Statistical Methods for Studying Emergence Curves in Weed Science

Autor: Miguel Ángel Reyes Cortés

Tese de doutoramento UDC / 2015

Directores: Ricardo Cao Abad e Mario Francisco Fernández.

Departamento de Matemáticas[1]

UNIVERSIDADE DA CORUÑA

[1] Programa regulado polo RD 1393/2007: Programa de doutoramento en Estatística e Investigación Operativa.

STATISTICAL METHODS FOR STUDYING EMERGENCE CURVES IN WEED SCIENCE

Thesis submitted at the Universidade da Coruña
in fulfilment of the requirements for
the degree of PhD in Statistics and Operations Research

by

Miguel Ángel Reyes Cortés
Department of Mathematics
Universidade da Coruña

September, 2015

A la memoria de mi padre
A la memoria de mi abuela materna
A mi madre
A Alma

Esta tese é o resultado de anos de traballo e interacción con outros. Quero agradecer ás persoas que formaron parte da miña rede humana nesta etapa. Que a propósito ou sen querelo influíron en min. Foron varias en diferentes momentos, así que non as mencionarei, mais elas saben quen son. Son aquelas que compartiron penas e glorias, reflexivas charlas de café ou o pracer catártico de compartir unhas *Estrella Galicia*, bos viños, boa comida, bo ambiente. Ou unha simple *gominola da inspiración*. Ata sempre.

Esta tesis es el resultado de años de trabajo e interacción con otros. Quiero agradecer a las personas que formaron parte de mi red humana en esta etapa. Que a propósito o sin quererlo influyeron en mí. Han sido varias en diferentes momentos, así que no las mencionaré, pero ellas saben quiénes son. Son esas que compartieron penas y glorias, reflexivas charlas de café o el placer catártico de compartir unas *Estrella Galicia*, buenos vinos, buena comida, buen ambiente. O una simple *gominola de la inspiración*. Hasta siempre.

This thesis is the result of years of work and interaction with others. I'd like to thank the people who were part of my human network at this stage. That on purpose or unwittingly influenced me. There have been several at different times, so I'm not going to mention any names, but they know who they are. They are those who shared sorrows and joys, deep and wide coffee conversations, or the cathartic pleasure of sharing some *Estrella Galicia*, fine wines, good food, good atmosphere. Or just a simple *inspirational jelly candy*. Farewell.

En consecuencia: el que quiera tener acierto sin error, orden sin desorden, es que no entiende los principios del cielo y la tierra. No sabe cómo encajan las cosas.

Chuang Tzu-*Grande y Pequeño*

Un cuento taoísta habla de un viejo que cayó accidentalmente en los rápidos de un río que llevaban a una elevada y peligrosa cascada. Los testigos de la escena temieron por su vida. Milagrosamente, salió vivo e ileso del fondo de la cascada, corriente abajo. La gente le preguntó cómo se las había arreglado para sobrevivir. "Me acomodé al agua en vez de acomodar el agua a mí. Sin pensarlo, me dejé amoldar por ella. Sumergiéndome en el torbellino, me desembaracé de él. Así fue como sobreviví."

Alan Watts

Negar la sucesión temporal, negar el yo, negar el universo astronómico, son desesperaciones aparentes y consuelos secretos. Nuestro destino […] no es espantoso por irreal; es espantoso porque es irreversible […]. El tiempo es la sustancia de que estoy hecho. El tiempo es un río que me arrebata, pero yo soy el río; es un tigre que me destroza, pero yo soy el tigre; es un fuego que me consume, pero yo soy el fuego. El mundo, desgraciadamente, es real.

Jorge Luis Borges- *Otras inquisiciones*

Contents

Acknowledgements

I would like to thank very strongly the directors of this thesis, Ricardo Cao Abad and Mario Francisco Fernández, for their infinite patience and for guiding me in the world of research. My permanent gratitude for all your support.

Let my appreciations go to Paul Janssen and Noël Veraverbeke for receiving me so warmly at the Center for Statistics (CenStat) of the Hasselt University. Their comments and contributions have been invaluable and enlightening. Very special thanks to Paul Janssen, for having always been on the lookout during my stay at CenStat, checking that everything was in order both personally and professionally. Also, I want to thank the Modeling, Optimization and Statistical Inference (MODES) group of the Universidade da Coruña (UDC), for allowing me to be a part of it all these years.

Finally, this long-term project was made possible thanks to different funding sources. First, the Spanish Agency for International Development Cooperation (AECID), from which I received financial support during my master's studies and during the early years of my doctoral training. Likewise, at different moments, the following projects also partially funded this research:

- *Inferencia estadística no paramétrica: aplicaciones en análisis térmico, riesgo de crédito, malherbología y genómica.*
  -Ministerio de Ciencia e Innovación (MTM2008-00166)

- *Inferencia estadística para datos complejos de alta dimensión: aplicaciones en análisis térmico, fiabilidad naval, genómica, malherbología, neurociencia y oncología.*
  -Ministerio de Ciencia e Innovación (MTM2011-22392)

- *Métodos estadísticos para la fiabilidad y el control de la calidad en buques.*
  -Consellería de Economia e Industria-Xunta de Galicia (10DPI105003PR)

- *Axudas para a consolidación e estruturación de unidades de investogación competitivas do Sistema Universitario de Galicia. Modalidade de grupos de referencia competitiva.*
  -Consellería de Cultura, Educación e Ordenación Universitaria-Xunta de Galicia (CN 2012/130)

- *Axudas para estadías predoutorais INDITEX-UDC, 2014.*
  -Convenio de colaboración entre a UDC e Inditex S.A. para a internacionalización dos estudos de doutoramento.

Resumo

Esta tese trata o problema da estimación da función de densidade e de distribución cando os datos se presentan agrupados. Para este propósito, considérase o estimador núcleo da densidade e proponse unha modificación para usalo con datos agrupados. Sempre que se cumpran os supostos axeitados, demóstrase que o coñecido selector plug-in $AMISE$ óptimo da ventá pode usarse satisfactoriamente con estes datos, o que na práctica leva a definir o concepto de agrupación lixeira. Para escenarios de agrupación pesada, proponse un selector bootstrap. Mediante estudos de simulación móstrase o bo desempeño do estimador cando se usa axeitadamente o selector plug-in ou o selector bootstrap, dependendo do grao de agrupación dos datos. Con base no estimador núcleo da densidade para datos agrupados, derívase un estimador núcleo da distribución para este tipo de datos. Obtéñense formalmente as súas propiedades asintóticas e estúdase o seu desempeño en diferentes escenarios de agrupación usando un selector plug-in adecuado. Finalmente, mediante aplicacións a datos reais, móstrase a efectividade dos métodos non paramétricos propostos nesta disertación, os mesmos que nalgúns casos superan o desempeño dalgúns métodos paramétricos habitualmente usados en malherboloxía para estimar a probabilidade de emerxencia das malas herbas.

Resumen

Esta tesis trata el problema de la estimación de la función de densidad y de distribución cuando los datos se presentan agrupados. Para este propósito, se considera el estimador núcleo de la densidad y se propone una modificación para usarlo con datos agrupados. Siempre que se cumplan los supuestos adecuados, se demuestra que el conocido selector plug-in $AMISE$ óptimo de la ventana puede usarse satisfactoriamente con estos datos, lo que en la práctica lleva a definir el concepto de agrupación ligera. Para escenarios de agrupación pesada, se propone un selector bootstrap. Mediante estudios de simulación se muestra el buen desempeño del estimador cuando se usa adecuadamente el selector plug-in o el selector bootstrap, dependiendo del grado de agrupación de los datos. Con base en el estimador núcleo de la densidad para datos agrupados, se deriva un estimador núcleo de la distribución para este tipo de datos. Se obtienen formalmente sus propiedades asintóticas y se estudia su desempeño en diferentes escenarios de agrupación usando un selector plug-in adecuado. Finalmente, mediante aplicaciones a datos reales, se muestra la efectividad de los métodos no paramétricos propuestos en esta disertación, mismos que en algunos casos superan el desempeño de algunos métodos paramétricos habitualmente usados en malherbología para estimar la probabilidad de emergencia de las malas hierbas.

