptions of Lemma 3.2 and assuming that B_2 and V_2 exist finite, the asymptotic bias and the variance of the bandwidth that minimizes* (44) *are:*

$$\begin{aligned}\mathrm{E}\left(\tilde{h}_{CV,n}\right)-\tilde{h}_{n0} &= \mathcal{B}n^{-3/5}+o\left(n^{-3/5}\right),\\ \operatorname{var}\left(\tilde{h}_{CV,n}\right) &= Vn^{-3/5}+o\left(n^{-3/5}\right),\end{aligned}$$

where

$$\begin{aligned}\mathcal{B} &= \frac{6B_2C_0^5+V_2-A_1C_0^5-A_2}{12B_1C_0^2+2V_1C_0^{-3}},\\ V &= \frac{R_1C_0^2+R_2C_0^{-3}}{\left(12B_1C_0^2+2V_1C_0^{-3}\right)^2}.\end{aligned}$$

The constants B_2 and V_2 were defined in p. 8 of the main paper, and the constant C_0 was defined in p. 5 of the main paper.

Proof. From equation (43), it follows that, up to first order,

$$\begin{aligned} \mathrm{E}\left(\tilde{h}_{CV,n}\right) - \tilde{h}_{n0} &= \frac{\tilde{M}_n'(\tilde{h}_{n0}) - \mathrm{E}\left\{\widetilde{CV}_n'(\tilde{h}_{n0})\right\}}{\tilde{M}_n''(\tilde{h}_{n0})}, && (45)\\ \mathrm{var}\left(\tilde{h}_{CV,n}\right) &= \frac{\mathrm{var}\left\{\widetilde{CV}_n'(\tilde{h}_{n0})\right\}}{\tilde{M}_n''(\tilde{h}_{n0})^2}. && (46)\end{aligned}$$

Since the first-order terms of $\tilde{M}_n'(\tilde{h}_{n0})$ and $\mathrm{E}\left\{\widetilde{CV}_n'(\tilde{h}_{n0})\right\}$ coincide, we must consider the second-order terms of $\tilde{M}_n'(\tilde{h}_{n0})$ and $\mathrm{E}\left\{\widetilde{CV}_n'(\tilde{h}_{n0})\right\}$ for the bias of $\tilde{h}_{CV,n}$, while for the variance, it will suffice to consider the first-order term of var $\left\{\widetilde{CV}_n'(\tilde{h}_{n0})\right\}$. Therefore, to proof Theorem 3.1, we only have to plug the results of Lemma 3.1 and Lemma 3.1 into (45) and (46).

Corollary 3.1. *Under assumptions of Theorem 3.1, the asymptotic distribution of the bandwidth that minimizes the modified version of the cross-validation criterion, given in equation* (11) *of the main paper, satisfies*

$$n^{3/10}\left(\tilde{h}_{CV,n} - \tilde{h}_{n0}\right) \xrightarrow{d} \mathrm{N}(0, V),$$

where the constant V was defined in Theorem 3.1.

Proof. Using the Cramér-Wold device (Cramér and Wold, 1936) and an argument similar to that followed in Barreiro-Ures et al. (2020), it is possible to derive the asymptotic normality of the statistic of interest, namely $n^{3/10}\left(\tilde{h}_{CV,n} - \tilde{h}_{n0}\right)$. The mean and variance of the asymptotic distribution of this statistic are an immediate consequence of Theorem 3.1.

Remark 3.1. *Under suitable assumptions:*

$$\tilde{h}_{CV,n} - \tilde{h}_{n0} = \hat{h}_{CV,n} - h_{n0} + O_p\left(n^{-2/5}\right).$$

Proof. We shall begin the sketch of the proof by showing that it stands to reason that the following expressions hold:

$$\begin{aligned} M_n(h) - \tilde{M}_n(h) &= O\left(h^8 + n^{-1}h^2 + n^{-2}\right),\\ CV_n(h) - \widetilde{CV}_n(h) &= O_p\left(h^6 + n^{-1/2}h^{7/2} + n^{-1}\right),\\ \tilde{h}_{n0} - h_{n0} &= O\left(n^{-4/5}\right).\end{aligned}$$

