Nonparametric density and regression estimation for samples of very large size

Daniel Barreiro Ures

PhD Thesis

2021

Nonparametric density and regression estimation for samples of very large size

Author: Daniel Barreiro Ures
Supervisors: Ricardo Cao Abad, Mario Francisco Fernández

PhD Thesis in Statistics and Operations Research
Department of Mathematics
2021

UNIVERSIDADE DA CORUÑA

The undersigned, Ricardo Cao Abad and Mario Francisco Fernández, certify that they are the advisors of the Doctoral Thesis entitled 'Nonparametric density and regression estimation for samples of very large size', developed by Daniel Barreiro Ures at the University of A Coruña (Department of Mathematics), as part of the interuniversity PhD program (UDC, USC and UVigo) of Statistics and Operations Research, and hereby gives his consent to the author to proceed with the thesis presentation and the subsequent defense.

Los abajo firmantes, Ricardo Cao Abad y Mario Francisco Fernández, hacen constar que son los directores de la Tesis Doctoral titulada 'Nonparametric density and regression estimation for samples of very large size', realizada por Daniel Barreiro Ures en la Universidade da Coruña (Departamento de Matemáticas) en el marco del programa interuniversitario (UDC, USC y UVigo) de doctorado en Estadística e Investigación Operativa, dando su consentimiento para que el autor proceda a su presentación y posterior defensa.

Os abaixo asinantes, Ricardo Cao Abad e Mario Francisco Fernández, fan constar que son os directores da Tese de Doutoramento titulada 'Nonparametric density and regression estimation for samples of very large size', desenvolta por Daniel Barreiro Ures na Universidade da Coruña (Departamento de Matemáticas) no marco do programa interuniversitario (UDC, USC e UVigo) de doutoramento en Estatística e Investigación de Operacións, dando o seu consentimento para que o autor proceda á súa presentación e posterior defensa.

A Coruña, October 5th, 2021.

Advisor:	Advisor:	PhD Student:
Ricardo Cao Abad	Mario Francisco Fernández	Daniel Barreiro Ures

The public defense of the Doctoral Thesis entitled "Nonparametric density and regression estimation for samples of very large size", developed by Daniel Barreiro Ures and supervised by Dr. Ricardo Cao Abad and Dr. Mario Francisco Fernández, will be held on 10th December, 2021, at the Faculty of Computer Sciences at the University of A Coruña, with the examining committee:
Dr. María Dolores Martínez Miranda (President)
Dr. Rubén Fernández Casal (Secretary)
Dr. Philippe Vieu (Board member)

A Coruña, December 10th, 2021.

PhD committee:

President	Secretary	Board member
Advisor:	Advisor:	PhD Student:
Ricardo Cao Abad	Mario Francisco Fernández	Daniel Barreiro Ures

Agradecimientos personales

Quisiera agradecer a mis directores de tesis, Ricardo Cao Abad y Mario Francisco Fernández, por su ayuda y dedicación, sin las cuales esta tesis no hubiese salido adelante.

Aprovecho también para mostrar mi agraedcimiento a Jeff Hart por el excelente trato recibido durante mi estancia en la Texas A&M University y por sus múltiples aportaciones a la tesis.

Por último, agradezco a mis padres y amigos por el indispensable apoyo recibido a lo largo de todos estos años.

Institutional acknowledgements

This research has been supported by MINECO Grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015, ED431C-2020-14, Centro Singular de Investigación de Galicia ED431G/01 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF (European Regional Development Fund). Additionally, this work has been partially carried out during a visit to the Texas A&M University, College Station, financed by INDITEX, with reference INDITEX-UDC 2019.

The author is grateful to the Centro de Coordinación de Alertas y Emergencias Sanitarias for kindly providing the COVID-19 hospitalization dataset.

Abstract

This dissertation mainly deals with the problem of bandwidth selection in the context of nonparametric density and regression estimation for samples of very large size. Some bandwidth selection methods have the disadvantage of high computational complexity. This implies that the number of operations required to compute the bandwidth grows very rapidly as the sample size increases, so that the computational cost associated with these algorithms makes them unsuitable for samples of very large size. In the present thesis, this problem is addressed through the use of subagging, an ensemble method that combines bootstrap aggregating or bagging with the use of subsampling. The latter reduces the computational cost associated with the process of bandwidth selection, while the former is aimed at achieving significant reductions in the variability of the bandwidth selector. Thus, subagging versions are proposed for bandwidth selection methods based on widely known criteria such as cross-validation or bootstrap. When applying subagging to the cross-validation bandwidth selector, both for the Parzen–Rosenblatt estimator and the Nadaraya–Watson estimator, the proposed selectors are studied and their asymptotic properties derived. The empirical behavior of all the proposed bandwidth selectors is shown through various simulation studies and applications to real datasets.

Resumen

Esta disertación aborda principalmente el problema de la selección de la ventana en el contexto de la estimación no paramétrica de la densidad y de la regresión para muestras de gran tamaño. Algunos métodos de selección de la ventana tienen el inconveniente de contar con una elevada complejidad computacional. Esto implica que el número de operaciones necesarias para el cálculo de la ventana crece muy rápidamente a medida que el tamaño muestral aumenta, de manera que el coste computacional asociado a estos algoritmos los hace inadecuados para muestras de gran tamaño. En la presente tesis, este problema se aborda mediante el uso del subagging, un método de aprendizaje conjunto que combina el bootstrap aggregating o bagging con el uso de submuestreo. Este último reduce el coste computacional asociado al proceso de selección de la ventana, mientras que el primero tiene como objetivo conseguir reducciones significativas en la variabilidad del selector de la ventana. Así, se proponen versiones subagging para métodos de selección de la ventana basados en criterios ampliamente conocidos, como la validación cruzada o el bootstrap. Al aplicar subagging al selector de la ventana de tipo validación cruzada, tanto para el estimador de Parzen–Rosenblatt como para el estimador de Nadaraya–Watson, se estudian los selectores propuestos y se derivan sus propiedades asintóticas. El comportamiento empírico de todos los selectores de la ventana propuestos se muestra mediante varios estudios de simulación y aplicaciones a conjuntos de datos reales.

Resumo

Esta disertación aborda o problema da selección da ventá no contexto da estimación non paramétrica da densidade e da regresión para mostras de gran tamaño. Algúns métodos de selección da ventá teñen o inconveniente de contar cunha alta complexidade computacional. Isto implica que o número de operacións necesarias para o cálculo da ventá crece moi rápidamente a medida que aumenta o tamaño muestral, polo que o coste computacional asociado a estes algoritmos fainos inadecuados para mostras de gran tamaño. Na presente tese, este problema abórdase mediante o uso do subagging, un método de aprendizaxe conxunta que combina o bootstrap aggregating ou bagging co uso de submostraxe. Este último reduce o custo computacional asociado ao proceso de selección da ventá, mentres que o primeiro ten como obxectivo conseguir reducións significativas na variabilidade do selector da ventá. Así, propóñense versións subagging para métodos de selección da ventá baseados en criterios amplamente coñecidos, como a validación cruzada ou o bootstrap. Ao aplicar subagging ao selector da ventá de tipo validación cruzada, tanto para o estimador de Parzen–Rosenblatt como para o estimador de Nadaraya–Watson, estúdanse os selectores propostos e derívanse as súas propiedades asintóticas. O comportamento empírico de todos os selectores da ventá propostos móstrase mediante varios estudos de simulación e aplicacións a conxuntos de datos reais.

Preface

The scope of this dissertation is the nonparametric estimation of density and regression functions. In particular, the focus is on the problem of bandwidth selection for the Parzen–Rosenblatt and Nadaraya–Watson estimators, mainly in the context of samples of very large size. Dealing with very large sample sizes, in conjunction with the fact that some of the most popular and widely used bandwidth selection methods have high computational complexity, make the adaptation of such bandwidth selection methods to the context of large sample sizes an imperative task.

The dissertation is structured as follows: in Chapter 1 the line of research followed in this dissertation is put in context and the motivations behind it are presented.

Chapter 2 provides an introduction to the field of nonparametric density and regression estimation, with special emphasis on the problem of bandwidth selection. In addition, bootstrap, bagging and subagging techniques are described and discussed, highlighting their applicability in the context of bandwidth selection.

Chapter 3 is devoted to the application of bagging in the process of bandwidth selection for the kernel density estimator. The classical results concerning cross-validation as a bandwidth selection method are presented. Then, a bagging version of this selector is described and its asymptotic properties are derived. In addition, bagging selectors are proposed for an error criterion other than cross-validation (bootstrap), as well as for situations where the rate of convergence to zero of the optimal bandwidth is not known, or where taking into account the second order terms of the optimal bandwidth may seem desirable. Finally, the empirical behavior of the techniques studied throughout the chapter is shown by means of simulation studies and applications to real datasets.

Chapter 4 deals with the application of bagging for the selection of the bandwidth

of the Nadaraya–Watson estimator. The first part of the chapter presents the classical results on the cross-validation bandwidth selector, which are further developed for the second part of the chapter, that deals with the application of bagging to the cross-validation selector. In this second part, the bagging selector of the bandwidth of the Nadaraya–Watson estimator is described and its asymptotic properties are derived. As in the previous chapter, practical behavior of the techniques proposed and studied throughout the chapter is shown through various simulation studies and an application to a real dataset.

Concluding remarks, as well as a glimpse on future lines of work are provided in Chapter 5.

In addition, the proofs of the lemmas, theorems, corollaries and other results presented throughout the dissertation are included in the corresponding appendices. Specifically, the proofs of the results concerning Chapter 3 are included in Appendices A and B, while the proofs of the results regarding Chapter 4 are included in Appendix C. Also, manuals for the `baggedcv` (Barreiro-Ures et al., 2019) and `baggingbwsel` (Barreiro-Ures et al., 2021b) R packages developed, among others, by the author of this dissertation, are included in Appendices D and E, respectively.

Finally, in the sections focused on the numerical application of the techniques proposed throughout the dissertation, whenever calculations have been performed in parallel, they have been carried out using an Intel Core i5-8600K 3.6GHz CPU.

Contents

Chapter 1

Motivation

The amount of data around the world is increasing every second, as is the speed at which it is generated and the sources from which it comes. Additionally, data can be structured, unstructured, in motion or stored. Big Data contains valuable knowledge, capable of transforming a business through analytical techniques (Business Analytics). By applying these techniques to this wide variety of data a clearer picture of the business and the variables that affect it can be obtained. Currently, in the statistical-computational literature there are proposals for the application of computational techniques based on the construction of parallel executable algorithms on CPU or GPU, generally through cluster computing platforms such as Hadoop or Spark (Meng et al., 2015). On the other hand, subsampling-based methods such as leveraging (Ma and Sun, 2015) and algorithms that allow the application of bootstrap in the context of large sample sizes were proposed (Kleiner et al., 2012). Furthermore, Farrash (2016) studied the application of ensemble methods in Big Data and, in particular, the selection of the size of the subsamples generated by these methods.

It is worth mentioning the proposals of Politis and Romano (1994) and Politis et al. (1999) to combine the use of the bootstrap method and subsampling. In particular, Politis and Romano (1994) show, under minimal assumptions, the good behavior of the bootstrap method when considering resamples of a size smaller than that of the original sample.

In the context of nonparametric density estimation, there are proposals based on the bootstrap aggregating or bagging technique for the selection of the bandwidth of

the Parzen–Rosenblatt density estimator (Hall and Robinson, 2009). In the case of regression estimation, it will be of interest to consider modifications of the Nadaraya–Watson proxy estimator similar to those proposed in Barbeito (2020) in order to facilitate the theoretical analysis of new methods inspired by the use of subsampling and bagging in the context of density estimation.

The main objective of the thesis is the proposal, study and application of computationally efficient estimation techniques, mainly in the context of Big Data, and in particular when working with very large sample sizes. Special importance is given to the problem of bandwidth selection in the context of nonparametric density and regression estimation. Both in the context of density and regression estimation, the thesis focuses on the theoretical and practical study of bandwidth selectors based on cross-validation and bootstrap criteria and the use of bagging. The thesis project is framed within the thematic of the research project entitled "Flexible statistical inference for complex, large-volume, high-dimensional data" (code MTM2017-82724-R), financed by the Ministry of Economy and Competitiveness.

The objectives of this doctoral thesis are focused on two aspects: methodological and computational. When formulating most of the new proposed techniques, the aim is to follow the guidelines of other simpler (or more particular) models were such techniques have already been established. For instance, one could start from well-known bandwidth selection methods such as cross-validation or bootstrapping. Since these methods are very computationally expensive, their adaptation to the context of large sample sizes could be done by using subsampling. However, the major difficulty lies in proving that these procedures work as expected or behave adequately compared to other existing ones. Both issues are addressed by trying to obtain asymptotic properties of the estimators, by carrying out simulation studies and by applying the proposed methods to real data and comparing the results with those obtained by other existing methods. In general, as a consequence of these analyses, either the proposed method behaves as expected, or we obtain indications as to why this is not the case (the causes of non-optimal behavior may appear even when the method works acceptably well). Thus, in many cases there exists the possiblity of modifying the proposed method in order to improve its behavior (such as bias correction or changes in the resampling process in bootstrap methods). The

importance of computer simulation studies, which truly are "test laboratories" for statistical methods, should be emphasized at this point. Simulation studies allow us to assess the performance of the proposed methods independently of the theoretical analysis, and with much less effort. Thus, these studies could provide clues that justify the rejection of certain proposals or, on the contrary, positive results that encourage further theoretical study of the methods in question. Regarding the computational aspects of the different objectives of the thesis, R software (R Core Team, 2021) is used to carry out the relevant simulation studies (an invaluable complement to methodological deepening). This software is made available to the international statistical (and, in general, scientific) community through the CRAN of R project. Taking into account the high computational demand of classical statistical methods for large sample sizes, an important aspect is the optimization of the algorithms to be implemented for the use of the techniques to be studied in the thesis. Most of the methods proposed throughout the thesis are included in the `baggedcv` (Barreiro-Ures et al., 2019) and `baggingbwsel` (Barreiro-Ures et al., 2021b) R packages.

Chapter 2

Introduction

This chapter is intended to provide an introduction to the field of nonparametric density and regression estimation, with special emphasis on the problem of bandwidth selection (Wand and Jones, 1995). In addition, bootstrap and bagging techniques are described and discussed, highlighting their applicability in the context of bandwidth selection.

2.1 Density estimation

Density estimation studies the relationship between the values a certain random variable can take and their probability, the latter being almost always unknown in practice. In other words, density estimation encompasses all those methods whose purpose is the estimation of the underlying density function of a random variable. These methods, in turn, can be classified into two broad groups: parametric and nonparametric density estimation methods. Parametric methods are based on the assumption that the underlying density function belongs to a certain parametric family of functions depending on some parameters, so that the problem of estimating the density is reduced to the estimation of these parameters. Nonparametric methods, on the other hand, do not impose such constraints, but try to estimate the underlying density from the data itself, which makes this group of methods more flexible than their parametric counterpart. A particular class of nonparametric methods is that of kernel methods (Silverman, 1986), which seek to estimate the density func-

tion as a locally weighted average, using a kernel function as a weighting function. Aside from the kernel function, these methods are highly dependent on the choice of a free parameter called the bandwidth or smoothing parameter which determines the amount of smoothing performed by the estimator, which in turn determines the trade-off between the bias and the variance of the estimator. The problem of bandwidth selection is therefore crucial and intrinsic to kernel methods. Multiple ways of addressing it have been proposed and studied over time, these including cross-validation (Rudemo, 1982; Bowman, 1984; Hall and Marron, 1987), bootstrapping (Cao, 1993) or plug-in methods (Sheather and Jones, 1991).

In this section, the kernel density estimator is presented and the problem of selecting its bandwidth is addressed. We will mainly focus on the cross-validation bandwidth selector, a perfect candidate for the application of bagging[1] due to its high variability (Park and Marron, 1990) and review its well known theoretical properties. Although they will not be studied theoretically, different techniques for bandwidth selection other than cross-validation, such as bootstrapping[2], will also be discussed.

2.1.1 Kernel density estimation

Let us begin by considering a sample of size n, $X_1, \ldots, X_n$, where the observations are independent and identically distributed to the continuous random variable X, whose probability density function is denoted by f. Instead of assuming that the underlying density function belongs to a certain parametric family of functions, as parametric methods do, nonparametric density estimation methods do not impose such a restriction on f, but rather aim to capture the main features of f from the data itself. This allows us to state that, in general, nonparametric methods have greater flexibility when compared to their parametric competitors. To illustrate this point we have simulated a sample of size 10^4 drawn from a normal mixture density

$$f_{NM}(x) = \sum_{i=1}^{3} w_i \frac{1}{\sigma_i} \phi\left(\frac{x - \mu_i}{\sigma_i}\right),$$

[1]The bagging technique will be discussed in detail in Section 2.4.
[2]The bootstrap method will be discussed in detail in Section 2.3.

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ denotes the standard normal density function and the vectors of means, variances and weights are given by

NM: $\mu = (5, 6, 8)$, $\sigma^2 = (1, 0.5, 1)$, $w = (0.2, 0.5, 0.3)$.

Figure 2.1 shows the parametric (log-normal, gamma and Weibull) fits as well as the kernel density estimate obtained for the simulated sample. As we can see, the kernel density estimator is the only one that manages to capture the most important features of the underlying density. Moreover, as can be observed in Figure 2.2, which shows a Gamma fit as well as a nonparametric estimate for a sample of size 10^4 drawn from a Gamma distribution, the nonparametric estimate can still keep up with the parametric estimate even when the latter's assumptions on the target density hold.

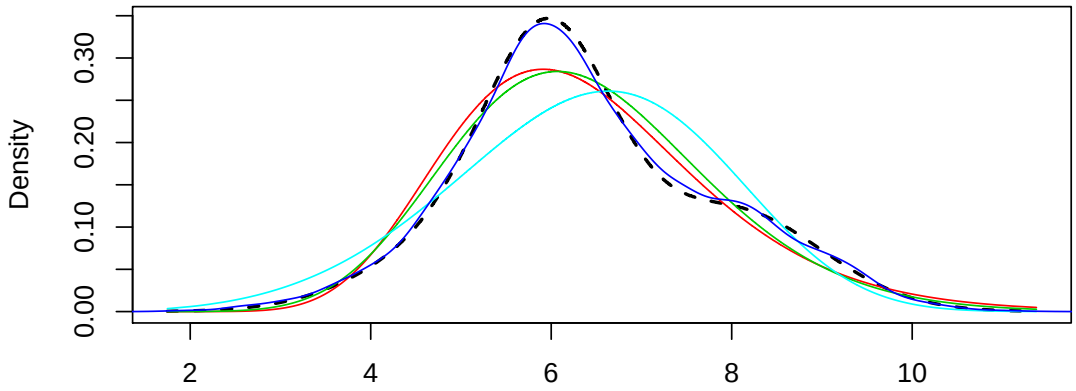

Figure 2.1: Parametric and nonparametric fits for density NM. The target density (dashed black line), the log-normal (red), gamma (green) and Weibull (light blue) fits as well as the kernel density estimate with bandwidth $h = 0.183$ (dark blue) are shown.

Before discussing the kernel density estimator in detail, we will discuss what could be considered the simplest method for estimating a density function, namely the histogram. The reason for starting with the histogram is that the kernel density estimator can be considered a generalization of what is usually referred to as the

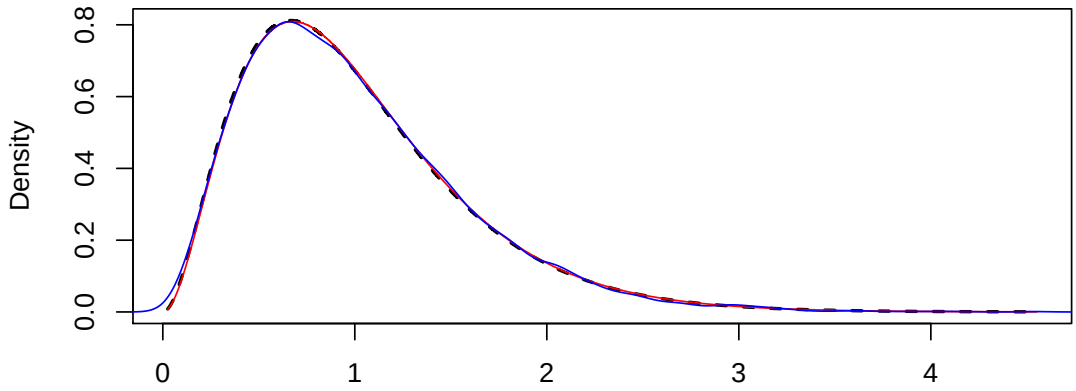

Figure 2.2: Gamma fit (red line) and kernel density estimate (dark blue line) for a sample of size 10^4 drawn from a gamma distribution (dashed black line).

moving histogram. The idea of the histogram is to aggregate the observations in intervals of the form $[a, b)$. Then, for every $x \in [a, b]$ the estimate of $f(x)$ would be

$$\frac{1}{n(b-a)} \sum_{i=1}^{n} 1_{[a,b)}(X_i),$$

where $1_S(\cdot)$ denotes the indicator function of the set S, that is,

$$1_S(x) = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{if } x \notin S \end{cases}$$

In more detail, given an origin x_0 and a bandwidth $h > 0$, let us define the constant-length intervals (also called bins) $I_j = [x_j, x_{j+1}]$, with $x_j = x_0 + hj$ and $j \in \mathbb{Z}$. Then, the histogram at a point x such that $x \in I_j$ can be defined as

$$\hat{f}_{hist}(x; x_0, h) = \frac{1}{nh} \sum_{i=1}^{n} 1_{I_j}(X_i). \tag{2.1}$$

Note that the choice of the parameter h has a significant impact on the behavior of the histogram defined in (2.1), as Figure 2.3 illustrates. There, histograms with different number of bins are shown for a sample of size $n = 5000$ drawn from the following mixture density (Marron and Wand, 1992):

D1: (claw density) with parameters $\mu = (0, -1, -0.5, 0, 0.5, 1)$,
$\sigma = (1, 0.1, 0.1, 0.1, 0.1, 0.1)$ and $w = (0.5, 0.1, 0.1, 0.1, 0.1, 0.1)$.

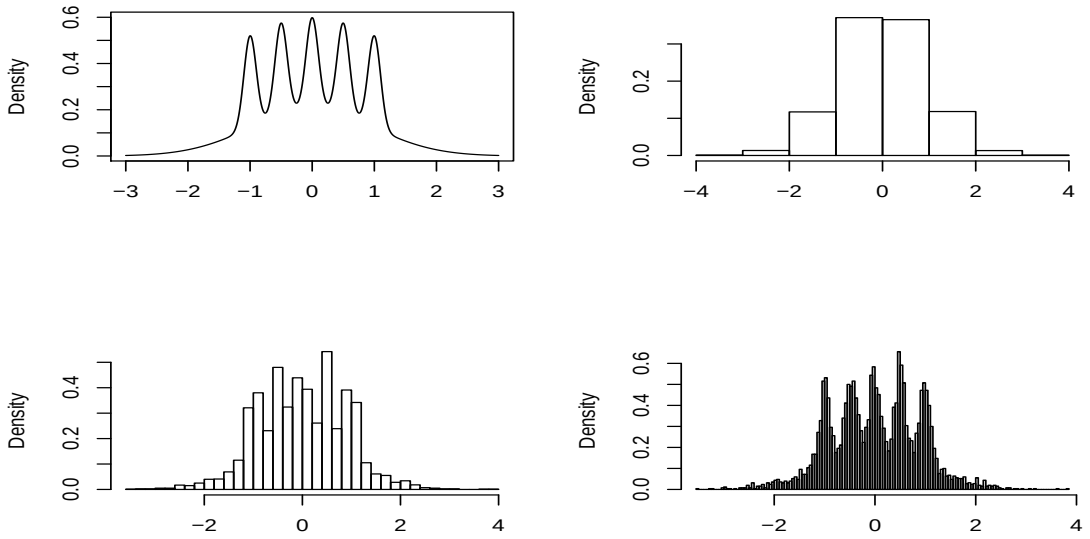

Figure 2.3: Target density (top left) and histograms built with 10 (top right), 50 (bottom left) and 250 (bottom right) bins for a sample of size 5000 drawn from density D1.

The idea of the moving histogram (also called the naïve density estimator) arose in order to remedy the problem of the histogram's dependence on the origin, x_0. This is achieved by making the previously defined intervals dependent on the point x at which the estimator is to be evaluated. In other words, given a bandwidth $h > 0$ the moving histogram at a point x can be defined as

$$\hat{f}_{MH}(x; h) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{(x-h, x+h)}(X_i),$$

which can be rewritten as

$$\hat{f}_{MH}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.2}$$

where $K(u) = \frac{1}{2}1_{(-1,1)}(u)$ is the uniform density function for the interval $(-1, 1)$. The kernel density estimator can be thought of as a generalization of the moving histogram defined in (2.2) such that the function K is not limited to the uniform density but rather it can be chosen from a wide range of density functions. Thus, the kernel density estimator or Parzen-Rosenblatt estimator (Parzen, 1962; Rosenblatt, 1956) has the following expression:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.3}$$

where K is usually assumed to be a symmetric kernel function, that is, a non-negative function such that $K(x) = K(-x)$ and

$$\int_{-\infty}^{\infty} K(x)\,dx = 1.$$

There is a wide range of possible kernel functions, including, among others, the Gaussian, uniform, triangular or Epanechnikov kernel functions (see Table 2.1 and Figure 2.4 for more information on these and other kernel functions). Using the notation $K_h(u) = \frac{1}{h}K(u/h)$, then (2.3) can be rewritten as

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

so that $\hat{f}_h(x)$ can be interpreted as a locally weighted average, with K being the weighting function and where $h > 0$ controls the influence that the observations which are close to x have on the estimate. Thus, it is easy to see the important role that the parameter h, usually called the bandwidth or smoothing parameter, plays in $\hat{f}_h$, and how making a good choice of the bandwidth is crucial to obtaining a good density estimate. In fact, the choice of the kernel function is of secondary

importance compared to the problem of bandwidth selection (Wand and Jones, 1995), as illustrated in Figure 2.5.

Let us now denote by $g_1 * g_2$ the convolution of the functions g_1 and g_2, that is,

$$(g_1 * g_2)(x) = \int_{-\infty}^{\infty} g_1(u) g_2(x-u)\, du.$$

Due to the aforementioned importance of selecting an appropriate bandwidth for (2.3), we are faced with the need to find a suitable optimality criterion for the bandwidth. In order to do so, one must study the properties of the kernel density estimator. The bias and variance of (2.3) have been studied by several authors and their expressions are well known (Parzen, 1962), namely:

$$\begin{aligned} \mathrm{E}\left[\hat{f}_h(x)\right] - f(x) &= (K_h * f)(x) - f(x), \\ \mathrm{var}\left[\hat{f}_h(x)\right] &= \frac{1}{n}\left\{\left[(K_h)^2 * f\right](x) - (K_h * f)^2(x)\right\} \end{aligned}$$

Therefore, the mean squared error of $\hat{f}_h(x)$ can be written as

$$\mathrm{MSE}\left[\hat{f}_h(x)\right] = \frac{1}{n}\left\{\left[(K_h)^2 * f\right](x) - (K_h * f)^2(x)\right\} + \left[(K_h * f)(x) - f(x)\right]^2. \quad (2.4)$$

Name	K(x)	Support
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	$\mathbb{R}$
Uniform	$\frac{1}{2}$	$[-1, 1]$
Triangular	$1 - \lvert x\rvert$	$[-1, 1]$
Epanechnikov	$\frac{3}{4}(1 - x^2)$	$[-1, 1]$
Quartic	$\frac{15}{16}(1 - x^2)^2$	$[-1, 1]$
Triweight	$\frac{35}{32}(1 - x^2)^3$	$[-1, 1]$

Table 2.1: Commonly used univariate kernel functions.

Note that (2.4) is not a random variable since it does not depend on the sample. However, it still depends on the particular value of x, and so (2.4) cannot work as a global optimality criterion. An oft-used criterion for defining an optimal bandwidth

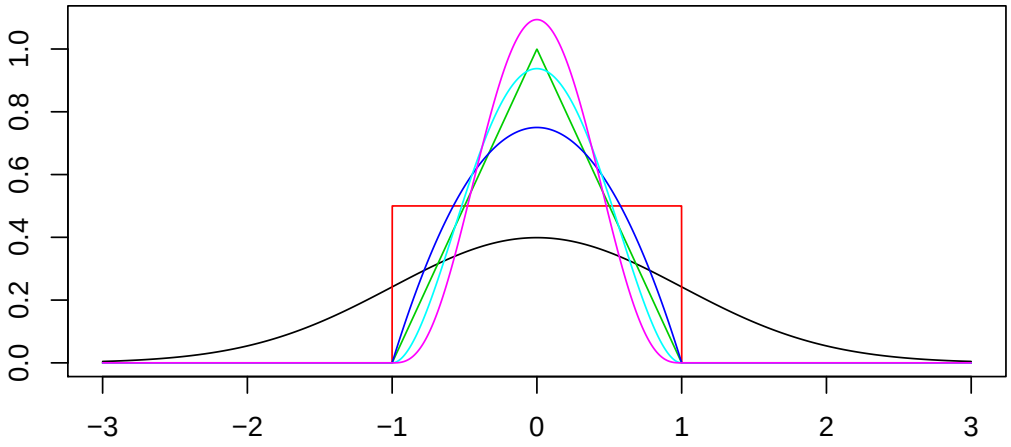

Figure 2.4: Gaussian (black), uniform (red), triangular (green), Epanechnikov (dark blue), quartic (light blue) and triweight (pink) kernel functions.

is based on mean integrated squared error or MISE, defined as

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

that is, an integrated version of (2.4). Suppose that f has two continuous derivatives. As shown by, for example, Silverman (1986), the minimizer, h_{n0}, of $M_n(h)$ with respect to h is asymptotic to

$$h_{na} = C_0 n^{-1/5} \tag{2.5}$$

as $n \to \infty$, where

$$C_0 = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{1/5}, \tag{2.6}$$

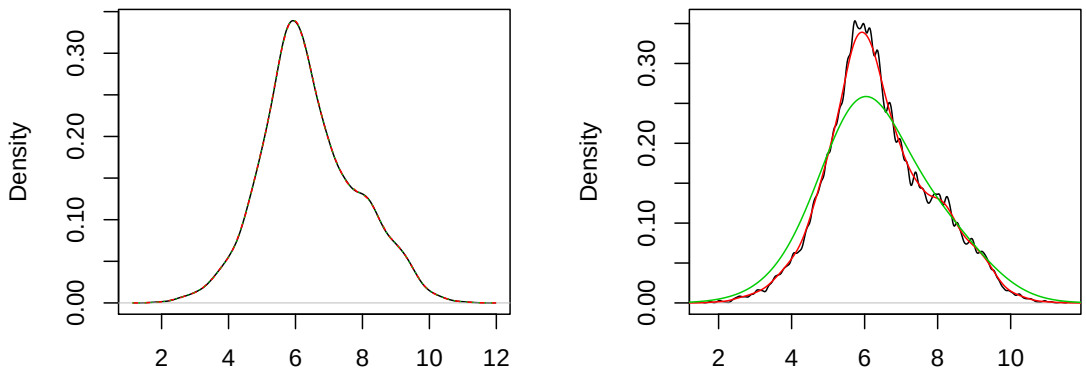

Figure 2.5: Left panel: kernel density estimates considering $h = 0.2$ and different kernel functions, namely Gaussian (black), Epanechnikov (red) and triangular (green). Right panel: kernel density estimates considering a Gaussian kernel and different values for the bandwidth, namely $h = 0.05$ (black), $h = 0.2$ (red) and $h = 0.8$ (green).

with $R(g) = \int g(x)^2\,dx$ and $\mu_j(g) = \int x^j g(x)\,dx$ $(j = 0, 1, \dots)$, provided that these integrals exist finite. Ideally, one would use h_{n0} as a bandwidth in (2.3), but of course h_{n0} depends on f and so this is not feasible. However, it can be estimated and to that effect numerous bandwidth selection methods, these including cross-validation (Rudemo, 1982; Bowman, 1984; Hall and Marron, 1987), bootstrap (Cao, 1993) and plug-in (Sheather and Jones, 1991) methods, have been proposed and studied over time. In the following section, we will focus on the study of the leave-one-out cross-validation criterion and the asymptotic properties of the cross-validation selector for the bandwidth of the Parzen-Rosenblatt estimator defined in (2.3).

2.1.2 Cross-validation method for bandwidth selection

Cross-validation is a rough-and-ready method of model selection that predates an early exposition of the method by Stone (1974). In its simplest form, cross-validation consists of dividing one's dataset into two parts, using one part to build one or more

models, and then predicting the date in the second part with the models so-built. In this way one can objectively compare the predictive ability of different models. The leave-one-out version of cross-validation 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.

While leave-one-out cross-validation is a very useful method, due in no small part to its wide applicability, it does have its drawbacks. In the context of smoothing parameter selection for function estimation, it has been regarded skeptically for many years owing to its large variability; see, e.g., Park and Marron (1990). A number of modified versions of cross-validation have been proposed in an effort to produce more stable smoothing parameter selectors. These include partitioned cross-validation (Marron, 1987; Bhattacharya and Hart, 2016), proposals of Stute (1992) and Feluch and Koronacki (1992), smoothed cross-validation (Hall et al., 1992), one-sided cross-validation (Hart and Yi, 1998; Martínez-Miranda et al., 2011), a bagged version of cross-validation (Hall and Robinson, 2009; Barreiro-Ures et al., 2021a), indirect cross-validation (Savchuk et al., 2010) and DO-validation (Mammen et al., 2011).

The cross-validation criterion is derived from the expression of the integrated squared error or ISE of $\hat{f}_h$,

$$\mathrm{ISE}\left(\hat{f}_h\right) = \int_{-\infty}^{\infty} \left[\hat{f}_h(x) - f(x)\right]^2 dx,$$

which is a random variable and can be rewritten as

$$\mathrm{ISE}\left(\hat{f}_h\right) = R(\hat{f}_h) - 2\int_{-\infty}^{\infty} \hat{f}_h(x) f(x)\, dx + R(f). \tag{2.7}$$

Note that the third summand in (2.7), $R(f)$, does not depend on the bandwidth and can therefore be ignored when constructing the cross-validation criterion. Now, it can be easily shown that the first summand, $R(\hat{f}_h)$, is equal to $R(K)/(nh)$ while the second summand, $\int \hat{f}_h f$, can be interpreted as the expected value of $\hat{f}_h$. A naïve

way of estimating the latter would be to replace it by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}_h(X_i).$$

However, by estimating the second summand in this way we would be using the same sample both to construct $\hat{f}_h$ and to estimate the summand in question. This would be problematic as it would lead to overfitting, thus producing an inconsistent empirical optimality criterion. Thus, the leave-one-out cross-validation or least-squares cross-validation criterion can be written as (Scott and Terrell, 1987)

$$CV(h) = \frac{R(K)}{nh} - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_h^{(-i)}(X_i), \quad h > 0, \tag{2.8}$$

where $\hat{f}_h^{(-i)}$ is a kernel estimate computed with the $n-1$ observations other than X_i, that is,

$$\hat{f}_h^{(-i)}(x) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} K_h\left(x - X_j\right), \quad i = 1, \ldots, n.$$

It is easily shown that $CV(h)$ is an unbiased estimator of $M_n(h) - R(f)$ for each $h > 0$ (Scott and Terrell, 1987). In turn, (2.8) admits the following expression (Scott and Terrell, 1987):

$$CV(h) = \frac{R(K)}{nh} + \frac{2}{n^2 h}\sum\sum_{i<j}\left(K * K - 2K\right)\left(\frac{X_i - X_j}{h}\right).$$

It seems natural then to estimate h_{n0} by $\hat{h}_n$, the minimizer of $CV(h)$. Hall and Marron (1987) show that[3]

$$n^{1/10}\left(\frac{\hat{h}_n - h_{n0}}{h_{n0}}\right) \xrightarrow{d} Z, \tag{2.9}$$

[3]Hereinafter, the notation $Z_n \xrightarrow{d} Z$ ($Z_n \xrightarrow{p} Z$) will be used to denote the fact that the sequence of random variables, Z_n, converges in distribution (probability) to the random variable Z.

where Z is normally distributed with mean 0. The good news here is that the relative error $(\hat{h}_n - h_{n0})/h_{n0}$ converges to 0 in probability as $n \to \infty$. The bad news is that the rate of convergence is very slow, $n^{-1/10}$, which confirms the large variability of cross-validation alluded to previously.

Apart from cross-validation, there are many other bandwidth selection methods for the kernel density estimator. Some of them are described below.

2.1.3 Other bandwidth selection methods

Pseudo-likelihood cross-validation

Inspired by the leave-one-out device, Habbema et al. (1974) proposed estimating the optimal bandwidth by means of the maximum likelihood procedure, that is, by maximizing

$$L(h) = \prod_{i=1}^{n} \hat{f}_h^{(-i)}(X_i),$$

where $\hat{f}_h^{(-i)}$ denotes the Parzen-Rosenblatt estimator constructed without the i-th observation. Despite its apparent appeal, the pseudo-likelihood bandwidth has a major drawback (Broniatowski et al., 1989), namely its bad behavior when working with heavy-tailed densities.

Smoothed cross-validation

Despite its name, the smoothed cross-validation selector has more to do with the bootstrap bandwidth selectors than with cross-validation (see Cao et al., 1994). This bandwidth selector was proposed in Hall et al. (1992) and is based on the bias-variance decomposition of the MISE:

$$M_n(h) = \int \operatorname{var}\left[\hat{f}_h(x)\right] dx + \int \left\{\mathrm{E}\left[\hat{f}_h(x)\right] - f(x)\right\}^2 dx.$$

As we have already seen, the dominant term of the variance, $R(K)/(nh)$, does not depend on f, while the dominant term of the bias does. Thus, the smoothed

cross-validation method proceeds by estimating the squared-bias term by

$$\hat{B}(h;g) = \int \left\{ \left[\int K_h(x-y)\hat{f}_g(y)\,dy \right] - \hat{f}_g(x) \right\}^2 dx,$$

where $g > 0$ is a pilot bandwidth. Then, the smoothed cross-validation bandwidth can be defined as

$$\hat{h}_{SCV,n} = \arg\min_{h>0} \left[\frac{R(K)}{nh} + \hat{B}(h;g) \right].$$

The smoothed cross-validation bandwidth has been shown to have better rates of convergence than the ordinary, least-squares cross-validation bandwidth. In particular, for certain choices of g, Jones et al. (1991) show that

$$\frac{\hat{h}_{SCV,n} - h_{n0}}{h_{n0}} = O_p\left(n^{-1/2}\right),$$

which is a much better rate than the well-known $n^{-1/10}$ rate of least-squares cross-validation.

Direct plug-in bandwidth

While some bandwidth selectors, such as those based on cross-validatory criteria, try to directly estimate the optimal bandwidth, h_{n0}, plug-in bandwidth selectors aim to estimate the asymptotically optimal bandwidth, h_{na}, defined in (2.5). That is, plug-in bandwidth selectors address the problem of estimating

$$h_{na} = C_0 n^{-1/5}$$

by replacing the constant C_0, which was defined in (2.6), by an estimate, C_0, leading to

$$\hat{h}_{n,dpi} = \hat{C}_0 n^{-1/5}. \tag{2.10}$$

The bandwidth selector defined by (2.10) is usually called the direct plug-in bandwidth. Note that the only unknown term in C_0 is $R(f'')$ so it is necessary

to address the issue of estimating integrated squared density derivatives, that is, functionals of the form $R\left[f^{k)}\right]$. Let us start by noting that

$$R\left[f^{k)}\right] = \int f^{k)}(x)^2\, dx$$

can be expressed as (Wand and Jones, 1995)

$$R\left[f^{k)}\right] = (-1)^k \int f^{2k)}(x) f(x)\, dx$$

under sufficient differentiability assumptions on f. Thus, we can limit ourselves to the study of functionals of the form

$$\psi_k = \int f^{k)} f(x)\, dx. \tag{2.11}$$

The fact that

$$\psi_k = \mathrm{E}\left[f^{k)}(X)\right]$$

motivates estimating (2.11) by (Wand and Jones, 1995)

$$\hat{\psi}_{k,g} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_g^{k)}(X_i - X_j), \tag{2.12}$$

where g is usually referred to as the pilot bandwidth, L is a symmetric kernel function of order t (that is, a kernel function whose first non-zero moment has degree t) and both g and L can be different from h and K, respectively. Again, the choice of g has a significant effect on the behavior of the estimator defined in (2.12) and for this reason and under certain regularity conditions an expression for the optimal value of the pilot bandwidth is provided in Wand and Jones (1995), namely,

$$g_{0,k} = \left[\frac{t! L^{(k)}(0)}{-\mu_t(L)\psi_{k+t} n}\right]^{1/(k+t+1)}.$$

However, when estimating (2.11) by means of $\hat{\psi}_{k,g_{0,k}}$ we are again confronted

with the problem of estimating a higher order functional, namely ψ_{k+t}. Naturally, if we wanted to estimate this new functional by means of $\hat{\psi}_{k+t,g_{0,k+t}}$ then we would need to select a new pilot bandwidth, whose optimal expression will in turn depend on another functional of an even higher order, and this process would continue indefinitely. The usual way to solve this problem is to stop the process after a predetermined number of steps and, in the last step, to estimate the corresponding functional by means of an automatic estimation which generally involves imposing some type of parametric assumptions on f, typically a normal scale final step.