Abstract

This thesis deals with the problem of estimating the density and distribution functions when the data at hand are grouped. For this, the classical kernel density estimator is considered and a suitable modification is proposed for using it with that type of data. Likewise, whenever the appropriate assumptions are met, it is formally proved that the well-known $AMISE$ optimal plug-in bandwidth selector can be successfully used in the presence of grouped data, which in practice leads to define the concept of light grouping. For scenarios of heavy grouping, an alternative bootstrap bandwidth selector is proposed. By means of simulation studies, it is shown the good performance of the estimator when adequately using either the plug-in or the bootstrap bandwidth selector, depending on the degree of grouping. Based on the kernel density estimator for grouped data, a kernel distribution estimator for grouped data is derived. Its asymptotic properties are formally obtained, and its performance is studied in different grouping scenarios using a suitable plug-in selector. Finally, applications to real data coming from weed science show the effectiveness of the nonparametric methods proposed in this dissertation, which in some cases outperform the typical parametric methods used by weed scientists for estimating seedling emergence probabilities.

Preface

The subject of this thesis arises from a practical problem posed by people from the Institute of Sustainable Agriculture (CSIC), in Córdoba, Spain, to the Modeling, Optimization and Statistical Inference group (MODES) of the Universidade da Coruña. The problem was on how to accurately predict weed seedling emergence when the data at hand is grouped.

Weeds are problematic both in agricultural and nonagricultural areas. Not only do they interfere in crops, but also they may cause economic losses or negative social impacts. Thus, to control weeds is of major importance, and accurate predictions of seedling emergence is crucial for making right decisions on the use of weed management strategies.

Traditionally, weed scientists have tackled the problem of modeling weed seedling emergence by fitting parametric nonlinear regression models to cumulative emergence patterns. However, these models have been questioned due to several major limitations. For example, sometimes parametric models are not flexible enough to capture complex features in the observed cumulative emergence values. Also, when obtained from consecutive monitoring times, the cumulative emergence values are not statistically independent, leading to positive autocorrelation of the residuals. This is an issue, since the construction of confidence intervals and hypothesis testing in standard nonlinear regression depend on uncorrelated residuals. Moreover, although the choice of a parametric model is based on experience, if the model is not appropriate, then there is a risk of getting wrong conclusions from the analysis. Those problems have not been explicitly considered in the weed science literature, where fitting the model has been the main goal, regardless of whether the statistical analysis is proper or not.

The objective of this monograph is to provide some other tools and approaches for predicting weed seedling emergence. A straightforward alternative to parametric nonlinear regression models is the nonparametric approach, which does not consider any specific model on the random variables under consideration. The basic idea is to impose minimum assumptions to get useful information from data, or as it is usally said, to let the data "to speak for themselves". This way, nonparametric techniques may provide more flexible estimations and, in some cases, more reliable results than parametric nonlinear regression methods.

From the statistical standpoint, the problem of seedling emergence can be viewed as that of finding structure in data. Hence, the problem can be addressed either by estimating the density or the distribution function. For this, and from the nonparametric front, a classic tool is kernel smoothing, whether for density or distribution estimation. However, as it was mentioned earlier, available data on seedling emergence typically come grouped, and the kernel density or distribution estimators cannot be used with such type of data. This forces to somehow redefine these estimators in order to get density or distribution estimations from grouped data. This is what this monograph is mainly about.

Besides starting with some basic weed science and kernel smoothing concepts, the first approach to the problem is on how to estimate the density function from grouped data. For this, a suitable modification of the kernel density estimator is proposed and its asymptotic properties are formally derived. A key element in kernel smoothing is the right choice of the bandwidth. Thus, two bandwidth selectors are proposed. The first one is a plug-in selector, directly obtained from the asymptotic properties. The second one is a bootstrap based bandwidth selector. By simulation studies, the performance of the modified kernel density estimator is examined when using those bandwidth selectors under different grouping scenarios. Facing applications, this allows for practical guidelines on when to conveniently use each of the bandwidth selectors.

Subsequently, a kernel distribution estimator for grouped data is derived from the previous estimator. Then, the procedure is quite parallel: its asymptotic properties are formally derived and a proper bandwidth selector is obtained. By means of simulation studies, its performance is examined under different grouping conditions, highlighting its main differences with respect to the kernel density estimator for grouped data.

This monograph is completed with a chapter of applications to real data on weed seedling emergence. The already obtained practical guidelines are tested, confirming its usefulness in practice. Moreover, the kernel distribution estimator is also tested versus some typical nonlinear regression models used in weed science. The results show that the nonparametric methods proposed are not only valid to describe weed seedling emergence, but also its flexibility allows them to better describe complex distributions that, due to its rigidity, parametric models tend to ovsersimplify.

Finally, given the general way in which kernel density and distribution estimators for grouped data were defined and studied, it is important to stress that they can be potentially applied to find structure in data not only in weed science, but in any grouped data set, regardless of the discipline they come from. Hopefully, more research on this topic will take place, trying to improve the results obtained by considering more elaborated modifications of the kernel density and distribution estimators to deal with grouped data, or by considering approaches like the nonparametric isotonic regression or nonhomogeneous Poisson processes.

Chapter 1

Introduction

The aim of this work is to study the nonparametric estimation of the density and distribution functions when the data at hand are grouped. Grouped data appear whether continuous random variables are measured or used in binned or rounded form or in systems in which the observation time is periodic. These type of data are common in areas like engineering, economics, health and life sciences, agriculture and many more.

The motivational problem of this work comes from a branch of agriculture called weed science. In this area, random variables based on humidity or temperature (or both) are very important for predicting weed emergence. In some weed science experiments the observation time is periodic, so researchers are unable to observe the exact values of those variables; instead, they obtain a data set consisting in counts between variable consecutive monitoring times. Moreover, indirect studies only allow access to data expressed as proportions of emerged seedlings.

In the context of weed science, knowing statistical emergence patterns is essential for an efficient application of herbicides and techniques that help erradicate weeds. For that, density estimation is of primary importance, and grouped data poses interesting challenges in implementing existing density estimators that require an adequate modification to be used with this type of data.

This chapter gives a brief introduction about grouped data and some central concepts on weed science and its environmental and public importance. Also, a short presentation of classical nonparametric estimators of the density and distribution functions will be given, showing its shortcomings. This motivates to propose the use of kernel density estimation and a suitable modification for dealing with the density and distribution estimation with grouped data.

1.1 Grouped data

In the experimental sciences, data usually come from measurements of continuous variables such as temperature, mass, weight, time, length, etc. However, measurements are observed and obtained in finite precision due to multiple factors such as limitations of the measurement instruments, imperfections in our senses and the almost incalculable variables of the physical world around us. Therefore, the true values of a continuous variable are not achievable, since there is always an error or a degree of uncertainty attached in any measurement. All continuous variables are at some point rounded or coarsened; in a very basic sense, all measurements are grouped.

Nonetheless, it is fair to say that sometimes there are systems in which an unlimited measurement accuracy is really not required. In those cases, either for any sort of convenience or because of the nature of the system, the data are measured or used in binned or rounded form. This type of data are usually known as *grouped data*, which also appear in systems where the observation time is not continuous but periodical, as in a medical follow up, in which the doctor checks a patient or a group of patients not continuosly, but from time to time. In such cases, the experimental conditions limit the researcher to observe time to event data distributed along a set of consecutive intervals, not knowing the exact values of the random variable of interest but only knowing the number of observations at each bin. Systems like this appear very frequently in a diversity of areas such as engineering, economics, social sciences, epidemiology and many more (Coit and Dey, 1999; Minoiu and Reddy, 2009; Guo, 2005; Pipper and Ritz, 2007). Especially in those kind of situations, the uncertainty in the measurements is not negligible at all and it should be taken into account to avoid serious mistakes when making inferences.

1.2 About Weed Science

The term *weed* is generally refered to undesirable invasive plants that may have a negative impact on the economy, ecology, human health or urban areas, such as reduce crop yields, activate allergies, stifle waterways, disrupt the habitats of other plants or animals, create safety risks or reduce aesthetic and property values[1].

Weed science is, then, the study of vegetation not only in agriculture but in areas in which plants need to be managed. Yet, it is not just about controlling plants, but the study of these plants and its genetics, and this labour is complex enough to include several other disciplines like statistics, ecology, physiology, biology and chemistry. The importance of weed science research becomes evident as it influences the development and assessment of weed control regulations, which brings together different actors on the decision making, such as academia itself, private industry and goverment and policy makers.