Recall that the Nadaraya–Watson estimator, $\hat{m}_h$, and its quadratic approximation, $\tilde{m}_h$, can be expressed as

$$\begin{aligned}\hat{m}_h(x) &= T+E+F,\\ \tilde{m}_h(x) &= T,\end{aligned}$$

where $T=A+B+C+D$ and A,B,C,D,E and F were defined just after equation (8) of the main paper. From the proof of Lemma 3.1 and the fact that

$$\begin{aligned}\mathrm{E}(\hat{a}\hat{e}^2) &= \mathrm{E}\left\{n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}Y_iK_h(x-X_i)K_h(x-X_j)K_h(x-X_k)\right\}\\ &= n^{-3}\left[n\mathrm{E}\left\{Y_1K_h(x-X_1)^3\right\}+n(n-1)\mathrm{E}\left\{Y_1K_h(x-X_1)K_h(x-X_2)^2\right\}\right.\\ &+ 2n(n-1)\mathrm{E}\left\{Y_1K_h(x-X_1)^2K_h(x-X_2)\right\}\\ &+ n(n-1)(n-2)\mathrm{E}\left\{Y_1K_h(x-X_1)K_h(x-X_2)K_h(x-X_3)\right\}]\\ &= 3R(K)m(x)f(x)^2n^{-1}h^{-1}+m(x)f(x)^3\\ &+ \left\{\mu_2(K)m(x)f(x)^2f''(x)+\mu_2(K)\varphi_1(x)f(x)^3\right\}h^2\\ &+ \left\{\frac{1}{4}\mu_2(K)^2m(x)f(x)f''(x)^2+\mu_4(K)f(x)^3\varphi_2(x)\right.\\ &+ \left.\mu_2(K)^2f(x)^2\varphi_1(x)f''(x)\right\}h^4\\ &+ O(h^6+n^{-1}),\end{aligned}$$

it follows that

$$\begin{aligned}\mathrm{E}(E) &= O\left(h^6+n^{-1}\right),\\ \mathrm{var}\,(E) &= O\left(n^{-1}h^7\right),\end{aligned}$$

and the same could be said of F. Then,

$$\begin{aligned}\left[\mathrm{E}\left\{\hat{m}_h(x)\right\}-m(x)\right]^2 &= \left[\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x)+\mathrm{E}\,(E+F)\right]^2\\ &= \left[\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x)\right]^2+\mathrm{E}\,(E+F)^2\\ &+ 2\mathrm{E}\,(E+F)\left[\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x)\right]\\ &= \left[\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x)\right]^2+O\left(h^8+n^{-1}h^2+n^{-2}\right),\end{aligned}$$

where we have used the fact that both $\mathrm{E}\,(E)$ and $\mathrm{E}\,(F)$ are $O\left(h^6+n^{-1}\right)$ and

$$\mathrm{E}\left\{\tilde{m}_h(x)\right\}-m(x)=O\left(h^2\right).$$

Also,

$$\begin{aligned}\mathrm{var}\left\{\hat{m}_h(x)\right\} &= \mathrm{var}\left\{\tilde{m}_h(x)\right\}+\mathrm{var}\,(E+F)+2\mathrm{cov}\left\{\tilde{m}_h(x),E+F\right\}\\ &= \mathrm{var}\left\{\tilde{m}_h(x)\right\}+O\left(n^{-1}h^3\right),\end{aligned}$$

where we have used the fact that both $\operatorname{var}(E)$ and $\operatorname{var}(F)$ are $O\left(n^{-1}h^7\right)$ and

$$\begin{aligned}
\operatorname{var}(E+F) &= \operatorname{var}(E)+\operatorname{var}(F)+2\operatorname{cov}(E,F)\\
&\leq \operatorname{var}(E)+\operatorname{var}(F)+2\sqrt{\operatorname{var}(E)\operatorname{var}(F)}\\
&= O\left(n^{-1}h^7\right),\\
\operatorname{cov}\{\tilde{m}_h(x),E+F\} &\leq \sqrt{\operatorname{var}\{\tilde{m}_h(x)\}\operatorname{var}(E+F)} = O\left(n^{-1}h^3\right).
\end{aligned}$$

Thus, it follows that

$$M_n(h)=\tilde{M}_n(h)+O\left(h^8+n^{-1}h^2+n^{-2}\right).$$

To avoid confusion, the functions E and F will be denoted below as $E_n(x)$ and $F_n(x)$, respectively, to indicate the fact that E and F depend on n and x. Now, there exist functions α_E, β_E and γ_E such that

$$\begin{aligned}
\mathrm{E}\{E_n(x)\} &= \alpha_E(x)h^6+\beta_E(x)n^{-1}+o\left(h^6+n^{-1}\right),\\
\operatorname{var}\{E_n(x)\} &= \gamma_E(x)n^{-1}h^7+o\left(n^{-1}h^7\right)
\end{aligned}$$