Normal scale bandwidth

The normal scale bandwidth selector is probably the simplest bandwidth selection method as it is based on the assumption that the underlying density, f, is a normal with variance σ^2. In this case, the asymptotically optimal bandwidth defined in (2.5) can be expressed as

$$h_{na} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5} \sigma.$$

The normal scale bandwidth, $\hat{h}_{n,NS}$, simply replaces σ by an estimate, $\hat{\sigma}$, leading to

$$\hat{h}_{n,NS} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5} \hat{\sigma},$$

where σ is usually estimated by means of the sample standard deviation, s_n, or the standardised interquartile range,

$$\hat{\sigma}_{IQR} = \frac{F_n^{-1}(0.75) - F_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

where Φ denotes the cumulative distribution function of the standard normal and, hence, Φ^{-1} denotes its quantile function. When K is the Gaussian kernel and σ is estimated by

$$\hat{\sigma} = \min\{s_n, \hat{\sigma}_{IQR}\},$$

then the normal scale bandwidth is usually referred to as the rule-of-thumb bandwidth, which can be expressed as

$$\hat{h}_{n,RT} = \left(\frac{4}{3n}\right)^{1/5} \hat{\sigma}.$$

Although the normal scale bandwidth selector provides a fast and automatic way of selecting the bandwidth of the kernel density estimator, it is well known that it leads to oversmoothed estimates when f deviates from normality. This is illustrated in Figure 2.6, where the kernel density estimates with bandwidths chosen by direct plug-in and rule-of-thumb methods and obtained for a sample of size $n = 5000$ drawn from the mixture density D1 are shown. As we can see, while the direct plug-in selector ($h = 0.081$) produces an adequate estimate that captures the most important features of the target density, the rule-of-thumb bandwidth ($h = 0.17$) leads to an oversmoothed estimate.

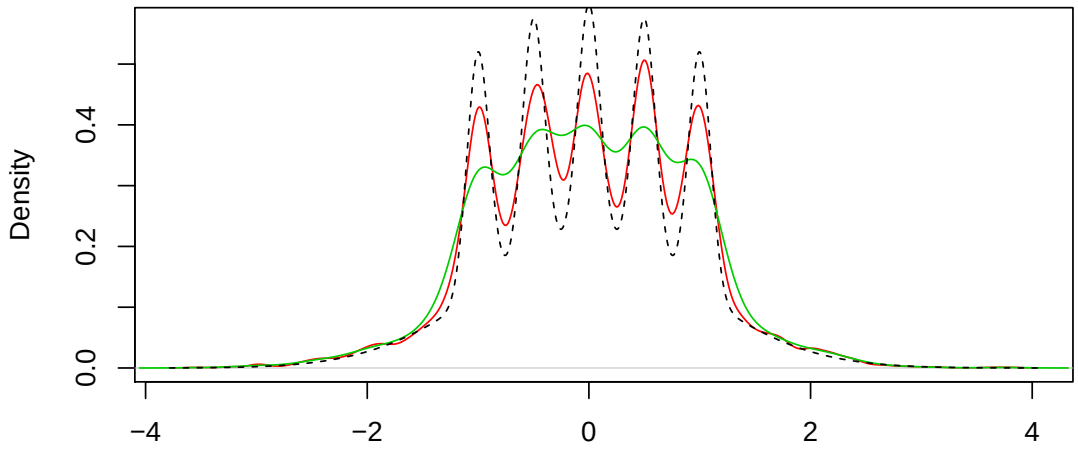

Figure 2.6: Target density (dashed black line) and kernel density estimates considering the direct plug-in (solid red line) and rule-of-thumb bandwidth (solid green line) for a sample of size $n = 5000$ drawn from density D1.

Bootstrap bandwidth

Consider a sample of size n, $X_1, \ldots, X_n$. Under the assumption that $h \to 0$ and $nh \to \infty$, it is well known (Parzen, 1962) that the Parzen-Rosenblatt estimator, $\hat{f}_h(x)$, has the following limit distribution,

$$\sqrt{nh}\left[\hat{f}_h(x) - f(x)\right] \xrightarrow{d} \mathrm{N}(b_0, v_0),$$

where

$$\begin{aligned} b_0 &= \frac{1}{2} c_0^{5/2} \mu_2(K) f''(x), \\ v_0 &= R(K) f(x), \\ c_0 &= \left[\frac{R(K) f(x)}{\mu_2(K)^2 f''(x)^2}\right]^{1/5}. \end{aligned}$$

If we were interested in approximating the sampling distribution of the previous statistic, the bootstrap resampling plan would be as follows (Cao, 1990):

Step 1. Consider a pilot bandwidth, g, and compute $\hat{f}_g$ from the original sample.

Step 2. Draw bootstrap resamples of size n, $X_1^*, \ldots, X_n^*$ from a population whose density function is given by $\hat{f}_g$. This can be done as follows:

(a) Generate a sample of size n, $U_1, \ldots, U_n$, where U_i is drawn from a discrete uniform distribution defined in $\{1, \ldots, n\}$, for every $i \in \{1, \ldots, n\}$.

(b) Generate a sample of size n, $Z_1, \ldots, Z_n$, where Z_i is drawn from the density function K, for every $i \in \{1, \ldots, n\}$.

(c) For every $i \in \{1, \ldots, n\}$, define $X_i^* = X_{U_i} + gZ_i$.

Step 3. Construct the bootstrap version of the Parzen-Rosenblatt estimator, $\hat{f}_h^*$, where

$$\hat{f}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right).$$

Step 4. Approximate the sampling distribution of $\sqrt{nh}\left[\hat{f}_h(x) - f(x)\right]$ by the resampling distribution of $\sqrt{nh}\left[\hat{f}_h^*(x) - \hat{f}_g(x)\right]$.

The resampling plan above does not aim at selecting the bandwidth for the kernel density estimator. However, it is still valid to bootstrap estimate the MISE function and provide a selector for the bandwidth of the Parzen–Rosenblatt estimator. Thus, if instead of the sampling distribution of $\hat{f}_h$, we are actually interested in the MISE of $\hat{f}_h$, the resampling procedure would consist of repeating steps 1–3 above and replacing step 4 with the following:

Step 4. Define the bootstrap version of the MISE as

$$M_n^*(h;g) = \mathrm{E}^* \left\{ \int \left[\hat{f}_h^*(x) - \hat{f}_g(x) \right]^2 dx \right\}, \tag{2.13}$$

where E^* denotes the fact that the previous expectation is evaluated for a random variable belonging to the bootstrap population and whose density function is given by $\hat{f}_g$, that is, E^* denotes the expectation in the resampling mechanism, conditionally on the original sample.

It should be noted that the bootstrap version of the MISE depends on the original sample but not on the resamples, and since all the terms that appear in (2.13) are known, it is not really necessary to draw any resample and approximate (2.13) by Monte Carlo. Cao (1993) gives a closed expression for (2.13),

$$M_n^*(h;g) = V_n^*(h;g) + B_n^*(h;g), \tag{2.14}$$

where

$$\begin{aligned} V_n^*(h;g) &= n^{-1}h^{-1}R(K) + n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(K_h * K_g) * (K_h * K_g)\right](X_i - X_j), \\ B_n^*(h;g) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(K_h * K_g - K_g) * (K_h * K_g - K_g)\right](X_i - X_j). \end{aligned}$$

Furthermore, if we assume that K is the Gaussian kernel and using the fact that $K_{h_1} * K_{h_2}$ is the density function of a normally distributed random variable with zero mean and variance given by $h_1^2 + h_2^2$ for any $h_1, h_2 > 0$, then (2.14) can be rewritten

as

$$\begin{aligned} M_n^*(h;g) &= \frac{R(K)}{nh} - \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\sqrt{2h^2+2g^2}}(X_i - X_j) \\ &+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[K_{\sqrt{2h^2+2g^2}}(X_i - X_j) \\ &- 2K_{\sqrt{h^2+2g^2}}(X_i - X_j) + K_{\sqrt{2g^2}}(X_i - X_j)\Big]. \end{aligned}$$

Since the problem of choosing the pilot bandwidth is closely linked to that of estimating the curvature of f, a sensible optimality criterion for the pilot bandwidth would be

$$g_0 = \arg\min_{g>0} \mathrm{E}\left\{\left[\int \hat{f}_g''(x)^2\,dx - \int f''(x)^2\,dx\right]^2\right\}. \tag{2.15}$$

Cao (1993) also provides an expression for the dominant term of the solution of the minimization problem defined by (2.15),

$$g_0 = \left[\frac{R\left(K''\right)}{n\mu_2(K)R\left(f'''\right)}\right]^{1/7} + o\left(n^{-1/7}\right).$$

Thus, we can directly calculate the bootstrap MISE bandwidth, h_{n0}^*, by minimizing (2.14), that is,

$$h_{n0}^* = \arg\min_{h>0}\left[V_n^*(h;g_0) + B_n^*(h;g_0)\right]. \tag{2.16}$$

2.2 Regression estimation

Regression analysis studies the relationship between an independent random variable or explanatory variable, X, and a dependent random variable or response, Y. The relationship between these variables is encapsulated in the regression function, $m(x) = \mathrm{E}\left(Y \mid X = x\right)$, whose estimation is often the main task in regression analysis. In this regard, most statistical methods for estimating m can be classified into two categories depending on their assumptions on m: parametric and nonparamet-

ric regression methods. Parametric methods are based on the assumption that the unknown regression function belongs to a certain parametric family of functions, depending on some parameters, and, therefore, the problem of estimating m is reduced to estimating these parameters. On the other hand, nonparametric methods do not assume any parametric form for the relationship between the variables, but rather estimate it from the data itself, which makes them more flexible than parametric methods. A particular class of nonparametric methods that are of interest to us is that of kernel regression methods, which seek to estimate m as a locally weighted average, using a kernel function as a weighting function. As in the case of kernel density estimation (see Section 2.1.1), these methods are highly dependent on the choice of a free parameter called bandwidth or smoothing parameter which determines the amount of smoothing performed by the estimator, which in turn determines the trade-off between the bias and the variance of the estimator. The problem of bandwidth selection is therefore crucial and intrinsic to kernel methods and multiple ways of addressing it have been proposed and studied over time, these including cross-validation (Härdle et al., 1988), bootstrapping (Cao and González-Manteiga, 1993) or plug-in methods (Ruppert et al., 1995).

2.2.1 Kernel regression estimation

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

$$Y = m(X) + \varepsilon,$$

where $m(x) = \mathrm{E}\,(Y \mid X = x)$ denotes the regression function and ε denotes the error term, which in turn satisfies the following conditions:

$$\begin{aligned} \mathrm{E}\,(\varepsilon \mid X = x) &= 0, \\ \mathrm{E}\left(\varepsilon^2 \mid X = x\right) &= \sigma^2(x). \end{aligned}$$

Analogously to what was presented in Section 2.1.1, which was dedicated to ker-

nel density estimation, kernel regression methods attempt to estimate the unknown regression function without imposing parametric constraints on m. Depending on the type of design, that is, depending on whether we are working in the context of fixed or random design, different types of kernel regression methods have been proposed and studied in the literature. In the case of fixed design, it is worth mentioning the Priestley-Chao estimator (Priestley and Chao, 1972) and the Gasser-Müller estimator (Gasser and Müller, 1979). Under the assumption of random design (although these methods can also be applied to the fixed design case), a particular and widely used class of kernel regression methods is that of local polynomial kernel estimators (Stone, 1977; Cleveland, 1979; Fan, 1992), which estimate $m(x)$ by locally fitting a d-th degree polynomial to the sample via weighted least squares. In other words, at a given point x these methods estimate $m(x)$ by fitting the polynomial

$$P_d(u) = \beta_0 + \beta_1(u - x) + \cdots + \beta_p(u - x)^d$$

to the sample and weighing the i-th observation, (X_i, Y_i), by $K_h(X_i - x)$, that is, the weight of the i-th observation depends on the proximity of X_i to x, which in turn is measured by the rescaled kernel K_h. Thus, the vector of estimates, $(\hat{\beta}_0, \ldots, \hat{\beta}_d)$, can be obtained as

$$(\hat{\beta}_0, \ldots, \hat{\beta}_d) = \underset{(\beta_0,\ldots,\beta_d)}{\arg\min} \sum_{i=1}^{n} \left[Y_i - P_d(X_i)\right]^2 K_h\left(X_i - x\right).$$

Then, the local polynomial kernel estimate of degree d at x can be defined as

$$\hat{m}_h(x; d) = \hat{\beta}_0.$$

The Nadaraya–Watson estimator or local constant estimator (Nadaraya, 1964; Watson, 1964) is a particular case of the local polynomial kernel estimator which corresponds to $d = 0$, that is, fitting 0-degree polynomials or local constants. The

Nadaraya–Watson estimator admits the following expression:

$$\hat{m}_h(x;0) = \frac{\sum_{i=1}^{n} K_h\left(x - X_i\right) Y_i}{\sum_{i=1}^{n} K_h\left(x - X_i\right)}, \tag{2.17}$$

where $h > 0$ denotes the bandwidth and K denotes the kernel function. For the sake of simplicity, in what follows the Nadaraya–Watson estimator will be denoted by $\hat{m}_h$. On the other hand, the local linear estimator, which corresponds to the case $d = 1$ can be expressed as

$$\hat{m}_h(x;1) = \boldsymbol{e_1'}\left(\boldsymbol{X_{x,1}'}\boldsymbol{W_{x,h}}\boldsymbol{X_{x,1}}\right)^{-1}\boldsymbol{X_{x,1}'}\boldsymbol{W_{x,h}}\boldsymbol{Y},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{W_{x,h}} = \operatorname{diag}\left[K_h(X_1 - x), \ldots, K_h(X_n - x)\right]$, $\boldsymbol{e_j}$ is a column vector having 1 in its j-th entry and zeros elsewhere and

$$\boldsymbol{X_{x,d}} = \begin{bmatrix} 1 & X_1 - x & \ldots & (X_1 - x)^d \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & \ldots & (X_n - x)^d \end{bmatrix}.$$

As in the case of density estimation, 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, as illustrated in Figure 2.7. Hence, we are faced with the problem of bandwidth selection.

Optimal bandwidths ofter refer to smoothing parameter values that minimize 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, our focus is on 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).

Taking this comment into account, a possible (global) criterion for optimality is that of the mean integrated squared error or MISE which, in the case of regression

estimation, is defined as

$$M_{n;d}(h) = \mathrm{E}\left\{ \int \left[\hat{m}_h(x;d) - m(x)\right]^2 f(x)\, dx \right\}, \tag{2.18}$$

where f denotes the marginal density function of the explanatory variable, X. The bandwidth that minimizes (2.18) is called the MISE bandwidth and is denoted by $h_{n0;d}$, that is,

$$h_{n0;d} = \arg\min_{h>0} M_{n;d}(h). \tag{2.19}$$

In the case of the Nadaraya-Watson estimator, the MISE function and its minimizer are denoted by $M_n(h)$ and h_{n0}, respectively. Although the same notation was used in Section 2.1.1 on kernel density estimation, we will keep it anyway since there will be no possibility of confusion in any case.

The MISE bandwidth depends on m and f and since in practice both of these functions are almost always unknown, h_{n0} cannot be directly calculated. However, it can be estimated and to that effect numerous bandwidth selection methods, these including cross-validation (Härdle et al., 1988), bootstrap (Cao and González-Manteiga, 1993) and plug-in (Ruppert et al., 1995) methods, have been proposed and studied over time. In the following section the leave-one-out cross-validation selector for the bandwidth of the kernel regression estimator is described.

2.2.2 Cross-validation method for bandwidth selection

Cross-validation is a method that provides an optimality criterion for the selection of the bandwidth of the kernel regression estimator, $\hat{m}_h(x;d)$, which works as an empirical analogue of the MISE and so it allows us to estimate $h_{n0;d}$. While in the case of density estimation the cross-validation criterion is based on the integrated squared error or ISE, in the case of regression estimation the cross-validation criterion seeks to minimize the prediction error of our estimator. In this sense, it would seem intuitive to try to find the bandwidth that minimizes the residual sum of squares of

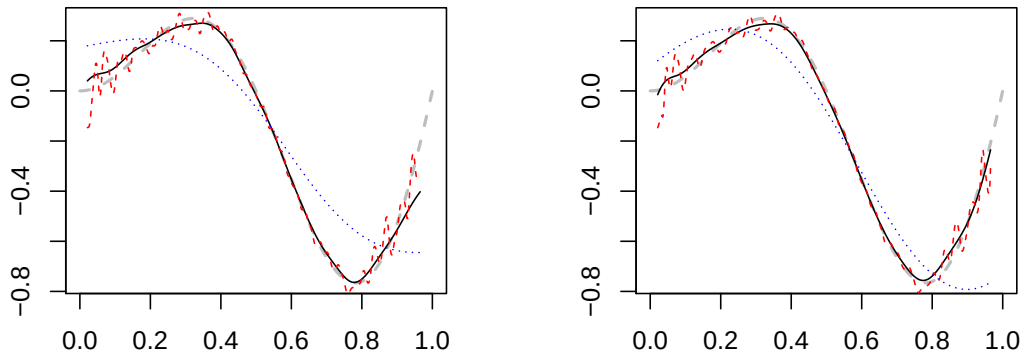

Figure 2.7: Local constant (left) and local linear (right) estimators for a sample of size 5000 drawn from the model $Y = m(X) + \varepsilon$, with X drawn from a Beta(3,3) distribution, ε drawn from a N(0,0.3) distribution and $m(x) = x\sin(2\pi x)$ (thick dashed gray line). A Gaussian kernel was considered and both estimators are plotted considering different values for the bandwidth, namely $h = h_{n0}$ (continuous black line), $h = h_{n0}/2$ (dashed red line) and $h = 2h_{n0}$ (dotted blue line). Note that the value of h_{n0} is different for the local constant and local linear estimators.

the corresponding estimator, given by

$$\sum_{i=1}^{n} \left[\hat{m}_h(X_i; d) - Y_i\right]^2 . \tag{2.20}$$

However, a bandwidth selector based on the minimization of (2.20) would lead to overfitting since, in the expression of the residual sum of squares, the same sample, $\mathcal{X}$, is used both to estimate m and to check the goodness-of-fit of $\hat{m}_h(\cdot; d)$. To avoid this problem, the cross-validation criterion proposes a modification of (2.20) so that the prediction error at X_i is estimated by the kernel regression estimator constructed without the i-th observation, (X_i, Y_i). Therefore, in the context of regression esti-

mation the leave-one-out cross-validation function can be defined as

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

where $\hat{m}_h^{(-i)}(\cdot;d)$ denotes the kernel regression estimator constructed using $\mathcal{X}\backslash\{(X_i,Y_i)\}$, that is, leaving out the i-th observation. Hence, the cross-validation bandwidth, $\hat{h}_{CV,n;d}$, can be defined as the bandwidth that minimizes $CV_n(\cdot;d)$, that is,

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

In the case of the Nadaraya–Watson estimator, the following notation will be used:

$$CV_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - Y_i\right]^2, \tag{2.21}$$

$$\hat{h}_{CV,n} = \arg\min_{h>0} CV_n(h). \tag{2.22}$$

As in the case of the kernel density estimator, there are a multitude of bandwidth selection methods for the local polynomial kernel regression estimator other than cross-validation. Some of them are described below.

2.2.3 Other bandwidth selection methods

Rule-of-thumb bandwidth

Similarly to the normal scale bandwidth discussed in the case of kernel density estimation, the so-called rule-of-thumb bandwidth selector proposed in Fan and Gijbels (1996) offers a quick and automatic way to select the bandwidth of the local polynomial kernel estimator. For the sake of simplicity, we will describe this bandwidth selector for the local linear estimator. In this case, the asymptotically optimal bandwidth admits the following expression,

$$h_{na} = \left[\frac{R(K)\int \sigma^2(x)\,dx}{\mu_2(K)^2\theta_{22}n}\right]^{1/5}, \tag{2.23}$$

where

$$\theta_{22} = \int m''(x)^2 f(x)\, dx.$$

Fan and Gijbels (1996) propose to replace the unknown quantities in (2.23) by parametric ordinary least squares estimates. Specifically, they suggest fitting a quartic polynomial,

$$P_q(x) = \sum_{i=0}^{4} \alpha_i x^i,$$

to the sample and estimate its parameters by ordinary least squares. Then, $m''(x)$ can be estimated by

$$\hat{P}_q''(x) = 2\hat{\alpha}_2 + 6\hat{\alpha}_3 x + 12\hat{\alpha}_4 x^2$$

and so

$$\hat{\theta}_{22} = \frac{1}{n} \sum_{i=1}^{n} \hat{P}_q''(X_i)^2,$$

where we have used the fact that

$$\theta_{22} = \mathrm{E}\left[m''(X)^2\right].$$

As for the other unknown quantity in (2.23), namely the integrated conditional variance, it is estimated by assuming homoscedasticity and replacing σ^2 by

$$\hat{\sigma}_q^2 = \frac{1}{n-5} \sum_{i=1}^{n} \left[Y_i - \hat{P}_q(X_i)\right]^2.$$

Furthermore, if one assumes that the conditional variance function is zero outside the support of X, then the integrated conditional variance can be estimated by

$$\left[X_{(n)} - X_{(1)}\right] \hat{\sigma}_q^2,$$

where $X_{(k)}$ denotes the k-th order statistic. In this case, the resulting bandwidth is called the rule-of-thumb bandwidth and admits the following expression,

$$\hat{h}_{n,RT} = \left\{ \frac{R(K)\left[X_{(n)} - X_{(1)}\right]\hat{\sigma}_q^2}{\mu_2(K)^2 \hat{\theta}_{22} n} \right\}^{1/5}.$$

Direct plug-in bandwidth

As in kernel density estimation, the main idea of plug-in bandwidth selectors is to replace the unknown terms appearing in the expression of the asymptotically optimal bandwidth which, in the case of the local linear estimator, is given by (2.23). As previously mentioned, the only unknown quantities in (2.23) are the density-weighted curvature of m,

$$\theta_{22} = \int m''(x)^2 f(x)\, dx$$

and the integrated conditional variance,

$$\int \sigma^2(x)\, dx.$$

The properties of functionals of the form

$$\theta_{st} = \int m^{s)}(x) m^{t)}(x) f(x)\, dx, \quad s, t \geq 0, \quad s + t \text{ even}$$

were studied in Ruppert et al. (1995), where they propose to estimate θ_{22} by

$$\hat{\theta}_{22,g} = \frac{1}{n} \sum_{i=1}^{n} \widehat{m''_g}(X_i; 3), \tag{2.24}$$

where

$$\widehat{m_h^{s)}}(x; d) = s! \boldsymbol{e}'_{\boldsymbol{s+1}} \left(\boldsymbol{X}'_{\boldsymbol{x,d}} \boldsymbol{W}_{\boldsymbol{x,h}} \boldsymbol{X}_{\boldsymbol{x,d}} \right)^{-1} \boldsymbol{X}'_{\boldsymbol{x,d}} \boldsymbol{W}_{\boldsymbol{x,h}} \boldsymbol{Y}, \quad s = 0, \ldots, d$$

and g is called the pilot bandwidth. Ruppert et al. (1995) also show that the optimal value of g, in the sense of minimizing the conditional asymptotic mean squared error

of (2.24), admits the following expression under the assumptions that the errors are homoscedastic with common variance σ^2 and the support of the density function f is $[a, b]$:

$$g_{AMSE} = C_2(K) \left[\frac{\sigma^2 (b-a)}{|\theta_{24}| n} \right]^{1/7},$$

where the quantity $C_2(K)$ only depends on the kernel function K and is defined in Ruppert et al. (1995). As for the common variance σ^2, Ruppert et al. (1995) propose to estimate it by

$$\hat{\sigma}^2_\lambda = \frac{1}{\nu} \sum_{i=1}^{n} \left[Y_i - \hat{m}_\lambda(X_i; 1) \right]^2, \tag{2.25}$$

where

$$\nu = n - 2 \sum_{i=1}^{n} w_{ii} + \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^2$$

and

$$w_{ij} = \boldsymbol{e_1'} \left(\boldsymbol{X'_{X_i,1}} \boldsymbol{W_{X_i,\lambda}} \boldsymbol{X_{X_i,1}} \right)^{-1} \boldsymbol{X'_{X_i,1}} \boldsymbol{W_{X_i,\lambda}} \boldsymbol{e_j}.$$

Ruppert et al. (1995) show that the bandwidth that minimizes the conditional asymptotic mean squared error of (2.25) verifies

$$\lambda_{AMSE} = C_3(K) \left[\frac{\sigma^4 (b-a)}{\theta_{22}^2 n^2} \right]^{1/9},$$

where the quantity $C_3(K)$ only depends on the kernel function K and is defined in Ruppert et al. (1995). As usual in plug-in methods, other unknown terms appear in the expression of the optimal pilot bandwidth, in this case of the form θ_{st}. Therefore, the process of selecting the optimal pilot bandwidth must stop at some point. This is done by estimating the corresponding functional, θ_{st}, by some automatic estimation method, typically a normal scale final step.

Bootstrap bandwidth

An alternative to the bandwidth selection methods reviewed so far would be to approximate the $M_n(h)$ function itself by resampling. Cao and González-Manteiga (1993) proposed a resampling plan to approximate the sampling distribution of the Nadaraya-Watson estimator. The main idea of this resampling plan can be used to select the bandwidth of the Nadaraya-Watson estimator by minimizing a certain bootstrap version of the MISE. The resampling plan is as follows:

Step 1. Select a pilot bandwidth, $g > 0$.

Step 2. Draw bootstrap samples, $\mathcal{X}^* = \{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)\}$, from the distribution function

$$\hat{F}(x, y) = \frac{1}{n} \sum_{i=1}^{n} 1_{S_y}(Y_i) \int_{-\infty}^{x} K_g(u - X_i)\, du,$$

where $S_y = \{Y_j \in \{Y_1, \ldots, Y_n\} \mid Y_j \leq y\}$. In order to do this, generate a sample of size n, $U_1, \ldots, U_n$, where U_i is drawn from a discrete uniform distribution defined in $\{1, \ldots, n\}$, for every $i = 1, \ldots, n$. Also, generate a sample of size n, $Z_1, \ldots, Z_n$, where Z_i is drawn from the density K, for every $i = 1, \ldots, n$. Then, define

$$X_i^* = X_{U_i} + gZ_i, \quad i = 1, \ldots, n.$$

Finally, simulate Y_i^*, $i = 1, \ldots, n$, from the discrete distribution defined in $\{Y_1, \ldots, Y_n\}$ that assigns to each Y_i probability w_i, with

$$w_i = \frac{K_g(X_i^* - X_i)}{\sum_{j=1}^{n} K_g(X_i^* - X_j)}, \quad i = 1, \ldots, n.$$

Step 3. Consider the Nadaraya–Watson estimator constructed with the bootstrap sam-

ple $\mathcal{X}^* = \{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)\}$,

$$\hat{m}_h^*(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i^*) Y_i^*}{\sum_{i=1}^{n} K_h(x - X_i^*)}.$$

Step 4. Repeat the previous steps B times and estimate $M_n(h)$ by a Monte Carlo approximation of

$$M_n^*(h; g) = \mathrm{E}^* \left\{ \int \left[\hat{m}_h^*(x) - \hat{m}_g(x) \right]^2 \hat{f}_g(x)\, dx \right\}.$$

Step 5. Repeat Steps 1–4 for a large number of values of h and define the bootstrap bandwidth as

$$h_{n0}^* = \arg\min_{h>0} M_n^*(h; g). \tag{2.26}$$

When defining the bootstrap bandwidth, other resampling plans could be considered instead of the one described above. For example, instead of just smoothing the explanatory variable, a fully smoothed bootstrap could be considered by modifying Step 2 above so that the resamples are now drawn from the distribution

$$\tilde{F}_g(x, y) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{y} K_g(v - Y_i)\, dv \int_{-\infty}^{x} K_g(u - X_i)\, du.$$

On the other hand, although the previous plan mimics heteroscedasticity, if the data was indeed heteroscedastic, it would make sense to consider the resampling plan known as wild bootstrap (Wu, 1986; Härdle and Marron, 1991).

2.3 Bootstrapping

Resampling refers to a class of statistical methods whose main idea is to generate new samples from one's own data with the objective of drawing certain conclusions

about a population parameter or sample statistic. Resampling methods, among which bootstrapping (Efron, 1979), jackknife (Quenouille, 1949; Tukey, 1958) and permutation tests (Fisher, 1935) should be highlighted, have multiple applications such as estimating the precision of a sample statistic, hypothesis testing or model validation.

Bootstrapping is a resampling method that allows us to estimate the sampling distribution of a certain statistic. This is done by assuming that the sample at hand is representative of the population from which it was drawn and sampling with replacement from the sample itself. In this way, we are imitating the original sampling process but with the advantage of working with a new, fully known (bootstrap) population. A simple version of the bootstrap method, generally known as naïve bootstrap, consists of drawing the resamples from the empirical distribution function, but this approach is known to fail in several situations. A more elaborate way to carry out the resampling process is that of the smooth bootstrap, in which the bootstrap population is not characterized by the empirical distribution function, but instead by a smooth estimate of the unknown density function. In particular, the resamples would belong to a population whose density function is given by a kernel estimate of the unknown density function, namely $\hat{f}_g$, where g is often referred to as the pilot bandwidth. Of course, this way of proceeding depends largely on the choice of g and, therefore, it is necessary to establish some optimality criterion for the pilot bandwidth.

In more detail, let $\mathcal{X}$ be a simple random sample of size n drawn from a population whose distribution function is given by F and suppose that we are interested in making inference about some population parameter $\theta = \theta(F)$. To do so, it is necessary to know the sampling distribution of a certain statistic $R(\mathcal{X}, F)$, which in many cases can take the form

$$R(\mathcal{X}, F) = \theta(F_n) - \theta(F),$$

where F_n denotes the empirical distribution function of $\mathcal{X}$. Naturally, the sampling distribution of $R(\mathcal{X}, F)$ is almost always unknown. The bootstrap approach starts by replacing F with an estimate, $\hat{F}$. From $\hat{F}$ and conditionally to the sample $\mathcal{X}$ we can draw resamples of size n, $\mathcal{X}^*$, which are usually called bootstrap samples. The

idea of the bootstrap method then lies in approximating the sampling distribution of $R(\mathcal{X}, F)$ by the resampling distribution or bootstap distribution of

$$R(\mathcal{X}^*, \hat{F}) = \theta(F_n^*) - \theta(\hat{F}),$$

where F_n^* denotes the empirical distribution of the bootstrap sample, $\mathcal{X}^*$. Again, the bootstrap distribution of $R(\mathcal{X}^*, \hat{F})$ is usually not computable in practice and must be approximated by Monte Carlo.

As an example, let us imagine that we have a sample of size n, $\mathcal{X} = (X_1, \ldots, X_n)$, drawn from a $\mathrm{N}(\mu, \sigma^2)$ and we are interested in approximating the bias of the sample mean by means of the naïve bootstap. Then we have

$$\begin{aligned}
\theta(F) &= \mu = \int x \, dF(x), \\
\theta(F_n) &= \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \int x \, dF_n(x), \\
\theta(F_n^*) &= \bar{X}_n^* = \frac{1}{n} \sum_{i=1}^{n} X_i^* = \int x \, dF_n^*(x)
\end{aligned}$$

and so we want to approximate the sampling distribution of

$$R(\mathcal{X}, F) = \bar{X}_n - \mu$$

by the bootstrap distribution of

$$R(\mathcal{X}^*, F_n) = \bar{X}_n^* - \bar{X}_n.$$

In this case we can apply the central limit theorem to obtain the distributions of both $R(\mathcal{X}, F)$ and $R(\mathcal{X}^*, F_n)$, namely

$$\begin{aligned}
R(\mathcal{X}, F) &\stackrel{d}{\simeq} \mathrm{N}(0, \sigma^2/n), \\
R(\mathcal{X}^*, F_n) &\stackrel{d}{\simeq} \mathrm{N}(0, s_n^2/n),
\end{aligned}$$

where s_n^2 denotes the sample variance. Figure 2.8 shows the sampling distribution of $R(\mathcal{X}, F)$ and the bootstrap distribution of $R(\mathcal{X}^*, F_n)$ along with a Monte Carlo

approximation of the latter for a sample of size $n = 1000$ drawn from a $\mathrm{N}(0, 1)$.

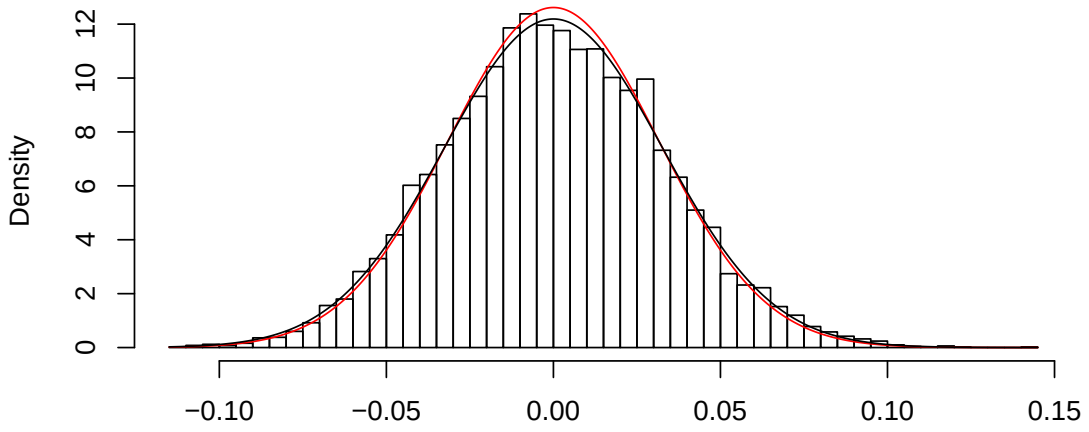

Figure 2.8: The sampling distribution of $\bar{X}_n - \mu$ (red) and the bootstrap distribution of $\bar{X}_n^* - \bar{X}_n$ (black) for a sample of size $n = 1000$ drawn from a $\mathrm{N}(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$ are shown. A Monte Carlo approximation of the bootstrap distribution of $\bar{X}_n^* - \bar{X}_n$ constructed with 10^4 bootstrap samples is also shown, using a histogram.

2.4 Bagging

Ensemble methods (Opitz and Maclin, 1999) are a family of techniques that combine the estimates or predictions of several base estimators or base models with the objective of producing a new estimator or predictor with better statistical properties. One of the most popular and widely used ensemble methods is bootstrap aggregating (Breiman, 1996a), also known as bagging, which is a resampling technique whose main purpose is to reduce the variability of a given base estimator. It is best suited for high-variance low-bias estimators. In the case of estimators which are nonlinear in the observations, such as decision trees or neural networks, it has been shown (Friedman and Hall, 2007) that bagging can lead to substantial reductions in the

variability of these estimators. This fact motivates us to study the application of bagging to the cross-validation bandwidth selector, a statistic that is not linear in the observations and has a high variability.

More precisely, let $\mathcal{X}$ denote a sample of size n drawn from the distribution P, $\hat{h} = \hat{h}(\mathcal{X})$ a base selector of the optimal bandwidth, h_{n0}, and

$$\hat{h}_A = \mathrm{E}_P\left[\hat{h}\left(\mathcal{X}\right)\right]$$

the "aggregated bandwidth" (following the notation of Breiman, 1996a), where E_P denotes the expectation over all samples of size n drawn from the distribution P. Then we have that

$$\begin{aligned}
\mathrm{E}_P\left\{\left[h_{n0} - \hat{h}(\mathcal{X})\right]^2\right\} &= h_{n0}^2 - 2h_{n0}\hat{h}_A + \mathrm{E}_P\left[\hat{h}(\mathcal{X})^2\right] \\
&\geq \left\{\mathrm{E}_P\left[h_{n0} - \hat{h}(\mathcal{X})\right]\right\}^2 \\
&= h_{n0}^2 - 2h_{n0}\hat{h}_A + \hat{h}_A^2 \\
&= \left(h_{n0} - \hat{h}_A\right)^2,
\end{aligned}$$

where we have used the fact that $\mathrm{E}\left[Z^2\right] \geq \mathrm{E}\left[Z\right]^2$ for any random variable Z. In other words, the squared error of the aggregated bandwidth, $\hat{h}_A$, is lower than that of the base selector, $\hat{h}$, and the difference between the two depends on how unequal the two sides of

$$\mathrm{E}_P\left[\hat{h}(\mathcal{X})^2\right] \geq \left\{\mathrm{E}_P\left[\hat{h}(\mathcal{X})\right]\right\}^2$$

are, that is, the larger $\mathrm{var}_P[\hat{h}(\mathcal{X})]$ (or, in the sense of Breiman, 1996b, the more unstable) the base selector, the greater the decrease in squared error by the aggregated bandwidth. However, while the aggregated bandwidth depends on the distribution, P, from which $\mathcal{X}$ was drawn, the bagging selector, $\hat{h}_{bag}$, actually depends on the distribution $P_{\mathcal{X}}$ which assigns mass $1/n$ to each observation belonging to $\mathcal{X}$ (although smooth estimates of P may be considered). In other words, the bagging bandwidth

is defined as

$$\hat{h}_{bag} = \mathrm{E}_{P_{\mathcal{X}}}\left[\hat{h}(\mathcal{X}^*)\right],$$

that is, $\hat{h}_{bag}$ is the bootstrapped version of $\hat{h}_A$.

Moreover, there is a point between maximum instability and maximum stability at which $\hat{h}_{bag}$ stops improving on $\hat{h}$ in terms of error and, in fact, starts to underperform the base selector.

More generally, given a sample of size n, $\mathcal{X}$, and denoting by $\hat{\theta}_n = \hat{\theta}_n(\mathcal{X})$ the base estimator, the bagging procedure can be summarized as follows:

Step 1. Generate a bootstrap sample of size n, $\mathcal{X}^*$, by sampling with replacement from $\mathcal{X}$.

Step 2. Compute the bootstrap estimate, $\hat{\theta}^*_n(\mathcal{X}^*)$.

Step 3. Define the bagged estimate as

$$\hat{\theta}_{n,bag} = \mathrm{E}^*\left[\hat{\theta}^*_n(\mathcal{X}^*)\right], \tag{2.27}$$

where E^* denotes the expectation over all the bootstrap samples of size n drawn with replacement from $\mathcal{X}$.

It should be noted that sometimes the bagged estimate defined in Step 3 above cannot be computed in practice and we must resort to a Monte Carlo approximation. In this case, once Steps 2 and 3 above have been repeated B times, the bagged estimate would be approximated by

$$\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_n(\mathcal{X}^*_i),$$

where $\mathcal{X}^*_i$ denotes the i-th bootstrap sample. If $\hat{\theta}_n$ is an estimator of a certain population parameter θ then it follows immediately from (2.27) that

$$\hat{\theta}_{n,bag} - \theta = \hat{\theta}_n - \theta + \left[\mathrm{E}^*\left(\hat{\theta}^*_n\right) - \hat{\theta}_n\right].$$

Thus, it seems reasonable to think that, in some situations, the bagging estimator may turn out to be more biased than $\hat{\theta}_n$. However, the rationale for bagging is that this increase in bias is offset by an even greater reduction in variance. This phenomenon is illustrated in Figure 2.9, where 500 samples of size $n = 100$ were generated from a normal mixture density with $\mu = (0, 0)$, $\sigma = (1, 0.1)$ and $w = (0.1, 0.9)$ as vectors of means, standard deviations and weights, respectively. For each simulated sample, the standard leave-one-out cross-validation bandwidth, $\hat{h}$, and a smoothed bagged version of it, $\hat{h}_{bag}$, were computed. The latter was computed using 100 resamples drawn from a smooth estimate of the underlying density, namely $\hat{f}_g$, that is, the kernel density estimator with bandwidth $g = \hat{h}$. It is clear from Figure 2.9 that the bagged estimator has a significantly lower variability than the standard, non-bagged estimator. In fact, the relative reduction in variance achieved by the bagged estimator with respect to the non-bagged estimator (59.7%) more than offsets the relative increase in bias (238%), thus achieving a 53.6% reduction in mean squared error.