[1] More information can be found at the Weed Science Society of America, http://wssa.net/weed/

1.2.1 Modeling seedling emergence

One of the central aims of weed science is to model seedling emergence, which is considered the most important phenological factor that influences the success of annual plants (Fernández-Quintanilla et al., 1986; Forcella et al., 2000). Modeling seedling emergence enables prediction of future weed appearance, leading to efficiently implement strategies for eradicating or controlling weeds, particularly in crop management (Leblanc et al., 2003).

Although prediction of weed emergence can be done by means of indexes (Naylor, 1981; Hunter et al., 1984), most modern approaches consider modeling techniques (Colbach et al., 2005). For modeling weed emergence, it is neccessary to take into account some of the main factors that influence the phenomenon. In this sense, temperature and water potential have been identified as the most important ones (Izquierdo et al., 2009). Consequently, emergence models typically use random variables based on temperature alone, or based on temperature and humidity at the same time. The former random variable is called *thermal time* (TT), and it is based on soil temperature above a reference temperature; the latter is the so-called *hydrothermal time* (HTT), which uses a combination of thermal and hydro-time over a water potential reference (Forcella et al., 2000; Bradford, 2002; Grundy, 2003). The evidence indicates that models based on HTT are more accurate in describing weed emergence than those just based on TT (Leguizamón et al., 2005; McGiffen et al., 2008), showing that water potential is an important factor that contributes to triggering emergence.

In trying to assess the relationship between weed emergence and HTT, weed scientists have tackled the problem from a regression point of view (Grundy, 2003). At a first glance, it may seem natural to use parametric models like the Gompertz or Logistic (Haj Seyed Hadi and Gonzalez-Andujar, 2009), where cumulative HTT (CHTT) is considered as the explanatory variable and the cumulative emergence as the response variable. However, this approach has some issues: first, parametric models are not always flexible enough to capture complex details in the HTT distribution, like abrupt bumps, thin spikes or heavy tails. Second, to be a reasonable model for fitting model emergence, care must be taken for the regression function to be between 0 and 1. Last, CHTT data are not statistically independent, which usually is a theoretical condition for classical regression models. But a change of perspective can be helpful: since cumulative emergence is an increasing function between 0 and 1, it makes sense to model it from a density or distribution estimation point of view, using just one random variable (CHTT at emergence), instead of two, as in parametric regression models.

1.2.2 Measuring HTT

To know the relationship between seedling emergence and environmental variables is useful for predicting weed emergence. As stated before, among the several factors that influence weed emergence, temperature and humidity are perhaps the most important. Therefore, another problem in weed science is determining the best way to measure both of them, as they lead to the final synthesis variable HTT (Schutte et al., 2008).

By *the best way* to measure temperature and humidity we mean *the best depth.* Given a sample of seedlings, each one is at its own depth, so a trivial answer would be *to measure it right at the position at which they are* (Royo-Esnal et al., 2010). Nevertheless, estimations of temperature and humidity at different soil layers may lead to very different values of HTT; so, it is natural to ask which of those different depths is the best one for improving prediction of weed emergence.

The construction of indices for helping to decide which of the soil depths is the best for improving prediction entails somehow measuring the spread of the underlying density of HTT. The flatter the density, the more spread it is, which improves prediction tasks. In other words, we are interested in knowing which of those different depths gives the most convenient distribution of HTT, so that prediction of weed emergence is more accurate.

The spread of a distribution can be measured by classical indices like the coefficient of variation or kurtosis. But also, it can be calculated by means of statistical functionals for measuring the roughness of the density, whether in the slope, as in

$$\sigma^3 \int f'^2 (x) \, dx,$$

or in the curvature, as in

$$\sigma^5 \int f''^2 (x) \, dx,$$

where σ stands for the population standard deviation and f stands for the unknown probability density function[2]. In any case, since in the weed science experiments the monitoring time is discontinuous, HTT are obtained as grouped data; i.e., as counts bewteen monitoring times. That incompleteness of data drives to propose adapted versions of the empirical estimates, whether for the case of the coefficient of variation or kurtosis, or for the case of density functionals, for which it is necessary to somehow estimate the density function f using grouped data.

In Section 1.3, some basic nonparametric estimators of the density and the distribution are presented. It will be clear that they have some drawbacks that encourage to prefer kernel density and distribution estimation.

1.3 Basics of density estimation

A very fundamental problem in statistics is the estimation of the *probability density function*, as it provides a description of the distribution of a continuous random variable, and

[2] For invariance convenience, both functionals appear multiplied by two suitable powers of σ.

whose integral across an interval gives the probability that the value of the variable lies within the same interval.

Basically, there are two approaches for density estimation. On the one hand, the *parametric approach* consists in assuming that data come from a known parametric family of distributions. Once a particular form for the underlying density has been specified, the problem of estimating the density is equivalent to estimating the parameters, which are substituted into the parametric formula. The main problem with this approach is its rigidity. Also, if the model is not correct, inferences may lead to erroneous interpretations of the data.

On the other hand, the nonparametric approach consists in making no assumptions about what the form of the density would be; if any, just the relatively weak assumption that the density is a smooth curve. This approach is appropiate when there is no information about the functional form of the density.

To nonparametrically estimate a curve f, data should be smoothed in some way and to some extent. *Smoothing* a data set means to approximate a function that attempts to capture important structure features, while leaving out other fine-scale structures. Once f is estimated, it is needed a criterion to determine how good the estimate $\hat{f}$ is with respect to f.

This section shows a brief revision of general concepts about smoothing as well as some typical nonparametric density and distribution estimators and its properties, such as the histogram and the empirical distribution function.

1.3.1 A discrepancy measure

To evaluate how good an estimate $\hat{f}$ is with respect to an objective function f, a measure of difference between them is needed. Let us consider first the estimation at a point x, $\hat{f}_n(x)$ [3]. A very popular error measure is the squared error, $SE(x) = \left[\hat{f}_n(x) - f(x)\right]^2$ along with its expected value, the mean squared error, $MSE(x) = \mathbb{E}\left[\hat{f}_n(x) - f(x)\right]^2$. These quantities let us locally evaluate the quality of the estimate $\hat{f}_n(x)$.

Expanding the squared term, it is easy to prove that the MSE can be decomposed into two parts,

$$MSE(x) = \mathbb{E}\left[\hat{f}_n(x) - f(x)\right]^2 = \mathbb{B}\left[\hat{f}_n(x)\right]^2 + \mathbb{V}\left[\hat{f}_n(x)\right], \qquad (1.1)$$

where $\mathbb{B}\left[\hat{f}_n(x)\right]$ stands for the bias of the estimator and $\mathbb{V}\left[\hat{f}_n(x)\right]$ is its variance.

When smoothing, the main challenge is to decide how much to smooth. A first idea is that the amount of smoothing should be such that the MSE is minimum, but minimizing the MSE entails to somehow minimize the sum of the squared bias and the variance. If the

[3] Since the estimate depends on each sample, an n subscript has been added to $\hat{f}$.

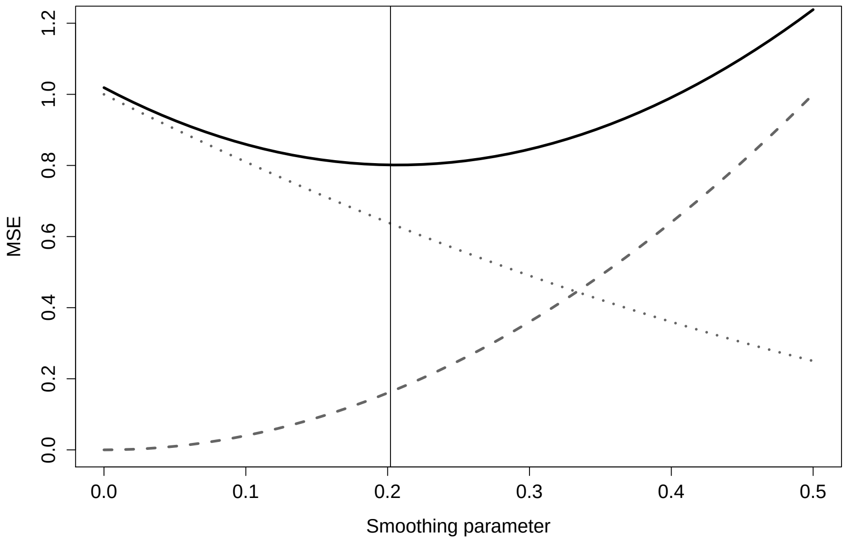

Figure 1.1: The bias-variance tradeoff. As the amount of smoothing increases, the bias (dashed line) increases and the variance (dotted line) decreases. An equilibrated amount of smoothing is indicated by the vertical line, where the MSE (solid line) is minimum.

data is oversmoothed, the variance term is small but the bias is large. When undersmoothing, occurs the opposite. So, minimizing the MSE entails balancing bias and variance, which is called the *bias-variance tradeoff*.