and so

$$\begin{aligned}
\mathrm{E}\left\{E_n(X_1)^2\right\} &= \mathrm{E}\left[\mathrm{E}\left\{E_n(X_1)^2 \mid X_1\right\}\right] = \mathrm{E}\left[\mathrm{E}\left\{E_n(X_1)\mid X_1\right\}^2+\operatorname{var}\{E_n(X_1)\mid X_1\}\right]\\
&= \mathrm{E}\left[\left\{\alpha_E(X_1)h^6+\beta_E(X_1)n^{-1}+o\left(h^6+n^{-1}\right)\right\}^2+\gamma_E(X_1)n^{-1}h^7\right.\\
&+ \left. o\left(n^{-1}h^7\right)\right]\\
&= h^{12}\int\alpha_E^2 f+n^{-2}\int\beta_E^2 f+n^{-1}h^7\int\left(2\alpha_E\beta_E+\gamma_E\right)f\\
&+ o\left(h^{12}+n^{-1}h^7+n^{-2}\right)
\end{aligned}$$

determines the order in probability of $E_n(X_1)^2$ due to $E_n(X_1)^2$ being a random variable that only takes positive values.

Since similar results can be obtained for $\mathrm{E}\left\{F_n(X_1)^2\right\}$ and $\mathrm{E}\{E_n(X_1)F_n(X_1)\}$ (using the Cauchy–Schwarz inequality), it can be stated that

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i)-\tilde{m}_h^{(-i)}(X_i)\right\}^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{E_{n-1}(X_i)+F_{n-1}(X_i)\right\}^2\right]\\
&= \mathrm{E}\left[\left\{E_{n-1}(X_1)+F_{n-1}(X_1)\right\}^2\right]\\
&= \mathrm{E}\left\{E_{n-1}(X_1)^2+F_{n-1}(X_1)^2+2E_{n-1}(X_1)F_{n-1}(X_1)\right\}\\
&= O_p\left(h^{12}+n^{-1}h^7+n^{-2}\right).
\end{aligned}$$

Then, since the random variable $\frac{1}{n}\sum\limits_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i)-\tilde{m}_h^{(-i)}(X_i)\right\}^2$ only takes positive values, its order in probability is given by its expected value and, hence,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i)-\tilde{m}_h^{(-i)}(X_i)\right\}^2 = O_p\left(h^{12}+n^{-1}h^7+n^{-2}\right).$$

Therefore, using the Cauchy–Schwarz inequality,

$$\begin{aligned}
CV_n(h) - \widetilde{CV}_n(h) &= \frac{1}{n}\left[\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) - Y_i\right\}^2 - \sum_{i=1}^{n}\left\{\tilde{m}_h^{(-i)}(X_i) - Y_i\right\}^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) - \tilde{m}_h^{(-i)}(X_i)\right\}\left\{\hat{m}_h^{(-i)}(X_i) + \tilde{m}_h^{(-i)}(X_i) - 2Y_i\right\} \\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) - \tilde{m}_h^{(-i)}(X_i)\right\}^2\right. \\
&\qquad \left.\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) + \tilde{m}_h^{(-i)}(X_i) - 2Y_i\right\}^2\right]^{1/2} \\
&= O_p\left(h^6 + n^{-1/2}h^{7/2} + n^{-1}\right),
\end{aligned}$$

where we have used

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_h^{(-i)}(X_i) + \tilde{m}_h^{(-i)}(X_i) - 2Y_i\right\}^2 = O_p(1).$$

Proceeding in a similary manner, albeit with more tedious calculations, it can be argued that

$$CV_n'(h_n^*) - \widetilde{CV}_n'(h_n^*) = O_p\left(n^{-4/5}\right),$$

for any bandwidth h_n^* that tends to zero at the optimal rate $n^{-1/5}$.

Finally, by means of a Taylor expansion, we have

$$0 = M_n'(h_{n0}) = M_n'(\tilde{h}_{n0}) + M_n''(\bar{h}_{n0})\left(h_{n0} - \tilde{h}_{n0}\right),$$

for some $\bar{h}_{n0}$ between h_{n0} and $\tilde{h}_{n0}$. Then, using the fact that $\tilde{M}_n'(\tilde{h}_{n0}) = 0$,

$$M_n'(\tilde{h}_{n0}) = \tilde{M}_n'(\tilde{h}_{n0}) + O\left(n^{-6/5}\right) = O\left(n^{-6/5}\right)$$

and

$$M_n''(\bar{h}_{n0}) = L_0 n^{-2/5} + o\left(n^{-2/5}\right),$$

for some constant L_0, we have

$$h_{n0} - \tilde{h}_{n0} = -\frac{M_n'(\tilde{h}_{n0})}{M_n''\left(\bar{h}_{n0}\right)} = O\left(n^{-4/5}\right).$$