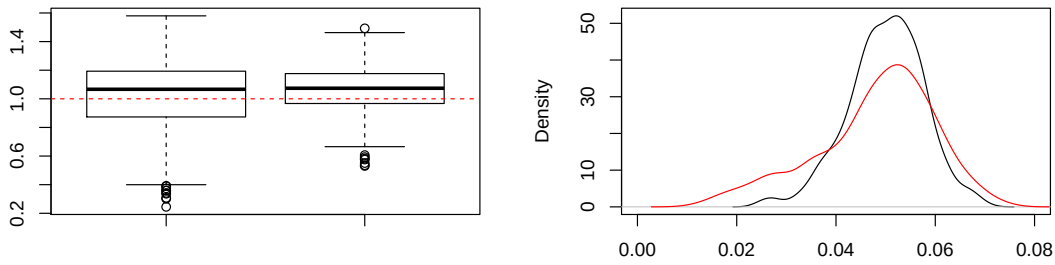

Figure 2.9: Sampling distribution of $\hat{h}/h_{n0}$ (left panel) and kernel density estimates of $\hat{h}$ (right panel), where $\hat{h}$ denotes the ordinary leave-one-out cross-validation bandwidth (left boxplot, red line) and a smoothed bagged version of it (right boxplot, black line). Both were approximated by 500 samples of size $n = 100$ generated from a normal mixture density with $\mu = (0, 0)$, $\sigma = (1, 0.1)$ and $w = (0.1, 0.9)$ as vectors of means, standard deviations and weights, respectively. For each simulated sample, the bagged estimator was computed using 100 resamples drawn from a smooth estimate of the underlying density.

Bagging has become a widely used technique especially in the field of machine

learning and multiple variants of the method (see, for example, Bühlmann and Yu, 2002) have been proposed over time such as bootstrap robust aggregating (bragging), subsample aggregating (subagging) and BagBoosting. One such variant of particular interest to us is subagging, which uses subsampling to achieve reductions not only in the variability of the estimator but also in computational time. Given a sample of size n, $\mathcal{X}$, and a base estimator, $\hat{\theta}_n = \hat{\theta}_n(\mathcal{X})$, subagging proceeds as follows:

Step 1. Randomly draw a subsample of size $r < n$, $\mathcal{X}^*$, by sampling without replacement[4] from $\mathcal{X}$.

Step 2. Compute the subsample estimate, $\hat{\theta}^*_r(\mathcal{X}^*)$.

Step 3. Define the subagging estimate as

$$\hat{\theta}_{n,SB(r)} = \binom{n}{r}^{-1} \sum_{(i_1,\ldots,i_r)\in\mathcal{I}} \hat{\theta}^*_r\left[\mathcal{X}_{(i_1,\ldots,i_r)}\right], \tag{2.28}$$

where $\mathcal{I}$ is the set of r-tuples whose elements in $\{1,\ldots,n\}$ are all distinct and $\mathcal{X}_{(i_1,\ldots,i_r)}$ denotes the subsample of size r made up of the elements in $\mathcal{X}$ whose indices are $i_1,\ldots,i_r$.

The subagging estimate defined in (2.28) averages the values of the subsample estimates obtained for the $\binom{n}{r}$ possible subsamples of size r generated by sampling without replacement from $\mathcal{X}$.

Furthermore, given a sample $(X_1,\ldots,X_n)$, a base estimator $\hat{\theta}_n = \hat{\theta}_n(X_1,\ldots,X_n)$ which is symmetric in the data and assuming that $r \leq n$ and $\mathrm{E}|\hat{\theta}^*_r(X_1,\ldots,X_r)|^2 < \infty$, Bühlmann and Yu (2002) prove that the subagging estimator defined in (2.28) satisfies:

$$\begin{aligned} \mathrm{E}\left[\hat{\theta}_{n,SB(r)}\right] &= \mathrm{E}\left[\hat{\theta}^*_r(X_1,\ldots,X_r)\right], \\ \mathrm{var}\left[\hat{\theta}_{n,SB(r)}\right] &\leq \frac{r}{n}\mathrm{var}\left[\hat{\theta}^*_r(X_1,\ldots,X_r)\right]. \end{aligned}$$

In other words, the subagging estimator, $\hat{\theta}_{n,SB(r)}$, has the same bias as the corresponding subsample estimator, $\hat{\theta}^*_r$, while the ratio of the variance of the subagging

[4]Sampling without replacement makes sense because we are interested in selecting random subsets from the data rather than mimicking the original sampling mechanism (bootstrapping).

estimator to that of the subsample estimator is bounded by r/n. Thus, the smaller r is with respect to n, the greater the possible reduction in variability achieved by the subagging estimator.

The subagging technique has the potential to obtain not only improvements in the variability of certain estimators but also significant reductions in computational time due to the use of subsampling. Therefore, the use of subagging is especially convenient in the context of bandwidth selection for both the Parzen–Rosenblatt and Nadaraya–Watson estimators. Naturally, the use of bagging in conjunction with binning[5] can provide even greater reductions in computation time. Hereinafter, the terms bagging and subagging will be used interchangeably to refer to the latter, unless otherwise specified.

[5]Binning is a pre-processing technique that consists of discretizing the data so that the observations that lie in a certain interval or bin are replaced by a value representative of that bin, usually its midpoint.

Chapter 3

Bagging bandwidth selection for the Parzen–Rosenblatt estimator

Kernel density estimation and the crucial problem of bandwidth selection were described in Section 2.1.1. In addition, the usefulness of bagging when working with highly variable estimators was discussed in Section 2.4. Thus, this chapter is devoted to the theoretical and empirical study of the bagged cross-validation bandwidth selector for the kernel density estimator defined in (2.3). The asymptotic properties of the proposed bandwidth selector are obtained and its better performance is shown, in terms of both rates of convergence and computational agility, with respect to the ordinary cross-validation bandwidth selector. Finally, the behavior of the proposed bagged bandwidth is illustrated by means of various simulation studies as well as by applications to real datasets. Many of the results presented in this chapter are included in Barreiro-Ures et al. (2021a).

3.1 Bagging cross-validation bandwidth selection

Let us denote by $X_1, \ldots, X_n$ a sample of size n whose observations are independent and identically distributed with density f. Consider a random sample of size $r < n$, $X_1^*, \ldots, X_r^*$, drawn without replacement from $X_1, \ldots, X_n$. This subsample is used to calculate a leave-one-out cross-validation bandwidth, $\hat{h}_r$. Assuming that f has two

continuous derivatives, a rescaled version of $\hat{h}_r$,

$$\tilde{h}_r = \left(\frac{r}{n}\right)^{1/5} \hat{h}_r,$$

is a reasonable estimator of the optimal MISE bandwidth, h_{n0}, for (2.3). Indeed, one can write

$$\hat{h}_r = h_{r0} + o_p\left(r^{-1/5}\right) = C_0 r^{-1/5} + o_p\left(r^{-1/5}\right),$$

where h_{r0} denotes the optimal MISE bandwidth for a sample of size r, and so

$$\tilde{h}_r = C_0 n^{-1/5} + o_p\left(n^{-1/5}\right) = h_{n0} + o_p\left(n^{-1/5}\right).$$

Bagging consists of repeating the resampling independently N times, leading to N rescaled bandwidths, $\tilde{h}_{r,1}, \ldots, \tilde{h}_{r,N}$. The bagging bandwidth is then defined to be

$$\hat{h}(r, N) = \frac{1}{N} \sum_{i=1}^{N} \tilde{h}_{r,i}. \tag{3.1}$$

This approeach was proposed and studied by Hall and Robinson (2009), although they focused on the unpractical case of $N = \infty$.

It is worth mentioning that 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.

3.1.1 Asymptotic results

In this section, the asymptotic properties of the bagged bandwidth defined in (3.1) will be derived and discussed. In particular, we will obtain asymptotic expressions for the bias and variance of the bagging bandwidth (3.1). Hall and Robinson (2009) studied this selector only in the case $N = \infty$. Unfortunately, the expression they

gave for the variance of (3.1), when $N = \infty$, is in error. We will provide a correct expression for this variance (see Appendix B), and, more importantly, will study the case of finite N, since there is an important interplay between the values of r and N. Obviously, in practice it is not possible to use $N = \infty$, and indeed there is a computational motivation for limiting the size of N. We will show that if N is, for example, of order n, then the rate of convergence of the variance to 0 is different than in the case $N = \infty$. This is a new result that does not arise from the method of proof used in Hall and Robinson (2009).

Regarding the bias of the bagging bandwidth (3.1), it is clear that

$$\mathrm{E}\left[\hat{h}(r,N)\right] = \mathrm{E}\left[(r/n)^{1/5}\hat{h}_r\right].$$

Therefore, we wish to know the bias of $(r/n)^{1/5}\hat{h}_r$ as an estimator of h_{n0}. We have

$$\mathrm{E}\left[(r/n)^{1/5}\hat{h}_r\right] - h_{n0} = B_{\mathrm{rescale}}(r,n) + (r/n)^{1/5}B_{\mathrm{CV}}(r),$$

where

$$B_{\mathrm{rescale}}(r,n) = (r/n)^{1/5}h_{r0} - h_{n0} \quad \text{and} \quad B_{\mathrm{CV}}(r) = \mathrm{E}\left(\hat{h}_r\right) - h_{r0}.$$

This bias due to rescaling, $B_{\mathrm{rescale}}(r,n)$, is well-understood. In fact, Marron (1987) shows that

$$B_{\mathrm{rescale}}(r,n) = \mu_{\mathrm{rescale}}r^{-2/5}n^{-1/5} + o\left(r^{-2/5}n^{-1/5}\right),$$

where

$$\mu_{\mathrm{rescale}} = \frac{R(K)^{3/5}R(f''')\mu_4(K)}{20R(f'')^{8/5}}.$$

Hall and Robinson (2009) also provide an expression for $B_{\mathrm{rescale}}(r,n)$, although their rate is in error.

The other bias component, B_{CV}, is the bias inherent to cross-validation itself. In

establishing (2.9), Hall and Marron (1987) write

$$\hat{h}_n - h_{n0} = \xi_n + e_n, \tag{3.2}$$

where $\mathrm{E}(\xi_n) = 0$ and $e_n = o_p(\xi_n)$, and hence $B_{\mathrm{CV}}(n)$ is lost in the term e_n. Doing so is acceptable in the case of ordinary cross-validation because of the fact that $\mathrm{var}(\xi_n)$ is so large. In the case of bagging, however, when $\mathrm{var}\left[\hat{h}(r, N)\right]$ becomes sufficiently small, one should no longer ignore $B_{\mathrm{CV}}(r)$, although this seems to be what both Marron (1987) and Hall and Robinson (2009) did.

In Appendix A, as part of the proof of the theorem stated below, we prove that $n^{2/5}e_n$ converges in distribution to a random variable with the following mean:

$$\mu_{\mathrm{CV}} = -\frac{8R(f)\int \mathcal{V}(u)W(u)\,du}{25R(K)^{8/5}R(f'')^{2/5}}, \tag{3.3}$$

where $\mathcal{V}$ and W are functions determined completely by K with $\int \mathcal{V}(u)W(u)\,du = 0.1431285$ in the case of the standard normal kernel.

The asymptotic bias and variance of (3.1) are stated in Theorem 3.1, whose proof is included in Appendix A. The following assumptions are needed:

A1 As $n \to \infty$, $r \to \infty$, $r = o(n)$ and N tends to a positive constant or ∞.

A2 K is a symmetric and twice differentiable density function and, without loss of generality, with variance 1.

A3 As $u \to \infty$, both $K(u)$ and $K'(u)$ are $o\left[\exp\left(-a_1 u^{a_2}\right)\right]$ for positive constants a_1 and a_2.

A4 The first three derivatives of f exist and are bounded and continuous.

Theorem 3.1 *Under assumptions A1–A4, the asymptotic bias of the bagged bandwidth defined in* (3.1) *is*

$$\mathrm{E}\left[\hat{h}(r, N)\right] - h_{n0} = r^{-1/5}n^{-1/5}\left(\mu_{\mathrm{CV}} + \mu_{\mathrm{rescale}}r^{-1/5}\right) + o\left(r^{-1/5}n^{-1/5}\right) \tag{3.4}$$

and its asymptotic variance is

$$\begin{aligned} \operatorname{var}\left[\hat{h}(r,N)\right] &= A_0 C_0^2 r^{-1/5} n^{-2/5} \left[\frac{1}{N} + \left(\frac{r}{n}\right)^2\right] \\ &+ o\left(\frac{r^{-1/5}n^{-2/5}}{N} + r^{9/5}n^{-12/5}\right), \end{aligned} \tag{3.5}$$

where C_0 and A_0 are constants given in (2.6) *and* (A.9)*, respectively.*

Theorem 3.2 *From Theorem 3.1 and its proof it follows that the asymptotic distribution of the bagged bandwidth* (3.1) *satisfies*

$$\frac{r^{1/10}n^{1/5}}{\sqrt{\frac{1}{N} + \left(\frac{r}{n}\right)^2}} \left[\hat{h}(r,N) - h_{n0}\right] \xrightarrow{d} \mathrm{N}(0, A_0 C_0^2).$$

In particular, if we assume that $r = o\left(n/\sqrt{N}\right)$, then

$$r^{1/10}n^{1/5}\sqrt{N}\left[\hat{h}(r,N) - h_{n0}\right] \xrightarrow{d} \mathrm{N}(0, A_0 C_0^2).$$

From (3.5), one could intuitively state that the optimal value of N is, precisely, $N = \infty$. However, there are computational reasons for limiting the value of N and, of course, in practice it is not possible to generate an infinite number of subsamples. In addition, the way in which the subsampling process was carried out imposes an upper bound on the value of N, namely, $\binom{n}{r}$. This is because $\binom{n}{r}$ is precisely the maximum number of distinct subsamples of size r that one could generate by sampling without replacement from a sample of size n.

From (3.1), it is interesting to observe that at $N = \infty$, the asymptotic variance of the bagged bandwidth is completely determined by the covariance between bandwidths for two different resamples. Furthermore, to first order, as derived in Bhattacharya and Hart (2016), the asymptotic correlation between bagged bandwidths from different resamples is independent of f and equal to $(r/n)^2$. This correlation is smaller when r is smaller, which is due to the fact that two resamples will usually have fewer data values in common when r is smaller. In fact, taking $N = \infty$ yields

the approximation

$$\operatorname{var}\left[\hat{h}(r,\infty)\right] = A_0 C_0^2 r^{9/5} n^{-12/5} + o\left(r^{9/5} n^{-12/5}\right), \tag{3.6}$$

which matches precisely one of the two summands in expression (13) of Hall and Robinson (2009). It can be shown that the other summand, rather than being the dominant term, as claimed in Hall and Robinson (2009), is actually negligible in comparison to (3.6) (see Appendix B for more information).

It is easily verified that the choice of r that minimizes the main term of (3.5) is asymptotic to $n/(3\sqrt{N})$. Therefore, if $N = n$, say, then the fastest rate at which $\operatorname{var}\left[\hat{h}(r,N)\right]/h_{n0}^2$ can converge to 0 is $n^{-11/10}$. In contrast, when $N = \infty$, the rate of convergence of $\operatorname{var}\left[\hat{h}(r,\infty)\right]/h_{n0}^2$ can be arbitrarily close to n^{-2} by allowing r to increase sufficiently slowly with n. This makes it clear that the properties of the bagged bandwidth are profoundly affected by how many resamples are taken, and hence it is not a good idea to analyze the bagged bandwidth by setting $N = \infty$.

It is remarkable how much stability bagging can provide. Whether N is ∞ or merely tending to ∞, $\operatorname{var}\left[\hat{h}(r,N)\right]/h_{n0}^2$ can converge to 0 faster than the usual parametric rate of n^{-1}. This is in stark contrast to the extremely slow rate of $n^{-1/5}$ for ordinary cross-validation. Unfortunately, this extreme stability cannot be fully taken advantage of since the bagged bandwidth is more biased than the ordinary cross-validation bandwidth. The largest reductions in variance are associated with small values of r, but it turns out that small r yields the largest bias.

As can be observed in (3.4), the bias source, B_{CV}, that has been ignored to date is actually of a larger order than the rescaling bias. This and the fact that $\mu_{\mathrm{CV}} < 0$ suggest that the bagged bandwidth would tend to be smaller than the optimal bandwidth h_{n0}. However, our experience in numerous simulations is that the bagged bandwidth actually tends to be larger than h_{n0}. The explanation for this phenomenon is simple: $\mu_{\mathrm{rescale}} > 0$ and μ_{rescale} is larger than $|\mu_{\mathrm{CV}}|$ in every case we have checked. Indeed, we have not found a case where $\mu_{\mathrm{rescale}}/|\mu_{\mathrm{CV}}|$ is less than 2, and it appears that there is no limit to how large this ratio can be.

Table 3.1 provides the constants μ_{rescale} and μ_{CV} for several densities. Two patterns are apparent here: (i) the heavier the tail of the density, the more dominant is

the rescaling bias, and (ii) rescaling bias is more dominant for multimodal mixtures of normals than for the normal itself. It is worth noting that the ratio $\mu_{\text{rescale}}/\mu_{\text{CV}}$ is invariant to location and scale, and hence the values of r_{crit} (defined as the smallest resample size at which the asymptotic mean of the bagged bandwidth is not larger than the optimal MISE bandwidth, h_{n0}) for any normal, logistic or Cauchy distribution are the same as in Table 3.1. Except in the case of the Beta$(5,5)$ and normal densities, the values of r_{crit} are very large, especially considering (as we shall subsequently see) that a good choice for r is usually much smaller than n. So, in spite of what the asymptotics suggest, it will often be the case that the bagged bandwidth is larger on average than the optimal bandwidth. This is a classic case of asymptotics not "kicking in" until the sample size is extremely large.

Density	μ_{rescale}	μ_{CV}	r_{crit}
Beta$(5,5)$	0.06554	-0.03070	45
Standard normal	0.44565	-0.18216	88
Standard logistic	0.92556	-0.25787	596
Bimodal mixture of two normals	0.31898	-0.05856	4795
Standard Cauchy	1.24349	-0.09793	$330,154$
Claw	0.22774	-0.00766	$> 10^7$

Table 3.1: Bias constants and critical r (r_{crit}) for the Gaussian kernel. The claw density (Marron and Wand, 1992) is a symmetric mixture of six normals and has five modes.

3.1.2 Choosing an optimal subsample size

In practice, an important step of our approach is, for fixed n and N, choosing the optimal subsample size, r_0. This optimal parameter can be selected by minimizing the asymptotic mean squared error of $\hat{h}(r,N)$, as a function of r:

$$\begin{aligned} \text{AMSE}\left[\hat{h}(r,N)\right] &= A_0 C_0^2 r^{-1/5} n^{-2/5} \left[\frac{1}{N} + \left(\frac{r}{n}\right)^2\right] \\ &+ r^{-2/5} n^{-2/5} \left(\mu_{\text{CV}} + \mu_{\text{rescale}} r^{-1/5}\right)^2. \end{aligned} \tag{3.7}$$

Since μ_{rescale}, μ_{CV}, A_0 and C_0 are unknown, we propose the following method to

estimate

$$r_0 = \arg\min_{r>1} \operatorname{AMSE}\left[\hat{h}(r, N)\right].$$

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

2. For each of these subsamples, fit a normal mixture model. To select the number of components of the mixture, the EM algorithm initialized by hierarchical model-based agglomerative clustering is used. Then, the optimal model is selected using the BIC. In practice, this process is performed employing the R package `mclust` (see Scrucca et al. (2016)).

3. Use $R(\hat{f}_i)$, $R(\hat{f}_i'')$ and $R(\hat{f}_i''')$ to estimate A_0, C_0, μ_{CV} and μ_{rescale}, where $\hat{f}_i$ denotes the density function of the normal mixture fitted to the i-th subsample. Denote these estimates by $\hat{A}_{0,i}$, $\hat{C}_{0,i}$, $\hat{\mu}_{\mathrm{CV,i}}$ and $\hat{\mu}_{\mathrm{rescale,i}}$.

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

$$\hat{D} = \frac{1}{s}\sum_{i=1}^{s} \hat{D}_i,$$

   where $\hat{D}_i$ can be $\hat{A}_{0,i}$, $\hat{C}_{0,i}$, $\hat{\mu}_{\mathrm{CV,i}}$ or $\hat{\mu}_{\mathrm{rescale,i}}$, and obtain $\widehat{\operatorname{AMSE}}\left[\hat{h}(t, N)\right]$ by plugging these bagged estimates into (3.7).

5. Finally, estimate r_0 by

$$\hat{r}_0 = \arg\min_{r>1} \widehat{\operatorname{AMSE}}\left[\hat{h}(r, N)\right].$$

Regarding the selection of s and t in Step 1, we have performed some empirical tests and observed that the estimation of h_{n0} by $\hat{h}(\hat{r}_0, N)$ is quite robust to the values of these parameters. In particular, values of $s \approx 50$ and $t \approx 0.01n$ have provided, in general, good results.

3.1.3 Simulation studies

To test the behavior of the bagged cross-validation bandwidth (3.1), some simulation studies were performed considering different density functions, sample sizes (n), subsample sizes (r) and number of subsamples (N). We present the results obtained for two normal mixture densities[6], although similar results were obtained for other densities. Let $\mu = (\mu_1, \ldots, \mu_k)$, $\sigma = (\sigma_1, \ldots, \sigma_k)$ and $w = (w_1, \ldots, w_k)$ denote the mean, standard deviation and weight vectors, respectively, of a mixture with density function

$$f(x) = \sum_{i=1}^{k} w_i \phi_{\mu_i, \sigma_i}(x),$$

where ϕ_{μ_i,σ_i} denotes the density function of a $\mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$. In addition to density D1, already defined in Section 2.1.1, the following density was considered:

D2: (mixture of two normals) with parameters $\mu = (0, 1.5)$, $\sigma = (1, 1/3)$ and $w = (0.75, 0.25)$.

In this experiment, 1000 samples of size $n = 10^5$ were simulated from the previous densities and the bagged, $\hat{h}(r, N)$, and standard leave-one-out cross-validation, $\hat{h}_n$, bandwidths were computed. The bagged bandwidths were calculated using $N = 500$ subsamples and considering four values for the size of the subsamples, r, including the optimal values, $r_0 = 13081$ and $r_0 = 20326$, for densities D2 and D1, respectively. For each sample, we also computed the estimated r_0 using the algorithm presented in Subsection 3.1.2 with values $s = 50$ and $t \in \{500, 1000, 5000\}$. The Gaussian kernel was used throughout the study. Furthermore, the R (R Core Team, 2021) package `baggedcv` (Barreiro-Ures et al., 2019), developed by the author of this dissertation, was employed to carry out the simulations.

To compute the different cross-validation bandwidths involved in this simulation ($\hat{h}_n$ and $\hat{h}_{r,i}$, $i = 1, \ldots, N$), the R function `bw.ucv` was employed. This function uses a binned implementation and, therefore, it is extremely fast. However, when the number of bins, `nb`, is significantly smaller than the sample size, `bw.ucv` has

[6]Since the densities considered in this section are normal mixtures, the exact value of the optimal MISE bandwidth, h_{n0}, was employed.

the disturbing tendency to choose the very smallest bandwidth allowed. This is illustrated in Listing 3.1, where we show the output of the `bw.ucv` function applied to a sample of size $n = 10^6$ drawn from a standard normal and the number of bins set to its default value of `nb` = 1000. In this case, the true cross-validation bandwidth is approximately 0.06, while `bw.ucv` returned a much smaller value (the lower bound of the search interval).

```
set.seed(1)
x = rnorm(10^6)
bw.ucv(x,lower=0.001,upper=1)

[1] 0.001045393
Warning message:
In bw.ucv(x, lower = 0.001, upper = 1) :
  minimum occured at one end of the range
```

Listing 3.1: Bad behavior of `bw.ucv` when using the default number of bins

For some densities, `bw.ucv` works fine with `nb` being relatively small with respect to the sample size. However, for more complex (heavy-tailed, multimodal, ...) densities, `nb` needs to be quite close to the sample size for `bw.ucv` to give sensible results. This limits the computational gain that binned cross-validation could in principle achieve. Even when `nb` is equal to the sample size, `bw.ucv` returns a wrong value in a small proportion of cases. In spite of this, in practice, we recommend using `bw.ucv` with `nb` close to the sample size. Taking this suggestion into account, if `nb` $= r$ at the subsample level for $\hat{h}(r, N)$, we found that the average of the bagged bandwidths obtained using `bw.ucv` is usually quite close to the results obtained employing the more accurate non-binned version of $\hat{h}(r, N)$. Moreover, by using `bw.ucv` in the implementation of $\hat{h}(r, N)$, its runtime can be significantly reduced, being even somewhat shorter than the time needed for the computation of the binned cross-validation bandwidth, specially for large sample sizes and certain values of r and N. This can be observed in Table 3.2, which shows the computing time for binned standard cross-validation and the bagged bandwidth selector for different values of n, r and N. For $\hat{h}(r, N)$, we considered `nb` $= r$ at the subsample level and the code was run in parallel using the function `bagcv` from the R package `baggedcv` (Barreiro-Ures et al., 2019). In the case of binned standard cross-validation, the

number of bins was also set equal to n to provide a fair comparison of both methods. As we can see, the bagged bandwidth can achieve a significant reduction in computing time with respect to binned standard cross-validation for samples of considerable size.

		Bagged CV		
		$r = 1000$	$r = 5000$	$r = 10,000$
n	bw.ucv($\cdot$, nb=n)	$N = 500$	$N = 500$	$N = 500$
10^5	3.1	1.1	2.0	4.3
10^6	367	1.3	2.2	4.4

Table 3.2: CPU elapsed time (seconds) for binned leave-one-out cross-validation and the bagged bandwidth selector. Computing time for bagged cross-validation depends on r, N and the number of CPU cores employed.

In addition to the substantial reduction in computing time, the bagged cross-validation bandwidth showed greater statistical precision. This can be observed in Figure 3.1, where the sampling distributions of $\log(\hat{h}_n/h_{n0})$ and $\log[\hat{h}(r,N)/h_{n0}]$, for different values of r, for models D2 (left panel) and D1 (right panel) are presented. Specifically, we considered, for D2, the values of r: $5000, 13081$ $(r_0), 20000$, and $\hat{r}_0$ computed with $s = 50$ and $t = 500, 1000, 5000$, while, for D1, the values of r employed were: $5000, 20326$ $(r_0), 25000$, and $\hat{r}_0$ computed with $s = 50$ and $t = 500, 1000, 5000$, using the function `mopt` from the R package `baggedcv`. It is clear that the bagged bandwidth achieves an important reduction in the mean squared error with respect to the standard leave-one-out cross-validation selector. Namely, the bagged bandwidth with $r = r_0$ produced a mean squared error which is 95.3% and 92.2% lower than that of the standard leave-one-out cross-validation bandwidth for models D2 and D1, respectively. This significant reduction is also observed (in general) when using $r = \hat{r}_0$ for each simulated sample. In that case, for $t = 1000$ ($t = 5000$), the mean squared error reduction with respect to standard leave-one-out cross-validation is 95.9% (95.9%) for model D2 and 92.3% (93.5%) for model D1. Additionally, Figure 3.2 shows the sampling distribution of $\text{ISE}[\hat{h}(r,N)]/\text{ISE}(\hat{h}_n)$ for both models and the same values of r considered in Figure 3.1. In this case, outliers were omitted in order to be able to appreciate the differences between the various boxplots. The means of $\text{ISE}[\hat{h}(r,N)]/\text{ISE}(\hat{h}_n)$ for the combinations of r and N and densities considered in

Figure 3.2 and the proportion of values of $\hat{h}(r, N)$ whose ISE is lower than that of $\hat{h}_n$ are shown in Table 3.3.

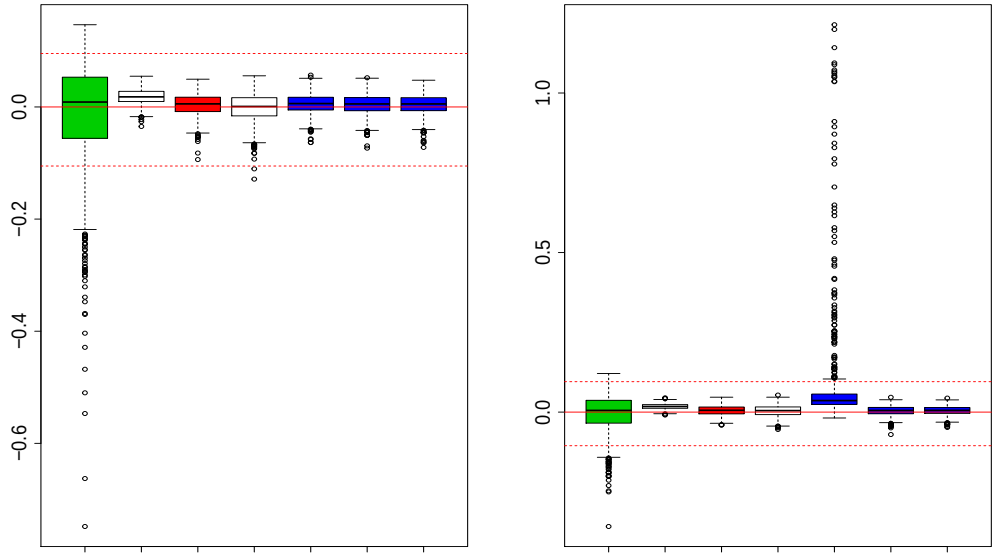

Figure 3.1: Sampling distribution of $\log\left(\hat{h}/h_{n0}\right)$, with $\hat{h}$ denoting the leave-one-out cross-validation (green) and the bagged bandwidths for different values of r. For the bagged bandwidths, we considered $N = 500$ and $r \in \{5000, 13081 \text{ (red)}, 20000, \hat{r}_0 \text{ (blue)}\}$, for density D2 (left panel); and $r \in \{5000, 20326 \text{ (red)}, 25000, \hat{r}_0 \text{ (blue)}\}$, for density D1 (right panel). The two white boxes correspond, from left to right, to $r = 5000$ and 20000, for D2 (left panel); and to $r = 5000$ and 25000, for D1 (right panel). The three blue boxes correspond, from left to right, to $t = 500, 1000, 5000$. Red dotted lines are plotted at values 0.9 and 1.1 for reference.

The mean squared error of the bagged bandwidth using $r = \hat{r}_0$ may be larger than the one for leave-one-out cross-validation for density D1 using $t = 500$ (left blue boxplot on the right panel in Figure 3.1). These results are somewhat misleading because the final behavior of the kernel density estimator with the bagged bandwidth selector (denoted by $\hat{h}$, for simplicity) is still very good in this setting. The distribution of $\hat{h}$ is biased upward, and there are numerous extremely large values of $\hat{h}$. However, it turns out that even the largest of these bandwidths produce very effective density estimates, as observed in Figure 3.3. Consider, for example, $\log(\hat{h}/h_{n0}) = 1$, which means that $\hat{h} \simeq 2.72 h_{n0}$. In Figure 3.3 we provide the claw

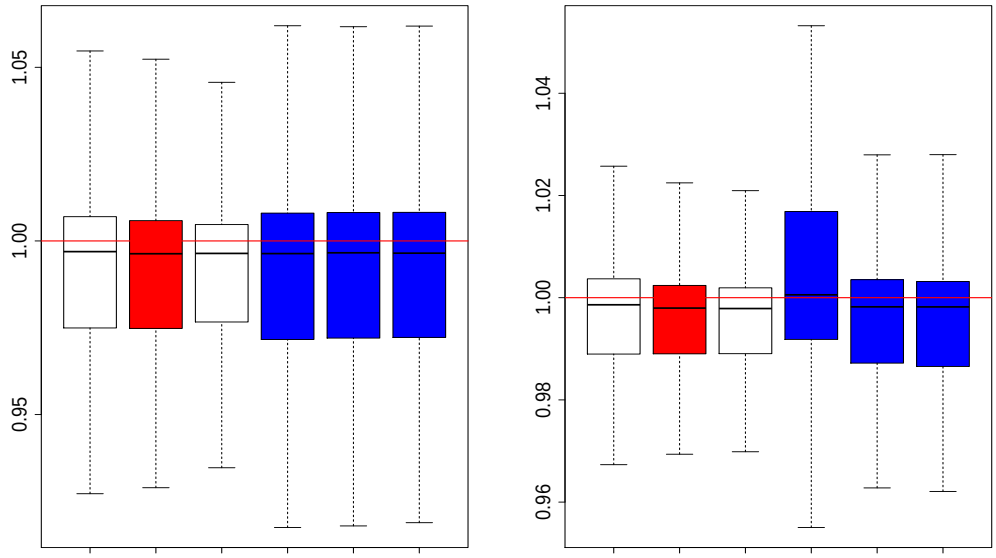

Figure 3.2: Sampling distribution of $ISE(\hat{h})/ISE(h_{n0})$, with $\hat{h}$ denoting the bagged bandwidths for different values of r. For the bagged bandwidths, we considered $N = 500$ and $r \in \{5000, 13081 \text{ (red)}, 20000, \hat{r}_0 \text{ (blue)}\}$, for density D2 (left panel); and $r \in \{5000, 20326 \text{ (red)}, 25000, \hat{r}_0 \text{ (blue)}\}$, for density D1 (right panel). The two white boxes correspond, from left to right, to $r = 5000$ and 20000, for D2 (left panel); and to $r = 5000$ and 25000, for D1 (right panel). The three blue boxes correspond, from left to right, to $t = 500, 1000, 5000$.

density and two kernel estimates from a sample of size 10^5. The bandwidths of the two estimates are $h_{n0} = 0.031$ and $2.72h_{n0} \simeq 0.084$. The kernel estimate with larger bandwidth captures the five modes and has better tail behavior than the estimate based on the MISE bandwidth. Figure 3.3 illustrates the fact that integrated squared error (ISE) loss is not always ideal. One might well prefer an estimate with larger than optimal ISE, as long as it captures all the important features of the underlying density and is smoother than the ISE optimal estimate.

In Figure 3.4, the sampling distribution of $\hat{r}_0/r_0$ is shown. It can be observed that the mean squared error of $\hat{r}_0$ is reduced as t increases. Furthermore, the bias of the estimator depends on the complexity of the target density. For small values of t, in spite of the high variability of $\hat{r}_0$, the sampling distribution of the bagged bandwidth, considering $r = \hat{r}_0$, is virtually unchanged with respect to the case $r = r_0$

Means

Density	B_1	B_2	B_3	B_4	B_5	B_6
D2	0.98533	0.98505	0.98512	0.98448	0.98428	0.98537
D1	0.99624	0.99594	0.99530	1.23742	0.99539	0.99494

Proportions

Density	B_1	B_2	B_3	B_4	B_5	B_6
D2	0.606	0.603	0.609	0.590	0.593	0.590
D1	0.584	0.622	0.637	0.461	0.604	0.599

Table 3.3: Means of ISE$[\hat{h}(r,N)]$/ISE$(\hat{h}_n)$ for the combinations of r and N and densities considered in Figure 3.2 and proportion of values of $\hat{h}(r,N)$ whose ISE is lower than that of $\hat{h}_n$. B_i refers to the i-th boxplot in order of appearance in Figure 3.2.

for densities that are not very complex, such as D2. For more complex densities, such as D1, the effect that the variability of $\hat{r}_0$ has on the bagged bandwidth is more noticeable for small values of t, translating into a more biased bandwidth. More importantly, when we compare the errors in Figure 3.4 and Figure 3.1, it is clear that there is a large range of values for r around its optimal value, r_0, such that the effect the error of $\hat{r}_0$ has on the sampling distribution of $\hat{h}(\hat{r}_0,N)$ is very small.

3.2 Bagging bootstrap bandwidth

Due to its quadratic complexity, computing the bootstrap bandwidth defined in (2.16) can become too computationally expensive very quickly as the sample size increases. A possible solution to this problem is to consider a subagged version of the bootstrap bandwidth and to take advantage of the computational benefits of working with subsamples of size $r < n$ rather than with the entire sample of size n. Note that, as in the case of cross-validation, (2.14) is a second-order U-statistic and hence the situation we are dealing with is very similar to the one studied in Section 2.1.2. To compute the subagged bootstrap bandwidth we propose the following procedure:

Step 1. Independently generate N subsamples of size $r < n$ by sampling without re-

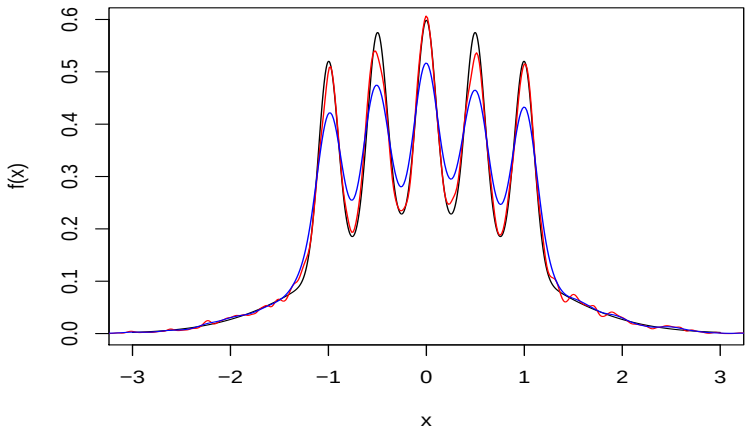

Figure 3.3: Claw density (black line) and kernel estimates (red and blue lines). The kernel estimates are computed from a sample of size 10^5. The red estimate uses the MISE optimal bandwidth of 0.031 and the blue one uses bandwidth 0.084.

placement from $X_1, \ldots, X_n$.

Step 2. For $i \in \{1, \ldots, N\}$, estimate the optimal pilot bandwidth, g_0, for example by fitting a mixture of normals to the corresponding subsample[7]. Denote these estimates by $\hat{g}_{0,1}, \ldots, \hat{g}_{0,N}$.

Step 2. For each of the subsamples, compute the bootstrap bandwidths

$$h^*_{r0,i} = \arg\min_{h>0} \left[V^*_{r,i}(h; \hat{g}_{0,i}) + B^*_{r,i}(h; \hat{g}_{0,i}) \right], \quad i = 1, \ldots, N.$$

Step 3. Compute the bagged bandwidth as the mean of the rescaled bootstap bandwidths,

$$\hat{h}^*(r, N) = \frac{1}{N} \left(\frac{r}{n} \right)^{1/5} \sum_{i=1}^{N} h^*_{r0,i}.$$

It should be noted that in the case of the bootstrap bandwidth, bagging has less

[7]For this purpose, the function `Mclust` from R package `mclust` was employed.

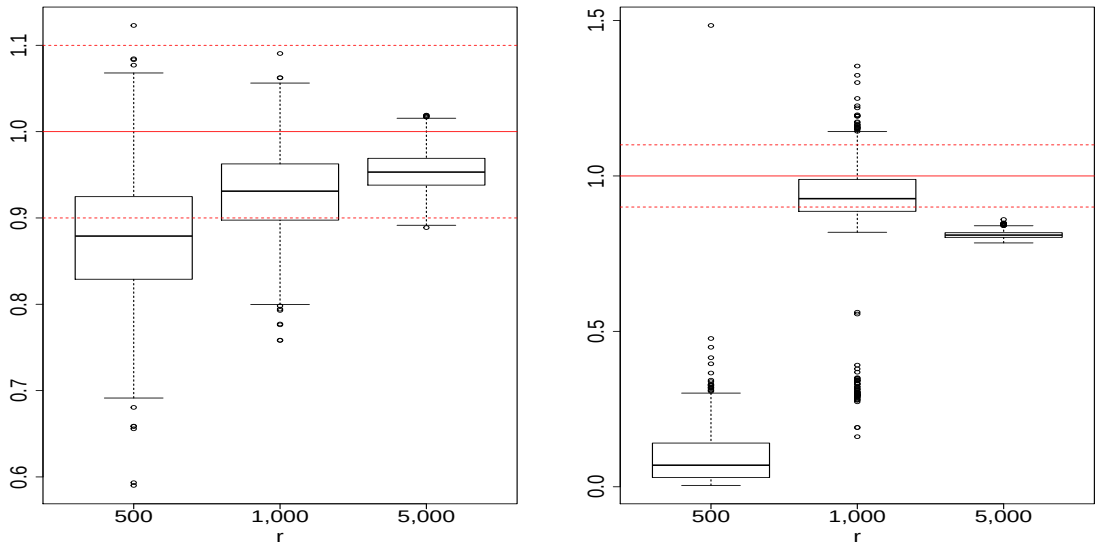

Figure 3.4: Sampling distribution of $\hat{r}_0/r_0$, with $\hat{r}_0$ denoting the estimator of the optimal subsample size, r_0, as defined in Section 3.1.2, for densities D1 (left panel) and D2 (right panel). The values chosen for the parameters of the estimator were $s = 50$ and (from left to right) $t \in \{500, 1000, 5000\}$. Red dotted lines are plotted at values 0.9 and 1.1 for reference.

room for improvement in terms of variance reduction when compared to the cross-validation bandwidth. Specifically, while $\hat{h}_n - h_{n0}$ converged to a normal distribution with zero mean and constant variance at the rate $n^{-3/10}$, in the case of the bootstrap bandwidth Cao (1993) showed that $h^*_{n0} - h_{n0}$ converges to a normal distribution with zero mean and constant variance at a faster rate, namely $n^{-39/70}$, where h^*_{n0} denotes the bootstrap bandwidth defined in (2.16). In this sense it is clear that the cross-validation bandwidth selector is a much better candidate for the application of bagging than the bootstrap bandwidth precisely because of the higher variability of the former. This implies that in the case of the subagged bootstrap bandwidth, little can be expected from the use of subagging in terms of variance reduction, and its benefits are expected to be purely computational. Hence, the number of subsamples, N, may be kept at moderate to low values and the size of the subsamples, r, should be chosen according to the cost, as a loss in statistical precision, that the user is willing to pay. To illustrate the effect r has on the computing time, Figure 3.5 shows the observed CPU elapsed time for both the ordinary bootstrap

bandwidth and its subagged version as a function of the sample size, n, which took the values $n = 10^4, 10^5, 10^6$. Both bootstrap bandwidths were computed employing the `hboot_bag` function from the `baggingbwsel` R package. In accordance with the above, a low and fixed value of N was considered, namely $N = 25$. The size of the subsamples, r, was chosen as $r = n^p$, with $p = 0.5, 0.6, 0.7, 0.8, 0.9$.