Usually, a global accuracy measure over the entire interval of definition of f is needed. A global measure of accuracy can be obtained by integrating the SE , leading to the integrated squared error,

$$ISE\left[\hat{f}_n\right] = \int \left[\hat{f}_n\left(u\right) - f\left(u\right)\right]^2 du, \tag{1.2}$$

although this measure is a random quantity, since it depends on each sample. For overcoming this situation, the mean of the ISE is also of interest as a global measure of accuracy:

$$MISE\left[\hat{f}_n\right] = \mathbb{E}\left\{\int \left[\hat{f}_n\left(u\right) - f\left(u\right)\right]^2 du\right\}. \tag{1.3}$$

By considering (1.1), the $MISE$ can be also expressed as

$$MISE\left[\hat{f}_n\right] = \int \mathbb{B}\left[\hat{f}_n(x)\right]^2 dx + \int \mathbb{V}\left[\hat{f}_n(x)\right] dx. \tag{1.4}$$

The $MISE$ is a widely used measure of overall discrepancy between $\hat{f}_n$ and f. As before, minimizing the $MISE$ entails balancing integrated squared bias and integrated variance. Virtually, all nonparametric density estimators have associated a parameter, a *smoothing parameter,* for controling the amount of smoothing on the data. Figure 1.1 shows the role of the smoothing parameter in balancing squared bias and variance for minimizing the MSE.

It is important to stress that the popularity of the ISE and the $MISE$ is just due to its mathematical simplicity, but there are other error measures that may be more appropiate in some contexts or may have some good interesting properties or interpretations. Let us define the L_p measure as

$$L_p = \left\{ \int \left| \hat{f}_n(x) - f\,(x) \right|^p dx \right\}^{1/p}.$$

In general, results obtained when working with a generic L_p are not greatly different than those working with L_2, although L_1 has received some special focus as it has shown to be outlier resistant, invariant under monotone transformations and having a nice interpretation (Devroye and Györfi, 1985). Nonetheless, the analysis of this measure is quite more complicated. Another interesting measure, especially in the machine learning context, is that of the Kullback-Leibler loss:

$$L_{KL} = \int f\,(x) \ln \left[\frac{f(x)}{\hat{f}_n(x)} \right] dx.$$

However, it is generally not recomended to use it in nonparametric density estimation, since it is extremely sensitive to the tails of the distribution (Kullback and Leibler, 1951; Hall, 1987).

Another appealing perspective on how to measure the difference between $\hat{f}_n$ and f is the visual error criteria (Marron and Tsybakov, 1995). The main argument of these authors is that the usual norms on function spaces measure something different from what we perceive in a plot. These norms basically measure vertical distances between the estimate and the target function, while the eye uses both vertical and horizontal information.

Sometimes, the MSE and $MISE$ may depend on the smoothing parameter in such a complicated way that makes it difficult to understand the influence of this parameter on the performance of the estimator. For overcoming this situation, a very useful approach is to consider a large sample approximation of the MSE and $MISE$ considering asymptotic expansions of the bias and variance, and ultimately analizing just the leading terms, which are called asymptotic MSE and $MISE$ ($AMSE$ and $AMISE$, respectively) . For this, the asymptotic notation and Taylor expansions are of great importance (see Appendix B).

1.3.2 Histograms

The histogram is perhaps the oldest and simplest to use nonparametric density estimator. It is very popular for summarizing large data sets and for giving a general impression of the shape and spread of the distribution. The histogram is usually formed by dividing the data range into equally sized bins, and then dividing the proportion of observations at each bin by the binwidth.

Formally, and without loss of generality, let us suppose that f is defined on the interval $[0,1]$. Let k be an integer and define bins $B_1 = \left[0, \frac{1}{k}\right)$, $B_2 = \left[\frac{1}{k}, \frac{2}{k}\right)$, ... , $B_k = \left[\frac{k-1}{k}, 1\right]$, for which the bindwidth is defined as $h = 1/k$. Let n_i be the number of observations in the i-th bin, such that $\sum_{i=1}^{n} n_i = n$. Let also $\hat{p}_i = n_i/n$ represent the observed proportion of data in the i-th bin and $p_i = \int_{B_i} f(u)du$. Then, the histogram estimator is defined as

$$\hat{f}_H(x) = \begin{cases} \frac{\hat{p}_1}{h} & x \in B_1 \\ \frac{\hat{p}_2}{h} & x \in B_2 \\ \vdots & \vdots \\ \frac{\hat{p}_k}{h} & x \in B_k \end{cases},$$

or, succintly,

$$\hat{f}_H(x) = \sum_{i=1}^{k} \frac{\hat{p}_i}{h} I_{B_i}(x),$$

where $I_B(x)$ is the indicator function[4].

To get an insight of the motivation of the histogram, note that following the definition and considering $x \in B_i$, its expectation is

$$\mathbb{E}\left[\hat{f}_H(x)\right] = \frac{\mathbb{E}\left[\hat{p}_i\right]}{h} = \frac{p_i}{h} = \frac{\int_{B_i} f(u)du}{h}.$$

Now, for a small binwidth, $\int_{B_i} f(u)du \approx f(x)h$, so that

$$\mathbb{E}\left[\hat{f}_H(x)\right] \approx \frac{f(x)h}{h} = f(x).$$

Changing the binwidth h (or alternatively, the number of bins k) will have an effect on how smooth the histogram looks. Figure 1.2 shows this effect: a large binwidth (just a few bins), as in (a), produce histograms not so variable that in general tend to be flat. On the opposite side, a small binwidth (a large number of bins), as in (c), produce highly variable histograms with a lot of bumps. That is to say, the binwidth is the histogram's *smoothing parameter*, since it controls the amount of smoothing on the data. From Figure 1.2, it seems that (b) could be an equilibrated choice of binwidth.

Another important consideration when constructing a histogram is the location of the interval limits (also called *breaks*), as it usually affects the shape of the estimation curve. A comparison of Figure 1.2 and Figure 1.3 shows how variable the histogram can be just

[4] $I_B(x) = 1$ if $x \in B$, and 0 otherwise.

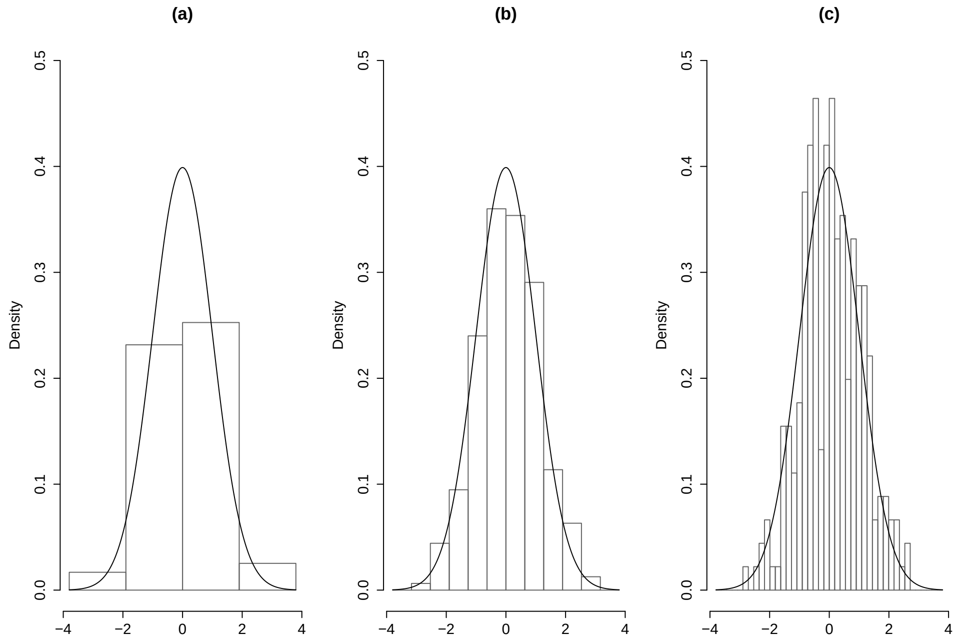

Figure 1.2: Histograms of a random sample of size 250 from a $N(0,1)$ (solid line), based on (a) 4 bins, (b) 12 bins, (c) 42 bins.

by changing the placement of the interval limits, suggesting different shapes even when using the same sample and the same binwidth.