Now, a Taylor expansion yields

$$0 = CV_n'(\hat{h}_{CV,n}) = CV_n'(\tilde{h}_{CV,n}) + CV_n''(h^*)\left(\hat{h}_{CV,n} - \tilde{h}_{CV,n}\right),$$

for some h^* between $\hat{h}_{CV,n}$ and $\tilde{h}_{CV,n}$. Note that

$$\tilde{M}_n''(h^*) - \tilde{M}_n''(C_0 n^{-1/5}) = \tilde{M}_n'''(h^{**})\left(h^* - C_0 n^{-1/5}\right) = o_p\left(n^{-2/5}\right),$$

for some h^{**} between h^* and the asymptotically optimal bandwidth, $C_0 n^{-1/5}$, where we have used $\tilde{M}_n'''(h^{**}) = O_p\left(n^{-1/5}\right)$ and $h^* - C_0 n^{-1/5} = o_p\left(n^{-1/5}\right)$. Then, since the order in probability of $\widetilde{CV}_n''(h^*)$ is given by its expected value, that is, $\tilde{M}_n''(h^*)$, we have:

$$\widetilde{CV}_n''(h^*) = L_0 n^{-2/5} + o_p\left(n^{-2/5}\right).$$

Consequently,

$$\hat{h}_{CV,n} - \tilde{h}_{CV,n} = -\frac{CV_n'(\tilde{h}_{CV,n})}{CV_n''(h^*)} = \frac{O_p\left(n^{-4/5}\right)}{L_0 n^{-2/5} + o_p\left(n^{-2/5}\right)} = O_p\left(n^{-2/5}\right),$$

where we have used the fact that $\widetilde{CV}_n'(\tilde{h}_{CV,n}) = 0$ and so

$$CV_n'(\tilde{h}_{CV,n}) = \widetilde{CV}_n'(\tilde{h}_{CV,n}) + O_p\left(n^{-4/5}\right) = O_p\left(n^{-4/5}\right).$$

Moreover, since $h_{n0} - \tilde{h}_{n0} = O\left(n^{-4/5}\right)$, we can also write

$$\hat{h}_{CV,n} - h_{n0} = \tilde{h}_{CV,n} - \tilde{h}_{n0} + O_p\left(n^{-2/5}\right).$$

Finally, expressions for the asymptotic bias and the variance of the bagged cross-validation bandwidth,

$$\tilde{h}(r,N) = \frac{1}{N}\left(\frac{r}{n}\right)^{1/5}\sum_{i=1}^{N}\tilde{h}_{CV,r,i}. \tag{47}$$

are given in Theorem 4.1.

Theorem 4.1. *Under assumptions* A1–A5*, the asymptotic bias and the variance of the bagged cross-validation bandwidth defined in* (47) *are:*

$$\begin{aligned} \mathrm{E}\left\{\tilde{h}(r,N)\right\} - \tilde{h}_{n0} &= (\mathcal{B} + C_1) r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right), \\ \mathrm{var}\left\{\tilde{h}(r,N)\right\} &= V r^{-1/5} n^{-2/5}\left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\} + o\left(\frac{r^{-1/5} n^{-2/5}}{N} + r^{9/5} n^{-12/5}\right), \end{aligned}$$

where the constants $\mathcal{B}$ and V were defined in Theorem 3.1 and the constant C_1 is defined in (48).

Proof. If we define

$$C_1 = -\frac{6B_2 C_0^5 + V_2}{12 B_1 C_0^2 + 2 V_1 C_0^{-3}}, \tag{48}$$

then we have

$$\begin{aligned}\tilde{h}_{r0} &= C_0 r^{-1/5} + C_1 r^{-3/5} + o\left(r^{-3/5}\right), \\ \left(\frac{r}{n}\right)^{1/5} \tilde{h}_{r0} &= C_0 n^{-1/5} + C_1 r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right)\end{aligned}$$

and

$$\begin{aligned}\left(\frac{r}{n}\right)^{1/5} \tilde{h}_{r0} - \tilde{h}_{n0} &= C_1 \left(r^{-2/5} n^{-1/5} - n^{-3/5}\right) + o\left(r^{-2/5} n^{-1/5} + n^{-3/5}\right) \\ &= C_1 r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right),\end{aligned}$$

where we have used the fact that $r = o(n)$. Therefore,

$$\begin{aligned}\mathrm{E}\left\{\tilde{h}(r, N)\right\} - \tilde{h}_{n0} &= \mathrm{E}\left\{\left(\frac{r}{n}\right)^{1/5} \tilde{h}_{CV,r,1}\right\} - \tilde{h}_{n0} \\ &= \left(\frac{r}{n}\right)^{1/5} \mathrm{E}\left(\tilde{h}_{CV,r,1} - \tilde{h}_{r0}\right) + \left\{\left(\frac{r}{n}\right)^{1/5} \tilde{h}_{r0} - \tilde{h}_{n0}\right\} \\ &= (\mathcal{B} + C_1) r^{-2/5} n^{-1/5} + o\left(r^{-2/5} n^{-1/5}\right).\end{aligned}$$