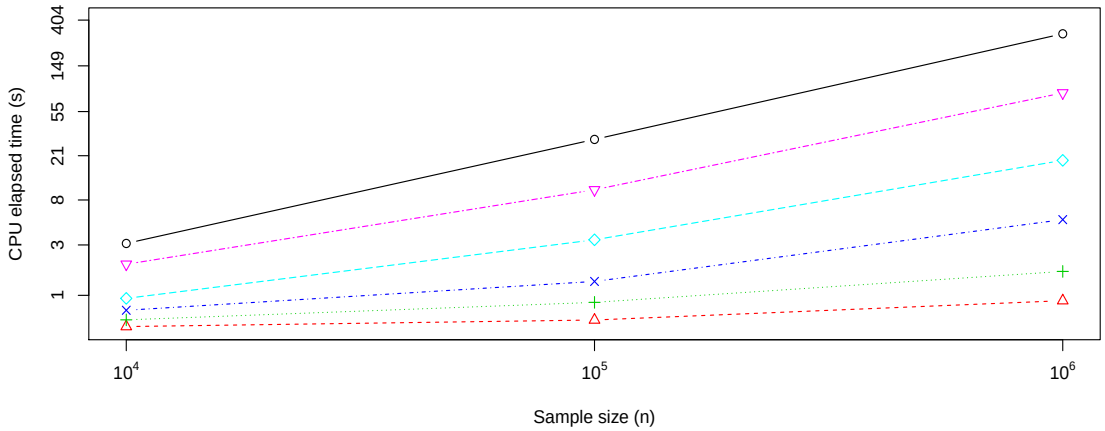

Figure 3.5: CPU elapsed time (seconds) for ordinary and bagged bootstrap bandwidths as a function of the sample size, $n = 10^4, 10^5, 10^6$. Variables are shown in logarithmic scale. For the subagged bootstrap bandwidth, the value of N was set to $N = 25$ and the subsample size, r, was chosen as $r = n^p$, with $p = 0.5$ (triangle point up), 0.6 (plus), 0.7 (cross), 0.8 (diamond), 0.9 (triangle point down). A binned implementation of the bandwidth selectors was considered, using $0.1n$ bins for the ordinary bootstrap bandwidth (circle) and $0.1r$ bins for the subagged bootstrap bandwidth.

Now, to assess the loss in statistical precision due to the use of subagging, we simulated samples of size $n = 10^5$ from different density functions. In addition to the density mixtures D1 and D2, defined in Sections 2.1.1 and 3.1.3, respectively, we considered the mixture density D3, with parameters $\mu = 0$, $\sigma = 1$ and $w = 1$, which corresponds to the density of the standard normal. Densities D3, D2 and D1 can be seen as representing low, medium and high "complexity" densities, respectively. Figure 3.6 shows the sampling distribution of $\hat{h}/h_{n0}$ and $M_n(\hat{h})/M_n(h_{n0})$, with $\hat{h}$ denoting both the ordinary bootstrap bandwidth and the subagged bootstrap bandwidth, for densities D1, D2 and D3. The number of subsamples was set to $N = 1$

and the size of the subsamples was chosen as $r = n^p$, with $p = 0.5, 0.6, 0.7, 0.8, 0.9$. From Figures 3.5 and 3.6 and as a rule-of-thumb, one may conclude that a sensible choice of r, in the sense of offering a certain balance between statistical precision and computational agility, would be $r = n^{0.7}$. As for the number of subsamples, N, as argued above, it should be kept at low values given the already low variability of the bootstrap bandwidth selector.

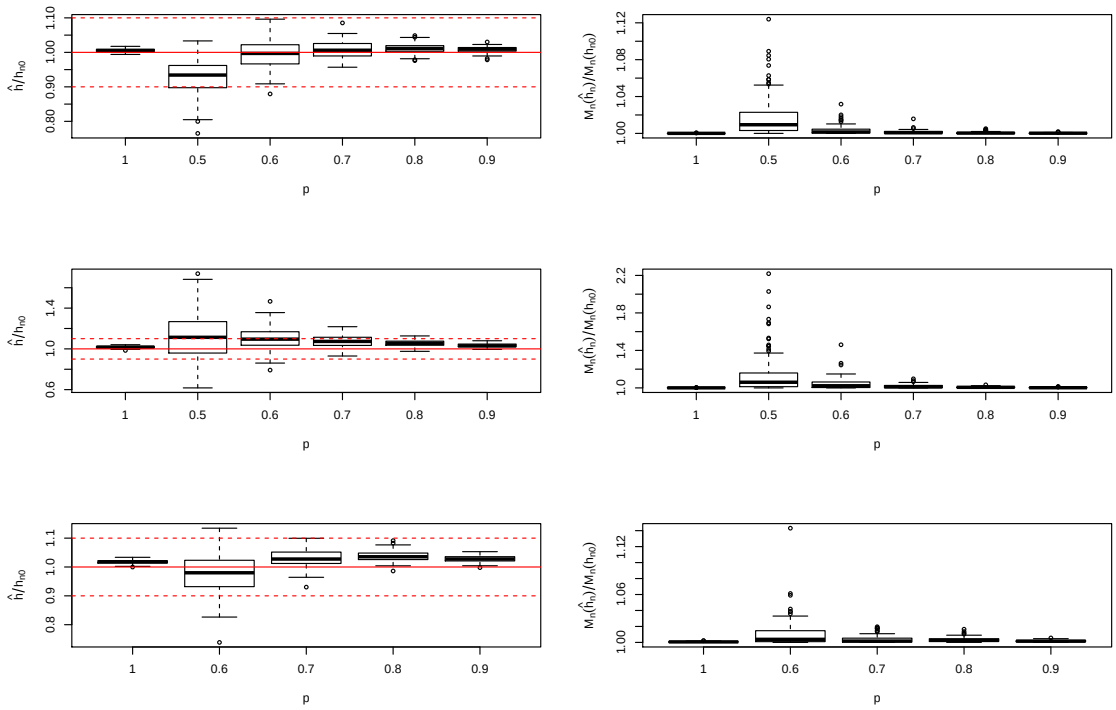

Figure 3.6: Sampling distribution of $\hat{h}/h_{n0}$ (left panels) and $M_n(\hat{h}_n)/M_n(h_{n0})$ (right panels), with $n = 10^5$ and $\hat{h}$ denoting both the ordinary bootstrap bandwidth (first boxplots) and the subagged bootstrap bandwidth (second to last boxplots), for densities D3 (top), D2 (center) and D1 (bottom). The number of subsamples was set to $N = 1$ and the size of the subsamples was chosen as $r = n^p$, with $p = 0.5, 0.6, 0.7, 0.8, 0.9$. The case $p = 1$ corresponds to the ordinary bootstrap bandwidth. For density D1, the case $p = 0.5$ was omitted because the bandwidths obtained were too large and altered the scale of the plots.

3.3 Bagging when the asymptotics of the optimal bandwidth are unknown

So far we have applied bagging under the assumption that the rate of convergence to zero of the optimal bandwidth is known. Specifically, it made sense for the bandwidths obtained for each subsample of size r to be rescaled by a factor $(r/n)^{1/5}$ precisely because we knew that, asymptotically, the optimal bandwidth tends to zero at the rate $n^{-1/5}$. However, there may be cases when this assumption does not hold and rescaling the bandwidths by a factor $(r/n)^{1/5}$ may lead to inconsistent estimators. For instance, one may be dealing with some overly complex estimator for which asymptotic theory has not yet been developed. For this kind of situations where the rate of convergence to zero of the optimal bandwidth is unknown, we have come up with a modified version of the usual bagging procedure that addresses the problem of how to rescale the bandwidths from a regression analysis perspective. Let us begin by assuming that, asymptotically, the optimal bandwidth, which in this case we will denote by h_{n1}, converges to zero at the rate n^{-p_1}, that is,

$$h_{n1} = p_0 n^{-p_1} + o\left(n^{-p_1}\right), \tag{3.8}$$

where both p_0 and $p_1 > 0$ are unknown constants. We may linearize (3.8) by taking logarithms, and so it would be sensible to approach the problem from a linear regression perspective. That is, we could consider the following linear regression model,

$$Y = \beta_0 + \beta_1 Z,$$

where $Z = \log(n)$, $Y = \log(\bar{h}_n)$ and $\bar{h}_n$ denotes some selector for h_{n1}. Thus, the problem of estimating p_0 and p_1 is equivalent to estimating β_0 and β_1. In fact, if we denote by $\hat{\beta}_0$ and $\hat{\beta}_1$ our estimates of β_0 and β_1, then

$$\begin{aligned} \hat{p}_0 &= e^{\hat{\beta}_0}, \\ \hat{p}_1 &= -\hat{\beta}_1. \end{aligned}$$

Thus, in this case the bagging bandwidth could be defined according to the following scheme:

Step 1. Consider a grid of subsample sizes, $r_1, \ldots, r_s$, with $r_i < n$ for $i \in \{1, \ldots, s\}$.

Step 2. For each $i \in \{1, \ldots, s\}$, compute bandwidth selectors for $h_{r_i 1}$. Denote these selectors by $\bar{h}_{r_1}, \ldots, \bar{h}_{r_s}$.

Step 3. Solve the ordinary least-squares problem given by

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{(\beta_0, \beta_1)}{\arg\min} \sum_{i=1}^{s} \left[\log(\bar{h}_{r_i}) - \beta_0 - \beta_1 \log(r_i)\right]^2.$$

Step 4. Estimate h_{n1} by

$$\hat{h}_{n1} = \hat{p}_0 n^{-\hat{p}_1}, \tag{3.9}$$

where $\hat{p}_0 = e^{\hat{\beta}_0}$ and $\hat{p}_1 = \hat{\beta}_1$.

Naturally, the fact that in this case we are estimating an additional parameter, namely p_1, will cause the variability of the bagging estimator to be greater than that of the case in which the rate of convergence to zero of the optimal bandwidth is known, thus making it difficult to produce improvements in statistical precision with respect to the ordinary, non-bagged estimator. However, improvements in computation time are still possible, although they may not be as significant as in the case in which p_1 is known given that it is necessary to consider a grid of subsample sizes large enough to carry out the regression and prevent the estimation error of $\hat{p}_1$ from being too large.

To test the behavior of the generalized bagging bandwidth defined in (3.9), 1000 samples of size $n = 10^5$ drawn from densities D1, D2 and D3 were simulated. Figure 3.7 shows the sampling distribution of $\hat{h}_n/h_{n0}$, $\hat{h}(r, N)/h_{n0}$ and $\hat{h}_{n1}/h_{n0}$. For $\hat{h}(r, N)$, the value of r was set at $r = 3000$. To compute $\hat{h}_{n1}$, the function `hsss_dens()` from the R package from `baggingbwsel` was employed, and the subsample sizes were selected as $(1000, 2000, 3000)$ and $(5000, 7500, 10^4)$. For both bagging selectors, the number of subsamples was set at $N = 100$. As can be seen, $\hat{h}_{n1}$ cannot compete with

$\hat{h}(r, N)$ in terms of statistical precision, the former showing a behavior very similar to that of $\hat{h}_n$. However, as already mentioned, the strength of $\hat{h}_{n1}$ lies in its capacity to obtain bandwidths for selectors that may be too complex and whose asymptotic properties are not known. Moreover, $\hat{h}_{n1}$ can still outperform the standard cross-validation bandwidth selector in terms of computational agility.

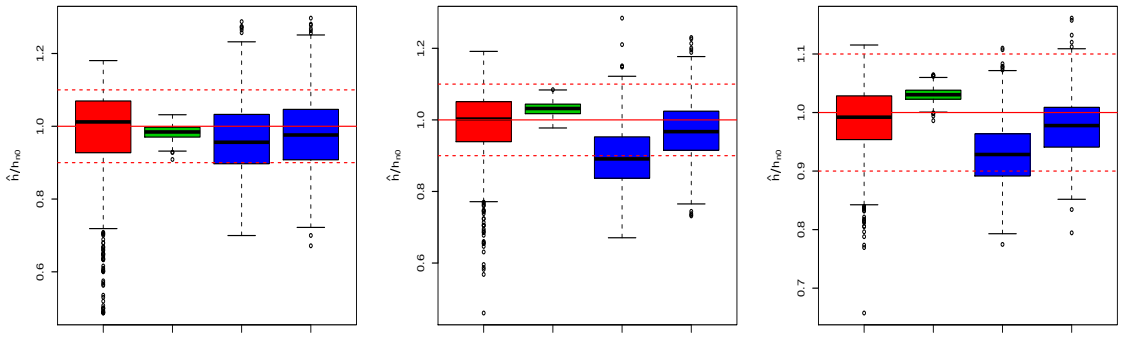

Figure 3.7: Sampling distribution of $\hat{h}/h_{n0}$ for 1000 samples of size $n = 10^5$ drawn from densities D3 (left panel), D2 (center panel) and D1 (right panel), where $\hat{h}$ denotes both the standard cross-validation bandwidth (red), the bagging bandwidth (green) defined in (3.1) and the generalized bagging bandwidth (blue) defined in (3.9). For $\hat{h}(r, N)$, the value of r was set at $r = 3000$. For $\hat{h}_{n1}$, the subsample sizes were selected as $(1000, 2000, 3000)$ and $(5000, 7500, 10^4)$. For both bagging selectors, the number of subsamples was set at $N = 100$.

3.4 Bagging with higher-order terms

So far we have only considered the dominant term of the optimal bandwidth when applying bagging and that is why it made sense to rescale the bandwidths obtained for each subsample of size r by a factor $(r/n)^{1/5}$. That is, we limited ourselves to working with

$$h_{n0} = C_0 n^{-1/5} + o\left(n^{-1/5}\right),$$

where the constant C_0 was defined in (2.6). The disadvantage of proceeding in this way is that, unless the subsamples are of a sufficient size, the behavior of the bandwidths obtained for them may diverge from what the asymptotics tell us. However, it is not unreasonable to think that by taking into account the second-order term of the optimal bandwidth, considering subsamples of smaller size may become feasible. In other words, we plan to work with the first and second order terms of the optimal bandwidth,

$$h_{n0} = C_0 n^{-1/5} + C_1 n^{-3/5} + o\left(n^{-3/5}\right), \tag{3.10}$$

and, therefore, we are forced to modify the bagging procedure accordingly. Note that it is not necessary to know the terms that appear in the expression of the constant C_1, which depends on the first and second order terms of the bias as well as the first and third order terms of the variance (the second-order term of the variance does not depend on the bandwidth). This is due to the fact that it is not necessary to estimate C_1 when applying bagging but instead we only need to know the order of the term it accompanies, namely $n^{-3/5}$. To apply bagging in this situation, it will not be enough to consider a single subsample size but instead we will need two, which we will denote by $r_1 < n$ and $r_2 < n$. This is due to the fact that two unknown terms appear in (3.10), namely C_0 and C_1, and so we will need to solve a system of two equations. In particular, this system of linear equations is given by

$$\begin{cases} h_{r_1 0} = C_0 r_1^{-1/5} + C_1 r_1^{-3/5}, \\ h_{r_2 0} = C_0 r_2^{-1/5} + C_1 r_2^{-3/5}. \end{cases} \tag{3.11}$$

Its solution is given in Proposition 3.1.

Proposition 3.1 *The solution of the system of linear equations given by* (3.11) *is*

$$\begin{aligned} C_0 &= \frac{h_{r_1 0} r_1^{3/5} - h_{r_2 0} r_2^{3/5}}{r_1^{2/5} - r_2^{2/5}}, \\ C_1 &= \frac{h_{r_1 0} r_1^{1/5} - h_{r_2 0} r_2^{1/5}}{r_1^{-2/5} - r_2^{-2/5}}. \end{aligned}$$

Now, if we denote by $\hat{h}(r_1, N)$ and $\hat{h}(r_2, N)$ the bagged cross-validation bandwidths computed from N subsamples of size r_1 and r_2, respectively, then our bandwidth selector based on (3.10) can be defined as

$$\hat{h}(r_1, r_2, N) = \hat{C}_0 n^{-1/5} + \hat{C}_1 n^{-3/5}, \tag{3.12}$$

where

$$\begin{aligned} \hat{C}_0 &= \frac{\hat{h}(r_1, N) r_1^{3/5} - \hat{h}(r_2, N) r_2^{3/5}}{r_1^{2/5} - r_2^{2/5}}, \\ \hat{C}_1 &= \frac{\hat{h}(r_1, N) r_1^{1/5} - \hat{h}(r_2, N) r_2^{1/5}}{r_1^{-2/5} - r_2^{-2/5}}. \end{aligned}$$

The bandwidth selector defined in (3.12) can also be generalized to the case studied in Section 3.3, where we assumed that the asymptotic properties of the optimal bandwidth are unknown. To do so, let us start by considering k subsample sizes, $r_1, \ldots, r_k$, of each of which we have generated N subsamples. Let us also denote by $\hat{h}(r_i, N)$ the bagged cross-validation bandwidth computed from the N subsamples of size r_i, with $i \in \{1, \ldots, k\}$. Now, let us assume that, asymptotically, the optimal bandwidth, h_{n2}, verifies

$$h_{n2} = p_0 n^{-p_1} + q_0 n^{-q_1} + o\left(n^{-q_1}\right), \tag{3.13}$$

where p_0, p_1, q_0 and q_1 are unknown positive constants and $p_1 < q_1$ so that $p_0 n^{-p_1}$ and $q_0 n^{-q_1}$ define the first and second order terms of h_{n1}, respectively. Let us now consider the nonlinear regression model

$$\bar{h}_n = \beta_0 n^{-\beta_1} + \beta_2 n^{-\beta_3},$$

where $\bar{h}_n$ denotes some selector for the optimal bandwidth. Then, if we define

$$\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right) = \underset{(\beta_0, \beta_1, \beta_2, \beta_3)}{\arg\min} \sum_{i=1}^{k} \left[\hat{h}(r_i, N) - \beta_0 r_i^{-\beta_1} - \beta_2 r_i^{-\beta_3}\right]^2,$$

the bagged selector based on (3.13) would be

$$\hat{h}(r_1, r_2, N) = \hat{\beta}_0 n^{-\hat{\beta}_1} + \hat{\beta}_2 n^{-\hat{\beta}_3}.$$

To test the behavior of the generalized bagging bandwidth defined in (3.12) and compare it with that of the bagging bandwidth defined in (3.1), 500 samples of size $n = 10^4$ drawn from the Claw density were simulated. Figure 3.8 shows the sampling distribution of $\hat{h}(r_1, r_2, N)/h_{n0}$, which was computed using `tss_dens()`, and $\hat{h}(r, N)/h_{n0}$. For the former, subsample sizes were selected as $r = n^p$, with $p = 0.7, 0.8, 0.9$. For the latter, subsample sizes were selected as $(r_1, r_2) = (n^p, n^q)$, with $p = (0.5, 0.6, 0.7, 0.8, 0.6, 0.7)$ and $q = (0.6, 0.7, 0.8, 0.9, 0.8, 0.9)$. For both selectors, the number of subsamples was set at $N = 100$. These results are not too encouraging for the generalized bagging bandwidth, as no improvement in statistical precision is observed with respect to the bagging bandwidth defined in (3.1).

3.5 Real data examples

To show the performance of the proposed bagged bandwidth selector, firstly, we considered the public dataset "On-Time: Reporting carrier On-Time Performance"[8] corresponding to the year 2017. In particular, we were interested in the variable `ArrDelay`, which measures the difference in minutes between the scheduled and actual arrival time (note that early arrivals show negative numbers). Since the values are reported in integer numbers of minutes, the dataset contains many ties and in order to avoid problems when performing cross-validation[9], we decided to remove the ties by jittering the data. In particular, we worked with the sample of size $n = 5,579,346$ which results from adding a random sample of size n, drawn from a continuous uniform distribution defined on the interval $(-0.5, 0.5)$, to the original dataset.[10]

To estimate the optimal subsample size, r_0, for the bagged bandwidth and con-

[8]The dataset is available at `https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ`

[9]The presence of many ties in the data usually causes the estimator to undersmooth, and this phenomenon becomes more severe as the sample size increases, if one keeps the percentage of ties constant.

[10]This way one can safely assume that the jittered sample comes from a continuous distribution.

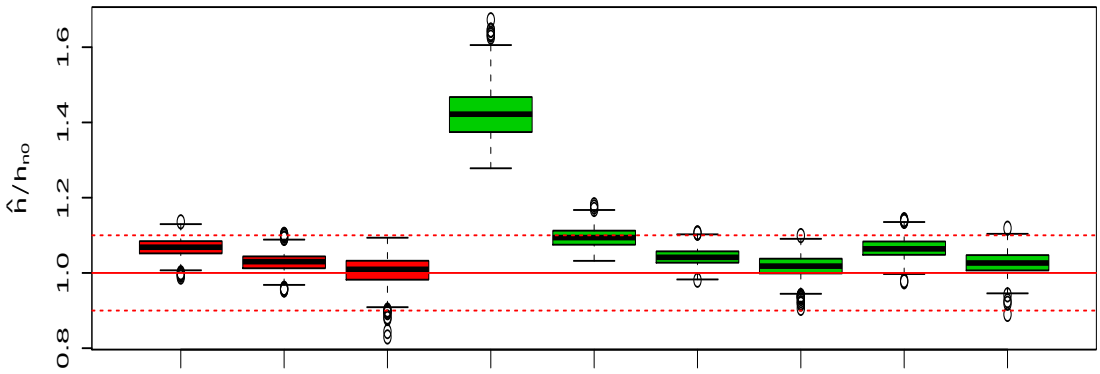

Figure 3.8: Sampling distribution of $\hat{h}/h_{n0}$ for 500 samples of size $n = 10^4$ drawn from density D1, where $\hat{h}$ denotes both the bagging bandwidth (red) defined in (3.1) and the generalized bagging bandwidth (green) defined in (3.12). For the former, subsample sizes were selected as $r = n^p$, with $p = 0.7, 0.8, 0.9$. For the latter, subsample sizes were selected as $(r_1, r_2) = (n^p, n^q)$, with $p = (0.5, 0.6, 0.7, 0.8, 0.6, 0.7)$ and $q = (0.6, 0.7, 0.8, 0.9, 0.8, 0.9)$. For both selectors, the number of subsamples was set at $N = 100$.

sidering $N = 100$ subsamples, we used the procedure described in Section 3.1.2. In particular, using $t = 1000$ and $s = 500$ yielded the estimate $\hat{r}_0 = 272,222$. The process of estimating r_0 with those parameters took 32 seconds. Then, our estimated bandwidth was $\hat{h}(r, N) = \hat{h}(272222, 100) = 0.490$, the calculation of which took 63 seconds. It should also be noted that the calculation of both $\hat{r}_0$ and $\hat{h}(r, N)$ were executed in parallel. Figure 3.9 shows the kernel density estimates obtained when considering the bagged bandwidth and the bandwidth produced by the R function `bw.ucv(·)`, using the same number of bins and search interval as in the case of the bagged bandwidth, that is, $h = \hat{h}(\hat{r}_0, 100)$ and $h =$ `bw.ucv(·, nb=1e5, lower=0.01, upper=1)`. As we can see, even with those parameters `bw.ucv` simply returns the lower bound of the search interval thus producing a heavily under-

smoothed estimate of the underlying density.

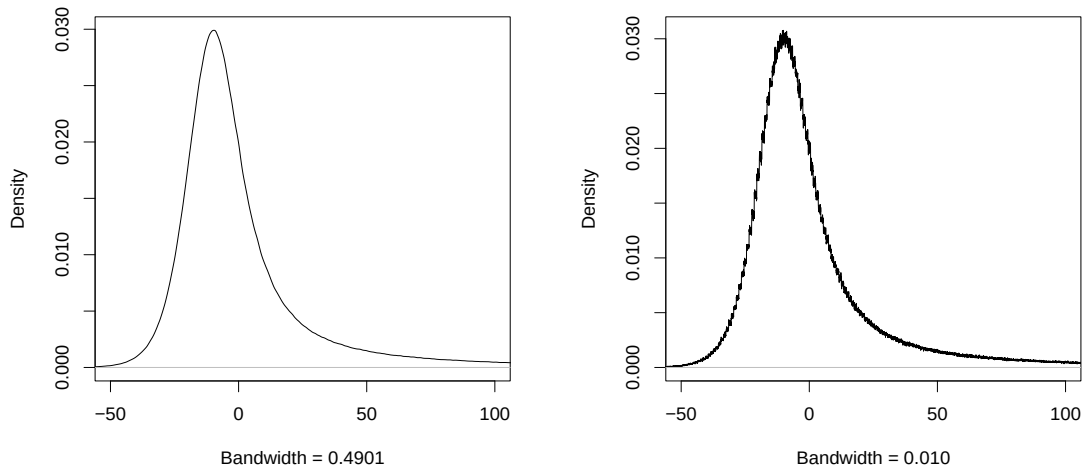

Figure 3.9: Kernel density estimates with bandwidths $h = \hat{h}(\hat{r}_0, N = 100)$ (left) and $h =$ `bw.ucv(·, nb=1e5, lower=0.01, upper=1)` (right).

Computing the standard cross-validation bandwidth for the whole sample is not an option due to the enormous time it would require, even with a binned implementation, as in the case of the R function `bw.ucv` (note that for this function to produce sensible results the number of bins must be very close to n, as previously discussed). Therefore, to predict the value of the cross-validation bandwidth for the original sample size, n, and also the time required for its computation, we used appropriate regression models. We repeated these experiments considering binned and non-binned cross-validation bandwidths. The predicted cross-validation bandwidth for the whole sample is practically identical whether or not one uses binning, with a large enough number of bins, and hence we just describe the experiment when using a binned implementation. Nevertheless, the predicted time is obviously much higher when binning is not used. Specifically, we selected 100 subsamples of sizes 557, 5579 and $55,793$ from the whole dataset. For each size and subsample, we computed the binned version of the leave-one-out cross-validation bandwidth (see Figure 3.10). Finally, we considered following the parametric regression model:

$$Y_i = \beta_0 n_i^{\beta_1}, \tag{3.14}$$

where $n_i \in \{557, 5579, 55793\}$ and $Y_i \in \{1.352, 2.129, 3.606\}$ denotes the mean of the binned cross-validation bandwidth using the subsamples of size n_i. Taking logarithms in (3.14) we get a linearized version of (3.14),

$$\log Y_i = \log \beta_0 + \beta_1 \log n_i,$$

which can be seen as a linear regression model with parameters $\log \beta_0$ (intercept) and β_1 (slope). We obtained the following least-squares estimates for the parameters of model (3.14):

$$\begin{aligned} \hat{\beta}_0 &= 13.69, \\ \hat{\beta}_1 &= -0.213. \end{aligned}$$

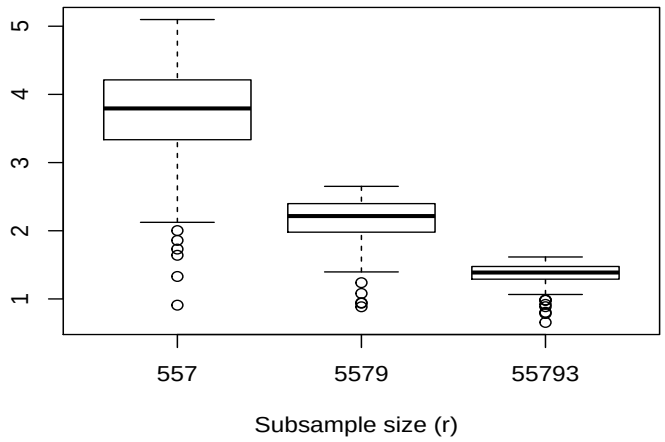

Figure 3.10: Boxplots of $\hat{h}_r$ for subsamples of size $r \in \{557, 5579, 55793\}$.

With these values of $\hat{\beta}_0$ and $\hat{\beta}_1$, the predicted value of the leave-one-out cross-validation bandwidth for the original sample size is $\hat{h}_n = 0.501$, very close to the value produced by the bagged bandwidth, $\hat{h}(r, N) = \hat{h}(272222, 100) = 0.490$. Figure 3.11 shows the fitted values for the nonlinear model defined in (3.14). Analogously, we considered a model similar to the one described in (3.14) to predict the time required to compute a binned version of the ordinary cross-validation bandwidth for the orig-

inal sample. Fitted values for the model are shown in Figure 3.12. As previously, we employed the R function `bw.ucv` with `nb` (number of bins) equal to the corresponding sample size to compute the different cross-validation bandwidths. In this case and under the same notation as in (3.14), we considered $n_i \in \{5579, 55793, 557934\}$ and $Y_i \in \{0.0102, 0.959, 103.08\}$, with Y_i now denoting the elapsed time (in seconds) needed to compute `bw.ucv(·, nb=`n_i`)`, that is, the binned cross-validation bandwidth for a sample of size n_i with the number of bins set to n_i. Again, using the same notation as in (3.14), we obtained the following estimates for the model parameters:

$$\begin{aligned} \hat{\beta}_0 &= 3.14 \times 10^{-10}, \\ \hat{\beta}_1 &= 2.002. \end{aligned}$$

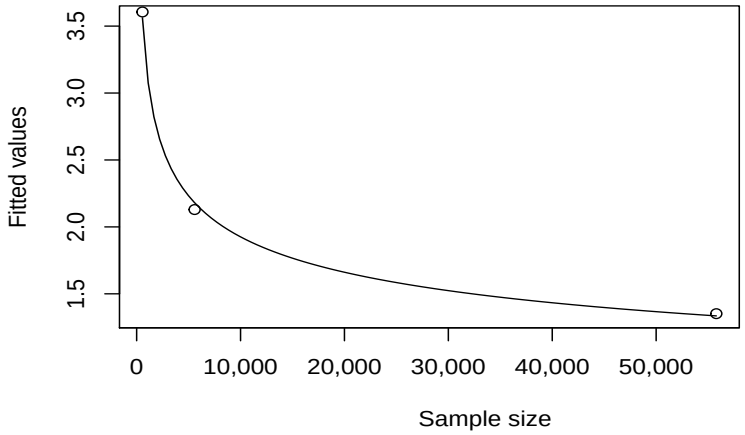

Figure 3.11: Fitted values for the regression model defined in (3.14). White dots correspond to the observations used to fit the model.

This means that the time needed to compute the binned cross-validation bandwidth for the original sample is predicted to be approximately 2.8 hours. Analogously, we repeated the experiment to predict the time required to compute a non-binned leave-one-out cross-validation bandwidth for the whole sample and this predicted time turned out to be 5.1 years. Fitted values for the model are shown in Figure 3.13.

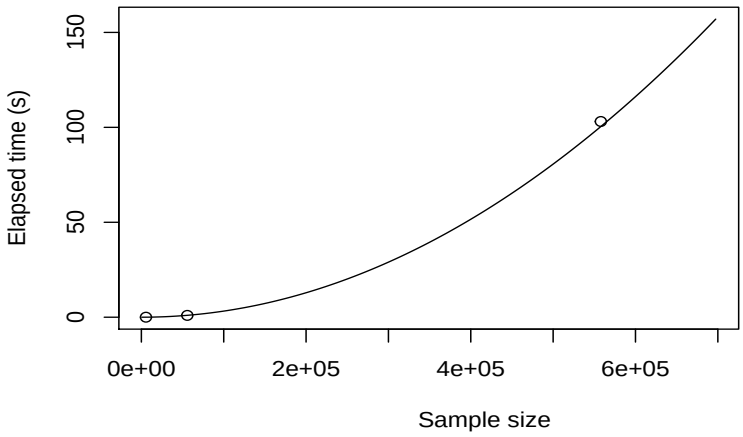

Figure 3.12: Fitted values for the regression model that relates the elapsed time needed to compute the binned cross-validation bandwidth to the sample size. White dots correspond to the observations used to fit the model.

The techniques discussed in this chapter were also applied to a dataset related to the current COVID-19 pandemic. This dataset consists of a sample of size $n = 105,235$ which contains the age and the hospitalization time of people infected with COVID-19 in Spain from January 1, 2020 to December 20, 2020. Due to the high number of ties present in the data 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 "age" sample and $U_2 - U_3$ to the original "hospitalization time" sample. For both of these samples, kernel density estimates were computed using different methods of

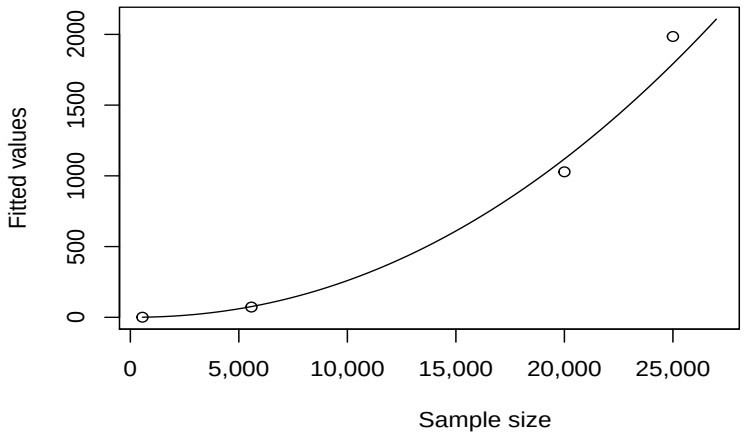

Figure 3.13: Fitted values for the regression model that relates the elapsed time needed to compute the standard non-binned cross-validation bandwidth to the sample size. White dots correspond to the observations used to fit the model.

bandwidth selection, namely, ordinary cross-validation, bagged cross-validation and bagged bootstrap (this bandwidth selector will be described below). These estimates are shown in Figure 3.14. In order to avoid boundary effects and alleviate the effect of outliers, both samples were first transformed by means of the Box-Cox family:

$$\begin{aligned} T_{age}(x) &= \frac{x^{1.4}}{1.4}, \\ T_{time}(y) &= \frac{y^{0.1}}{0.1}. \end{aligned}$$

Let us denote these transformed samples by $X_{age} = (X_1, \ldots, X_n)$ and $Y_{time} = (Y_1, \ldots, Y_n)$. The bandwidths, h_{age} and h_{time}, were then computed for these transformed samples and finally the results were detransformed and returned to their original scale by means of the kernel density estimators

$$\hat{f}_{age}(x) = \frac{1}{nh_{age}} \sum_{i=1}^{n} x^{0.4} \phi \left(\frac{x^{1.4}/1.4 - X_i}{h_{age}} \right)$$

and

$$\hat{f}_{time}(y) = \frac{1}{nh_{time}} \sum_{i=1}^{n} x^{-0.9} \phi \left(\frac{x^{0.1}/0.1 - Y_i}{h_{time}} \right)$$

The subagged bootstrap bandwidths obtained for the transformed samples relative to the age of people hospitalized after being infected with COVID-19 and the hospitalization time were, respectively, $\hat{h}^*(r, N) = 7.47$ and $\hat{h}^*(r, N) = 0.098$. In both cases, the number of subsamples was set to $N = 100$, the size of the subsamples was chosen as $r = \lfloor n^{0.7} \rfloor = 3277$ and the number of bins used to compute the bandwidths was $\lfloor 0.1r \rfloor = 327$. As for the other bandwidth selectors, the values obtained for the transformed sample relative to the age (hospitalization time) were, respectively for the ordinary and bagging cross-validation bandwidths, $\hat{h}_n = 0.8$ (0.072) and $\hat{h}(r, N) = \hat{h}(3277, 100) = 5.96$ (0.087). For comparative purposes, direct plug-in bandwidths were also computed for the two transformed samples and their values turned out to be $\hat{h}_{n,dpi} = 5.91$ (0.086).

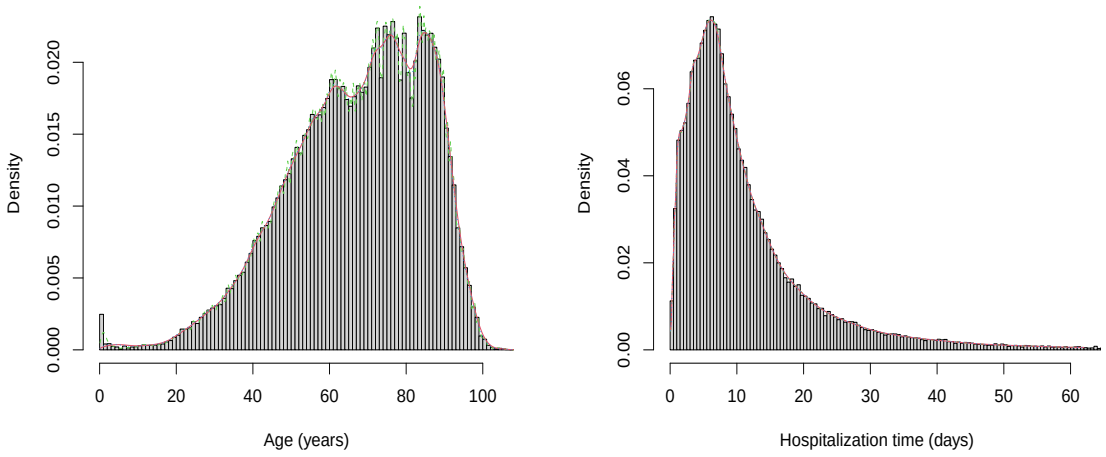

Figure 3.14: Histograms and kernel density estimates for the age (left panel) and hospitalization time (right panel) of people infected with COVID-19 in Spain from January 1, 2020 to December 20, 2020. For the kernel density estimates, the ordinary cross-validation (dashed green), bagged cross-validation (dotted blue) and bagged bootstrap bandwidth (solid red) were considered.

Chapter 4

Bagging bandwidth selection for the Nadaraya–Watson estimator

Kernel regression estimation and the crucial problem of bandwidth selection were presented in Section 2.2.1. In addition, the usefulness of bagging when working with highly variable estimators was discussed in Section 2.4. Thus, this Chapter is devoted to the theoretical and empirical study of the bagged cross-validation bandwidth selector for the Nadaraya–Watson estimator defined in (2.17). First, the hitherto ignored second-order asymptotics of the ordinary cross-validation selector of the bandwidth of the Nadaraya–Watson estimator is studied, this being necessary to proceed with the theoretical analysis of the bagged bandwidth. Then, the asymptotic properties of the proposed bagged bandwidth selector are obtained and its better performance is shown, in terms of both rates of convergence and computational agility, in relation to the ordinary cross-validation bandwidth selector. Finally, the behavior of the proposed bagged bandwidth is illustrated by means of various simulation studies as well as by an application to a real dataset related to the current COVID-19 pandemic.

4.1 Cross-validation bandwidth selection

Before moving on to the study of the asymptotic properties of the cross-validation bandwidth defined in (2.22), it would be of interest to take a closer look at the cross-validation function defined in (2.21) in order to elucidate its relationship with the

MISE and other error criteria.

Consider a sample $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn from the nonparametric regression model outlined at the beginning of Section 2.2.1. Let (X_0, Y_0) be an observation independent of the sample $\mathcal{X}$ and drawn from the same distribution. We are interested in studying the following error criteria and finding their most appropriate empirical analogues:

$$\mathrm{E}\left\{[m(X_0) - \hat{m}_h(X_0)]^2\right\}, \tag{4.1}$$

$$\mathrm{E}\left\{[Y_0 - \hat{m}_h(X_0)]^2\right\}, \tag{4.2}$$

where the estimator $\hat{m}_h$ was constructed using the sample $\mathcal{X}$. We have that

$$\begin{aligned}
\mathrm{E}\left\{[m(X_0) - \hat{m}_h(X_0)]^2\right\} &= \mathrm{E}\left(\mathrm{E}\left\{[m(X_0) - \hat{m}_h(X_0)]^2 \mid \mathcal{X}\right\}\right) \\
&= \mathrm{E}\left\{\int [m(x) - \hat{m}_h(x)]^2 f(x)\, dx\right\} \\
&= M_n(h),
\end{aligned}$$

where we have used the law of iterated expectations, which states that, for any two random variables Z_1 and Z_2 defined on the same probability space, it is satisfied that

$$\mathrm{E}(Z_1) = \mathrm{E}\left[\mathrm{E}(Z_1 \mid Z_2)\right].$$

Thus, we have that the MISE is an equivalent error criterion to the one defined in (4.1), which may be interpreted as the (out-of-sample) mean squared estimation error of $\hat{m}_h$. On the other hand, the error criterion given by (4.2) can be seen as the mean squared prediction error of $\hat{m}_h$. In addition, both error criteria are closely related. Indeed, using $\mathrm{E}(Y_0^2 \mid X_0) = m(X_0)^2 + \sigma^2(X_0)$ and $\mathrm{E}(Y_0 \mid X_0) = m(X_0)$, we have:

$$\begin{aligned}
\mathrm{E}\left\{[Y_0 - \hat{m}_h(X_0)]^2\right\} &= \mathrm{E}\left(\mathrm{E}\left\{[Y_0 - \hat{m}_h(X_0)]^2\right\} \mid \mathcal{X}, X_0\right) \\
&= \mathrm{E}\left[\mathrm{E}\left(Y_0^2 \mid X_0\right) - 2\hat{m}_h(X_0)\mathrm{E}\left(Y_0 \mid X_0\right) + \hat{m}_h(X_0)^2\right] \\
&= \mathrm{E}\left[m(X_0)^2 + \sigma^2(X_0) + [\hat{m}_h(X_0) - m(X_0)]^2 - m(X_0)^2\right] \\
&= M_n(h) + \int \sigma^2(x) f(x)\, dx.
\end{aligned}$$

Thus, both criteria have the same minimizer since the difference between the two is given by the expected value of the conditional variance, $\int \sigma^2 f$, which does not depend on the bandwidth.