The dependency of the histogram on the location of the breaks is one of its main disadvantages. Another drawback is that it estimates all densities by a step function, while most of them are smooth. A natural solution to the problem of the bin location is to average shifted histograms (Scott, 1985), although eventually it aproximates the kernel density estimator (Härdle, 1991; Härdle and Scott, 1992), which is smoother and uses the data more efficiently.

1.3.3 The empirical distribution function

The (cumulative) distribution function F is another point of view for describing structure in a data set. Although they are strictly different problems, density and distribution estimation are closely related as both functions are linked by the relationship $F' = f$ [5].

Let us consider a sample $(X_1, X_2, ..., X_n)$. A natural estimator of F is the empirical distribution function $\hat{F}_n$,

[5] A distribution function F is absolutely continuous if there is a function f such that $F(x) = \int_{-\infty}^{x} f(u)\,du$. The function $f(x)$ is called a probability density of the random variable X. Then, due to the properties of the integral, $F'(x) = f(x)$ at the points of continuity of f.

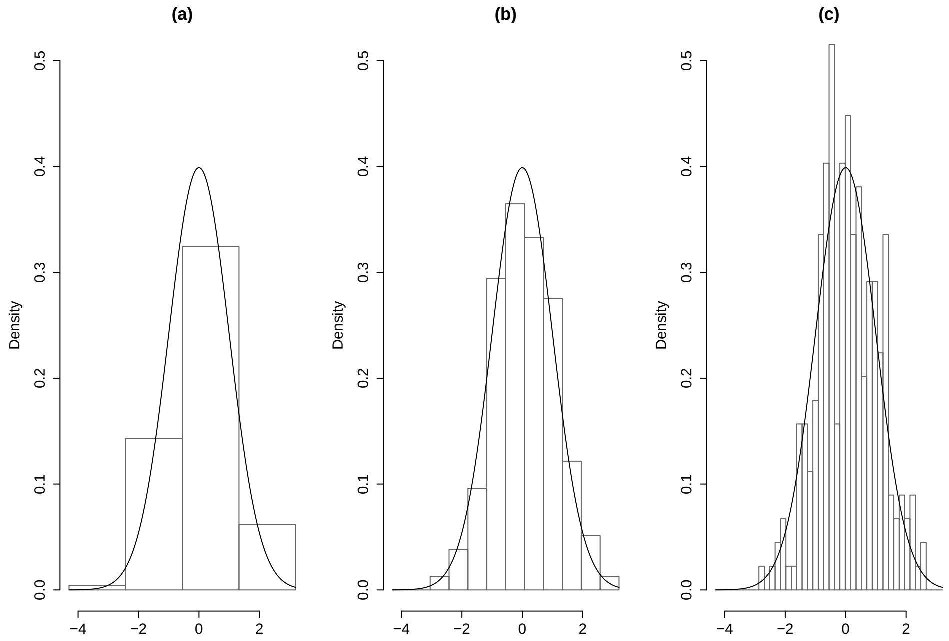

Figure 1.3: Histograms of a random sample of size 250 from a $N(0,1)$ (solid line) based on (a) 4 bins, (b) 12 bins, (c) 42 bins. The breaks were shifted by 0.5 units with respect to those in Figure 1.2.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i), \tag{1.5}$$

which is the distribution that puts mass $1/n$ on each data point. Figure 1.4 shows $\hat{F}_n(x)$ based on a random sample from a $N(0,1)$.

The empirical distribution function has some good properties. For example, it can be proved that at any fixed value x, the mean and variance of $\hat{F}_n$ is

$$\mathbb{E}\left[\hat{F}_n(x)\right] = F(x)$$

and

$$\mathbb{V}\left[\hat{F}_n(x)\right] = \frac{1}{n}F(x)\left[1 - F(x)\right],$$

respectively. Thus, the MSE goes to zero as n increases, and using Chebyshev's inequality, it can also be proved that $\hat{F}_n(x) \overset{p}{\to} F(x)$.

Although the last is an interesting and a desirable property, the Glivenko-Cantelli theorem goes further and gives a much stronger result, stating that

$$\sup_x \left|\hat{F}_n(x) - F(x)\right| \overset{a.s.}{\to} 0;$$

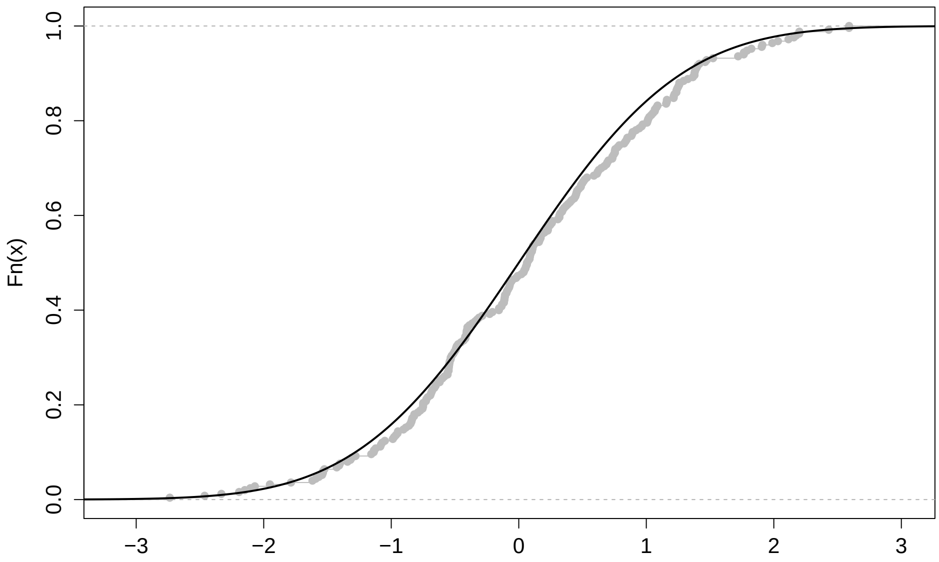

Figure 1.4: Emprical distribution function of a random sample of size 250 from a $N(0,1)$.

i.e., the empirical distribution function converges almost surely in probability to the true value everywhere, as the maximum gap between the two of them goes to zero as n increases. Despite those nice properties of $\hat{F}_n$, one disadvantage is that sometimes it does not translate well into a probability density f. Sorting the sample into increasing order, it assigns probability zero for values between consecutive observations. That could be right, of course, but we can always expect to have some new observations between the previous ones. So, at the end, even though $\hat{F}_n$ is already smooth to some extent, further degree of smoothing can be an advantage, and kernel estimation applied to distribution function gives that extra smoothing.

1.4 Summary

In this chapter, the problem that gave rise to this research was presented: to estimate the probability density function of weed emergence, considering that, by the experimental conditions and the very nature of the random variables used, the data collected are grouped. For better understanding the context, it was also given a basic review of some of the most important weed science concepts.

As it was referenced, this problem is not unique to weed science, but it is shared in various areas of knowledge. Therefore, in trying to find a solution to this problem, the statistical techniques that are proposed in this work will be quite general, so they will be applicable to any set of grouped data, regardless of their origin.

The available statistical tools for estimating the density and the distribution functions implicitly consider that the data at hand is accurate, in the sense that when gathering the data, the uncertainty attached to the measurement process was negligible. But, as it was seen in this chapter, sometimes that is not the case, and those available tools cannot be used with grouped data.

A quick review of some of the typical nonparametric density and distribution estimators, such as the histogram and the empirical distribution function, was given. While, by its simplicity, these estimators have certain implementation advantages, they also have special disadvantages, like providing estimates that are not smooth enough, or their inefficiency when using the data. In this sense, kernel estimators are an improvement in estimating the density and distribution functions, using data more efficiently and providing some extra smoothness, which is usually convenient. Moreover, kernel estimators are very intuitive and its mathematical treatment is relatively easy.