Regarding the variance, we have

$$\mathrm{var}\left\{\tilde{h}(r, N)\right\} = \frac{1}{N}\left(\frac{r}{n}\right)^{2/5} \left\{\mathrm{var}\left(\tilde{h}_{CV,r,1}\right) + (N-1)\mathrm{cov}\left(\tilde{h}_{CV,r,1}, \tilde{h}_{CV,r,2}\right)\right\} \tag{49}$$

and

$$\mathrm{cov}\left(\tilde{h}_{CV,r,1}, \tilde{h}_{CV,r,2}\right) \approx \tilde{M}_r''(\tilde{h}_{r0})^{-2} \mathrm{cov}\left\{\widetilde{CV}_1'(\tilde{h}_{r0}), \widetilde{CV}_2'(\tilde{h}_{r0})\right\}, \tag{50}$$

where for $q \in \{1, 2\}$,

$$\widetilde{CV}_q'(h) = \frac{2}{r(r-1)^2 h} \sum_{\substack{i,j,k \in I_q \\ j,k \neq i}} \left\{A_{ij} \alpha_{1h}(X_i - X_j) - h^{-1} B_{ijk} \beta_{1h}(X_i - X_j, X_i - X_k)\right\},$$

with $I_1, I_2 \sim U(\mathcal{P})$ and $\mathcal{P} = \{I \subset \{1, \ldots, n\} \mid \#I = r\}$.

Now,

$$\begin{aligned}\mathrm{cov}\left\{\widetilde{CV}_1'(h), \widetilde{CV}_2'(h)\right\} &= \mathrm{cov}\left[\mathrm{E}\left\{\widetilde{CV}_1'(h) \mid I_1, I_2\right\}, \mathrm{E}\left\{\widetilde{CV}_2'(h) \mid I_1, I_2\right\}\right] \\ &+ \mathrm{E}\left[\mathrm{cov}\left\{\widetilde{CV}_1'(h), \widetilde{CV}_2'(h) \mid I_1, I_2\right\}\right] \\ &= \mathrm{E}\left[\mathrm{cov}\left\{\widetilde{CV}_1'(h), \widetilde{CV}_2'(h) \mid I_1, I_2\right\}\right]\end{aligned}$$

since $\mathrm{E}\left\{\widetilde{CV}_q'(h) \mid I_1, I_2\right\}$, for $q \in \{1, 2\}$, does not depend on I_1, I_2 and, therefore, it is not random.

On the other hand,

$$\begin{aligned}\operatorname{cov}\left\{\widetilde{CV}_1'(h),\widetilde{CV}_2'(h)\mid I_1,I_2\right\} &= \frac{4}{r^2(r-1)^4h^2}\sum_{\substack{i,j,k\in I_1\\ l,s,t\in I_2\\ j,k\neq i\\ s,t\neq l}}\operatorname{cov}\left\{A_{ij}\alpha_{1h}(X_i-X_j)\right. \qquad (51)\\ &- h^{-1}B_{ijk}\beta_{1h}(X_i-X_j,X_i-X_k),A_{ls}\alpha_{1h}(X_l-X_s)\\ &- \left. h^{-1}B_{lst}\beta_{1h}(X_l-X_s,X_l-X_t)\right\}.\end{aligned}$$

Following the proof of Lemma 3.2, we only need to count the number of cases associated with C_{123124} and C_{123425}. If we define $M=\#\left(I_1\cap I_2\right)$, which is a random variable, then the number of times C_{123124} and C_{123425} appear in (51) is $M(M-1)(r^2-4r-M)=M^2r^2+o\left(M^2r^2\right)$ for C_{123124}, and $M^2r^3+o\left(M^2r^3\right)$ for C_{123425}.