Among the possible empirical analogues of (4.1) and (4.2) are, respectively, the mean average squared estimation error (MASEE) and the mean average squared prediction error (MASPE), given by

$$\begin{aligned} \mathrm{MASEE}(h) &= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h(X_i) - m(X_i)\right]^2\right\}, \qquad (4.3) \\ \mathrm{MASPE}(h) &= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h(X_i) - Y_i\right]^2\right\}. \qquad (4.4) \end{aligned}$$

Note that

$$\begin{aligned} \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[m(X_i) - Y_i\right]^2\right\} &= \mathrm{E}\left\{\left[m(X_1) - Y_1\right]^2\right\} \\ &= \mathrm{E}\left(\varepsilon_1^2\right) = \mathrm{E}\left[\mathrm{E}\left(\varepsilon_1^2 \mid X_1\right)\right] \\ &= \mathrm{E}\left[\sigma^2(X_1)\right] = \int \sigma^2(x) f(x)\, dx. \end{aligned}$$

Then, after a few simple calculations we obtain

$$\begin{aligned} \mathrm{MASPE}(h) &= \mathrm{MASEE}(h) + \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[m(X_i) - Y_i\right]^2\right\} \\ &+ 2\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h(X_i) - m(X_i)\right]\left[m(X_i) - Y_i\right]\right\} \\ &= \mathrm{MASEE}(h) + \int \sigma^2(x) f(x)\, dx \\ &+ 2\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h(X_i) - m(X_i)\right]\left[m(X_i) - Y_i\right]\right\}, \end{aligned}$$

while

$$\mathrm{E}\left\{\left[Y_0 - \hat{m}_h(X_0)\right]^2\right\} = M_n(h) + \int \sigma^2(x) f(x)\, dx,$$

where we have used the fact that

$$\mathrm{E}\left\{[Y_0 - m(X_0)]^2\right\} = \int \sigma^2(x) f(x)\, dx$$

and

$$\begin{aligned} & \mathrm{E}\left\{[Y_0 - m(X_0)]\,[m(X_0) - \hat{m}_h(X_0)]\right\} \\ = \; & \mathrm{E}\left(\mathrm{E}\left\{[m(X_0) - \hat{m}_h(X_0)]\,[Y_0 - m(X_0) \mid X_0, \mathcal{X}]\right\}\right) = 0. \end{aligned}$$

Note that (4.3) and (4.4) are not suitable empirical analogues of (4.1) and (4.2), respectively, since in (4.3) and (4.4) the same sample, $\mathcal{X}$, is used in both the construction of the estimator $\hat{m}_h$ and the evaluation of the fit, thus not correctly mimicking (4.1) and (4.2). With the above in mind, let us now define leave-one-out versions of the criteria given in (4.3) and (4.4),

$$\begin{aligned} \widetilde{\mathrm{MASEE}}(h) &= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - m(X_i)\right]^2\right\}, \\ \widetilde{\mathrm{MASPE}}(h) &= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - Y_i\right]^2\right\}, \end{aligned}$$

where $\hat{m}_h^{(-i)}$ denotes the Nadaraya–Watson estimator constructed without the i-th observation, that is,

$$\hat{m}_h^{(-i)}(x) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{n} K_h\,(x - X_j)\, Y_j}{\sum_{\substack{j=1 \\ j\neq i}}^{n} K_h\,(x - X_j)}.$$

Now we have

$$\begin{aligned}
\widetilde{\mathrm{MASPE}}(h) &= \widetilde{\mathrm{MASEE}}(h) + \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[m(X_i) - Y_i\right]^2\right\} \\
&+ 2\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - m(X_i)\right]\left[m(X_i) - Y_i\right]\right\} \\
&= \widetilde{\mathrm{MASEE}}(h) + \int \sigma^2(x) f(x)\, dx, \qquad (4.5)
\end{aligned}$$

since

$$\begin{aligned}
& \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - m(X_i)\right]\left[m(X_i) - Y_i\right]\right\} \\
= \; & \mathrm{E}\left(\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - m(X_i)\right]\left[m(X_i) - Y_i\right] \mid \mathcal{X}^{-i}\right\}\right) \\
= \; & \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - m(X_i)\right]\mathrm{E}\left[m(X_i) - Y_i \mid X_i\right]\right\} \\
= \; & 0,
\end{aligned}$$

where $\mathcal{X}^{-i} = \{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), X_i, (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n))\}$ and we have used the fact that

$$\mathrm{E}\left[m(X_i) - Y_i \mid X_i\right] = 0, \quad i = 1, \ldots, n.$$

The result shown in (4.5) is valid for the case where the bandwidth is not random, but rather it was fixed in advance and therefore does not depend on the sample. However, since our main goal is to study the properties of the cross-validation bandwidth selector, we are interested in knowing whether expression (4.5) still approximately holds for the case where the bandwidth of the Nadaraya-Watson estimator is random.

In particular, we conjecture that

$$\begin{aligned} & \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(-i)}_{\hat{h}_{CV,n}}(X_i)-Y_i\right]^2\right\} \\ = \; & \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(-i)}_{\hat{h}_{CV,n}}(X_i)-m(X_i)\right]^2\right\}+\mathrm{ECV}+o(1), \end{aligned} \tag{4.6}$$

where ECV denotes the expected value of the conditional variance function, that is,

$$\mathrm{ECV}=\int \sigma^2(x)f(x)\,dx.$$

To get an idea of the validity of expression (4.6), we have considered samples of size $n \in \{100, 1000, 20000\}$ drawn from models M1, M2 and M3, which will be defined in Section 4.4, and considered, for each of the simulated samples, the cross-validation bandwidth when constructing the Nadaraya–Watson estimator. Table 4.1 shows the relative error (RE), defined as

$$\mathrm{RE}=\frac{\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(-i)}_{\hat{h}_{CV,n}}(X_i)-Y_i\right]^2\right\}-\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}^{(-i)}_{\hat{h}_{CV,n}}(X_i)-m(X_i)\right]^2\right\}-\mathrm{ECV}}{\mathrm{ECV}},$$

for each of the sample sizes and models considered. As can be seen, the results are consistent with expression (4.6).

	M1	M2	M3
$n = 100$	-1.35	-2.17	-3.02
$n = 1000$	0.44	0.39	0.63
$n = 20000$	0.097	0.097	0.10

Table 4.1: Relative error (RE) for $n \in \{100, 1000, 20000\}$ and models M1, M2 and M3. Values are shown multiplied by 100.

Thus, the close connection between the MISE and cross-validation criteria follows from the fact that minimizing $\widetilde{\mathrm{MASPE}}(h) = \mathrm{E}\left[CV_n(h)\right]$ is equivalent to minimizing $\widetilde{\mathrm{MASEE}}(h)$, the latter being an empirical analogue of $M_n(h)$.

The idea behind the cross-validation bandwidth selector in kernel regression was briefly discussed in Section 2.2.2. Now, in order to derive the asymptotic properties

of (2.22) as an estimator of (2.19), studying certain moments of (2.21) and its derivatives will be required. However, the fact that the Nadaraya–Watson estimator has a random denominator makes this a very difficult task. To overcome this problem, it is useful to rewrite the Nadaraya–Watson estimator as

$$\hat{m}_h(x) = A + B + C + D + E + F, \tag{4.7}$$

where

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

and

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

Expression (4.7) 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 unobservable, modified version of the Nadaraya–Watson estimator given by $\tilde{m}_h(x) =$

$A + B + C + D$, that is:

$$\begin{aligned} \tilde{m}_h(x) &= m(x) + \frac{1}{n^2 f(x)^2} \sum_{j=1}^{n} \sum_{k=1}^{n} K_h\left(x - X_j\right) \left[Y_j - m(x)\right] \\ &\quad \left[2f(x) - K_h\left(x - X_k\right)\right], \end{aligned} \tag{4.8}$$

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". Moreover, (4.8) does not define an estimator but a theoretical approximation of (2.17). 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

$$\bar{m}_h(x) = m(x) + \frac{1}{nf(x)} \sum_{i=1}^{n} K_h\left(x - X_i\right) \left[Y_i - m(x)\right]. \tag{4.9}$$

Following this approach, (4.8) could be used to define a theoretical approximation of the MISE function defined in (2.18), namely

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

The bandwidth that minimizes $\tilde{M}_n(h)$ is denoted by $\tilde{h}_{n0}$. On the other hand, (4.8) can also be used to define a modified version of the cross-validation criterion,

$$\widetilde{CV}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left[\tilde{m}_h^{(-i)}(X_i) - Y_i\right]^2, \tag{4.10}$$

where $\tilde{m}_h^{(-i)}$ denotes the leave-one-out version of (4.8) without the i-th observation, that is,

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

The bandwidth that minimizes (4.10) is denoted by $\tilde{h}_{CV,n}$. Using Taylor expansions, we can obtain the following approximation:

$$\begin{aligned}\tilde{h}_{CV,n} - \tilde{h}_{n0} &\approx -\frac{\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})}{\tilde{M}''_n(\tilde{h}_{n0})} \\ &+ \frac{\left[\widetilde{CV}'_n(\tilde{h}_{n0}) - \tilde{M}'_n(\tilde{h}_{n0})\right]\left[\widetilde{CV}''_n(\tilde{h}_{n0}) - \tilde{M}''_n(\tilde{h}_{n0})\right]}{\tilde{M}''_n(\tilde{h}_{n0})^2}, \qquad (4.11)\end{aligned}$$

where the second term of (4.11) 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 $\mathrm{E}[\widetilde{CV}_n^{k)}(h)]$ and $\tilde{M}_n^{k)}(h)$ coincide for every $k \geq 1$, we need to calculate the second order terms of both $\mathrm{E}[\widetilde{CV}'_n(\tilde{h}_{n0})]$ 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 $\mathrm{var}[\widetilde{CV}'_n(\tilde{h}_{n0})]$ is enough, and so it is useful to work with the simpler, linear approximation of $\hat{m}_h(x)$ given by (4.9).

4.1.1 Asymptotic results

The asymptotic bias and variance of the cross-validation bandwidth minimizing (4.10) are derived in this section. For this, some previous lemmas are proved. The following asumptions are needed:

B1. K is a symmetric and differentiable kernel function.

B2. For every $j = 0, \ldots, 6$, the integrals $\mu_j(K)$, $\mu_j(K')$ and $\mu_j(K^2)$ exist and are finite.

B3. The functions m and f are eight times differentiable.

B4. The function σ^2 is four times differentiable.

Lemma 4.1 provides expressions for the first and second order terms of both the bias and the variance of (4.8).

Lemma 4.1 *Under assumptions* B1–B4*, the bias and the variance of the modified version of the Nadaraya–Watson estimator defined in* (4.8) *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

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

Assuming that

$$\begin{aligned}
B_1 &= \mu_2(K)^2\int\left[\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right]^2 f(x)\,dx, \\
V_1 &= R(K)\int\sigma^2(x)\,dx, \\
B_2 &= 2\mu_2(K)\int\left[\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right]\left\{\mu_4(K)\left[\frac{1}{24}m^{4)}(x) + \frac{1}{6}\frac{m'''(x)f'(x)}{f(x)}\right.\right. \\
&+ \left.\left.\frac{1}{4}\frac{m''(x)f''(x)}{f(x)} + \frac{1}{6}\frac{m'(x)f'''(x)}{f(x)}\right] - \mu_2(K)^2\frac{f''(x)}{f(x)}\left[\frac{1}{4}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right]\right\} \\
&\quad f(x)\,dx
\end{aligned}$$

and

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

exist finite, then, it follows from Lemma 4.1 that

$$\tilde{M}_n(h) = B_1 h^4 + V_1 n^{-1} h^{-1} + B_2 h^6 + V_2 n^{-1} h + O\left(h^8 + n^{-1}h^2 + n^{-2}h^{-2} + n^{-3}h^{-3}\right).$$

Lemma 4.2 provides expressions for the first and second order terms of both the expectation and variance of $\widetilde{CV}'_n(h)$.

Lemma 4.2 *Let us define*

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

Then, under assumptions B1–B4*, 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 4.1, which can be derived from (4.11), Lemma 4.1 and Lemma 4.2, provides asymptotic expressions for the bias and variance of the cross-validation bandwidth minimizing (4.10).

Theorem 4.1 *Under the assumptions of Lemma 4.2 and assuming that B_2 and V_2*

exist finite, the asymptotic bias and variance of the bandwidth that minimizes (4.10) *are:*

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

where

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

Corollary 4.1 *Under the assumptions of Theorem 4.1, the asymptotic distribution of the bandwidth that minimizes* (4.10) *is*

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

where the constant V was defined in Theorem 4.1.

Remark 4.1 *Although the results presented so far involve only the modified cross-validation bandwidth, defined as the bandwidth that minimizes* (4.10)*, it seems reasonable to think that these asymptotic results also apply to the standard cross-validation bandwidth defined in* (2.22)*, this being the rationale behind the decomposition of the Nadaraya–Watson estimator proposed in* (4.7)*. Despite the lack of a rigorous demonstration in this regard, the equation below allows us to assess how fast the two bandwidths approach each other as the sample size increases. Thus, under suitable assumptions, it can be proved that*

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

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

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

4.2 Bagging cross-validation in kernel regression estimation

Although the cross-validation method studied in the previous section is very useful when selecting bandwidths in nonparametric regression, it has the disadvantage of having a high computational cost when the sample size is even moderately large. This problem can be partially circumvented by using bagging in the bandwidth selection procedure. In this section, we explain how bagging may be applied to the cross-validation bandwidth selector in the context of nonparametric regression. Additionally, the asymptotic properties of the proposed bagged bandwidth are derived. Apart from the obvious reduction in computing time, we will see that the bagged cross-validation bandwidth also presents better theoretical properties than the standard, non-bagged cross-validation selector.

Let $\mathcal{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a simple random sample of size n drawn from the nonparametric regression model specified in Section 2.2.1 and let $\mathcal{X}^* = \{(X_1^*, Y_1^*), \ldots, (X_r^*, Y_r^*)\}$ be a random sample of size $r < n$ drawn without replacement from $\mathcal{X}$. This subsample is used to calculate a cross-validation bandwidth, $\hat{h}_{CV,r}$. A rescaled version of $\hat{h}_{CV,r}$, namely $(r/n)^{1/5}\hat{h}_{CV,r}$, can be seen as a feasible estimator of the optimal MISE bandwidth, h_{n0}, for $\hat{m}_h$. Bagging consists of repeating the resampling 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:

$$\hat{h}(r, N) = \frac{1}{N}\left(\frac{r}{n}\right)^{1/5} \sum_{i=1}^{N} \hat{h}_{CV,r,i}. \tag{4.12}$$

Although the same notation was used for the bagging cross-validation selector (see equation (3.1)) for the bandwidth of the kernel density estimator, we will keep it since there will be no possibility of confusion.

As pointed out in Chapter 3, in the case of kernel density estimation, both the asymptotic properties and the empirical behavior of this type of bandwidth selector have already been studied in Hall and Robinson (2009) for $N = \infty$ and generalized in Barreiro-Ures et al. (2021a), where the asymptotic properties of the bandwidth selector are derived for the more practical case of a finite N. Furthermore, as discussed

in Section 3.1, 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 (4.12) can be defined. This modified bagging bandwidth uses modified cross-validation (see (4.10)) bandwidths $\tilde{h}_{CV,r,i}$ instead of $\hat{h}_{CV,r,i}$, for $i = 1, \ldots, N$, and it is given by

$$\tilde{h}(r, N) = \frac{1}{N} \left(\frac{r}{n}\right)^{1/5} \sum_{i=1}^{N} \tilde{h}_{CV,r,i}. \tag{4.13}$$

In the next section, asymptotic expressions for the bias and variance of the bagging bandwidth defined in (4.13), as well as its limit distribution, are obtained when considering the Nadaraya-Watson estimator. From these results and considering Remark 4.1, it seems reasonable that similar results for (4.12) should be obtained.

4.2.1 Asymptotic results

Expressions for the bias and the variance of (4.13) are given in Theorem 4.2. In addition to assumptions B1–B4, assumption A1, stated in Section 3.1.1 will also be necessary.

Theorem 4.2 *Under assumptions B1–B4 and A1, the bias and variance of the bagged cross-validation bandwidth defined in* (4.13) *verify*

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

where the constants $\mathcal{B}$ and V were defined in Theorem 4.1 and the constant C_1 is defined in (C.42).

Corollary 4.2 *Under assumptions B1–B4 and A1, the asymptotic distribution of the bagged cross-validation bandwidth defined in* (4.13) *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 4.1. In particular, if we assume that $r=o\left(n/\sqrt{N}\right)$, then

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

Using Remark 4.1, it could be proved that similar results to those in Corollary 4.2 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 to a normal distribution with zero mean and constant variance at the rate $n^{-3/10}$, this result can be improved through the use of bagging and letting r and N tend to infinity at adequate rates. For example, if both r and N were to tend 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 $n^{-3/10}$.

4.3 Choosing an optimal subsample size

Analogously to what was pointed out in Section 3.1.2 in the case of the bagging cross-validation bandwidth selector for the kernel density estimator, 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 the variance of the bagging bandwidth would converge to zero at the rate

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

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

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

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

$$\text{AMSE}\left[\tilde{h}(r, N)\right] = (\mathcal{B} + C_1)^2 r^{-4/5} n^{-2/5} + V r^{-1/5} n^{-2/5}\left[\frac{1}{N} + \left(\frac{r}{n}\right)^2\right]. \quad (4.14)$$

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

$$r_0 = \arg\min_{r>1} \text{AMSE}\left[\tilde{h}(r, N)\right].$$

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 (through kernel density estimation, for example) and an estimate, $\hat{m}$, of the regression function (for instance by fitting a polynomial whose degree could be chosen by some criterion such as the AIC or BIC). 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 $i \in \{1, \ldots, s\}$ denotes the i-th

subsample.

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 $\widehat{\text{AMSE}}\left[\tilde{h}(r,N)\right]$ by plugging these bagged estimates into (4.14).

Step 5. Finally, estimate r_0 by

$$\hat{r}_0 = \arg\min_{r>1}\widehat{\text{AMSE}}\left[\tilde{h}(r,N)\right].$$

Additionally, if we assume that $r = o\left(n/\sqrt{N}\right)$, then we would have

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

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 as long as N tends to infinity at a rate faster than $n^{3/20}$. Furhtermore, 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 this way, the complexity of the algorithm will only equal that of standard cross-validation when N tends to infinity at the rate $n^{6/13}$.

4.4 Simulation studies

The behavior of the leave-one-out cross-validation bandwidths is evaluated by simulation in this section. To that end, the following regression models were considered:

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 4.1. The Gaussian kernel was used to compute the Nadaraya–Watson estimator throughout this section. Moreover, to reduce computing times in the simulations, we used binning to select the ordinary and bagged cross-validation bandwidths.

In a first step, we empirically checked how close the bandwidths that minimize the MISE of (4.8) and (2.17) are. For this, we simulated 100 samples of sizes 1000 and 5000 from models M1, M2 and M3 and computed the corresponding MISE curves for the standard Nadaraya–Watson estimator and for its modified version, given in (4.8). Figure 4.2 shows, for the previous models and each of the considered sample sizes, the MISE curves for (4.8) and the standard Nadaraya–Watson estimator. As it can be observed, the bandwidth that minimizes the MISE of (4.8) and the MISE of the standard Nadaraya–Watson estimator appear to be quite close even for moderately small sample sizes. Naturally, the distance between the minima of both curves tends to zero as the sample size increases. On the other hand, Figure 4.3 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.

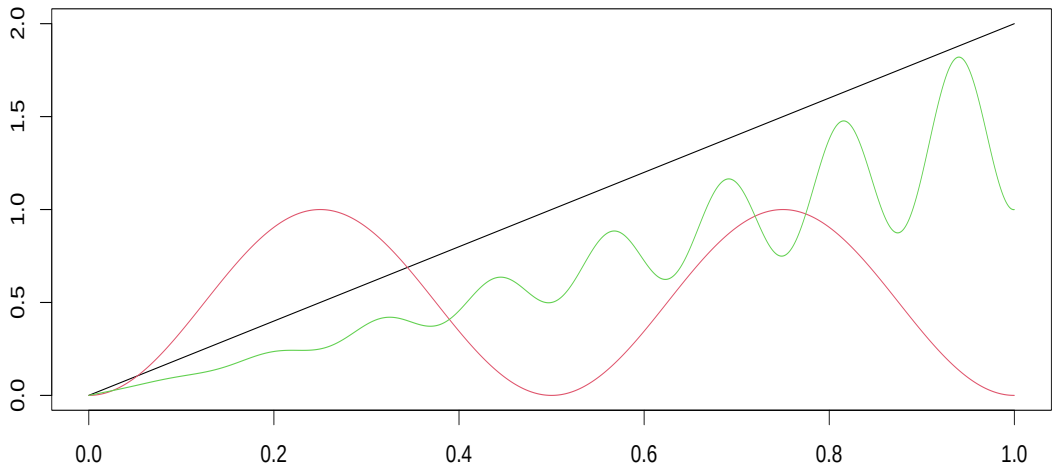

Figure 4.1: Regression function of models M1 (black), M2 (red) and M3 (green).

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 (the optimal MISE bandwidth was approximated by Monte Carlo simulations). Figure 4.4 shows the sampling distribution of both the standard cross-validation bandwidth, $h_{CV,n}$, and the statistic $S_n = n^{3/10}\left(\hat{h}_{CV,n} - h_{n0}\right)$, considering values of n between 50 and 5000 and samples drawn from model M2. Figure 4.4, in conjunction with Figure 4.5, which shows the kernel density estimates of S_n for each value of n considered, confirm the result reflected in Corollary 4.1, since the sampling distribution of S_n seems to tend to a normal distribution with zero mean and constant variance.

In the next part of the study, we focused on empirically analyzing the performance of the bagged cross-validation bandwidth $\hat{h}(r, N)$, defined in (4.13) and computed using the `bagreg` function from package `baggingbwsel`, for different values of n, r and N. Figure 4.6 shows the sampling distribution of $\hat{h}/h_{n0}$, where $\hat{h}$ denotes either the ordinary or the bagged cross-validation bandwidth. For this, 1000 samples of size $n = 10^5$ 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

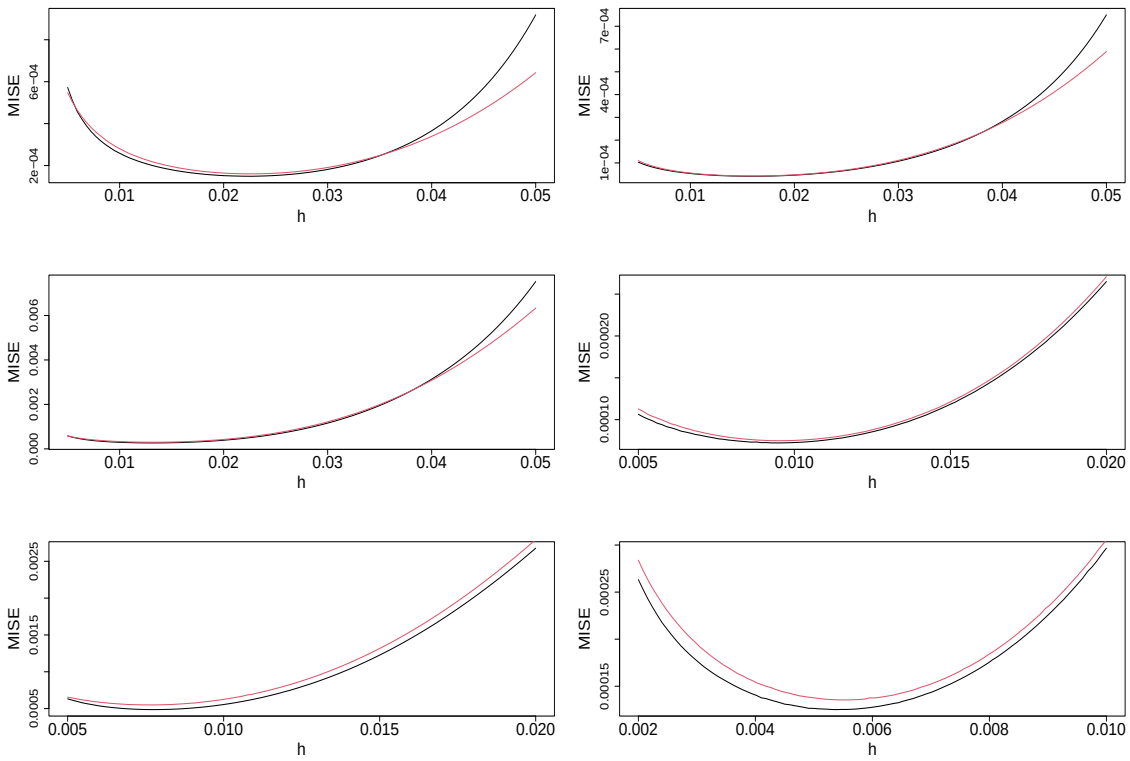

Figure 4.2: MISE curve for the standard Nadaraya–Watson estimator (black line) and its quadratic approximation (red line), defined in (4.8), with their minima (black and red points, respectively). First row: model M1, second row: model M2, third row: model M3. First column: $n = 1000$, second column: $n = 5000$.

models, it is observed how the squared 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 4.2, which shows the ratio of the mean squared errors of the bagged bandwidth and the ordinary cross-validation bandwidth, $\mathrm{MSE}\Big[\hat{h}(r,N)\Big]\,/\mathrm{MSE}\Big(\hat{h}_{CV,n}\Big)$, for all three models.

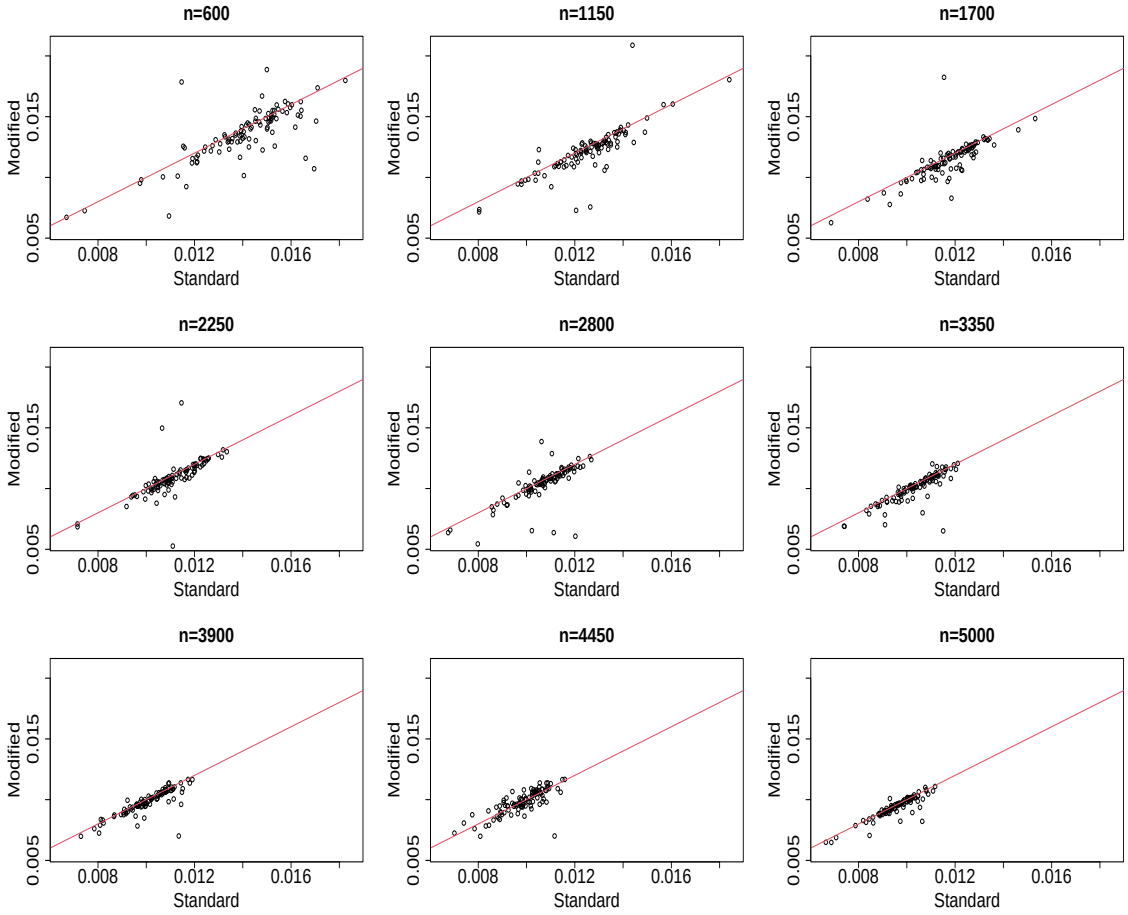

Figure 4.3: Cross-validation bandwidths using the standard Nadaraya–Watson estimator (x-axis) and its modified version (y-axis) for samples of sizes ranging from 600 to 5000 drawn from model M2.

Apart from the better statistical precision of the cross-validation bandwidths selected using bagging, another potential advantage of this approach is the reduction in computing time, especially when working with large sample sizes. To analyze this issue, Figure 4.7 shows the CPU elapsed times for the computation of the ordinary and bagged cross-validation bandwidths as a function of the sample size (n). 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. Different sample sizes, $n \in \{5000, 28750, 52500, 76250, 10^5\}$, and a fixed number of subsamples, $N = 25$,

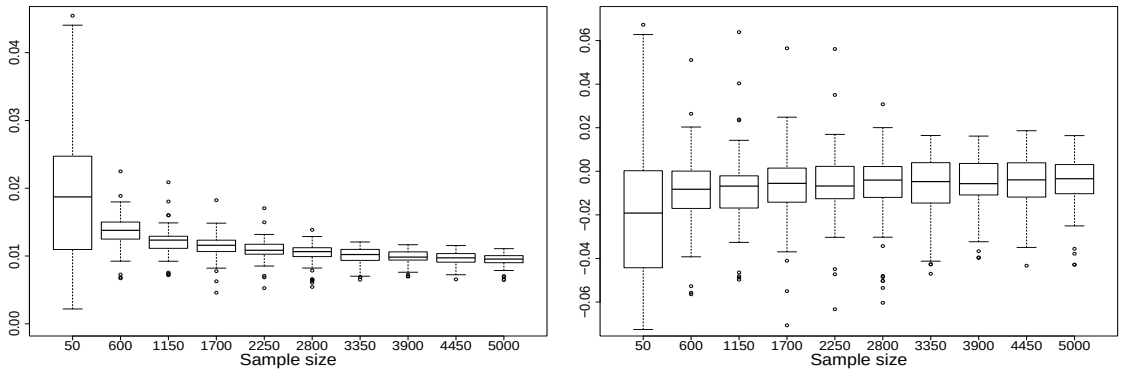

Figure 4.4: Sampling distribution of the standard cross-validation bandwidth (left), $\hat{h}_{CV,n}$, and sampling distribution of $n^{3/10}(\hat{h}_{CV,n} - h_{n0})$ (right), for samples drawn from model M2 and considering values of n between 50 and 5000.

	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

Table 4.2: Ratio of the mean squared errors of the bagged and the ordinary cross-validation bandwidth for models M1–M3. Different values of r and $N = 25$ were considered for a sample size of $n = 10^5$.

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.

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 4.7, the following log-linear model was used:

$$T(n) = \alpha n^{\beta}, \tag{4.15}$$

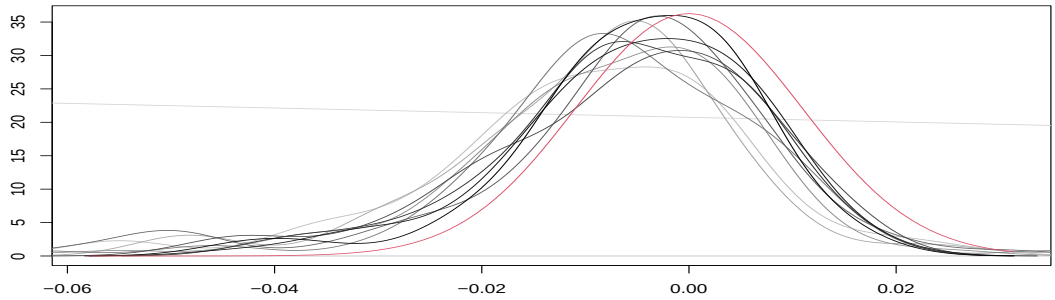

Figure 4.5: Kernel density estimates of $S_n = n^{3/10}(\hat{h}_{CV,n} - h_{n0})$, for n ranging from 50 to 5000 and samples drawn from model M2. The limit distribution of S_n is shown in red.

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 time 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 (4.15) estimating $\alpha, \beta > 0$ by least squares and, subsequently, to make predictions. Table 4.3 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 4.3 serve to illustrate the important reductions in computing time that bagging can provide for certain choices of r and N, especially when working with 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 4.7, Figure 4.8 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 4.7.

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

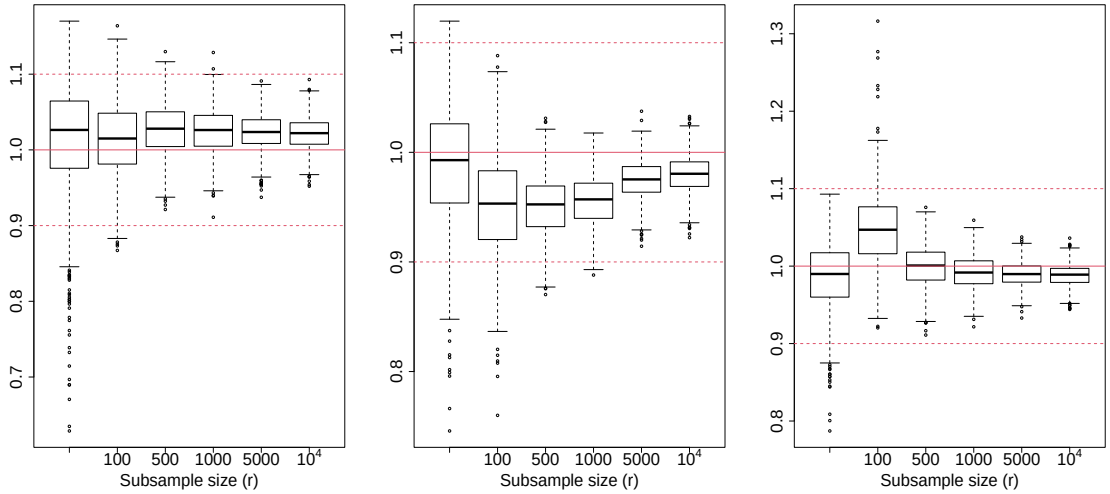

Figure 4.6: Sampling distribution of $\hat{h}_{CV,n}/h_{n0}$ (first boxplot on each panel) and $\hat{h}(r,N)/h_{n0}$ (second to sixth boxplots on each panel) for models M1 (left panel), M2 (central panel) and M3 (right panel), where the considered subsample sizes are $r \in \{100, 500, 1000, 5000, 10^4\}$ 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.

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 4.9 shows the sampling distribution of $\hat{h}/h_{n0}$, where $\hat{h}$ denotes either the ordinary or bagged cross-validation bandwidth. In the three scenarios, it can be observed that 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 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 for models M2 and M3, obtaining similar results.

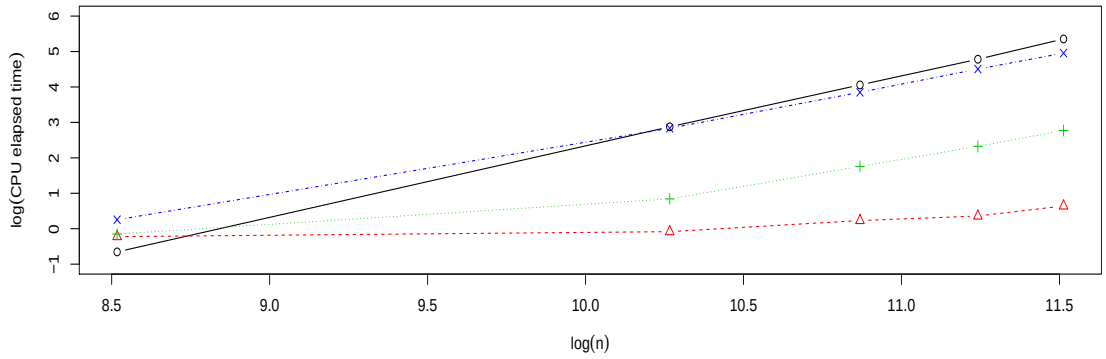

Figure 4.7: CPU elapsed time (seconds) as a function of the sample size of standard cross-validation (solid line-circles) and bagged cross-validation. Both variables are shown on a logarithmic scale. A fixed number of subsamples was used, $N = 25$. Three growth rates for r were considered, namely, $r = n^{0.7}$ (dashed line-triangles), $r = n^{0.8}$ (dotted line-pluses) and $r = n^{0.9}$ (dashed-dotted line-crosses).

Finally, it would also be of interest to check by means of simulations the validity of the results presented in Theorem 4.1. That is, we want to empirically check whether the rates of convergence to zero of the bias and variance of the ordinary cross-validation bandwidth, $\hat{h}_{CV,n}$, are both asymptotic to $n^{-3/5}$. We start by assuming

	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

Table 4.3: Predicted CPU elapsed time for the standard and the bagging cross-validation method using three different choices for the subsample size.