The next chapter will give an overview of kernel density and distribution estimation and its main features and advantages. This will lay the foundation for the main objective of this thesis: a modification of the kernel density estimator for grouped data, which will subsequently allow to obtain a kernel estimator for the distribution with grouped data.

Chapter 2

Kernel estimation of the density and distribution functions

As was shown in Section 1.3, tools like the histogram and the empirical distribution function are informative but have some limitations, like being not smooth or sensitive enough to local properties of the density f or the distribution F. Kernel estimation is an easy and attractive way to solve those problems, while its simplicity allows to mathematically study its properties in detail. This chapter gives a quick review of kernel density and distribution estimation. The important topic of bandwidth selection will also be discussed.

2.1 Kernel density estimation

Given a random sample $(X_1, X_2, ..., X_n)$ coming from an unknown density f, the *kernel density estimator* is defined as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(x - X_i\right), \tag{2.1}$$

where $K_h\left(u\right) = \frac{1}{h} K\left(\frac{u}{h}\right)$, K is a function called *kernel* and $h > 0$ is the smoothing parameter or *bandwidth*. The kernel K could be any smooth function, usually such that $K(x) \geqslant 0$ and

Condition 2.1. $\int K\left(x\right) dx = \mu_0\left(K\right) = 1$

Condition 2.2. $\int xK\left(x\right) dx = \mu_1\left(K\right) = 0$

Condition 2.3. $\int x^2 K\left(x\right) dx = \mu_2\left(K\right) = \iota^2 < \infty$

Requiring Condition 2.1 (i.e., that K must be a density function) assures that the resulting estimation is a density funtion as well, while requiring Condition 2.2 comes from implicitly assuming that the kernel is symmetric, i.e., $K\left(-x\right) = K\left(x\right)$. Condition 2.3 is just a finite second moment assumption.

The basic principles of kernel estimation date back to the 1950s, to the seminal works of Fix and Hodges (1951) and Akaike (1954). Nevertheless, Murray Rosenblatt and Emanuel Parzen are credited for kernel smoothing as it is in its current form (Rosenblatt, 1956; Parzen, 1962). Since then, there have been written several good books on the subject. See, for instance, Silverman (1986); Scott (1992); Wand and Jones (1995). Most of the mathematical derivation that is to come over the next pages can be verified in any of those references.

2.1.1 Exact MSE and MISE calculations

If one wants to evaluate the performance of an estimator, whether locally or globally, some error measure between the estimation and the target function is necessary. According to the arguments presented in Subsection 1.3.1, in this dissertation, expressions (1.1) and (1.3) wil be considered for evaluating the performance of (2.1).

Let us first consider the local case. To compute the MSE of (2.1), its bias and variance are needed. Applying the expectation operator to (2.1),

$$\mathbb{E}\left[\hat{f}_h\left(x\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K_h\left(x - X_i\right)\right] = \mathbb{E}\left[K_h\left(x - X\right)\right]. \tag{2.2}$$

By definition of expectation and using the convolution notation, the rightmost expression can be written as

$$\mathbb{E}\left[\hat{f}_h\left(x\right)\right] = \int K_h\left(x-u\right)f(u)du = \left(K_h * f\right)(x), \tag{2.3}$$

so the bias of (2.1) is

$$\mathbb{B}\left[\hat{f}_h\left(x\right)\right] = \mathbb{E}\left[\hat{f}_h\left(x\right)\right] - f\left(x\right) = \left(K_h * f\right)(x) - f\left(x\right). \tag{2.4}$$

Doing similar calculations, the variance of (2.1) is

$$\mathbb{V}\left[\hat{f}_h\left(x\right)\right] = \frac{1}{n}\left[\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right]. \tag{2.5}$$

Combining (2.4) and (2.5) it is obtained the MSE:

$$MSE\left[\hat{f}_h\left(x\right)\right] = \left[\left(K_h * f\right)(x) - f\left(x\right)\right]^2 + \frac{1}{n}\left[\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right]. \tag{2.6}$$

Now, considering the global case, note that changing the order of integration in (1.3) leads us to

$$MISE\left[\hat{f}_h\right] = \int \mathbb{E}\left[\hat{f}_h(x) - f(x)\right]^2 dx = \int MSE\left[\hat{f}_h\left(x\right)\right]dx,$$

and by (2.6), it follows that

$$MISE\left[\hat{f}_h\right] = \int \left[\left(K_h * f\right)(x) - f\left(x\right)\right]^2 dx + \frac{1}{n}\int \left[\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right] dx,$$

which can be modified to a more tractable form,

$$\begin{aligned} MISE\left[\hat{f}_h\right] &= \frac{1}{nh}\int K^2(x)\,dx + \left(1 - \frac{1}{n}\right)\int \left(K_h * f\right)^2(x)\,dx \\ &\quad -2\int \left(K_h * f\right)(x) f(x)\,dx + A(f), \end{aligned} \tag{2.7}$$

where, for any square integrable function ϱ, $A(\varrho) = \int \varrho^2(x)\,dx$.

An optimal bandwith h can be obtained by minimizing Equation (2.7). Nevertheless, although (2.7) is a nice and compact expression, it has the downside that it depends on the bandwidth h in a complicated way. For this, the large sample approximation of the $MISE$, the $AMISE$, commented in Subsection 1.3.1, is of great value, as it depends on h in a very simple form.

2.1.2 Asymptotic approximations of the MSE and MISE

In this subsection, a large sample approximation of the $MISE$ will be obtained. This approximation enables to observe a direct dependency of the $MISE$ on the bandwidth h, which is very helpful for choosing this parameter for an optimal performance of (2.1). Before starting, some assumptions are neccesary.

Assumption 2.1. *The density f is such that its second derivative f'' is continuous, square integrable and ultimately monotone.*

Assumption 2.2. *The bandwidth $h = h_n$ (in what follows just h) is a non-random sequence of positive numbers. Also, h approaches to zero slower than n goes to infinity, that is to say, $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.*

Assumption 2.3. *The kernel K is a probability density function such that it has finite fourth moment and it is symmetric about the origin.*

Using (2.3) and the change of variable $z = \frac{x-u}{h}$,

$$\mathbb{E}\left[\hat{f}_h(x)\right] = \int K(z) f(x - hz)\,dz.$$

By a Taylor series about x,

$$\mathbb{E}\left[\hat{f}_h(x)\right] = \int K(z)\left[f(x) - hzf'(x) + \frac{1}{2}h^2 z^2 f''(x) + o\left(h^2\right)\right] dz.$$

Using Assumptions 2.1 to 2.3,

$$\mathbb{E}\left[\hat{f}_h\left(x\right)\right] = f\left(x\right) + \frac{1}{2}h^2 f''\left(x\right)\mu_2\left(K\right) + o\left(h^2\right),$$

so the bias is

$$\mathbb{B}\left[\hat{f}_h\left(x\right)\right] = \frac{1}{2}h^2 f''\left(x\right)\mu_2\left(K\right) + o\left(h^2\right). \tag{2.8}$$

Assumption 2.2 ensures that Eq. (2.8) goes to zero as n increases; i.e., the kernel density estimator (2.1) is asymptotically unbiased.

Now, for obtaining an asymptotic expression for the variance, consider Eq. (2.5). Using a Taylor series about x,

$$\begin{aligned}\mathbb{V}\left[\hat{f}_h\left(x\right)\right] &= \frac{1}{nh}\int K^2\left(z\right) f\left(x - hz\right) dz - \frac{1}{n}\mathbb{E}\left[\hat{f}_h\left(x\right)\right]^2 \\ &= \frac{1}{nh}\int K^2(z)\left[f(x) + o(1)\right] dz - \frac{1}{n}\left[f(x) + o(1)\right]^2 \\ &= \frac{1}{nh} f(x) A\left(K\right) + o\left(\frac{1}{nh}\right). \end{aligned} \tag{2.9}$$

As the variance is an $O\left(\frac{1}{nh}\right)$, Assumption 2.2 guarantees that it asymptotically converges to zero.