Plugging these numbers into (51), we get

$$\operatorname{cov}\left\{\widetilde{CV}_1'(h),\widetilde{CV}_2'(h)\mid I_1,I_2\right\}=\frac{4}{r^2(r-1)^4h^2}\left(C_{123124}M^2r^2+C_{123425}M^2r^3\right)+Z,$$

where $Z=o_p\left(C_{123124}M^2r^{-4}+C_{123425}M^2r^{-3}\right)$.

To compute the expected value of the previous term we proceed by computing:

$$\begin{aligned}\mathrm{E}\left(M^2\mid I_1\right) &= \mathrm{E}\left[\left\{\sum_{i\in I_1}1_{I_2}(i)\right\}^2\mid I_1\right]=\mathrm{E}\left\{\sum_{i\in I_1}\sum_{j\in I_1}1_{I_2}(i)1_{I_2}(j)\mid I_1\right\}\\ &= \sum_{i\in I_1}\sum_{j\in I_1}\mathbb{P}(i,j\in I_2\mid I_1)=r\mathbb{P}(1\in I_2)+r(r-1)\mathbb{P}(1\in I_2)^2\\ &= r\frac{r}{n}+r(r-1)\frac{r^2}{n^2}\\ &= \frac{r^2\{n+r(r-1)\}}{n^2}\\ &= \mathrm{E}\left(M^2\right),\end{aligned}$$

where $1_{I_2}(\cdot)$ denotes the indicator function of I_2 and we have used the fact that $1_{I_2}(i)$ is a Bernoulli distribution with parameter r/n. Therefore,

$$\begin{aligned}\operatorname{cov}\left\{\widetilde{CV}_1'(h),\widetilde{CV}_2'(h)\right\} &= R_1(n^{-1}r^{-1}h^2+rn^{-2}h^2)\\ &+ R_2n^{-2}h^{-3}+O\left(n^{-2}h^{-1}+n^{-1}r^{-1}h^4+n^{-2}rh^4\right)\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}\left\{\widetilde{CV}_1'(h_{r0}),\widetilde{CV}_2'(h_{r0})\right\} &= R_1C_0^2(n^{-1}r^{-7/5}+n^{-2}r^{3/5})\\ &+ R_2C_0^{-3}n^{-2}r^{3/5}+O\left(n^{-1}r^{-9/5}+n^{-2}r^{1/5}\right). \qquad (52)\end{aligned}$$

Now, plugging (52) into (50), we get

$$\operatorname{cov}\left(\tilde{h}_{CV,r,1},\tilde{h}_{CV,r,2}\right)=Vn^{-2}r^{7/5}+Wn^{-1}r^{-3/5}+O\left(n^{-2}r+n^{-1}r^{-1}\right), \qquad (53)$$

where

$$W = \frac{R_1 C_0^2}{\left(12 B_1 C_0^2 + 2 V_1 C_0^{-3}\right)^2}.$$

Finally, plugging (53) into (49) yields

$$\operatorname{var}\left\{\tilde{h}(r,N)\right\} = V r^{-1/5} n^{-2/5} \left\{\frac{1}{N} + \left(\frac{r}{n}\right)^2\right\} + o\left(r^{9/5} n^{-12/5}\right).$$

Corollary 4.1. *Under the assumptions of Thorem 4.1, the asymptotic distribution of the bagged cross-validation bandwidth defined in* (47) *satisfies:*

$$\frac{r^{1/10} n^{1/5}}{\sqrt{\frac{1}{N} + \left(\frac{r}{n}\right)^2}} \left\{\tilde{h}(r,N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0,V),$$

where the constant V *was defined in Theorem 3.1. In particular, if we assume that* $r = o\left(n/\sqrt{N}\right)$*, then,*

$$r^{1/10} n^{1/5} \sqrt{N} \left\{\tilde{h}(r,N) - \tilde{h}_{n0}\right\} \xrightarrow{d} \mathrm{N}(0,V).$$

Proof. The result is obtained immediately from Corollary 3.1 and Theorem 4.1.

2. Simulation studies

In this section, we complete the simulations presented in the main paper, adding two additional plots not included in the paper for reasons of space. In this simulation study, we considered the following regression models:

M1: $Y = m(X) + \varepsilon$, $m(x) = 2x$, $X \sim \mathrm{Beta}(3,3)$, $\varepsilon \sim \mathrm{N}(0, 0.1^2)$,

M2: $Y = m(X) + \varepsilon$, $m(x) = \sin(2\pi x)^2$, $X \sim \mathrm{Beta}(3,3)$, $\varepsilon \sim \mathrm{N}(0, 0.1^2)$,