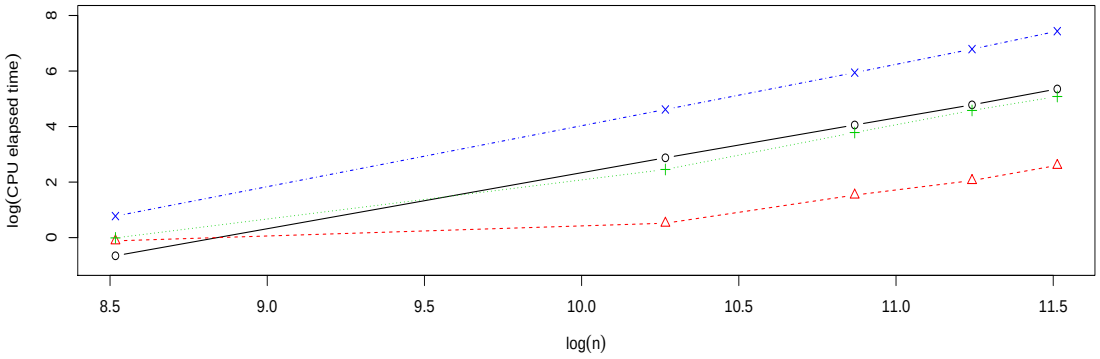

Figure 4.8: CPU elapsed time (seconds) as a function of the sample size of standard cross-validation (solid line-circles) and bagged cross-validation. Both variables are shown on a logarithmic scale. The number of subsamples grows with n at the rate $N = \sqrt{n}$. Three growth rates for r were considered, namely, $r = n^{0.7}$ (dashed line-triangles), $r = n^{0.8}$ (dotted line-pluses) and $r = n^{0.9}$ (dashed-dotted line-crosses).

that

$$\mathrm{E}\left(\hat{h}_{CV,n} - h_{n0}\right) \approx \eta_0 n^{-\eta_1}, \tag{4.16}$$

$$\mathrm{var}\left(\hat{h}_{CV,n}\right) \approx \zeta_0 n^{-\zeta_1}, \tag{4.17}$$

for certain constants η_0, ζ_0 and $\eta_1, \zeta_1 > 0$. By taking logarithms in (4.16) and (4.17) we get

$$\log\left[\mathrm{E}\left(\hat{h}_{CV,n} - h_{n0}\right)\right] \approx \log(\eta_0) - \eta_1 \log(n), \tag{4.18}$$

$$\log\left[\mathrm{var}\left(\hat{h}_{CV,n}\right)\right] \approx \log(\zeta_0) - \zeta_1 \log(n). \tag{4.19}$$

From (4.18) and (4.19) it would seem natural to consider the following linear regression models,

$$Z_{b,i} = C_b + r_b W_i, \tag{4.20}$$

$$Z_{v,i} = C_v + r_v W_i, \tag{4.21}$$

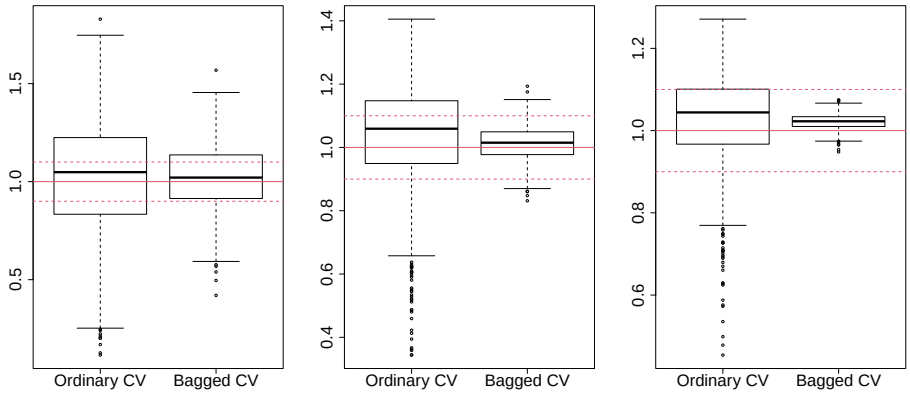

Figure 4.9: Sampling distribution of $\hat{h}/h_{n0}$, where $\hat{h}$ denotes either the ordinary or bagged cross-validation bandwidth, for samples of size $n = 50$ (left panel), $n = 500$ (central panel) and $n = 5,000$ (right panel) drawn from model M1. The values of r and N were chosen as $r = N = 4\sqrt{n}$. Dashed lines are plotted at values 0.9 and 1.1 for reference.

where

$$\begin{aligned} W_i &= \log(n_i), \\ Z_{b,i} &= \log\left[\mathrm{E}\left(\hat{h}_{CV,n_i} - h_{n_i 0}\right)\right], \\ Z_{v,i} &= \log\left[\mathrm{var}\left(\hat{h}_{CV,n_i}\right)\right]. \end{aligned}$$

Once models (4.20) and (4.21) have been fitted to $\{W_i, Z_{b,i}\}_{i=1}^{T}$ and $\{W_i, Z_{v,i}\}_{i=1}^{T}$, respectively, and thus obtained the estimates $\hat{C}_b$, $\hat{C}_v$, $\hat{r}_b$ and $\hat{r}_v$ for the parameters of both models, the final estimators are

$$\begin{aligned} \hat{\eta}_0 &= e^{\hat{C}_b}, \\ \hat{\zeta}_0 &= e^{\hat{C}_v}, \\ \hat{\eta}_1 &= -\hat{r}_b, \\ \hat{\zeta}_1 &= -\hat{r}_v. \end{aligned}$$

To implement this procedure, we simulated $T = 100$ samples of size n_i, with $n_i \in \{10^4, 5 \times 10^4, 10^5, 1.5 \times 10^5\}$, drawn from model M2. These samples were then

used to fit the linear models described in (4.20) and (4.21) (see Figure 4.10). Thus, the estimates obtained for the rates of convergence to zero of the bias and variance of the cross-validation bandwidth were

$$\begin{aligned} \hat{\eta}_1 &= 0.66, \\ \hat{\zeta}_1 &= 0.59, \end{aligned}$$

which are coherent with the rate of convergence to zero of both the bias and variance of the cross-validation bandwidth being asymptotic to $n^{-3/5} = n^{0.6}$, as stated in Theorem 4.1. Furthermore, the divergence between the theoretical convergence rates and the estimates is more than likely due to (i) Monte Carlo approximation error, (ii) the fact that the sample sizes considered may not have been large enough and (iii) the fact that a binned implementation of the cross-validation bandwidth was considered and a moderate number of bins (2000) were used to compute the bandwidth for each sample.

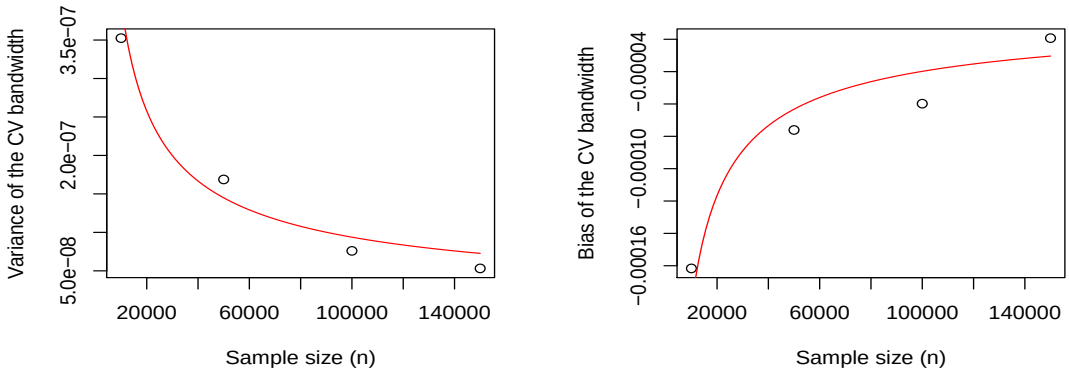

Figure 4.10: Log-linear fits for the bias (left) and variance (right) of the cross-validation bandwidth. 100 samples of sizes 10^4, 5×10^4, 10^5 and 1.5×10^5 were simulated from model M2.

4.5 Application to COVID-19 data

We shall now proceed to illustrate the performance of the techniques studied in the previous sections by applying them to a real dataset related to the current COVID-19 pandemic. This dataset, which contains the age (the explanatory variable) and the hospitalization time (the response variable) of people infected with COVID-19 in Spain from January 1, 2020 to December 20, 2020, was described and preprocessed as indicated in Section 3.5. The data was jittered as follows:

```
# xx: raw 'age' data
# yy: raw 'hospitalization time' data

set.seed(1)
noise1 = sample(runif(1e6,0,1), n, replace=FALSE)
nties = sum(duplicated(noise1))
ties = which(duplicated(noise1))
noise1_ = c(noise1[-ties], runif(nties,0,1))

noise2 = sample(runif(1e6,0,1), n, replace=FALSE)
nties2 = sum(duplicated(noise2))
ties2 = which(duplicated(noise2))
noise2_ = c(noise2[-ties2], runif(nties2,0,1))

noise3 = sample(runif(1e6,0,1), n, replace=FALSE)
nties3 = sum(duplicated(noise3))
ties3 = which(duplicated(noise3))
noise3_ = c(noise3[-ties3], runif(nties3,0,1))

x = xx + noise1_ # jittered 'age' data
y = yy + noise2_ - noise3_ # jittered 'hospitalization time' data
```

Figure 4.11 shows scatterplots for the complete sample as well as for 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 3000. The calculations associated with

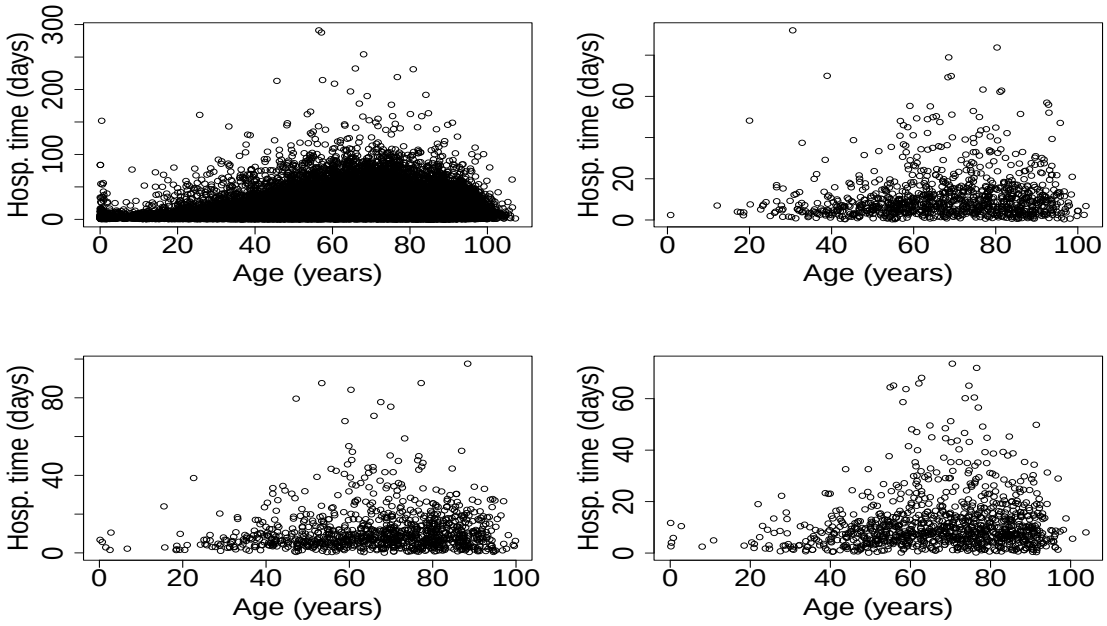

Figure 4.11: Full COVID-19 sample (top left panel) and three randomly chosen subsamples of size 1000.

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. Both the ordinary and bagging cross-validation bandwidths can be computed as follows, using the R packages `sm` and `baggingbwsel`:

```
# x: jittered age variable (numeric)
# y: jittered hospitalization time variable (numeric)

sm.mod::h.select(x,y,lower=log(0.1), upper=log(2), method="cv",
        poly.index=0, nbins=10000) # Ordinary, non-bagged CV bandwidth

baggingbwsel::bagreg(x, y, r=30000, s=10, h0=0.1, h1=2, nb=3000,
        ncores=5) # Bagging CV bandwidth
```

Actually, we had to employ a modified version of the `sm` package, `sm.mod`, so that it was possible to pass the limits of the search interval as arguments to the `h.select` function. Figure 4.12 shows the Nadaraya–Watson estimates with both standard and bagged cross-validation bandwidths. For comparative purposes, the local lin-

ear estimate with direct plug-in bandwidth Ruppert et al. (1995) is also shown. As can be seen, 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 transformation of the explanatory variable was carried out as follows:

```
# xx: raw 'age' variable

sxx = sort(xx)
xxx = ecdf(sxx)(sxx)
jumps = c(xxx[1], diff(ecdf(sxx)(unique(sxx))))

jumps.amp = NULL
for(i in 1:length(unique(sxx)))
{
  jumps.amp[which(sxx==unique(sxx)[i])] = jumps[i]
}

# Jittering the new samples:
set.seed(1)
xxxjit = xxx + runif(n, 0, jumps.amp)
yyy = yy[order(xx)]
yyyjit = yyy + runif(n) - runif(n)
```

The resulting estimators are shown in Figure 4.13, where the explanatory variable

was returned to its original scale by means of its empirical quantile function.

Figure 4.12: Kernel regression estimations for the COVID-19 data. The Nadaraya-Watson estimator with standard cross-validation bandwidth (dashed red line) and bagged cross-validation bandwidth (solid black line) as well as the local linear estimator with plug-in bandwidth (dotted blue line) 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 4.14. 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 4.14, 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 4.14.

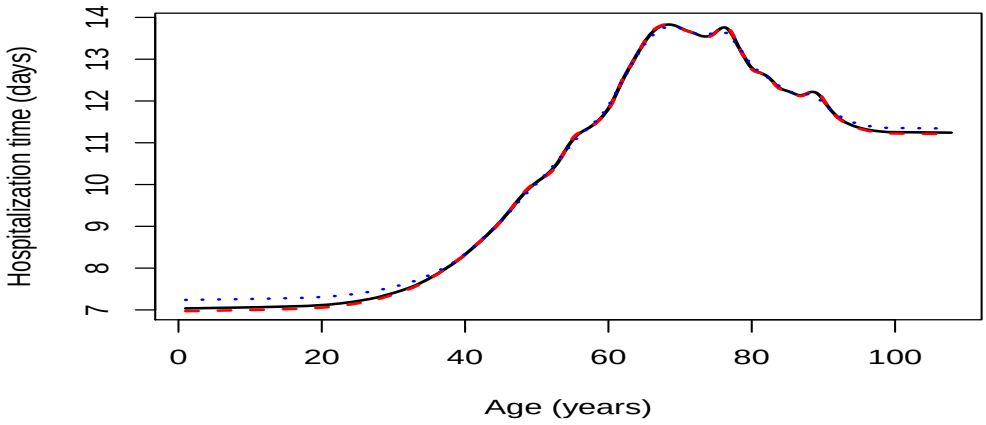

Figure 4.13: Kernel regression estimations for the COVID-19 data. To eliminate boundary effects, the explanatory variable was transformed by means of the empirical distribution function and then returned to its original scale by means of the empirical quantile function. The Nadaraya-Watson estimator with standard cross-validation bandwidth (dashed red line) and bagged cross-validation bandwidth (solid black line) as well as the local linear estimator with plug-in bandwidth (dotted blue line) are shown.

4.6 Bagging bootstrap bandwidth

As in the case of kernel density estimation and similarly to Section 3.2, the application of bagging to the bootstrap bandwidth selector for the Nadaraya–Watson estimator is considered.

Step 1. Select a pilot bandiwdth, g, and consider the Nadaraya–Watson estimator, $\hat{m}_g$, of m.

Step 2. Compute the model residuals

$$\hat{\varepsilon}_i = Y_i - \hat{m}_g(X_i), \quad i \in \{1, \dots, n\}.$$

Step 3. Generate a sample of size n, $U_1, \dots, U_n$, where U_i is drawn from a discrete uniform distribution defined in $\{1, \dots, n\}$ for every $i \in \{1, \dots, n\}$. Define the

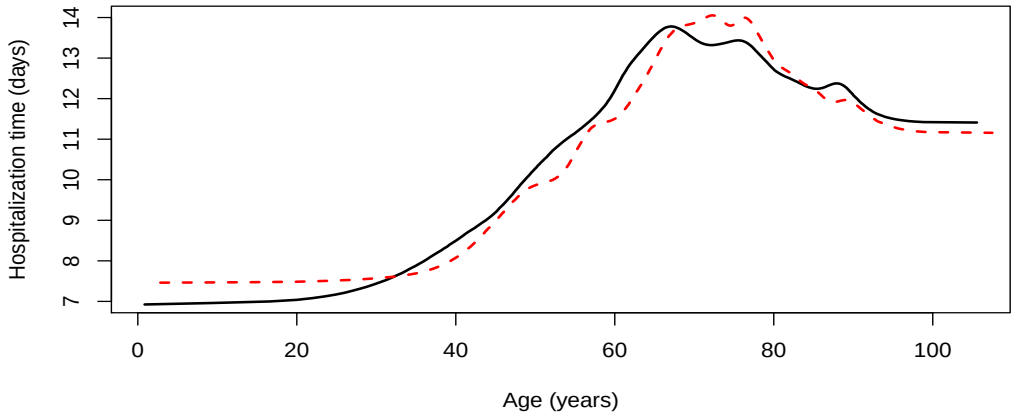

Figure 4.14: Kernel regression estimations for the COVID-19 data by gender, removing boundary effects. The Nadaraya–Watson estimators with bagged cross-validation bandwidths are shown for male (solid line) and female (dashed line) patients.

bootstrap responses by

$$Y_i^* = \hat{m}_g(X_i) + \hat{\varepsilon}_{U_i}.$$

Step 4. Consider the Nadaraya–Watson estimator constructed with the bootstrap sample $\{(X_1, Y_1^*), \ldots, (X_n, Y_n^*)\}$, that is,

$$\hat{m}_h^*(x) = \frac{\sum_{i=1}^{n} K_h\,(x - X_i)\, Y_i^*}{\sum_{i=1}^{n} K_h\,(x - X_i)}.$$

Step 5. Repeat the previous steps B times and approximate $M_n(h)$ by

$$M_n^*(h; g) = \mathrm{E}^* \left\{ \int \left[\hat{m}_h^*(x) - \hat{m}_g(x)\right]^2 \hat{f}_g(x)\, dx \right\}. \tag{4.22}$$

Step 6. Repeat Steps 1–5 for a large number of values of h and define the bootstrap

bandwidth as

$$h_{n0}^* = \arg\min_{h>0} M_n^*(h; g). \tag{4.23}$$

It is now straightforward to consider the bagging version of the bootstrap bandwidth defined in (4.23), namely:

Step 1. Independently generate N subsamples of size $r < n$ by sampling without replacement from $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Step 2. For $i \in \{1, \ldots, N\}$, select a pilot bandiwdth, g_i.

Step 3. For each of the subsamples, compute $h_{r0,i}^* = \arg\min_{h>0} M_r^*(h; g_i)$, where $M_n^*(h; g)$ was defined in (4.22). Denote these bandwidths by $h_{r0,1}^*, \ldots, h_{r0,N}^*$.

Step 4. Compute the bagged bandwidth as the mean of the rescaled bootstrap bandwidths,

$$\hat{h}^*(r, N) = \frac{1}{N} \left(\frac{r}{n}\right)^{1/5} \sum_{i=1}^{N} h_{r0,i}^*.$$

Unlike in the case of the bootstrap bandwidth selector for the kernel density estimator, the problem of pilot bandwidth selection has not been studied for the bootstrap bandwidth selector defined in (4.23).

Chapter 5

Conclusions and future work

Chapter 3 was devoted to analyzing the problem of bandwidth selection for the kernel density estimator defined in (2.3) using a bagging approach. In Section 3.1, the asymptotic properties of a bagged cross-validation bandwidth were studied in the case of a finite number of subsamples. The main results were established in Theorems 3.1 and 3.2. The former provides asymptotic expressions for the bias and variance of the proposed bagging cross-validation bandwidth, and its limit distribution is given in the latter. An automatic method for selecting the size of the subsamples based on the minimization of the mean squared error of the bagging cross-validation bandwidth was also proposed. The practical behavior of the bagging bandwidth was shown through different simulation studies and applications to real datasets. This bandwidth selector is an alternative to standard cross-validation, and it is able to achieve a large reduction in mean squared error due to a decrease in variance that greatly offsets its increase in bias. Furthermore, because of using subsampling, computing time can be significantly reduced with respect to using binned standard cross-validation. In the remaining sections of Chapter 3, other versions of the bagging bandwidth selector were proposed for certain types of scenarios: Section 3.2 includes a bagging bootstrap selector. Section 3.3 deals with situations where the rate of convergence to zero of the optimal bandwidth is not known, while Section 3.4 focuses on cases where it might be useful to incorporate second order terms in the bagging mechanism.

Although our proposed bandwidth selectors are mainly based on cross-validation

or bootstrapping, bagged versions of other bandwidth selection methods, such as plug-in, can be considered. While both cross-validation and bootstrap approaches try to estimate h_{n0}, plug-in bandwidths are estimators of the asymptotically optimal bandwidth, h_n, and hence only the estimation of $R(f'')$ is required. It is worth noting that there is a clear similarity between the three methods. Both cross-validation (Scott and Terrell, 1987) and bootstrap (Cao, 1993) bandwidths are the minimizers of criteria of the form

$$\sum_{(i,j)\in\mathcal{I}} H_{nhg}(X_i - X_j) + \frac{R(K)}{nh}, \tag{5.1}$$

where $\mathcal{I} \subset \{1,\ldots,n\} \times \{1,\ldots,n\}$ and H_{nhg} is a function which may depend on the sample size, n, the bandwidth, h, and a pilot bandwidth, g. Note that g plays a role only in the bootstrap criterion. Although plug-in bandwidths are not solutions to a minimization problem, the non-parametric estimation of $R(f'')$ using a pilot bandwidth, g, requires working with a U-statistic like the one given in (5.1), which would only depend on n and g. Due to the non-linearity of (5.1) with respect to the observations, it stands to reason that a bagged implementation of this method could reduce its variability, as in the case of cross-validation.

Chapter 4 focused on the problem of selecting the bandwidth for the Nadaraya–Watson estimator, defined in (2.17) in a random-design regression model. The asymptotic properties of the cross-validation bandwidth selector, $\tilde{h}_{CV,n}$ (the minimizer of (4.10)), based on the theoretical approximation of the Nadaraya–Watson proposed in (4.8) were studied in Section 4.1. The main results were established in Theorem 4.1 and Corollary 4.1. While the former provides asymptotic expressions for the bias and variance of $\tilde{h}_{CV,n}$, the limit distribution of $\tilde{h}_{CV,n}$ is given in the latter. Furthermore, in Remark 4.1 some reasons were offered as to why it would be expected that the asymptotic results mentioned above should also apply to the standard cross-validation bandwidth selector defined in (2.22). A more rigorous proof of this seemingly sensible intuition would be a line of future work to be considered.

In Section 4.2, a bagging cross-validation bandwidth selector for the Nadaraya–Watson estimator was proposed and its empirical behavior and asymptotic properties were subsequently studied. The main results were established in Theorem 4.2 and

Corollary 4.2. The former provides asymptotic expressions for the bias and variance of the proposed bagging cross-validation bandwidth, and its limit distribution is given in the latter. An interesting line of future work would involve the study of a bagging cross-validation bandwidth selector, analogous to the one defined in (4.13), for the local linear estimator. The main reason for studying such a bandwidth selector lies in the well-known fact that the local linear estimator has better statistical properties than the Nadaraya–Watson estimator (e.g., the local linear estimator is generally less biased and performs better at the boundary regions than the Nadaraya–Watson estimator). On the other hand, optimal pilot bandwidth choice for the bootstrap bandwidth selector defined in (2.26) remains to be studied. This would be a previous and necessary step to obtain the asymptotic properties of (2.26).

In addition to those already mentioned, other future lines of research to consider are the following: (i) extension of the proposed techniques to the case of multidimensional or dependent data, (ii) extension of the proposed techniques to bandwidth selection in classification problems, (iii) application of bagging to problems other than bandwidth selection, (iv) optimization of the code developed from the techniques proposed throughout the thesis, and (v) Python implementation of the proposed techniques (Python has an API for Apache Spark called PySpark which allows for parallel computing, very useful when dealing with large datasets or highly complex models).

List of Figures

List of Tables

Appendix A

Proofs of the results of Chapter 3

To prove Theorem 3.1, we establish one lemma in advance.

Lemma A.1 *Under assumptions A1–A4,*

$$n^{1/5} CV'''\left(\tilde{h}_n\right) = o_p(1),$$

where $\tilde{h}_n$ is a bandwidth between the cross-validation bandwidth $\hat{h}_n$ and the MISE minimizer h_{n0}.

Proof of Lemma A.1 First, we write

$$n^{1/5} CV'''(\tilde{h}_n) = \alpha_1 + \alpha_2, \tag{A.1}$$

with

$$\alpha_1 = n^{1/5} CV'''(h_{n0})$$

and

$$\alpha_2 = n^{1/5}\left[CV'''(\tilde{h}_n) - CV'''(h_{n0})\right].$$

To prove Lemma A.1 it is sufficient to show that $\alpha_1 = o_p(1)$ and $\alpha_2 = o_p(1)$. In order to study the term α_1, we first consider the asymptotic MISE of the Parzen–Rosenblatt estimator of the density function. It is well known that if K is a second

order symmetric kernel function and considering that K has variance 1, as stated in Assumption A2, the MISE is:

$$M_n(h) = \frac{R(K)}{nh} + \frac{1}{4}h^4 R(f'') + o\left[(nh)^{-1} + h^4\right],$$

and hence

$$M_n'''(h) = -\frac{6R(K)}{nh^4} + 6hR(f'') + o\left[(nh^4)^{-1} + h\right].$$

Since

$$-\frac{6R(K)}{nh_{na}^4} + 6h_{na}R(f'') = 0,$$

where h_{na} denotes the bandwidth minimizing the asymptotic MISE, it follows immediately that $n^{1/5}M_n'''(h_{n0})$ converges to zero. Now, we can write

$$n^{1/5}CV'''(h_{n0}) = n^{1/5}M_n'''(h_{n0}) + n^{1/5}\eta_n,$$

where $\eta_n = CV'''(h_{n0}) - M_n'''(h_{n0})$. Thus, to prove that $\alpha_1 = o_p(1)$, it is sufficient to prove that

$$\eta_n = o_p\left(n^{-1/5}\right),$$

or, by Markov's inequality, that

$$n^{2/5}\text{var}\left[CV'''(h_{n0})\right] = o(1).$$

It is easy to prove that, for every $r \geq 1$,

$$CV^{r)}(h) = M_n^{r)}(h) + \frac{1}{n(n-1)}\sum_{i\neq j}\bar{\gamma}_{nh}^{r)}(X_i - X_j), \tag{A.2}$$

where

$$\begin{aligned}
\gamma_n(u) &= \frac{n-1}{n} K * K(u) - 2K(u), \\
\gamma_{nh}(u) &= \gamma_n(u/h)/h, \\
\bar{\gamma}_{nh}(u) &= \gamma_{nh}(u) - \mathrm{E}\left[\gamma_{nh}(X_1 - X_2)\right], \\
\bar{\gamma}_{nh}^{r)}(u) &= \frac{d^r}{dh^r} \bar{\gamma}_{nh}(u).
\end{aligned}$$

Therefore,

$$\mathrm{var}\left[CV'''(h)\right] = \frac{1}{n^2(n-1)^2} \sum_{\substack{i,j,k,l=1 \\ i \neq j \\ k \neq l}}^{n} \mathrm{cov}\left[\Psi_3(X_i - X_j), \Psi_3(X_k - X_l)\right],$$

where

$$\Psi_3(u) = \frac{d^3}{dh^3}\gamma_{nh}(u) = -\left[\frac{6}{h^4}\gamma_n(u/h) + \frac{18u}{h^5}\gamma_n'(u/h) + \frac{9u^2}{h^6}\gamma_n''(u/h) + \frac{u^3}{h^7}\gamma_n'''(u/h)\right].$$

Counting the different possible cases, we get

$$\begin{aligned}
\mathrm{var}\left[CV'''(h)\right] &= \frac{1}{n^2(n-1)^2}\left\{4n(n-1)(n-2)\mathrm{cov}\left[\Psi_3(X_1 - X_2), \Psi_3(X_1 - X_3)\right]\right. \\
&+ \left. 2n(n-1)\mathrm{var}\left[\Psi_3(X_1 - X_2)\right]\right\}.
\end{aligned}$$

Let us now define $\tilde{\Psi}_3(u)$ as the function such that $\Psi_3(u) = \tilde{\Psi}_3(u/h)/h$. Consequently,

$$\tilde{\Psi}_3(u) = -\frac{1}{h^3}\left[6\gamma_n(u) + 18u\gamma_n'(u) + 9u^2\gamma_n''(u) + u^3\gamma_n'''(u)\right].$$

We shall now proceed to compute $\mu_j\left(\tilde{\Psi}_3\right)$, for $j = 0, 2, 4, 6$, and $\mu_j\left(\tilde{\Psi}_3^2\right)$, for $j = 0, 2$, since we will need these quantities later on. Note that $\mu_j\left(\tilde{\Psi}_3\right) = 0$ for every odd j, since $\tilde{\Psi}_3$ is symmetric.

For $j = 0$,

$$\mu_0\left(\tilde{\Psi}_3\right) = -\frac{1}{h^3}\left[6\mu_0(\gamma_n) + 18\mu_1(\gamma_n') + 9\mu_2(\gamma_n'') + \mu_3(\gamma_n''')\right].$$

Using integration by parts and the fact that $\mu_0(K) = \mu_0(K * K) = 1$, we get

$$\begin{aligned}
\mu_0(\gamma_n) &= -\frac{n+1}{n}, \\
\mu_1(\gamma_n') &= \frac{n+1}{n}, \\
\mu_2(\gamma_n'') &= -2\frac{n+1}{n}, \\
\mu_3(\gamma_n''') &= 6\frac{n+1}{n},
\end{aligned}$$

and hence

$$\mu_0\left(\tilde{\Psi}_3\right) = 0.$$

Now,

$$\mu_2\left(\tilde{\Psi}_3\right) = -\frac{1}{h^3}\left[6\mu_2(\gamma_n) + 18\mu_3(\gamma_n') + 9\mu_4(\gamma_n'') + \mu_5(\gamma_n''')\right].$$

Partial integration and the equality $\mu_2(K * K) = 2\mu_2(K)$ give

$$\begin{aligned}
\mu_2(\gamma_n) &= -\frac{2\mu_2(K)}{n}, \\
\mu_3(\gamma_n') &= \frac{6\mu_2(K)}{n}, \\
\mu_4(\gamma_n'') &= -\frac{24\mu_2(K)}{n}, \\
\mu_5(\gamma_n''') &= \frac{120\mu_2(K)}{n},
\end{aligned}$$

and, therefore,

$$\mu_2\left(\tilde{\Psi}_3\right) = 0.$$

We have

$$\mu_4\left(\tilde{\Psi}_3\right) = -\frac{1}{h^3}\left[6\mu_4(\gamma_n) + 18\mu_5(\gamma_n') + 9\mu_6(\gamma_n'') + \mu_7(\gamma_n''')\right].$$

Using integration by parts and the fact that $\mu_4(K * K) = 2\mu_4(K) + 6\mu_2(K)^2$, we

get

$$\begin{aligned}
\mu_4(\gamma_n) &= 6\mu_2(K)^2 - 2\mu_4(K)/n, \\
\mu_5(\gamma_n') &= -30\mu_2(K)^2 + 10\mu_4(K)/n, \\
\mu_6(\gamma_n'') &= 180\mu_2(K)^2 - 60\mu_4(K)/n, \\
\mu_7(\gamma_n''') &= -1260\mu_2(K)^2 + 420\mu_4(K)/n,
\end{aligned}$$

and, therefore,

$$\mu_4\left(\tilde{\Psi}_3\right) = \frac{144\mu_2(K)^2}{h^3} + O\left(\frac{1}{nh^3}\right).$$

Finally,

$$\mu_6\left(\tilde{\Psi}_3\right) = -\frac{1}{h^3}\left[6\mu_6(\gamma_n) + 18\mu_7(\gamma_n') + 9\mu_8(\gamma_n'') + \mu_9(\gamma_n''')\right].$$

Using integration by parts and the fact that $\mu_6(K*K) = 2\mu_6(K)+30\mu_2(K)\mu_4(K)$, we get

$$\begin{aligned}
\mu_6(\gamma_n) &= 30\mu_2(K)\mu_4(K) + O\left(\frac{1}{n}\right), \\
\mu_7(\gamma_n') &= -210\mu_2(K)\mu_4(K) + O\left(\frac{1}{n}\right), \\
\mu_8(\gamma_n'') &= 1680\mu_2(K)\mu_4(K) + O\left(\frac{1}{n}\right), \\
\mu_9(\gamma_n''') &= -15120\mu_2(K)\mu_4(K) + O\left(\frac{1}{n}\right),.
\end{aligned}$$

and so

$$\mu_6\left(\tilde{\Psi}_3\right) = \frac{3600\mu_2(K)\mu_4(K)}{h^3} + O\left(\frac{1}{nh^3}\right).$$

Analogously, it can be proved that

$$\begin{aligned}\mu_0\left(\tilde{\Psi}_3^2\right) &= O\left(\frac{1}{h^6}\right),\\ \mu_2\left(\tilde{\Psi}_3^2\right) &= O\left(\frac{1}{h^6}\right).\end{aligned}$$

On the other hand,

$$\operatorname{var}\left[\Psi_3(X_1 - X_2)\right] = I_1 - I_2^2$$

and

$$\operatorname{cov}\left[\Psi_3(X_1 - X_2), \Psi_3(X_1 - X_3)\right] = I_3 - I_2^2,$$

where

$$\begin{aligned}I_1 &= \int \Psi_3^2 * f(x) f(x)\, dx,\\ I_2 &= \int \Psi_3 * f(x) f(x)\, dx\end{aligned}$$

and

$$I_3 = \int \Psi_3 * f(x)^2 f(x)\, dx.$$

Simple algebra and Taylor expansions give

$$\begin{aligned}I_1 &= \frac{1}{h}\iint \tilde{\Psi}_3(u)^2 f(x)\left[f(x) + \frac{h^2u^2}{2} f''(\zeta)\right] dx\, du\\ &= \frac{1}{h}\left\{\mu_0\left(\tilde{\Psi}_3^2\right) R(f) + O\left[h^2\mu_2\left(\tilde{\Psi}_3^2\right)\right]\right\} = O\left(\frac{1}{h^7}\right),\\ I_2 &= \iint \tilde{\Psi}_3(u) f(x)\left[\frac{h^4u^4}{4!} f^{(4)}(x) + \frac{h^6u^6}{6!} f^{(6)}(\zeta)\right] dx\, du\\ &= \frac{h^4}{24}\mu_4\left(\tilde{\Psi}_3\right) R(f'') + O\left[h^6\mu_6\left(\tilde{\Psi}_3\right)\right] = 6\mu_2(K)^2 R(f'')h + O\left(h^3\right)\end{aligned}$$

and

$$\begin{aligned} I_3 &= \iint f(x) \left[\int \frac{1}{h} \tilde{\Psi}_3 \left(\frac{x-y}{h} \right) f(y)\, dy \right]^2 dx \\ &= \int f(x) \left[6\mu_2(K)^2 f^{(4)}(x) h + O\left(h^3\right) \right]^2 dx \\ &= 36\mu_2(K)^4 \int f^{(4)}(x)^2 f(x)\, dx\, h^2 + O\left(h^4\right). \end{aligned}$$

Therefore,

$$\operatorname{var}\left[\Psi_3(X_1 - X_2)\right] = O\left(\frac{1}{h^7}\right)$$

and

$$\operatorname{cov}\left[\Psi_3(X_1 - X_2), \Psi_3(X_1 - X_3)\right] = \mathcal{L}h^2 + O\left(h^4\right),$$

where

$$\mathcal{L} = 36\mu_2(K)^4 \left[\int f^{(4)}(x)^2 f(x)\, dx - R(f'')^2 \right].$$

Consequently,

$$\operatorname{var}\left[CV'''(h)\right] = O\left(\frac{1}{n^2 h^7} + \frac{h^2}{n}\right) \tag{A.3}$$

and

$$\operatorname{var}\left[CV'''(h_{n0})\right] = O\left(n^{-3/5}\right).$$

Therefore, as required,

$$\operatorname{var}\left[CV'''(h_{n0})\right] = o(n^{-2/5})$$

and so $\alpha_1 = o_p(1)$.

To handle the term α_2 in (A.1), we write

$$\alpha_2 = n^{1/5}\left[CV'''(\tilde{h}_n) - CV'''(h_{n0})\right] = n^{1/5}(\tilde{h}_n - h_{n0})CV^{(4)}(\overline{h}_n), \tag{A.4}$$

where $\overline{h}_n$ is an intermediate value between $\tilde{h}_n$ and h_{n0}. The results of Hall and Marron (1987) imply that $\tilde{h}_n - h_{n0} = O_p\left(n^{-3/10}\right)$. Thus, in view of (A.4), to prove $\alpha_2 = o_P(1)$ it is sufficient to show that

$$n^{-1/10} \sup_{h \in I(h_n, h_{n0})} |CV^{4)}(h)| = o_p(1),$$

where $I(h_n, h_{n0})$ is the interval with endpoints h_n and h_{n0}.

Let a be arbitrarily small but fixed, and such that $an^{-1/5} < h_{n0} < a^{-1}n^{-1/5}$. Without loss of generality, we suppose that $CV(h)$ is minimized over a finite set I_n having equally spaced points on the interval $(an^{-1/5}, a^{-1}n^{-1/5})$. It is assumed that the number of points in I_n is $n^{2/5-d}$, where $0 < d < 1/5$. Let h_n^* be the minimizer of $M_n(h)$ over I_n. Then optimizing CV over I_n suffices since $h_n^* - h_{n0}$ is of order $n^{-3/5+d}$, implying that this source of error is smaller than $n^{-2/5}$ and hence negligible for the current argument. It is enough to show that $n^{-1/10}\max_{h\in I_n}|CV^{4)}(h)|$ converges in probability to 0. Since

$$|CV^{4)}(h)| \leq |CV^{4)}(h) - E_n(h)| + |E_n(h)|,$$

where $E_n(h) = E\left[CV^{4)}(h)\right]$, it suffices to show that

$$\lim_{n\to\infty} n^{-1/10} \max_{h\in I_n} |E_n(h)| = 0$$

and

$$n^{-1/10} \max_{h\in I_n} |CV^{4)}(h) - E_n(h)| = o_p(1).$$

For any $\epsilon > 0$, we have

$$\begin{aligned} P\left[n^{-1/10}\max_{h\in I_n}|CV^{4)}(h) - E_n(h)| \geq \epsilon\right] &\leq P\left[\bigcup_{h\in I_n}\left\{n^{-1/10}|CV^{4)}(h) - E_n(h)| \geq \epsilon\right\}\right] \\ &\leq \sum_{h\in I_n} P\left[n^{-1/10}|CV^{4)}(h) - E_n(h)| \geq \epsilon\right] \\ &\leq \sum_{h\in I_n}\frac{\operatorname{var}\left[CV^{4)}(h)\right]}{n^{1/5}\epsilon^2} \\ &\leq \frac{n^{1/5-d}}{\epsilon^2}\max_{h\in I_n}\operatorname{var}\left[CV^{4)}(h)\right]. \end{aligned}$$

Let us now obtain uniform bounds for the expectation and variance of $CV^{4)}(h)$. It is straightforward to prove that

$$E_n(h) = M^{4)}(h) \sim 6\mu_2(K)^2R(f'') + 24R(K)n^{-1}h^{-5}$$

and, since $h_{n0} \sim h_{na} = C_0n^{-1/5}$, we have that $E_n(h_{n0}) \sim \mathcal{D}$, for some constant $\mathcal{D} > 0$. On the other hand, since $I_n \subset [an^{-1/5}, a^{-1}n^{-1/5}]$, we get

$$\max_{h\in I_n} E\left[CV^{(4)}(h)\right] = O(1). \tag{A.5}$$

To obtain a uniform bound for the variance, long and tedious calculations can be performed to get a similar expression to (A.3), but for the fourth derivative:

$$\operatorname{var}\left[CV^{4)}(h)\right] = O\left(\frac{1}{n^2h^9}\right).$$

Using again $h_{n0} \sim C_0n^{-1/5}$ and $I_n \subset [an^{-1/5}, a^{-1}n^{-1/5}]$, we obtain

$$\max_{h\in I_n}\operatorname{var}\left[CV^{4)}(h)\right] = O\left(n^{-1/5}\right). \tag{A.6}$$

Using expressions (A.5) and (A.6), it now follows that

$$\max_{h\in I_n} n^{-1/10}|CV^{4)}(h)| = o_p(1),$$

thus completing the proof of Lemma A.1.