Squaring (2.8) and adding (2.9),

$$MSE\left[\hat{f}_h\left(x\right)\right] = AMSE\left[\hat{f}_h\left(x\right)\right] + o\left(h^4 + \frac{1}{nh}\right), \tag{2.10}$$

where

$$AMSE = \frac{1}{4}h^4\mu_2(K)^2 f''\left(x\right)^2 + \frac{1}{nh}f(x)A(K)$$

is the so-called asymptotic MSE. Considering Assumption 2.1 and integrating (2.10),

$$MISE\left[\hat{f}_h\right] = AMISE\left[\hat{f}_h\right] + o\left(h^4 + \frac{1}{nh}\right),$$

where

$$AMISE\left[\hat{f}_h\right] = \frac{1}{4}h^4\mu_2(K)^2 A\left(f''\right) + \frac{1}{nh}A\left(K\right) \tag{2.11}$$

is the asymptotic $MISE$, a useful large sample approximation to the $MISE$ since, unlike expression (2.7), its dependency on the bandwidth h is pretty straightforward.

The bias-variance tradeoff is also clear from (2.11). On the one hand, the bandwidth h must be small to reduce the bias term; however, if h is small, the variance term increases, since it depends on $(nh)^{-1}$. Thus, to minimize the $AMISE$, h must be such that each of $AMISE$'s terms is smaller as n increases. This shows the important role of the bandwidth on the performance of (2.1).

Another strong point of the $AMISE$ is that it enables to easily find an asymptotically optimal choice for h. Differentiating (2.11) with respect to h, equating to zero and solving for h, it is obtained

$$h_{AMISE} = \left[\frac{A(K)}{\mu_2(K)^2 A(f'')\, n}\right]^{\frac{1}{5}}. \tag{2.12}$$

Besides that h_{AMISE} depends on the choice of the kernel K (which entirely depends on the user), it also inversely depends on $A(f'')$. This is a problem, since $A(f'')$ depends on f'', the second derivative of f, the unknown function to estimate. Despite that, $A(f'')$ allows to appreciate the effect of the curvature on the optimal bandwidth. The quantity $|f''(x)|$ measures the curvature at a point x and $A(f'')$ is a measure of the total curvature of f. That means that if the function f has little curvature, $A(f'')$ is small and large values for h_{AMISE} are required. If the function has too much curvature, then $A(f'')$ is large and small values for h_{AMISE} are then required. Some rules for estimating $A(f'')$ will be given later on.

Substituting the expression for the optimal bandwith h_{AMISE}, Eq. (2.12), in (2.11), it is obtained

$$\inf_h AMISE\left[\hat{f}_h\right] = \frac{5}{4}\left[\mu_2(K)^2 A(K)^4 A\left(f''\right)\right]^{\frac{1}{5}} n^{-\frac{4}{5}}, \tag{2.13}$$

which is the smallest $AMISE$ when using the kernel K for estimating f.

Summarizing and expressing the information from (2.12) and (2.13) in terms of the $MISE$, it may be said that

$$h_{MISE} \sim \left[\frac{A(K)}{\mu_2(K)^2 A(f'')\, n}\right]^{\frac{1}{5}} \tag{2.14}$$

and

$$\inf_h MISE\left[\hat{f}_h\right] \sim \frac{5}{4}\left[\mu_2(K)^2 A(K)^4 A\left(f''\right)\right]^{\frac{1}{5}} n^{-\frac{4}{5}}, \tag{2.15}$$

where h_{MISE} is the bandwidth that minimizes the $MISE$.

Expressions (2.14) and (2.15) give, as n increases, the rate of convergence to zero for the optimal bandwidth h_{MISE} and the minimum $MISE$. So, according to the stated assumptions, the best rate of convergence of the $MISE$ of (2.1) is $n^{-\frac{4}{5}}$.

It is in this sense that the kernel estimator is more efficient than the histogram, as it was already mentioned in Subsection 1.3.2 . If f' is absolutely continuous and $A(f') < \infty$, it can be shown that

Kernel	Form	Inefficiency
Epanechnikov	$K(x) = \frac{3}{4}\left(1-x^2\right) I_{[-1,1]}\left(x\right)$	1
Biweight	$K(x) = \frac{15}{16}\left(1-x^2\right)^2 I_{[-1,1]}(x)$	1.0061
Triweight	$K(x) = \frac{35}{32}\left(1-x^2\right)^3 I_{[-1,1]}(x)$	1.0135
Gaussian	$K\left(x\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right]$	1.0513
Uniform	$K\left(x\right) = \frac{1}{2} I_{[-1,1]}\left(x\right)$	1.0758

Table 2.1: Different types of kernel functions and their inefficiency.

$$AMISE\left[\hat{f}_H\right] = \frac{1}{12} h_H^2 A\left(f'\right) + \frac{1}{nh_H}, \tag{2.16}$$

$$h_{H_{MISE}} \sim \left[\frac{6}{A\left(f'\right)}\right]^{\frac{1}{3}} n^{-\frac{1}{3}} \tag{2.17}$$

and

$$\inf_{h_H} MISE\left[\hat{f}_H\right] \sim \frac{1}{4}\left[36A\left(f'\right)\right]^{\frac{1}{3}} n^{-\frac{2}{3}}, \tag{2.18}$$

where $AMISE\left[\hat{f}_H\right]$ is the $AMISE$ for the histogram and h_H represents the binwidth (Scott, 1979).

Apart from the fact that the integrated squared bias term of the histogram is $O\left(h_H^2\right)$, which is larger than $O\left(h^4\right)$ for the kernel estimator (compare (2.11) and (2.16)), it follows from (2.17) and (2.18) that choosing the binwidth optimally, the $MISE$ for the histogram decreases to zero at rate $n^{-\frac{2}{3}}$, while for the kernel estimator is $n^{-\frac{4}{5}}$. Thus, the kernel estimator is superior to the histogram in terms of asymptotic efficiency.

Compared with parametric estimates, in which we would expect for the $MISE$ a rate of convergence $O\left(n^{-1}\right)$, the rate $O\left(n^{-\frac{4}{5}}\right)$ is just a minor price to pay. Recall that the kernel estimator works for each density f that is twice continuously differentiable, while the parametric estimator fails if the true density simply does not belong to the assumed model. Moreover, it can be proved that, under the stated assumptions, the rate $O\left(n^{-\frac{4}{5}}\right)$ for the kernel estimator is the best possible (see, for instance, Chapter 24 of Van der Vaart (1998)).

2.1.3 Choosing the kernel K

As mentioned at the begining of Section 2.1, the kernel K is tipically chosen such that it is a probability density function and Conditions 2.2 and 2.3 hold. With these assumptions, it was proved that the best rate of convergence for the kernel estimator is $O\left(n^{-\frac{4}{5}}\right)$, outperforming the histogram. When K is supposed to be a density function, it necessarily follows Condition 2.3, that is, $\mu_2(K) < \infty$. However, it is possible to improve that convergence rate by allowing the kernel K to be negative for some values, making possible to build a