M3: $Y = m(X) + \varepsilon$, $m(x) = x + x^2 \sin(8\pi x)^2$, $X \sim \mathrm{Beta}(3,3)$, $\varepsilon \sim \mathrm{N}(0, 0.1^2)$,

whose regression functions are plotted in Figure 1 of the main paper. The R (R Development Core Team, 2021) package `baggingbwsel` (Barreiro-Ures et al., 2021) was employed to carry out the simulation experiments. In a first step, we empirically checked how close the bandwidths that minimize the MISE of the Nadaraya–Watson estimator and its modified version given, respectively, in equations (2) and (9) of the main paper are. For this, we simulated 100 samples of sizes 1000 and 5000 from models M1, M2 and M3 and compute the corresponding MISE curves for the standard Nadaraya–Watson estimator and for its modified version, using the Gaussian kernel. Figure 1 shows these curves. It can be observed that the bandwidth that minimizes the MISE of standard Nadaraya–Watson estimator and the MISE of its modified version appear to be quite close for both sample sizes, although the distance between the minima of both curves seems to tend to zero as the sample size increases. On the other hand, Figure 2 shows the standard and modified cross-validation bandwidths (using the standard and modified version of the Nadaraya–Watson estimator, respectively) obtained for samples of sizes ranging from 600 to 5000 drawn from model M2. It can be seen that both bandwidth selectors provide similar results, which in turn get closer as n increases.

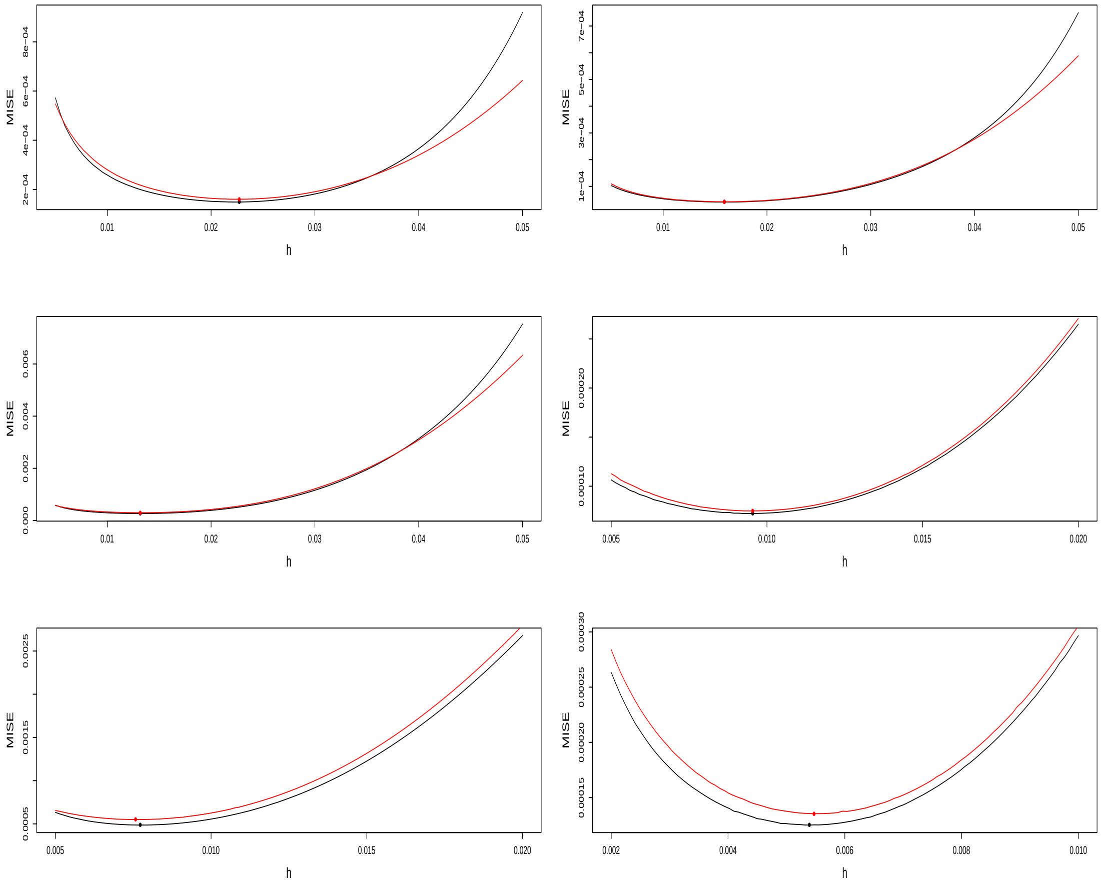

Figure 1. MISE curve for the standard Nadaraya–Watson estimator (black line) for its modified version (red line) and with their minima (red and black points, respectively). First row: model M1, second row: model M2, third row: model M3. First column: $n = 1000$, second column: $n = 5000$

References

Barbeito, I. (2020) *Exact bootstrap methods for nonparametric curve estimation*. Ph.D. thesis, Universidade da Coruña. `https://ruc.udc.es/dspace/handle/2183/26466`.