Proof of Theorem 3.1 The variance of the bagged bandwidth is:

$$\text{var}\left[\hat{h}(r,N)\right] = \frac{1}{N}\text{var}\left(\tilde{h}_{r,1}\right) + \frac{N-1}{N}\text{cov}\left(\tilde{h}_{r,1}, \tilde{h}_{r,2}\right). \tag{A.7}$$

The work of Hall and Marron (1987) provides an approximation to the variance of $\tilde{h}_{r,1}$:

$$\frac{\text{var}\left(\tilde{h}_{r,1}\right)}{h_{n0}^2} = A_0 r^{-1/5} + o\left(r^{-1/5}\right), \tag{A.8}$$

where

$$A_0 = \frac{8R(\mathcal{V})R(f)\mu_2(K)^{4/5}}{25R(K)^{9/5}R(f'')}, \tag{A.9}$$

the function $\mathcal{V}$ is defined in Bhattacharya and Hart (2016) and only depends on the kernel K. Bhattacharya and Hart (2016) derived the following approximation to the second term in (A.7):

$$\text{cov}\left(\tilde{h}_{r,1}, \tilde{h}_{r,2}\right) = \text{var}\left(\tilde{h}_{r,1}\right)\left(\frac{r}{n}\right)^2 + o\left(r^{9/5}n^{-12/5}\right). \tag{A.10}$$

Plugging (A.8) and (A.10) into (A.7), when N is either fixed or tending to ∞ with n, then,

$$\text{var}\left[\hat{h}(r,N)\right] \sim A_0 C_0^2 r^{-1/5} n^{-2/5}\left[\frac{1}{N} + \left(\frac{r}{n}\right)^2\right].$$

Regarding the bias of $\hat{h}(r,N)$, as explained in Subsection 3.1.1, we only have to focus on deriving the bias inherent to cross-validation itself. Let $\hat{h}_n$ be the ordinary cross-validation bandwidth for a sample of size n. Using the fact that $CV'(\hat{h}_n) = 0$, a Taylor expansion gives

$$\hat{h}_n - h_{n0} = -\frac{CV'(h_{n0})}{CV''(h_n)},$$

for h_n between $\hat{h}_n$ and h_{n0}. Now expand $1/CV''(h_n)$ in a Taylor series about $\Delta = M_n''(h_{n0})$, yielding

$$\hat{h}_n - h_{n0} = -\frac{CV'(h_{n0})}{\Delta} + \frac{CV'(h_{n0})[CV''(h_n) - \Delta]}{\hat{\Delta}^2},$$

where $\hat{\Delta}$ is between $CV''(h_n)$ and $M_n''(h_{n0})$. Using the notation in (3.2), $\xi_n = -CV'(h_{n0})/\Delta$ and

$$e_n = \frac{CV'(h_{n0})[CV''(h_n) - \Delta]}{\hat{\Delta}^2}.$$

The random variable $-CV'(h_{n0})/\Delta$ has mean 0 and is $O_p\left(n^{-3/10}\right)$, as shown by Hall and Marron (1987). We will show that $n^{2/5}e_n \xrightarrow{d} Y$, where $\mathrm{E}(Y) = \mu_{\mathrm{CV}} < 0$, with μ_{CV} given in (3.3), and $\mathrm{var}(Y) > 0$. Indeed, this will establish the first order bias of $\hat{h}_n$ as an estimator of h_{n0}. Using results of Hall and Marron (1987), $n^{4/5}\hat{\Delta}^2 \xrightarrow{p} \mathcal{E}^2 > 0$, where $\mathcal{E}$ is the limit of $n^{2/5}M_n''(h_{n0})$ as $n \to \infty$. It is sufficient then to consider

$$n^{6/5}CV'(h_{n0})[CV''(h_n) - \Delta] = n^{6/5}CV'(h_{n0})[CV''(h_{n0}) - \Delta + \delta_n], \qquad \text{(A.11)}$$

where $\delta_n = CV''(h_n) - CV''(h_{n0})$. Now,

$$\delta_n = (h_n - h_{n0})CV'''\left(\tilde{h}_n\right),$$

where $\tilde{h}_n$ is between h_n and h_{n0}. From Hall and Marron (1987), we know that $CV'(h_{n0}) = O_p\left(n^{-7/10}\right)$ and $h_n - h_{n0} = O_p\left(n^{-3/10}\right)$. It follows that

$$n^{6/5}CV'(h_{n0})\delta_n = O_p(1)n^{1/5}CV'''\left(\tilde{h}_n\right).$$

Considering Lemma A.1, in equation (A.11) we just need to investigate

$$n^{6/5}CV'(h_{n0})\left[CV'(h_{n0}) - \Delta\right].$$

Hall and Marron (1987) showed that

$$n^{7/10} CV'(h_{n0}) \xrightarrow{d} \mathrm{N}(0, \sigma_1^2).$$

As shown in Bhattacharya and Hart (2016), $h_{n0}\left[CV''(h_{n0}) - \Delta\right]$ is identical in structure to $CV'(h_{n0})$, and hence

$$n^{7/10} h_{n0}\left[CV''(h_{n0}) - \Delta\right] \sim C_0 \sqrt{n}\left[CV''(h_{n0}) - \Delta\right] \xrightarrow{d} \mathrm{N}(0, \sigma_2^2).$$

Using the Cramér-Wold device (Cramér and Wold, 1936), it follows that

$$\sqrt{n}\left[n^{1/5} CV'(h_{n0}), CV''(h_{n0}) - \Delta\right]$$

converges in distribution to a bivariate normal random variable (Y_1, Y_2) with mean vector 0 and some covariance matrix Σ, not defined here for the sake of brevity. Using Theorem B., p. 124 of Serfling (1980), we have

$$n^{6/5} CV'(h_{n0})\left[CV''(h_{n0}) - \Delta\right] \xrightarrow{d} Y_1 Y_2,$$

where (Y_1, Y_2) are bivariate normal with mean vector 0 and covariance matrix Σ. Bhattacharya and Hart (2016) show that $\mathrm{E}(Y_1 Y_2)$ is

$$-\frac{8R(f)R(f'')^{4/5}}{R(K)^{4/5}} \int \mathcal{V}(u) W(u)\, du,$$

where

$$\begin{aligned}
\mathcal{V}(u) &= K * K(u) - K * L(u) - K(u) + L(u), \\
W(u) &= 3K * K(u) + L * L(u) - 5K * L(u) + K * H(u) - 2K(u) + 3L(u) - H(u), \\
L(u) &= -uK'(u) \\
&\text{and} \\
H(u) &= -uL'(u).
\end{aligned}$$

Also, taking into account that (Bhattacharya and Hart, 2016)

$$M_n''(h_{n0}) \sim 5R(K)^{2/5}R(f'')^{3/5}n^{-2/5},$$

the limiting expectation of $n^{2/5}(\hat{h}_n - h_{n0})$ is

$$\frac{\mathrm{E}(Y_1Y_2)}{\mathcal{E}^2} = \mu_{\mathrm{CV}} = -\frac{8R(f)\int \mathcal{V}(u)W(u)\,du}{25R(K)^{8/5}R(f'')^{2/5}},$$

which completes the proof.

Proof of Theorem 3.2 The asymptotic normality of the statistic of interest, namely $n^{3/10}\left[\hat{h}(r,N) - h_{n0}\right]$, can be derived from the method of proof of Theorem 3.1. Furthermore, the mean and the variance of the asymptotic distribution of this statistic are an immediate consequence of Theorem 3.1.

Proof of Proposition 3.1 Multiplying the two equations in (3.11) by $r_1^{-3/5}$ and subtracting them we get

$$h_{r_1 1}r_2^{-3/5} - h_{r_2 1}r_1^{-3/5} = C_0\left(r_1^{-1/5}r_2^{-3/5} - r_1^{-3/5}r_2^{-1/5}\right)$$

and then

$$C_0 = \frac{h_{r_1 1}r_1^{3/5} - h_{r_2 1}r_2^{3/5}}{r_1^{2/5} - r_2^{2/5}}.$$

Similarly, multiplying the two equations in (3.11) by $r_2^{-1/5}$ and substracting them yields

$$h_{r_1 1}r_2^{-1/5} - h_{r_2 1}r_1^{-1/5} = C_1\left(r_1^{-3/5}r_2^{-1/5} - r_1^{-1/5}r_2^{-3/5}\right)$$

and so

$$C_1 = \frac{h_{r_1 1}r_1^{1/5} - h_{r_2 1}r_2^{1/5}}{r_1^{-2/5} - r_2^{-2/5}}.$$

Appendix B

Corrigendum to Theorem 1 of Hall and Robinson (2009)

In this Appendix we show that the approximation for the variance of the bagging bandwidth studied in Hall and Robinson (2009) is in error. A correct expression for this variance and the corresponding proof is provided.

The bagged bandwidth studied in Hall and Robinson (2009), $\hat{h}_{bagg}$, corresponds to the case where $N = \infty$, that is, $\hat{h}_{bagg} = \hat{h}(r, \infty)$, following the notation adopted in Section (3.1). From (3.5) it follows that

$$\text{var}\left(\hat{h}_{bagg}\right) = A_0 C_0^2 r^{9/5} n^{-12/5} + o\left(r^{9/5} n^{-12/5}\right), \tag{B.1}$$

which exactly matches the second term given in equation (13) of Hall and Robinson (2009). However, it is claimed in that paper that the dominant term would be of order $r^{4/5}n^{-7/5}$. We will prove that this last statement is wrong and that in fact the dominant term is precisely the one given in (B.1).

It can be easily proven that, for a sample of size n, we have

$$CV(h) = M_n(h) - R(f) + S(h), \tag{B.2}$$

where $S(h) = S_1(h) + S_2(h)$. Furthermore, using the same notation as in Hall and

Robinson (2009):

$$\begin{aligned} S_1(h) &= \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{1-n^{-1}}{h^2}\int \delta_i\mu - \frac{1}{h}\left[\mu(X_i) + \int K_i f - 2\int \mu f\right]\right\}, \\ S_2(h) &= \frac{2}{n(n-1)}\sum\sum_{1\le i<j\le n}\left(\frac{1-n^{-1}}{h^2}\int \delta_i\delta_j - \frac{2}{h}\delta_{ij}\right), \end{aligned}$$

where

$$\begin{aligned} K_i(x) &= K\left(\frac{x-X_i}{h}\right), \\ \mu(x) &= \mathrm{E}\left[K_i(x)\right], \\ \delta_i(x) &= K_i(x) - \mu(x), \\ \delta_{ij} &= K_i(X_j) - \mu(X_j) - \int K_i f + \int \mu f. \end{aligned}$$

From (B.2) it follows that, for any $k \in \mathbb{N}$,

$$\mathrm{var}\left[CV^{k)}(h)\right] = \mathrm{var}\left[S^{k)}(h)\right].$$

More importantly, finding the variance of the cross-validation bandwidth, whether bagged or ordinary, boils down to finding $\mathrm{var}\left[S'(h)\right]$. From (A.2) it follows that

$$\mathrm{var}\left[S'(h)\right] = \frac{1}{n^4h^2}\mathrm{var}\left[\sum_{i\neq j} H(X_i - X_j)\right], \tag{B.3}$$

where $H(u) = \gamma_{e,h}(u) + u\left(\gamma_{e,h}\right)'(u)$, $\gamma_{e,h}(u) = \gamma_e(u/h)/h$ and $\gamma_e(u) = \frac{n}{n-1}\gamma_n(u)$. Let us now define $\tilde{H}(u) = \gamma_e(u) + u\gamma_e'(u)$, so we have that $H(u) = \tilde{H}_h(u)$. Collecting similar cases and doing straightforward calculations we get

$$\mathrm{var}\left[\sum_{i\neq j} H(X_i - X_j)\right] = 4n(n-1)(n-2)C_b + 2n(n-1)C_c, \tag{B.4}$$

where $C_b = \mathrm{cov}\left[H(X_1 - X_2), H(X_1 - X_3)\right]$ and $C_c = \mathrm{var}\left[H(X_1 - X_2)\right]$. These terms

can be further decomposed into

$$C_b = C_{b1} - C_{b2}^2 \tag{B.5}$$

and

$$C_c = C_{c1} - C_{b2}^2, \tag{B.6}$$

where

$$C_{b1} = \int H * f(x)^2 f(x)\, dx,$$
$$C_{c1} = \int H^2 * f(x) f(x)\, dx$$

and

$$C_{b2} = \int H * f(x) f(x)\, dx.$$

Using the fact that $\tilde{H}$ is symmetric, $\mu_0\left(\tilde{H}\right) = 0$, $\mu_2\left(\tilde{H}\right) = 4\mu_2(K)/(n-1)$ and $\mu_4\left(\tilde{H}\right), \mu_6\left(\tilde{H}\right) = O(1)$, we have

$$\begin{aligned}
C_{b2} &= \iint \frac{1}{h}\tilde{H}\left(\frac{x-y}{h}\right) f(y) f(x)\, dx\, dy \\
&= \iint \tilde{H}(u) f(x-hu) f(x)\, dx\, du \\
&= \iint \tilde{H}(u)\left[f(x) - huf'(x) + \cdots - \frac{h^5u^5}{5!}f^{(5)}(x) + \frac{h^6u^6}{6!}f^{(6)}(\bar{x})\right] f(x)\, dx\, du \\
&= \int f(x)\left[\frac{h^2}{2}\mu_2\left(\tilde{H}\right) f''(x) + \frac{h^4}{4!}\mu_4\left(\tilde{H}\right) f^{(4)}(x) + O\left(h^6\right)\right] dx \\
&= O\left(\frac{h^2}{n} + h^4\right),
\end{aligned}$$

and, therefore,

$$C_{b2}^2 = O\left(\frac{h^4}{n^2} + h^8 + \frac{h^6}{n}\right). \tag{B.7}$$

For the term C_{b1},

$$\begin{aligned} C_{b1} &= \int f(x)\left[\int \frac{1}{h}\tilde{H}\left(\frac{x-y}{h}\right) f(y)\,dy\right]^2 dx \\ &= \int f(x)\left[\int \tilde{H}(u) f(x-hu)\,du\right]^2 dx \\ &= \int f(x)\left[\frac{1}{4}\mu_2(K)^2 f^{(4)}(x)h^4 + O\left(h^6\right)\right]^2 dx \\ &= \int f(x)\left[\frac{1}{16}\mu_2(K)^4 f^{(4)}(x)^2 h^8 + O\left(h^{10}\right)\right] dx \\ &= O\left(\frac{h^4}{n^2} + h^8 + \frac{h^6}{n}\right). \end{aligned} \tag{B.8}$$

The term C_{c1} can be handled in a similar way,

$$\begin{aligned} C_{c1} &= \frac{1}{h^2}\iint \tilde{H}\left(\frac{x-y}{h}\right)^2 f(y) f(x)\,dx\,dy \\ &= \frac{1}{h}\iint \tilde{H}(u)^2 f(x-hu) f(x)\,dx\,du \\ &= \frac{1}{h}\iint \tilde{H}(u)^2 f(x)\left[f(x) - huf'(x) + \frac{h^2u^2}{2}f''(\tilde{x})\right] dx\,du \\ &= \frac{R(f)R\left(\tilde{H}\right)}{h} + O\left(h\right), \end{aligned} \tag{B.9}$$

where we have used the fact that $\mu_1\left(\tilde{H}^2\right) = O\left(1\right)$.

Plugging (B.7), (B.8) and (B.9) into (B.5) and (B.6) yields, respectively,

$$C_b = O\left(\frac{h^4}{n^2} + h^8 + \frac{h^6}{n}\right), \tag{B.10}$$

$$C_c = \frac{R(f)R\left(\tilde{H}\right)}{h} + O\left(h\right). \tag{B.11}$$

Now, plugging (B.10) and (B.11) into (B.4) and then into (B.3), and using the fact that

$$2R(f)R\left(\tilde{H}\right) = A_3 + O\left(n^{-1}\right),$$

we get

$$\text{var}\left[S'(h)\right] = A_3\frac{1}{n^2h^3} + O\left(\frac{1}{n^2h}\right), \tag{B.12}$$

where A_3 is defined on p. 183 of Hall and Robinson (2009). Equation (B.12) is completely consistent with the results obtained in Hall and Marron (1987) and Scott and Terrell (1987). Now, taking variance in equation (A2) of Hall and Robinson (2009) and plugging (B.12) into that expression yields (B.1). Equation (B.12) is enough to show that expression (A3) of Hall and Robinson (2009) is wrong, which in turn explains the error in their equation (13) regarding the variance of the bagged bandwidth. Nonetheless, we will provide an asymptotic expression for $\text{var}\left[S_1'(h)\right]$, since that is where the error in Hall and Robinson (2009) comes from.

From the definition of $V_{nh}(X_i)$ and $S_1(h)$ given on p. 184 of Hall and Robinson (2009), it is easy to show that

$$V_{nh}(X_1) = \left(1 - n^{-1}\right)\tilde{z}_1^{(h)} - \tilde{T}_1^{(h)},$$

where

$$\tilde{z}_1^{(h)} = K_h * K_h * f(X_1) - \int K_h * f(x)^2\,dx$$

and

$$\tilde{T}_1^{(h)} = 2\left[K_h * f(X_1) - \int K_h * f(x)f(x)\,dx\right].$$

Let us define the functions ν and η:

$$\nu(x) = K(x) + xK(x)$$

and

$$\eta(x) = K * K(x) + x(K * K)'(x).$$

Then, we have that

$$\frac{d}{dh}\tilde{T}_1^{(h)} = -\frac{1}{h}\nu_h * f(X_1)$$

and

$$\frac{d}{dh}\tilde{z}_1^{(h)} = -\frac{1}{h}\left\{\eta_h * f(X_1) - \mathrm{E}\left[\eta_h * f(X_1)\right]\right\}.$$

Therefore,

$$\frac{d}{dh}V_{nh}(X_1) = \frac{1}{h}\left\{\tau_h * f(X_1) - \mathrm{E}\left[\tau_h * f(X_1)\right]\right\},$$

where

$$\tau(x) = 2K(x) + 2xK'(x) - \frac{n-1}{n}\left[K * K(x) + x(K * K)'(x)\right].$$

We have that

$$\mathrm{var}\left[\frac{d}{dh}V_{nh}(X_1)\right] = \frac{1}{h}\left\{\mathrm{E}\left[\tau_h * f(X_1)^2\right] - \mathrm{E}\left[\tau_h * f(X_1)\right]^2\right\}.$$

It is easy to show that

$$\begin{aligned}
\mu_0(\tau) &= 0, \\
\mu_2(\tau) &= -\frac{4}{n}\mu_2(K), \\
\mu_4(\tau) &= -\frac{8}{n}\mu_4(K) + 24\frac{n-1}{n}\mu_2(K)^2, \\
\mu_6(\tau) &= -\frac{12}{n}\mu_6(K) + 180\frac{n-1}{n}\mu_2(K)\mu_4(K).
\end{aligned}$$

Using standard calculations, it is easy to prove that

$$\begin{aligned}
\mathrm{E}\left[\tau_h * f(X_1)\right] &= \iint \tau(u) f(x) f(x-hu)\, dx\, du \\
&= \iint \tau(u) f(x) \left[f(x) - huf'(x) + \cdots - \frac{h^7 u^7}{7!} f^{(7)}(x) \right. \\
&+ \left. \frac{h^8 u^8}{8!} f^{(8)}(\tilde{x}) \right] dx\, du \\
&= -\frac{h^2}{2}\mu_2(\tau) R(f') + \frac{h^4}{24}\mu_4(\tau) R(f'') - \frac{h^6}{6!}\mu_6(\tau) R(f''') + O\left(h^8\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}\left[\tau_h * f(X_1)\right]^2 &= \frac{h^4}{4}\mu_2(\tau)^2 R(f')^2 - \frac{h^6}{24}\mu_2(\tau)\mu_4(\tau) R(f') R(f'') \\
&+ \frac{h^8}{24^2}\mu_4(\tau)^2 R(f'')^2 + \frac{h^8}{6!}\mu_2(\tau)\mu_6(\tau) R(f') R(f''') + O\left(h^{10}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathrm{E}\left[\tau_h * f(X_1)^2\right] &= \iiint \tau(u) f(x-hu) \tau(v) f(x-hv) f(x)\, dx\, du\, dv \\
&= \iiint \tau(u)\tau(v) f(x) \left[f(x) - huf'(x) + \cdots + O\left(h^{10}\right) \right] \\
&\quad \left[f(x) - hvf'(x) + \cdots + O\left(h^{10}\right) \right] dx\, du\, dv \\
&= \frac{h^4}{4}\mu_2(\tau)^2 J_2 + \frac{h^6}{24}\mu_2(\tau)\mu_4(\tau) J_3 + \frac{h^8}{6!}\mu_2(\tau)\mu_6(\tau) J_4 + \frac{h^8}{24^2}\mu_4(\tau)^2 J_1 \\
&+ O\left(h^{10}\right),
\end{aligned}$$

where

$$\begin{aligned}
J_1 &= \int f^{(4)}(x)^2 f(x)\, dx, \\
J_2 &= \int f(x) f''(x)^2\, dx, \\
J_3 &= \int f(x) f''(x) f^{(4)}(x)\, dx, \\
J_4 &= \int f(x) f''(x) f^{(6)}(x)\, dx.
\end{aligned}$$

So, we have that

$$\begin{aligned}\mathrm{var}\left[\frac{d}{dh}V_{nh}(X_1)\right] &= \frac{h^2}{4}\mu_2(\tau)^2\left[J_2 - R(f')^2\right] + \frac{h^4}{24}\mu_2(\tau)\mu_4(\tau)\left[J_3 + R(f')R(f'')\right] \\ &+ \frac{h^6}{6!}\mu_2(\tau)\mu_6(\tau)\left[J_4 - R(f')R(f''')\right] + \frac{h^6}{24^2}\mu_4(\tau)^2\left[J_1 - R(f'')^2\right] \\ &+ O\left(h^8\right).\end{aligned}$$

Finally, since

$$\mathrm{var}\left[S_1'(h)\right] = \frac{4}{n}\mathrm{var}\left[\frac{d}{dh}V_{nh}(X_1)\right],$$

it follows that

$$\mathrm{var}\left[S_1'(h)\right] = 4\mu_2(K)^4\left[J_1 - R(f'')^2\right]\frac{h^6}{n} + O\left(\frac{h^8}{n}\right).$$

This, in conjunction with (B.12), proves that $\mathrm{var}\left[S_1'(h)\right]$ is negligible with respect to $\mathrm{var}\left[S_2'(h)\right]$ and, in particular, that $\mathrm{var}\left[S_1'(h)\right]$ cannot be asymptotic to $A_2\,h^2/n$ as claimed in Hall and Robinson (2009).

Appendix C

Proofs of the results of Chapter 4

Proof of Lemma 4.1 Recall that

$$\tilde{m}_h(x) = A + B + C + D,$$

where A, B, C and D were defined in Section 4.1. Let us start by defining

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

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

Therefore,

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

Also,

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

and hence

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

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

$$\begin{aligned}
\mathrm{E}(\hat{a}\hat{e}) &= \mathrm{E}\left[n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i K_h(x-X_1)K_h(x-X_j)\right] = n^{-2}\left\{n\mathrm{E}\left[Y_1 K_h(x-X_1)^2\right]\right. \\
&+ \left. n(n-1)\mathrm{E}\left[Y_1 K_h(x-X_1)K_h(x-X_2)\right]\right\} \\
&= n^{-1}\mathrm{E}\left[Y_1 K_h(x-X_1)^2\right] + \frac{n-1}{n}\mathrm{E}\left[Y_1 K_h(x-X_1)\right]\mathrm{E}\left[K_h(x-X_1)\right] \\
&= R(K)m(x)f(x)n^{-1}h^{-1} + m(x)f(x)^2 \\
&+ h^2\mu_2(K)\left[\frac{1}{2}f''(x)m(x)f(x) + f(x)^2\varphi_1(x)\right] + h^4\left[\frac{1}{24}\mu_4(K)f^{4)}(x)m(x)f(x)\right. \\
&+ \left.\mu_4(K)f(x)^2\varphi_2(x) + \frac{1}{2}\mu_2(K)^2 f''(x)f(x)\varphi_1(x)\right] + O(h^6+n^{-1}).
\end{aligned}$$

Therefore,

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

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,

$$\begin{aligned} \mathrm{E}(D) &= R(K)m(x)f(x)^{-1}n^{-1}h^{-1} + \frac{1}{4}h^4\mu_2(K)^2 f''(x)^2 m(x)f(x)^{-2} \\ &+ O\left(h^6 + n^{-1}\right). \end{aligned} \tag{C.4}$$

Adding (C.1), (C.2), (C.3) and (C.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

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

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,

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

and

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

On the other hand,

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

and so

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

Straightforward calculations lead to

$$\begin{aligned}
\text{cov}(\hat{a},\hat{e}) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}\left[Y_i K_h(x-X_i), K_h(x-X_j)\right] \\
&= n^{-1}\text{cov}\left[Y_1 K_h(x-X_1), K_h(x-X_1)\right] \\
&= n^{-1}\left\{\text{E}\left[Y_1 K_h(x-X_1)^2\right] - \text{E}\left[Y_1 K_h(x-X_1)\right]\text{E}\left[K_h(x-X_1)\right]\right\} \\
&= R(K) m(x) f(x) n^{-1}h^{-1} - m(x) f(x)^2 n^{-1} \\
&+ \mu_2(K^2)\varphi_1(x) f(x) n^{-1}h + O\left(n^{-1}h^2\right)
\end{aligned}$$

and hence

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

Now,

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

and

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

Therefore,

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

and

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

Some auxiliary covariances are needed:

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

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

Hence,

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

and

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

Another covariance term is needed:

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

Therefore,

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

and

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

The following covariance is also needed:

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

Hence,

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

and

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

Hence,

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

and

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

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

are $O\left(n^{-1}h^2 + n^{-2}h^{-2} + n^{-3}h^{-3}\right)$.

Therefore, plugging (C.6)-(C.13) into (C.5) yields

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

Proof of Lemma 4.2 For the sake of simplicity, we will denote by "$Z(h,n) \stackrel{2}{=}$" the second order terms of a function $Z(h,n)$. For example, if $Z(h,n) = a_0 + a_1 h + a_2 h^3 + o(h^3)$, for some constants a_0, a_1 and a_2, then we would denote $Z(h,n) \stackrel{2}{=} a_1 h$.

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:

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

and so

$$\mathrm{E}\left[\widetilde{CV}'_n(h)\right] = \frac{2}{(n-1)^4 h}\mathrm{E}\left[\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\left(\Lambda_{11}^{j} + \Lambda_{12}^{jk}\right)\left(\Lambda_{21}^{l} + \Lambda_{22}^{ls}\right)\right], \quad \text{(C.14)}$$

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

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{21}^{l}\right) &= (n-1)^2\left[(n-1)\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{2}\right)\right. \\
&+ \left.(n-1)(n-2)\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{3}\right)\right]. \qquad \text{(C.15)}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^{2}\Lambda_{21}^{2}\right) &= -4\mathrm{E}\left\{f(X_1)^{-2}K_h\left(X_1 - X_2\right)\alpha_{1h}\left(X_1 - X_2\right)\left[Y_2 - m(X_1)\right]^2\right\} \\
&= -4\mathrm{E}\left(f(X_1)^{-2}K_h\left(X_1 - X_2\right)\alpha_{1h}\left(X_1 - X_2\right)\left\{\sigma^2(X_2)\right.\right. \\
&+ \left.\left.\left[m(X_2) - m(X_1)\right]^2\right\}\right) \\
&= -4h^{-1}\iint f(x_1)^{-1}K(u)\alpha_1(u)\left\{\sigma^2(x_1 - hu)\right. \\
&+ \left.\left[m(x_1 - hu) - m(x_1)\right]^2\right\}f(x_1 - hu)\,dx_1du \\
&\overset{2}{=} -4h^{-1}\iint f(x_1)^{-1}K(u)\alpha_1(u)h^2u^2f(x_1)\varphi_4(x_1)\,dx_1du \\
&= 2\mu_2\left(K^2\right)h\int\varphi_4 \qquad \text{(C.16)}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^2\Lambda_{21}^3\right) &= -4\mathrm{E}\left\{f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_3\right)\left[Y_2-m(X_1)\right]\right. \\
&\quad \left.\left[Y_3-m(X_1)\right]\right\} \\
&= -4\mathrm{E}\left\{f(X_1)^{-2}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_3\right)\left[m(X_2)-m(X_1)\right]\right. \\
&\quad \left.\left[m(X_3)-m(X_1)\right]\right\} \\
&= -4\iiint f(x_1)^{-1}K(u)\alpha_1(v)\left[m(x_1-hu)-m(x_1)\right] \\
&\quad \left[m(x_1-hv)-m(x_1)\right]f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} -4\iiint f(x_1)^{-1}K(u)\alpha_1(v)h^6\left(u^2v^4+u^4v^2\right)f(x_1)^2\varphi_6(x_1)\varphi_7(x_1) \\
&\quad dx_1dudv \\
&= 24\mu_2(K)\mu_4(K)h^6\int\varphi_6\varphi_7 f, \qquad \text{(C.17)}
\end{aligned}$$

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 (C.16) and (C.17) into (C.15) we get:

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^n\sum_{k=2}^n\sum_{l=2}^n\sum_{s=2}^n\Lambda_{11}^j\Lambda_{21}^l\right) &\overset{2}{=} 2\mu_2\left(K^2\right)n^3h\int\varphi_4+24n^4h^6\int\varphi_6\varphi_7 f \\
&+ O\left(n^2h+n^3h^6\right). \qquad \text{(C.18)}
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{21}^{l}\right) &= (n-1)\left\{(n-1)(n-2)(n-3)\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{4}\right)\right. \\
&+ (n-1)(n-2)\left[\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{21}^{3}\right)+\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{2}\right)\right. \\
&+ \left.\left.\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{3}\right)\right]\right\}+o\left(n^3h+n^4h^6\right). \qquad \text{(C.19)}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{4}\right) &= 2\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\alpha_{1h}\left(X_1-X_4\right)\right. \\
&\quad \left.\left[Y_2-m(X_1)\right]\left[Y_4-m(X_1)\right]\right\} \\
&= 2\iiiint f(x_1)^{-2}K(u)K(v)\alpha_1(w)\left[m(x_1-hv)-m(x_1)\right] \\
&\quad \left[m(x_1-hw)-m(x_1)\right]f(x_1-hu)f(x_1-hv)f(x_1-hw) \\
&\quad dx_1dudvdw \\
&\overset{2}{=} 2\iiiint f(x_1)^{-2}K(u)K(v)\alpha_1(w)h^6\left[(w^4v^2+w^2v^4)f(x_1)^3\right. \\
&\quad \left.\varphi_6(x_1)\varphi_7(x_1)+w^2u^2v^2\frac{1}{2}f''(x_1)f(x_1)^2\varphi_7(x_1)^2\right]f(x_1-hu) \\
&\quad f(x_1-hv)f(x_1-hw)\,dx_1dudvdw \\
&= -2h^6\left[6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+\mu_2(K)^3\int\varphi_7^2f''\right], \qquad \text{(C.20)}
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{21}^{2}\right) &= 2\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_2\right)\alpha_{1h}\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\right. \\
&\quad \left.\left[Y_2-m(X_1)\right]^2\right\} \\
&= 2h^{-1}\iint f(x_1)^{-2}K(u)\alpha_1(u)K(v)\left\{\sigma^2(x_1-hu)\right. \\
&+ \left.\left[m(x_1-hu)-m(x_1)\right]^2\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} 2h^{-1}\iiint f(x_1)^{-2}K(u)\alpha_1(u)K(v)h^2\left[u^2f(x_1)^2\varphi_4(x_1)\right. \\
&+ \left. v^2\frac{1}{2}\sigma^2(x_1)f(x_1)f''(x_1)\right]dx_1dudv \\
&= h\left[\frac{1}{2}R(K)\mu_2(K)\int\sigma^2f''f^{-1}-\mu_2\left(K^2\right)\int\varphi_4\right] \qquad \text{(C.22)}
\end{aligned}$$

and

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

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 (C.20), (C.21), (C.22) and (C.23) into (C.19) we get

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{21}^{l}\right) &\stackrel{2}{=} -2n^4h^6\left[6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+\mu_2(K)^3\int\varphi_7^2 f''\right] \\
&+ n^3h\left[\frac{1}{2}R(K)\mu_2(K)\int\sigma^2 f''f^{-1}\right. \\
&- \left.\mu_2\left(K^2\right)\int\varphi_4\right]. \qquad \text{(C.24)}
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{22}^{ls}\right) &= (n-1)^2(n-2)(n-3)\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{34}\right) \\
&+ (n-1)^2(n-2)\left[\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{23}\right)+\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{32}\right)\right. \\
&+ \left.\mathrm{E}\left(\Lambda_{11}^3\Lambda_{22}^{22}\right)\right]+o\left(n^3h+n^4h^6\right). \qquad \text{(C.25)}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{34}\right) &= 2h^{-1}\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_3,X_1-X_4\right)\right. \\
&\quad \left.\left[Y_2-m(X_1)\right]\left[Y_3-m(X_1)\right]\right\} \\
&= 2\iiiint f(x_1)^{-2}K(u)\Gamma_1(v,w)\left[m(x_1-hu)-m(x_1)\right] \\
&\quad \left[m(x_1-hv)-m(x_1)\right]f(x_1-hu)f(x_1-hv)f(x_1-hw) \\
&\quad dx_1dudvdw \\
&\stackrel{2}{=} 2\iiiint f(x_1)^{-2}K(u)\Gamma_1(v,w)h^6\left[(u^2v^4+u^4v^2)f(x_1)^3\varphi_6(x_1)\varphi_7(x_1)\right. \\
&+ \left.w^2u^2v^2\frac{1}{2}f''(x_1)f(x_1)^2\varphi_7(x_1)^2\right]dx_1dudvdw \\
&= -h^6\left[12\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f+4\mu_2(K)^3\int\varphi_7^2 f''\right], \qquad \text{(C.26)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{23}\right) &= 2h^{-1}\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_2,X_1-X_4\right)\right. \\
&\quad \left.[Y_2-m(X_1)]^2\right\} \\
&= 2h^{-1}\iiint f(x_1)^{-2}K(u)\Gamma_1(u,v)\left\{\sigma^2(x_1-hu)\right. \\
&+ \left.[m(x_1-hu)-m(x_1)]^2\right\}f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} 2h^{-1}\iiint f(x_1)^{-2}K(u)\Gamma_1(u,v)h^2\left[u^2f(x_1)^2\varphi_4(x_1)\right. \\
&+ \left.v^2\frac{1}{2}\sigma^2(x_1)f''(x_1)f(x_1)\right]dx_1dudv \\
&= -h\left[\mu_2\left(K^2\right)\int\varphi_4+\frac{3}{2}R(K)\mu_2(K)\int\sigma^2f''f^{-1}\right], \qquad \text{(C.27)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^2\Lambda_{22}^{32}\right) &= 2h^{-1}\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_2\right)\Gamma_{1h}\left(X_1-X_3,X_1-X_2\right)\right. \\
&\quad \left.[Y_2-m(X_1)]\,[Y_3-m(X_1)]\right\} \\
&= 2h^{-1}f(x_1)^{-2}K(u)\Gamma_1(v,u)\,[m(x_1-hu)-m(x_1)] \\
&\quad [m(x_1-hv)-m(x_1)]\,f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} O\left(h^3\right) \qquad \text{(C.28)}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{11}^3\Lambda_{22}^{22}\right) &= 2h^{-1}\mathrm{E}\left\{f(X_1)^{-3}K_h\left(X_1-X_3\right)\Gamma_{1h}\left(X_1-X_2,X_1-X_2\right)\right. \\
&\quad \left.[Y_2-m(X_1)]\,[Y_3-m(X_1)]\right\} \\
&= 2h^{-1}f(x_1)^{-2}K(u)\Gamma_1(v,v)\,[m(x_1-hu)-m(x_1)] \\
&\quad [m(x_1-hv)-m(x_1)]\,f(x_1-hu)f(x_1-hv)\,dx_1dudv \\
&\overset{2}{=} O\left(h^3\right), \qquad \text{(C.29)}
\end{aligned}$$

where we have used the fact that

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

Then, plugging (C.26), (C.27), (C.28) and (C.29) into (C.25) we get

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{11}^{j}\Lambda_{22}^{ls}\right) &\overset{2}{=} -n^4h^6\left[12\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f\right. \\
&+ \left. 4\mu_2(K)^3\int\varphi_7^2 f''\right] - n^3h\left[\mu_2\left(K^2\right)\int\varphi_4\right. \\
&+ \left.\frac{3}{2}R(K)\mu_2(K)\int\sigma^2 f'' f^{-1}\right]. \qquad \text{(C.30)}
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{22}^{ls}\right) &= (n-1)(n-2)(n-3)(n-4)\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{45}\right) \\
&+ (n-1)(n-2)(n-3)\left[\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{22}^{34}\right)+\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{24}\right)\right. \\
&+ \mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{42}\right)+\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{24}\right)+\mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{42}\right) \\
&+ \left.\mathrm{E}\left(\Lambda_{12}^{34}\Lambda_{22}^{22}\right)\right] + o\left(n^3h+n^4h^6\right). \qquad \text{(C.31)}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{45}\right) &= -h^{-1}\mathrm{E}\left\{f(X_1)^{-4}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\right. \\
&\quad \left.\Gamma_{1h}\left(X_1-X_4,X_1-X_5\right)\left[Y_2-m(X_1)\right]\left[Y_4-m(X_1)\right]\right\} \\
&= -\int\cdots\int f(x_1)^{-3}K(u)K(v)\Gamma_1(w,z)\left[m(x_1-hu)-m(x_1)\right] \\
&\quad \left[m(x_1-hw)-m(x_1)\right]f(x_1-hu)f(x_1-hv)f(x_1-hw)f(x_1-hz) \\
&\quad dx_1dudvdwdz \\
&\overset{2}{=} -\int\cdots\int f(x_1)^{-3}K(u)K(v)\Gamma_1(w,z)h^6\left[(u^4w^2+u^2w^4)f(x_1)^4\right. \\
&\quad \left.\varphi_6(x_1)\varphi_7(x_1)+(u^2w^2v^2+u^2w^2z^2)\frac{1}{2}f(x_1)^3f''(x_1)\varphi_7(x_1)^2\right] \\
&\quad dx_1dudvdwdz \\
&= h^6\left[6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7f+3\mu_2(K)^3\int\varphi_7^2f''\right] \qquad \text{(C.32)}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{24}\right) &= -h^{-1}\mathrm{E}\left\{f(X_1)^{-4}K_h\left(X_1-X_2\right)K_h\left(X_1-X_3\right)\right. \\
&\quad \left.\Gamma_{1h}\left(X_1-X_2,X_1-X_4\right)\left[Y_2-m(X_1)\right]^2\right\} \\
&= -\iiiint f(x_1)^{-3}K(u)K(v)\Gamma_1(u,w)\left\{\sigma^2(x_1-hu)\right. \\
&+ \left.\left[m(x_1-hu)-m(x_1)\right]^2\right\}f(x_1-hu)f(x_1-hv)f(x_1-hw)f(x_1-hz) \\
&\quad dx_1dudvdw \\
&\overset{2}{=} -h^{-1}\iiiint f(x_1)^{-3}K(u)K(v)\Gamma_1(u,w)h^2\left[(v^2+w^2)\frac{1}{2}f(x_1)^2\sigma^2(x_1)\right. \\
&\quad \left.f''(x_1)+u^2f(x_1)^3\varphi_4(x_1)\right]\,dx_1dudvdw \\
&= \frac{1}{2}h\left[R(K)\mu_2(K)\int\sigma^2f''f^{-1}+\mu_2\left(K^2\right)\int\varphi_4\right], \qquad \text{(C.33)}
\end{aligned}$$

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

$$\mathrm{E}\left(\Lambda_{12}^{22}\Lambda_{22}^{34}\right), \mathrm{E}\left(\Lambda_{12}^{23}\Lambda_{22}^{42}\right), \mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{24}\right), \mathrm{E}\left(\Lambda_{12}^{32}\Lambda_{22}^{42}\right), \mathrm{E}\left(\Lambda_{12}^{34}\Lambda_{22}^{22}\right) \overset{2}{=} O\left(h^3\right). \quad \text{(C.34)}$$

Then, plugging (C.32), (C.33) and (C.34) into (C.31) we get

$$\begin{aligned} \mathrm{E}\left(\sum_{j=2}^{n}\sum_{k=2}^{n}\sum_{l=2}^{n}\sum_{s=2}^{n}\Lambda_{12}^{jk}\Lambda_{22}^{ls}\right) &\overset{2}{=} n^4h^6\left[6\mu_2(K)\mu_4(K)\int\varphi_6\varphi_7 f + 3\mu_2(K)^3\int\varphi_7^2 f''\right] \\ &+ \frac{1}{2}n^3h\left[R(K)\mu_2(K)\int\sigma^2 f''f^{-1}\right. \\ &+ \left.\mu_2\left(K^2\right)\int\varphi_4\right]. \end{aligned} \quad \text{(C.35)}$$

Finally, plugging (C.18), (C.24), (C.30), and (C.35) into (C.14) yields:

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

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 in Lemma 4.2. 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 $\tilde{m}_h$, defined in (4.8), we can employ the simpler, linear approximation of $\hat{m}_h$, denoted by $\bar{m}_h$ and defined in (4.9). This linear approximation of the Nadaraya-Watson estimator was already proposed in Barbeito (2020) and it can be expressed as

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

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,

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

where

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

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

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

Among the previous covariances, it can be argued that the only ones that contribute to the dominant term of var $\left[\widetilde{CV}'_n(h)\right]$ are C_{123145}, C_{123245}, C_{123425} and C_{123124}.

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

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

Then, using the law of total covariance:

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

Finally,

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

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

$$\begin{aligned} & \operatorname{cov}\left\{[Y_j - m(X_i)][Y_k - m(X_i)], [Y_r - m(X_l)]\right. \\ & \left.[Y_s - m(X_l)] \mid X_i, X_j, X_k, X_l, X_r, X_s\right\} \\ = \; & 0. \end{aligned}$$

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

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

Making the following changes of variable,

$$\begin{cases} x_4 = x_2 - h u_4 \\ x_3 = x_1 - h u_3 \\ x_2 = x_1 - h u_2 \end{cases}$$

and using the fact that

$$\begin{aligned}
&\iiint \alpha_1(u_4) \beta_1(u_2, u_3) u_4^i u_2^j u_3^k \, du_4 du_2 du_3 = ik \mu_i(K) \mu_j(K) \mu_k(K) = 0 \\
&\iff i = 0 \text{ or } k = 0 \text{ or } (i, j \text{ or } k \text{ is an odd number}),
\end{aligned}$$

we obtain that

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

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

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

Therefore,

$$\begin{aligned}
C_{123245} \; = \; & -4\mu_2(K)^2h^4\int f(x)^{-1}\sigma^2(x)\left[\frac{1}{4}m''(x)^2f(x)^2 + m'(x)^2f'(x)^2\right. \\
& + \left. m'(x)m''(x)f(x)f'(x)\right]\,dx + O\left(h^6\right). \qquad \text{(C.36)}
\end{aligned}$$

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

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

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,

$$\begin{aligned}
C_{123425} &= 8R(K)^2 \mu_2(K)^2 h^4 \int \sigma^2 f(x)^{-1} \left[\frac{1}{4}(m'')^2 f^2 + m' m'' f f' + (m')^2 (f')^2\right] \\
&+ O\left(h^6\right). \qquad \text{(C.37)}
\end{aligned}$$

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

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

We have that

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

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,

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

Therefore,

$$C_{123124} = \mu_2\left[(K')^2\right] h^{-1}\int (\sigma^2)^2 + O\left(h\right). \tag{C.38}$$

Using similar arguments and calculations we get

$$\begin{aligned} C_{123145} &= 4\mu_2(K)^2 h^4 \int f^{-1}\sigma^2 \left[\frac{1}{4}(m'')^2 f^2 + (m')^2 (f')^2 + m'm''ff'\right] \\ &+ O\left(h^6\right). \end{aligned} \tag{C.39}$$

Finally, considering (C.36), (C.37), (C.38) and (C.39) we obtain the main term of the variance of $\widetilde{CV}'_n(h)$ given in Lemma 4.2.