Barreiro-Ures, D., Cao, R., Francisco-Fernández, M. and Hart, J. D. (2020) Bagging cross-validated bandwidths with application to big data. *Biometrika*. `https://doi.org/10.1093/biomet/asaa092`.

Barreiro-Ures, D., Hart, J. D., Cao, R. and Francisco-Fernández, M. (2021) *baggingbwsel: Bagging Bandwidth Selection in Kernel Density and Regression Estimation*. R package version 1.0. `https://cran.r-project.org/package=baggingbwsel`.

Cramér, H. and Wold, H. (1936) Some theorems on distribution functions. *Journal of*

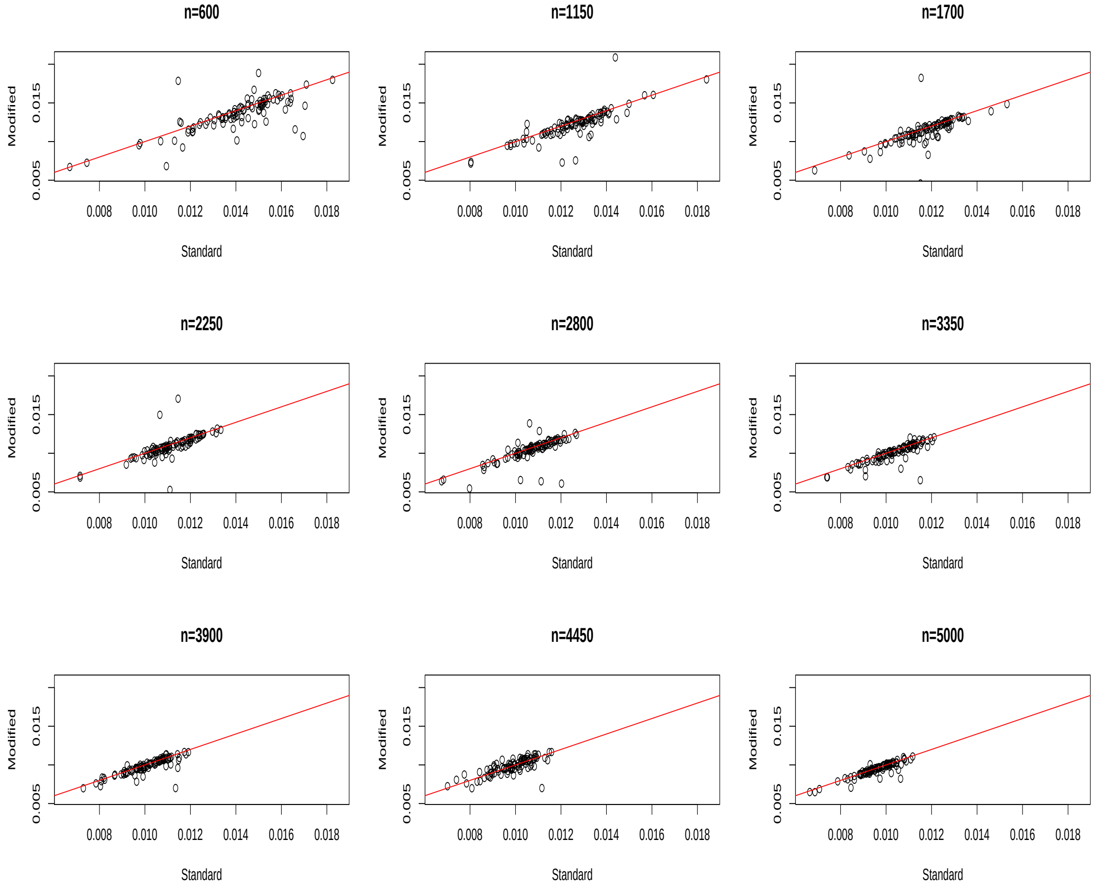

Figure 2. Cross-validation bandwidths using the standard Nadaraya–Watson estimator (x-axis) and its modified version (y-axis) for samples of sizes ranging from 600 to 5000 drawn from model M2.

the London Mathematical Society, **s1-11**, 290–294. URL: `https://doi.org/10.1112/jlms/s1-11.4.290`.

R Development Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.