Proof of Theorem 4.1 From equation (4.11), it follows that, up to first order,

$$\mathrm{E}\left(\tilde{h}_{CV,n}\right) - \tilde{h}_{n0} = \frac{\tilde{M}'_n(\tilde{h}_{n0}) - \mathrm{E}\left[\widetilde{CV}'_n(\tilde{h}_{n0})\right]}{\tilde{M}''_n(\tilde{h}_{n0})}, \tag{C.40}$$

$$\mathrm{var}\left(\tilde{h}_{CV,n}\right) = \frac{\mathrm{var}\left[\widetilde{CV}'_n(\tilde{h}_{n0})\right]}{\tilde{M}''_n(\tilde{h}_{n0})^2}. \tag{C.41}$$

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 $\mathrm{var}\left[\widetilde{CV}'_n(\tilde{h}_{n0})\right]$. Therefore, to proof Theorem 4.1, we only have to plug the results of Lemma 4.1 and Lemma 4.2 into (C.40) and (C.41).

Proof of Corollary 4.1 Using the Cramér–Wold device (Cramér and Wold, 1936) and an argument similar to that followed in the proof of Theorem 3.1, the asymptotic normality of the statistic of interest, namely $n^{3/10}\left(\tilde{h}_{CV,n} - \tilde{h}_{n0}\right)$, can be derived. The expressions for the mean and the variance of the asymptotic distribution of this statistic are an immediate consequence of Theorem 4.1.

Sketch of the proof of Remark 4.1 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 in Section 4.1. From the proof of Lemma 4.1 and the fact that

$$\begin{aligned}\mathrm{E}\left[Y_1 K_h(x - X_1)^3\right] &= O\left(h^{-2}\right),\\ \mathrm{E}\left(\hat{a}\hat{e}^2\right) &= \mathrm{E}\left[n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} Y_i K_h(x - X_i) K_h(x - X_j) K_h(x - X_k)\right]\\ &= n^{-3}\left\{n\mathrm{E}\left[Y_1 K_h(x - X_1)^3\right]\right.\\ &+ n(n-1)\mathrm{E}\left[Y_1 K_h(x - X_1) K_h(x - X_2)^2\right]\\ &+ 2n(n-1)\mathrm{E}\left[Y_1 K_h(x - X_1)^2 K_h(x - X_2)\right]\\ &+ \left. n(n-1)(n-2)\mathrm{E}\left[Y_1 K_h(x - X_1) K_h(x - X_2) K_h(x - X_3)\right]\right\}\\ &= 3R(K)m(x)f(x)^2 n^{-1}h^{-1} + m(x)f(x)^3\\ &+ \left[\mu_2(K)m(x)f(x)^2 f''(x) + \mu_2(K)\varphi_1(x)f(x)^3\right]h^2\\ &+ \left[\frac{1}{4}\mu_2(K)^2 m(x)f(x)f''(x)^2 + \mu_4(K)f(x)^3\varphi_2(x)\right.\\ &+ \left.\mu_2(K)^2 f(x)^2\varphi_1(x)f''(x)\right]h^4 + O\left(h^6 + n^{-1}\right)\end{aligned}$$

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,

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

where we have used the fact that both $\mathrm{E}\left(E\right)$ and $\mathrm{E}\left(F\right)$ 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,

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

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

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

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 by $E_n(x)$ 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

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

and so

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

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

Since similar results can be obtained for $\mathrm{E}\left[F(X_1)^2\right]$ and $\mathrm{E}\left[E(X_1)F(X_1)\right]$ (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 $\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_h^{(-i)}(X_i) - \tilde{m}_h^{(-i)}(X_i)\right]^2$ 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,

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

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 similar 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, the main term of $\tilde{M}_n''(h^*)$, 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

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

Proof of Theorem 4.2 If we define

$$C_1 = -\frac{6B_2C_0^5 + V_2}{12B_1C_0^2 + 2V_1C_0^{-3}}, \tag{C.42}$$

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

$$\mathrm{var}\left[\tilde{h}(r,N)\right] = \frac{1}{N}\left(\frac{r}{n}\right)^{2/5} \left[\mathrm{var}\left(\tilde{h}_{CV,r,1}\right) + (N-1)\mathrm{cov}\left(\tilde{h}_{CV,r,1}, \tilde{h}_{CV,r,2}\right)\right] \tag{C.43}$$

and

$$\mathrm{cov}\left(\tilde{h}_{CV,r,1}, \tilde{h}_{CV,r,2}\right) \approx M_r''\left(\tilde{h}_{r0}\right)^{-2} \mathrm{cov}\left[\widetilde{CV}'_1(\tilde{h}_{r0}), \widetilde{CV}'_2(\tilde{h}_{r0})\right], \tag{C.44}$$

where

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

for $q \in \{1,2\}$, $I_1, I_2 \sim U(\mathcal{P})$ and $\mathcal{P} = \{I \subset \{1,\ldots,n\} \mid \#I = r\}$.

Now,

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

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 is therefore not random.

On the other hand,

$$\begin{aligned}&\operatorname{cov}\left[\widetilde{CV}'_1(h), \widetilde{CV}'_2(h) \mid I_1, I_2\right] \\ = \;& \frac{4}{r^2(r-1)^4h^2} \sum_{\substack{i,j,k \in I_1 \\ l,s,t \in I_2 \\ j,k \neq i \\ s,t \neq l}} \operatorname{cov}\left[A_{ij}\alpha_{1h}(X_i - X_j)\right. \\ - \;& h^{-1}B_{ijk}\beta_{1h}(X_i - X_j, X_i - X_k), A_{ls}\alpha_{1h}(X_l - X_s) \\ - \;& \left. h^{-1}B_{lst}\beta_{1h}(X_l - X_s, X_l - X_t)\right]. \end{aligned} \tag{C.45}$$

Following the proof of Lemma 4.2, we need only count the number of cases associated with C_{123124} and C_{123425}. If we define $M = \#(I_1 \cap I_2)$, which, recall, is a random variable, then the number of times C_{123124} and C_{123425} appear in (C.45) is

$$\begin{aligned} C_{123124} &: M(M-1)(r^2 - 4r - M) = M^2r^2 + o\left(M^2r^2\right), \\ C_{123425} &: M^2r^3 + o\left(M^2r^3\right). \end{aligned} \tag{C.46}$$

Plugging (C.46) into (C.45) 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:

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

where $1_{I_2}(\cdot)$ denotes the indicator function of I_2 and we have used the fact that $1_{I_2}(i) \sim \mathrm{Ber}(r/n)$. Therefore,

$$\begin{aligned}
\mathrm{cov}\left[\widetilde{CV}'_1(h), \widetilde{CV}'_2(h)\right] &= R_1(n^{-1}r^{-1}h^2 + rn^{-2}h^2) + R_2 n^{-2}h^{-3} \\
&+ O\left(n^{-2}h^{-1} + n^{-1}r^{-1}h^4 + n^{-2}rh^4\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{cov}\left[\widetilde{CV}'_1(\tilde{h}_{r0}), \widetilde{CV}'_2(\tilde{h}_{r0})\right] &= R_1 C_0^2(n^{-1}r^{-7/5} + n^{-2}r^{3/5}) + R_2 C_0^{-3} n^{-2} r^{3/5} \\
&+ O\left(n^{-1}r^{-9/5} + n^{-2}r^{1/5}\right). \qquad \text{(C.47)}
\end{aligned}$$

Now, plugging (C.47) into (C.44) we get

$$\mathrm{cov}\left(\tilde{h}_{CV,r,1}, \tilde{h}_{CV,r,2}\right) = Vn^{-2}r^{7/5} + Wn^{-1}r^{-3/5} + O\left(n^{-2}r + n^{-1}r^{-1}\right), \qquad \text{(C.48)}$$

where

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

Finally, plugging (C.48) into (C.43) yields

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

Proof of Corollary 4.2 The result is obtained immediately from Corollary 4.1 and Theorem 4.2.

Appendix D

R package `baggedcv`

This appendix includes the documentation concerning the R package `baggedcv`, developed, among others, by the author of this dissertation. This R package implements the bagging version of the cross-validation bandwidth selector for the kernel density estimator (`bagcv`), studied in Section 3.1. In addition, a function (`mopt`) is also included to select the optimal size of the subsamples, as seen in Section 3.1.2.

Package 'baggedcv'

July 26, 2019

Type Package

Title Bagged Cross-Validation for Kernel Density Bandwidth Selection

Version 1.0

Date 2019-07-03

Author Daniel Barreiro Ures, Jeffrey D. Hart, Ricardo Cao, Mario Francisco-Fernandez

Maintainer Daniel Barreiro Ures <daniel.barreiro.ures@udc.es>

Description Bagged cross-validation for bandwidth selection
in kernel density estimation (Hall and Marron (1987) <doi:10.1007/BF00363516>). This bandwidth selector can achieve greater statistical precision than standard cross-validation while being computationally fast.

License GPL-3

Encoding UTF-8

Imports parallel, foreach, doParallel, mclust, kedd, stats

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-07-26 07:50:02 UTC

R topics documented:

`bagcv` *Bagged CV bandwidth selector*

Description

Bagged CV bandwidth selector

Usage

```
bagcv(x, r, s, h0, h1, nb = r, ncores = parallel::detectCores())
```

Arguments

x	Vector. Sample.
r	Positive integer. Size of the subsamples.
s	Positive integer. Number of subsamples.
h0	Positive real number. Range over which to minimize, left bound.
h1	Positive real number. Range over which to minimize, right bound.
nb	Positive integer. Number of bins to use in the `bw.ucv` function.
ncores	Positive integer. Number of cores with which to parallelize the computations.

Details

Bagged cross-validation bandwidth for kernel density estimation.

Value

Bagged CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^6)
bagcv(x, 5000, 100, 0.01, 1, 5000, 2)
```

---

mopt *Estimation of the optimal subsample size for bagged CV*

Description

Estimation of the optimal subsample size for bagged CV

Usage

```
mopt(x, N, r = 1000, s = 100, ncores = parallel::detectCores())
```

Arguments

x	Vector. Sample.
N	Positive integer. Number of subsamples for the bagged bandwidth.
r	Positive integer. Size of the subsamples.
s	Positive integer. Number of subsamples.
ncores	Positive integer. Number of cores with which to parallelize the computations.

Details

Estimates the optimal size of the subsamples for the bagged CV bandwidth selector.

Value

Estimate of the optimal subsample size.

Examples

```
set.seed(1)
x <- rt(10^5, 5)
mopt(x, 500, 500, 10, 2)
```

Appendix E

Rcpp package `baggingbwsel`

This appendix includes the documentation concerning the Rcpp package `baggingbwsel`, developed, among others, by the author of this dissertation. This Rcpp package is an extension of the previous package, `baggedcv`, to the case of nonparametric regression estimation, when considering the Nadaraya–Watson estimator. All the techniques studied in Chapters 3 and 4 are implemented in the package.

Package 'baggingbwsel'

July 12, 2021

Type Package

Title Bagging Bandwidth Selection in Kernel Density and Regression Estimation

Version 1.0

Date 2021-07-08

Description Bagging bandwidth selection methods for the Parzen-Rosenblatt and Nadaraya-Watson estimators. These bandwidth selectors can achieve greater statistical precision than their non-bagged counterparts while being computationally fast. See Barreiro-Ures et al. (2020) <doi:10.1093/biomet/asaa092> and Barreiro-Ures et al. (2021) <arXiv:2105.04134>.

License GPL-3

Encoding UTF-8

Depends mclust, foreach

Imports Rcpp (>= 1.0.3), parallel, doParallel, kedd, stats, sm, nor1mix, rgl, rpanel, tkrplot, misc3d

LinkingTo Rcpp

RoxygenNote 7.1.1

NeedsCompilation yes

Author Daniel Barreiro-Ures [aut, cre], Jeffrey Hart [aut], Ricardo Cao [aut], Mario Francisco-Fernandez [aut]

Maintainer Daniel Barreiro-Ures <daniel.barreiro.ures@udc.es>

Repository CRAN

Date/Publication 2021-07-12 07:20:02 UTC

R topics documented:

`baggingbwsel-package` *A short title line describing what the package does*

Description

A more detailed description of what the package does. A length of about one to five lines is recommended.

Details

This section should provide a more detailed overview of how to use the package, including the most important functions.

Author(s)

Your Name, email optional.

Maintainer: Your Name <your@email.com>

References

This optional section can contain literature or other references for background information.

See Also

Optional links to other man pages

Examples

```
## Not run:
   ## Optional simple examples of the most important functions
   ## These can be in \dontrun{} and \donttest{} blocks.

## End(Not run)
```

bagcv *Bagged CV bandwidth selector for Parzen-Rosenblatt estimator*

Description

Bagged CV bandwidth selector for Parzen-Rosenblatt estimator

Usage

```
bagcv(x, r, s, h0, h1, nb = r, ncores = parallel::detectCores())
```

Arguments

x	Vector. Sample.
r	Positive integer. Size of the subsamples.
s	Positive integer. Number of subsamples.
h0	Positive real number. Range over which to minimize, left bound.
h1	Positive real number. Range over which to minimize, right bound.
nb	Positive integer. Number of bins.
ncores	Positive integer. Number of cores with which to parallelize the computations.

Details

Bagged cross-validation bandwidth selector for the Parzen-Rosenblatt estimator.

Value

Bagged CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^6)
bagcv(x, 5000, 100, 0.01, 1, 1000, 2)
```

bagreg *Bagged CV bandwidth selector for Nadaraya-Watson estimator*

Description

Bagged CV bandwidth selector for Nadaraya-Watson estimator

Usage

```
bagreg(x, y, r, s, h0, h1, nb = r, ncores = parallel::detectCores())
```

Arguments

x	Covariate vector.
y	Response vector.
r	Positive integer. Size of the subsamples.
s	Positive integer. Number of subsamples.
h0	Positive real number. Range over which to minimize, left bound.
h1	Positive real number. Range over which to minimize, right bound.
nb	Positive integer. Number of bins to use in cross-validation.
ncores	Positive integer. Number of cores with which to parallelize the computations.

Details

Bagged cross-validation bandwidth selector for the Nadaraya-Watson estimator.

Value

Bagged CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^5)
y <- 2*x+rnorm(1e5,0,0.5)
bagreg(x, y, 1000, 10, 0.01, 1, 1000, 2)
```

hboot_bag *Bagging bootstrap bandwidth selector for Parzen-Rosenblatt estimator*

Description

Bagging bootstrap bandwidth selector for Parzen-Rosenblatt estimator

Usage

```
hboot_bag(
  x,
  m = n,
  N = 1,
  nb = 1000L,
  g,
  lower,
  upper,
  ncores = parallel::detectCores(logical = FALSE)
)
```

Arguments

x	Vector. Sample.
m	Positive integer. Size of the subsamples.
N	Positive integer. Number of subsamples.
nb	Positive integer. Number of bins.
g	Positive real number. Pilot bandwidth.
lower	Positive real number. Range over which to minimize, left bound.
upper	Positive real number. Range over which to minimize, right bound.
ncores	Positive integer. Number of cores with which to parallelize the computations.

Details

Bagging bootstrap bandwidth selector for the Parzen-Rosenblatt estimator.

Value

Bagged CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^5)
hboot_bag(x, 5000, 10, 1000, lower=0.001, upper=1, ncores=2)
```

`hsss_dens` *Generalized bagging CV bandwidth selector for Parzen-Rosenblatt estimator*

Description

Generalized bagging CV bandwidth selector for Parzen-Rosenblatt estimator

Usage

```
hsss_dens(x, r, s, nb = r, h0, h1, ncores = parallel::detectCores())
```

Arguments

`x`	Vector. Sample.
`r`	Positive integer. Size of the subsamples.
`s`	Positive integer. Number of subsamples.
`nb`	Positive integer. Number of bins.
`h0`	Positive real number. Range over which to minimize, left bound.
`h1`	Positive real number. Range over which to minimize, right bound.
`ncores`	Positive integer. Number of cores with which to parallelize the computations.

Details

Generalized bagging cross-validation bandwidth selector for the Parzen-Rosenblatt estimator.

Value

Bagged CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^5)
hsss_dens(x, 5000, 100, 1000, 0.001, 1, 2)
```

`mopt` *Estimation of the optimal subsample size for bagged CV bandwidth for Parzen-Rosenblatt estimator*

Description

Estimation of the optimal subsample size for bagged CV bandwidth for Parzen-Rosenblatt estimator

Usage

```
mopt(x, N, r = 1000, s = 100, ncores = parallel::detectCores())
```

Arguments

`x`	Vector. Sample.
`N`	Positive integer. Number of subsamples for the bagged bandwidth.
`r`	Positive integer. Size of the subsamples.
`s`	Positive integer. Number of subsamples.
`ncores`	Positive integer. Number of cores with which to parallelize the computations.

Details

Estimates the optimal size of the subsamples for the bagged CV bandwidth selector for the Parzen-Rosenblatt estimator.

Value

Estimate of the optimal subsample size.

Examples

```
set.seed(1)
x <- rt(10^5, 5)
mopt(x, 500, 500, 10, 2)
```

`tss_dens` *Second order bagging CV bandwidth selector for Parzen-Rosenblatt estimator*

Description

Second order bagging CV bandwidth selector for Parzen-Rosenblatt estimator

Usage

```
tss_dens(x, r, s, h0, h1, nb = 1000, ncores = 1)
```

Arguments

`x`	Vector. Sample.
`r`	Vector. The two subsample sizes.
`s`	Positive integer. Number of subsamples.
`h0`	Positive real number. Range over which to minimize, left bound.
`h1`	Positive real number. Range over which to minimize, right bound.
`nb`	Positive integer. Number of bins.
`ncores`	Positive integer. Number of cores with which to parallelize the computations.

Details

Second order bagging cross-validation bandwidth selector for the Parzen-Rosenblatt estimator.

Value

Second order bagging CV bandwidth.

Examples

```
set.seed(1)
x <- rnorm(10^5)
tss_dens(x, 5000, 10, 0.01, 1, 1000, 2)
```

Appendix F

Resumen en español

Se resumen aquí los estudios desarrollados, así como los resultados obtenidos, a lo largo del período de realización de la tesis doctoral. En esta se aborda el problema de la selección de la ventana en la estimación no paramétrica de las funciones de densidad y regresión, centrándose en contextos de muestras de gran tamaño. El coste computacional asociado a algunos de los métodos de selección de la ventana más conocidos como, por ejemplo, aquellos basados en criterios de tipo validación cruzada o bootstrap, los hace inadecuados para contextos de muestras de gran tamaño. En la tesis, este problema se aborda mediante el uso del subagging, un método de aprendizaje conjunto que combina el bootstrap aggregating o bagging con el submuestreo. A lo largo de la tesis se proponen versiones subagging de métodos de selección de la ventana para estimadores tipo núcleo de la densidad y de la regresión y basados en criterios ampliamente conocidos, como validación cruzada o el bootstrap. Tanto en el caso del estimador de Parzen–Rosenblatt, en estimación no paramétrica de la densidad, como en el caso del estimador de Nadaraya–Watson, en estimación no paramétrica de la regresión, los selectores de la ventana propuestos se estudian teórica (obtención de expresiones asintóticas para el sesgo, la varianza y la distribución límite) y empíricamente (mediante diversos estudios de simulación y aplicaciones a conjuntos de datos reales), obteniendo generalmente resultados muy positivos, tanto en términos de precisión estadística (disminuciones sustanciales en el error cuadrático medio y mejores tasas de convergencia) como de agilidad computacional (reducciones drásticas en los tiempos de computación), respecto a sus

análogos "clásicos".

Capítulo 1: Motivación

En el primer capítulo de la tesis se expone el tema objeto de estudio y se presentan las motivaciones que han dado lugar a la investigación desarrollada a lo largo de la tesis doctoral.

Vivimos en la era del Big Data, y la cantidad de datos que se generan en todo el mundo aumenta constantemente. Así, ofrecer soluciones al problema de la gestión y procesamiento de cantidades masivas de datos se torna prioritario. Actualmente, en la literatura estadístico-computacional existen diferentes propuestas dedicadas a la aplicación de técnicas computacionales basadas en el diseño de algoritmos paralelizables en CPU o GPU, por ejemplo mediante plataformas de computación en clúster como Apache Hadoop o Apache Spark. Por otra parte, existen también propuestas de agilización computacional basadas en el uso de submuestreo y métodos de aprendizaje conjunto.

En este sentido, el objetivo principal de la tesis es la propuesta, estudio y aplicación de técnicas de estimación computacionalmente eficientes en el contexto de muestras de gran tamaño, otorgando especial importancia al problema de la selección de la ventana en el campo de la estimación no paramétrica de las funciones de densidad y de regresión.

Capítulo 2: Introducción

En este segundo capítulo de la tesis se pretende introducir al lector en el campo de la estimación no paramétrica de las funciones de densidad y regresión, haciendo especial hincapié en el problema de la selección de la ventana de los estimadores de Parzen–Rosenblatt, en el caso de la densidad, y de Nadaraya–Watson, en el caso de la regresión. Además de discutir distintos métodos de selección de la ventana, como los basados en criterios de tipo validación cruzada, plug-in, etc., se ofrece una introducción al método bootstrap y a las técnicas de aprendizaje conjunto bagging (bootstrap aggregating) y subagging (subsample aggregating), señalando su aplica-

bilidad en selección de la ventana. La técnica del subagging, además, juega un papel central a lo largo de la tesis por su capacidad para aportar mejoras en agilidad computacional y, como se demuestra en capítulos subsiguientes, también en precisión estadística.

En particular, este capítulo introductorio se divide en cuatro secciones. La primera de ellas se dedica a la estimación de la función de densidad, se introduce al lector en el campo de la estimación tipo núcleo de la densidad, describiendo el estimador de Parzen–Rosenblatt de la densidad y presentando algunas de sus propiedades más destacables, así como discutiendo algunos de los métodos de selección de la ventana para dicho estimador más conocidos, haciendo énfasis en el caso del selector de validación cruzada. La segunda sección del capítulo trata la estimación de la función de regresión, se ofrece una introducción al campo de la estimación tipo núcleo de la regresión, describiendo el estimador de Nadaraya–Watson de la regresión y presentando algunas de sus propiedades más importantes, y de nuevo se discuten algunos de los métodos de selección de la ventana de tal estimador más populares, dando especial importancia al selector de validación cruzada. En la tercera sección se describe el método de remuestreo conocido como bootstrap y se ilustra su funcionamiento. En la cuarta y última sección de este capítulo se introduce al lector en el campo de los métodos de aprendizaje conjunto, particularmente en el método bagging y su variante conocida como subagging, ilustrando su funcionamiento y destacando su aplicabilidad a la hora de abordar el problema de la selección de la ventana y de reducir la variabilidad de ciertos selectores de la ventana.

Capítulo 3: Selección de la ventana del estimador de Parzen–Rosenblatt mediante bagging

En el tercer capítulo de la tesis se formaliza la aplicación del subagging al selector de validación cruzada de la ventana del estimador de Parzen–Rosenblatt, discutiendo sus similitudes y diferencias respecto a otras propuestas previas existentes en la literatura. En el caso del selector de validación cruzada, la ventana subagging puede calcularse como el promedio de las ventanas de validación cruzada reescaladas por el factor $(r/n)^{1/5}$ obtenidas para un cierto número de submuestras de tamaño $r < n$

(donde n denota el tamaño de la muestra original) generadas a partir de la muestra original mediante muestreo sin reemplazamiento. El hecho de tener que reescalar las ventanas se debe a que estas han sido obtenidas para submuestras de tamaño r y a que la ventana de validación cruzada, al igual que la ventana óptima, tiende a cero a la tasa $n^{-1/5}$.

En primer lugar se estudian algunas de las propiedades asintóticas del selector propuesto, obteniendo expresiones asintóticas para el sesgo y la varianza del selector, así como para la distribución límite del error del selector. Una de las implicaciones más destacables de estos resultados es que demuestran que el selector subagging puede presentar una precisión estadística notablemente superior a la del selector de validación cruzada estándar. Esta potencial superioridad estadística se muestra a través de mejores tasas de convergencia para el error del selector. Además, en este punto se señala la importancia de no reducir el análisis asintótico del selector propuesto al caso en el que el número de submuestras consideradas es infinito, tal y como ocurre en Hall and Robinson (2009). Al dar libertad a la tasa de crecimiento del número de submuestras, como parámetro del selector, se puede mostrar la estrecha relación que existe entre dicho parámetro y el tamaño de las submuestras. En particular, se demuestra cómo ciertas elecciones del tamaño y número de submuestras permite que el error del selector propuesto converja a su distribución límite a la tasa $n^{-1/2}$ (donde n denota el tamaño de la muestra original), es decir, más rápidamente que el error del selector de validación cruzada estándar, que tiende a su distribución límite a la tasa $n^{-3/10}$.

Una vez estudiadas las propiedades del selector subagging y mostrada la íntima relación entre el número y el tamaño de las submuestras consideradas, se propone un método automático para la selección del tamaño submuestral basado en la minimización del error cuadrático medio estimado del selector.

A continuación, el comportamiento empírico del selector propuesto se analiza mediante diversos estudios de simulación. En general, los resultados son muy positivos y dejan patente la capacidad del bagging para reducir la variabilidad del selector de validación cruzada. En muchos casos, esta drástica reducción en la varianza del selector subagging domina frente al incremento en sesgo experimentado a causa del uso de submuestreo, pudiendo obtenerse así reducciones significativas en el error

cuadrático medio del selector, en algunos casos de más del 90% respecto al de validación cruzada estándar. Además, en lo que respecta al error del estimador de Parzen–Rosenblatt con ventana subagging, se muestra que el error cuadrático integrado del estimador con ventana subagging fue generalmente menor (el 60% de las veces en los escenarios considerados) que el error cuadrático integrado del estimador con ventana de validación cruzada estándar. Más allá de las mejoras en precisión estadística, se realizaron también estudios de simulación con el objetivo de mostrar las ventajas de tipo computacional de emplear subagging.

Aunque el capítulo se centra en la aplicación del subagging al selector de validación cruzada de la ventana del estimador de Parzen–Rosenblatt, también se describe la aplicación del subagging a otros selectores de la ventana, como el basado en el método bootstrap. En este caso, debido al hecho de que el selector bootstrap es mucho menos variable que el selector de validación cruzada, la capacidad del bagging para obtener mejoras en la precisión estadística se ve mermada. Sin embargo, el uso del submuestreo sigue permitiendo reducir significativamente el tiempo de computación a cambio de un ligero aumento en el error cuadrático medio. Además, a la vista de los resultados obtenidos en diversos estudios de simulación, se ofrece una regla automática para la elección del tamaño submuestral que permite garantizar un balance entre precisión estadística y agilidad computacional.

Además, algunas de las secciones del capítulo se dedican a la generalización del subagging a situaciones en las que la tasa de convergencia a cero de la ventana óptima no sea conocida o en las que pudiese resultar de interés incorporar los términos de segundo orden de la ventana óptima en el mecanismo del subagging. En el primer caso, cabe destacar la capacidad de obtener selectores de la ventana en situaciones en las que todavía no se haya desarrollado teoría asintótica y con un coste bastante modesto en términos de precisión estadística. En cuanto al segundo caso, los estudios de simulación parecen indicar que la incorporación de los términos de segundo orden no permite incrementar la precisión estadística del selector, al menos en los escenarios considerados.

Finalmente, se estudia el comportamiento de algunos de los selectores propuestos aplicándolos a dos conjuntos de datos reales: en el primero de ellos se trata de estimar la función de densidad del tiempo de demora experimentado en vuelos es-

tadounidenses durante el año 2017; en el segundo, el objetivo pasa por estimar las funciones de densidad de la edad y el tiempo de hospitalización de pacientes infectados con COVID-19 en España durante el año 2020.

Capítulo 4: Selección de la ventana del estimador de Nadaraya–Watson mediante bagging

El cuarto capítulo de la tesis se dedica a la aplicación del subagging al selector de validación cruzada de la ventana del estimador de Nadaraya–Watson. De manera análoga a lo visto para el estimador de Parzen–Rosenblatt de la densidad, en el caso del selector de validación cruzada, la ventana subagging puede calcularse como el promedio de las ventanas de validación cruzada reescaladas por el factor $(r/n)^{1/5}$ obtenidas para un cierto número de submuestras de tamaño $r < n$ generadas a partir de la muestra original mediante muestreo sin reemplazamiento. De nuevo, la necesidad de reescalar las ventanas se debe a que estas han sido obtenidas para submuestras de tamaño r y a que la ventana de validación cruzada, al igual que la ventana óptima, tiende a cero a la tasa $n^{-1/5}$.

En la primera parte del capítulo se estudian las propiedades asintóticas del selector de validación cruzada estándar. El hecho de que el estimador de Nadaraya–Watson presenta un denominador aleatorio complica sustancialmente el análisis teórico, por lo que se utilizó una aproximación teórica (cuadrática) compuesta por los términos de primer y segundo orden del estimador de Nadaraya–Watson y que, además, tiene la ventaja de no presentar un denominador aleatorio. Se obtuvieron las propiedades asintóticas, tanto de esta aproximación teórica del estimador de Nadaraya–Watson, como del selector de validación cruzada definido a partir de dicha aproximación del estimador. Además, se ofrece un bosquejo de demostración que permite afirmar, con cierto rigor, que los resultados asintóticos obtenidos para el selector de validación cruzada basado en esta aproximación teórica del estimador de Nadaraya–Watson podrían extenderse al selector de validación cruzada basado en la versión exacta del estimador.

Una vez conocidas las propiedades del selector de validación cruzada modificado, es decir, aquel definido a partir de la aproximación del estimador de Nadaraya–

Watson, la siguiente parte del capítulo se dedica al análisis teórico de la versión subagging del selector de validación cruzada modificado de la ventana de este estimador. Se estudian algunas de las propiedades asintóticas del selector propuesto, obteniendo expresiones asintóticas para el sesgo y la varianza del mismo, así como para la distribución límite de su error. Tal y como ocurría en el capítulo anterior, una de las implicaciones más destacables de estos resultados es que demuestran que el selector subagging puede presentar una precisión estadística notablemente superior a la del selector de validación cruzada estándar. Esta potencial superioridad estadística se muestra a través de mejores tasas de convergencia para el error del selector. Al igual que en el caso de la densidad, el hecho de dar libertad a la tasa de crecimiento del número de submuestras, como parámetro del selector, permite mostrar la estrecha relación existente entre dicho parámetro y el tamaño de las submuestras. En particular, se demuestra cómo ciertas elecciones del tamaño y número de submuestras permite que el error del selector propuesto converja a su distribución límite a la tasa $n^{-1/2}$, es decir, más rápidamente que el error del selector de validación cruzada estándar, que tiende a su distribución límite a la tasa $n^{-3/10}$.

Una vez estudiadas las propiedades del selector subagging y mostrada la íntima relación entre el número y el tamaño de las submuestras consideradas, se proponen distintos criterios de optimalidad para la selección del tamaño submuestral.

A continuación, el comportamiento empírico del selector propuesto se analiza mediante diversos estudios de simulación. En general, tal y como ocurre en el caso de la densidad, los resultados son muy positivos y dejan patente la capacidad del bagging para reducir la variabilidad del selector de validación cruzada. En muchos casos, esta drástica reducción en la varianza del selector subagging domina frente al incremento en sesgo experimentado a causa del uso de submuestreo, pudiendo obtenerse así reducciones significativas en el error cuadrático medio del selector, en algunos casos de más del 90% respecto al de validación cruzada estándar. Además de analizar las mejoras en precisión estadística, se realizaron también estudios de simulación con el objetivo de mostrar las enormes ventajas computacionales del uso de subagging. En este sentido y a modo de ejemplo, se estimó que, frente a los 7 años que requeriría el cálculo de la ventana de validación cruzada estándar para una muestra de tamaño 10^8, el cálculo de la ventana subagging solamente necesitaría

alrededor de 16 horas, para ciertos valores del tamaño submuestral y el número de submuestras.

El comportamiento empírico del selector propuesto también se estudia aplicándolo al conjunto de datos del COVID-19 ya tratado en el capítulo anterior. En este caso, el objetivo pasa por estudiar la relación entre la edad y el tiempo de hospitalización experimentado por los pacientes infectados con COVID-19. De los resultados obtenidos se deduce que el tiempo de hospitalización esperado crece de forma no lineal para pacientes de menos de 70 años y que esta tendencia se revierte para pacientes con edades comprendidas entre los 70 y 100 años, pudiendo deberse esto último al hecho de que los pacientes de este grupo de edad tienen más probabilidades de fallecer a causa de la infección y, por lo tanto, finalizar su período de hospitalización de manera prematura. Además, pudo observarse que el tiempo de hospitalización esperado es generalmente menor en el caso de las mujeres, exceptuando los grupos de edad de menos de 30 años y de entre 65 y 85 años, aunque esta diferencia en los tiempos de hospitalización medios de hombres y mujeres no es realmente significativa ya que no suele ser mayor a un día.

Aunque el capítulo se centra en la aplicación del subagging al selector de validación cruzada de la ventana del estimador de Parzen–Rosenblatt, también se describe la aplicación del subagging al selector bootstrap de la ventana del estimador de Nadaraya–Watson.

Capítulo 5: Conclusiones y trabajos futuros

En este capítulo se exponen las conclusiones derivadas de la realización de la tesis y se perfilan futuras líneas de trabajo.

La principal conclusión que se desprende de la tesis es que el uso del subagging a la hora de abordar el problema de la selección de la ventana permite obtener mejoras sustanciales tanto en términos de mayor precisión estadística como de mayor agilidad computacional, especialmente cuando el subagging se aplica a selectores de la ventana con una alta variabilidad. Esta doble virtud del subagging lo hace particularmente interesante cuando se dispone de muestras de gran tamaño. Sin embargo, los estudios de simulación han mostrado que el uso del subagging pudiera ser aconsejable incluso

en contextos de muestras de tamaño moderado dada su capacidad para reducir el error cuadrático medio del selector base.

En cuanto a las posibles líneas de trabajo para el futuro, estas incluyen el estudio del selector subagging de la ventana del estimador local lineal de la función de regresión, selección de la ventana piloto óptima para el selector bootstrap de la ventana del estimador de Nadaraya–Watson, extensión de las técnicas propuestas a los casos multidimensional y de datos dependientes, extensión de las técnicas propuestas a problemas de clasificación o selección de modelos y optimización del código desarrollado a lo largo de la tesis e incorporado en los paquetes de R `baggedcv` y `baggingbwsel`, ambos publicados en el CRAN. Otra posible línea de trabajo para el futuro podría pasar por la utilización de plataformas de computación en clúster como Apache Hadoop o Apache Spark y la traducción del código desarrollado a otros lenguajes de programación como Python, muy popular entre la comunidad de Machine Learning. Una de las ventajas del lenguaje Python es que cuenta con una API (interfaz de programación de aplicaciones) de Apache Spark llamada PySpark que permite la computación en paralelo, algo de gran utilidad a la hora de tratar con datos de gran tamaño o modelos con una alta complejidad.

Apéndices

En los apéndices se encuentran las demostraciones de los resultados (teoremas, lemas, etc.) enunciados en los distintos capítulos de la tesis. En particular, el Apéndice A recoge las demostraciones de los resultados expuestos en el Capítulo 3, en el Apéndice B se incluye una correción del resultado principal de Hall and Robinson (2009) y en el Apéndice C se incorporan las demostraciones de los resultados presentados en el Capítulo 4. Por último, los Apéndices D y E contienen la documentación de los paquetes de R `baggedcv` y `baggingbwsel`, respectivamente, que han sido publicados en el CRAN e incluyen funciones para aplicar buena parte de las técnicas propuestas a lo largo de la tesis.

Bibliography

Barbeito, I. (2020). *Exact bootstrap methods for nonparametric curve estimation*. PhD thesis, Universidade da Coruña.

Barreiro-Ures, D., Cao, R., Francisco-Fernández, M., and Hart, J. D. (2021a). Bagging cross-validated bandwidths with application to big data. *Biometrika, to appear*.

Barreiro-Ures, D., Hart, J. D., Cao, R., and Francisco-Fernandez, M. (2019). *baggedcv: bagged cross-validation for kernel density bandwidth selection*. R package version 1.0. `https://cran.r-project.org/package=baggedcv`.

Barreiro-Ures, D., Hart, J. D., Cao, R., and Francisco-Fernández, M. (2021b). *baggingbwsel: bagging bandwidth selection in kernel density and regression estimation*. R package version 1.0. `https://cran.r-project.org/package=baggingbwsel`.

Bhattacharya, A. and Hart, J. D. (2016). Partitioned cross-validation for divide-and conquer density estimation. arXiv:1609.00065.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927–961.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

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

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Broniatowski, M., Deheuvels, P., and Devroye, L. (1989). On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate. *The Annals of Statistics*, 17(3):1070–1086.

Cao, R. (1990). Órdenes de convergencia para las aproximaciones normal y bootstrap en estimación no paramétrica de la función de densidad. *Trabajos de Estadística*, 5(2):23–32.

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, 45(1):137–160.

Cao, R., Cuevas, A., and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, 17(2):153–176.

Cao, R. and González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics*, 2(4):379–388.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on statistics and applied probability series. Chapman and Hall, Boca Raton, Florida.

Farrash, M. (2016). *Machine learning ensemble method for discovering knowledge from big data*. PhD thesis, University of East Anglia.

Feluch, W. and Koronacki, J. (1992). A note on modified cross-validation in density estimation. *Computational Statistics & Data Analysis*, 13(2):143–151.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Friedman, J. H. and Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In Gasser, T. and Rosenblatt, M., editors, *Smoothing Techniques for Curve Estimation*, pages 23–68, Berlin, Heidelberg. Springer Berlin Heidelberg.

Habbema, J. D. F., Hermans, J., and Van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation. In Bruckmann, G., editor, *Compstat 1974: Proceedings in Computational Statistics*, pages 101–110, Vienna.

Hall, P. and Marron, J. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, 74:567–581.

Hall, P., Marron, J., and Park, B. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, 92(1):1–20.

Hall, P. and Robinson, A. P. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96(1):175–186.

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *Journal of the American Statistical Association*, 93(442):620–631.

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(401):86–95.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19(2):778–796.

Jones, M., Marron, J., and Park, B. (1991). A simple root n bandwidth selector. *The Annals of Statistics*, 19(4):1919–1932.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. (2012). The big data bootstrap. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 2, pages 1787–1794, Edinburgh, Scotland.

Köhler, M., Schindler, A., and Sperlich, S. (2014). A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review*, 82(2):243–274.

Ma, P. and Sun, X. (2015). Leveraging for big data regression. *WIREs Computational Statistics*, 7(1):70–76.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P., and Sperlich, S. (2011). Do-validation for kernel density estimation. *Journal of the American Statistical Association*, 106(494):651–660.

Marron, J. S. (1987). Partitioned cross-validation. *Econometric Reviews*, 6(2):271–283.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736.

Martínez-Miranda, M. D., Nielsen, J. P., and Sperlich, S. A. (2011). One-sided cross-validation for density estimation with an application to operational risk. In Gregoriou, G. N., editor, *Operational Risk toward Basel III*, chapter 9, pages 177–195. John Wiley & Sons, Ltd, New Jersey.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M., Zadeh, R., Zaharia, M., and Talwalkar, A. (2015). Mllib: machine learning in apache spark. *Journal of Machine Learning Research*, 17(34):1–7.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11(1):169–198.

Park, B. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031 – 2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer New York.

Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B*, 34(3):385–392.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B*, 11(1):68–84.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

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(432):1257–1270.

Savchuk, O., Hart, J. D., and Sheather, S. J. (2010). Indirect cross-validation for density estimation. *Journal of the American Statistical Association*, 105(489):415–423.

Scott, D. and Terrell, G. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53(3):683–690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B*, 36(2):111–147.

Stute, W. (1992). Modified cross-validation in density estimation. *Journal of Statistical Planning and Inference*, 30(3):293–305.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (preliminary report). *The Annals of Mathematical Statistics*, 29(2):614–623.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295.