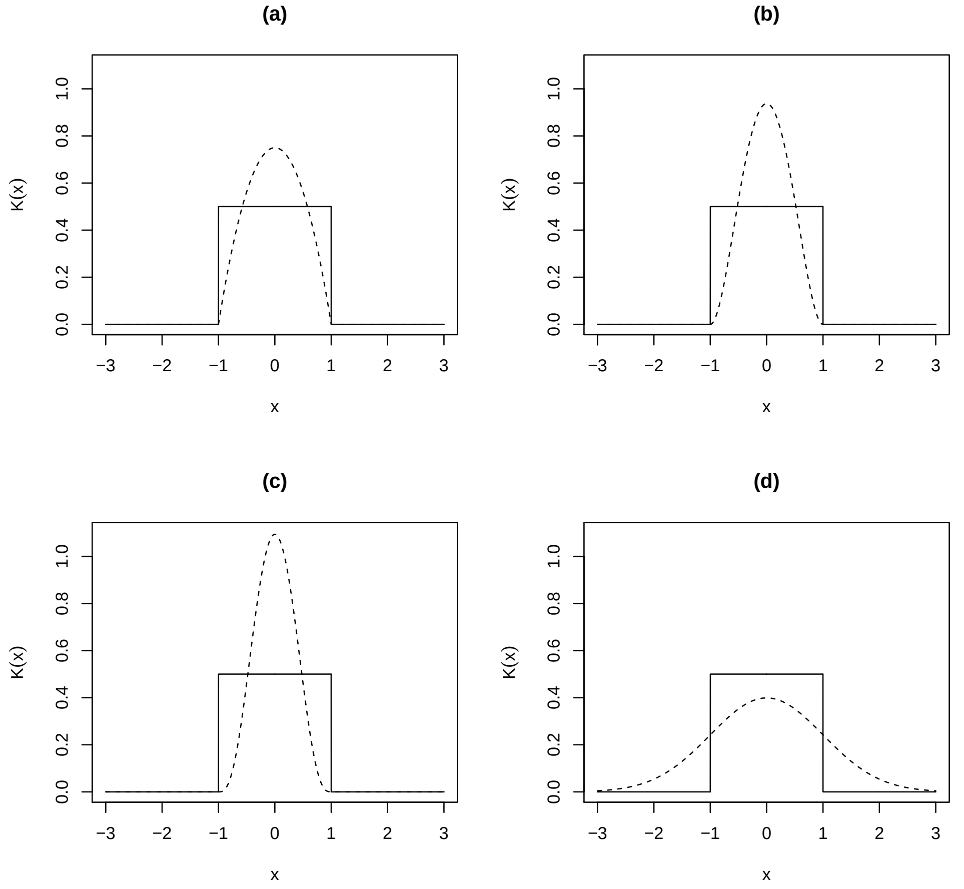

Figure 2.1: Plots for some commonly used kernels: (a) Epanechnikov, (b) Biweight, (c) Triweight, (d) Gaussian. As a reference, all four kernels are imposed over the box-shape uniform kernel.

kernel such that $\mu_2(K) = 0$, with the effect of reducing the bias.

The last idea can be generalized. K is said to be an r-th order kernel if

Condition 2.4. $\mu_0(K) = 1$

Condition 2.5. $\mu_j(K) = 0$ *for* $j = 1, 2, ..., r - 1$

Condition 2.6. $\mu_r(K) \neq 0$

Note that considering that K be symmetric implies that r is even.

Although improving the rate of convergence seems as a good idea, it is not recommended, since the density restriction for K ensures that the estimate will also be a density. For more details on these higher order kernels, see, for instance, Wand and Schucany (1990) and Wand and Jones (1995).

From Subsection 2.1.2, it is clear that the smallest $AMISE$ for the kernel estimator also depends on the choice of K (see Eq. (2.13)). As K is under control of the user, it is

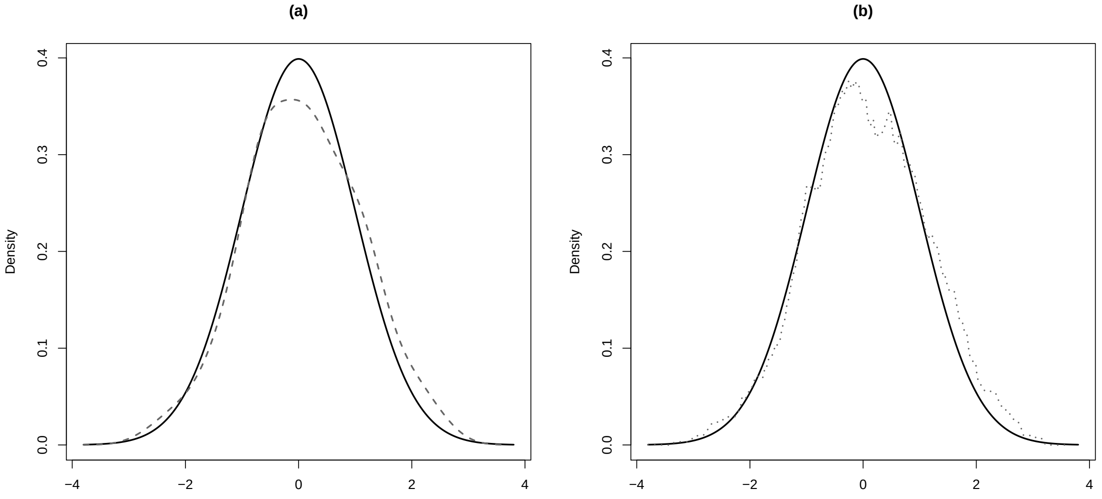

Figure 2.2: Kernel density estimation for a sample of size 250 from a $N(0,1)$ (solid line) using (a) the gaussian kernel, (b) the uniform kernel.

reasonable to ask how to choose K in the best way, in the sense that K should minimize $\kappa(K) = \left[\mu_2\left(K\right)^2 A\left(K\right)^4\right]^{\frac{1}{5}}$. It can be proved that the kernel that minimizes $\kappa\left(K\right)$ is

$$K_\iota^*\left(x\right) = \begin{cases} \frac{3}{4\sqrt{5}\iota}\left[1 - \frac{x^2}{5\iota^2}\right] & x \in \left[-\sqrt{5}\iota, +\sqrt{5}\iota\right] \\ 0 & x \notin \left[-\sqrt{5}\iota, +\sqrt{5}\iota\right] \end{cases},$$

where ι is an arbitrary scale parameter (Hodges and Lehman, 1956). The simplest version is the so-called Epanechnikov kernel, K_E, attained when $\iota^2 = 1/5$.

Since the Epanechnikov kernel is the most efficient one, the inefficiency of any other generic kernel K can be evaluated by comparing $\kappa\left(K\right)$ with $\kappa\left(K_E\right)$. This is typically done by means of the ratio $\left[\kappa\left(K\right)/\kappa\left(K_E\right)\right]^{\frac{5}{4}}$.

Some frequently used kernels and its inefficiency are shown in Table 2.1. Based on that information, it can be concluded that the choice of the kernel is not of much importance as they all perform about the same. That means that the choice of the kernel can be made based on other criteria such as ease of implementation. Apparently, the uniform kernel seems to be the simplest one; however, other kernels are prefered in practice. To see why, let us take a look at Figure 2.2, which shows the density estimation of a random sample from a $N(0,1)$. One of the estimations was made considering the uniform kernel and the other one using the Gaussian kernel, shown on the bottom right in Figure 2.1.

When using the uniform kernel, the resulting estimate is somewhat irregular; it does not look like the kind of function that intuitively it would be called smooth. On the other hand, the estimate using the Gaussian kernel is noticeably smoother. This is because the resulting estimate inherits the continuity and differentiability of the kernel used. Moreover,

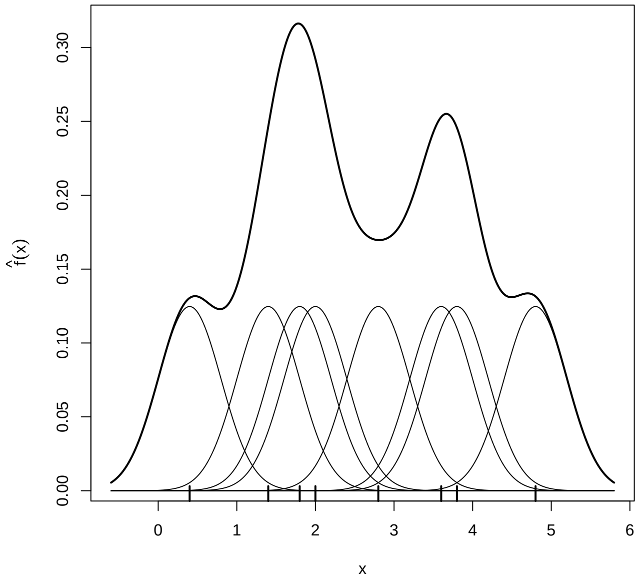

Figure 2.3: Kernel estimation showing the contributions of Gaussian kernels at each data point of the data set (0.4, 1.4, 1.8, 2.0, 2.8, 3.6, 3.8, 4.8). The data set was artificially created only to exemplify how the kernel estimator works. An arbitrary bandwidth $h = 0.4$ was also used.

although the Epanechnikov kernel has some good theoretical properties, it also has a practical disadvantage: it is not everywhere differentiable. This would entail that the estimation would not be everywhere differentiable as well. Thus, in practice, it is preferable not to use the uniform nor the Epanechnikov kernel, but some other smoother kernel. In that sense, the Gaussian kernel is typically the most used.

Figure 2.3 shows the way in which the kernel estimator operates. The estimator is a sum of bumps located at each observation. The kernel used determines the specific form of the bumps, while the bandwidth h determines its amplitude. Figure 2.3 also shows the density estimation as a result of adding all individual Gaussian bumps.

2.1.4 On how difficult a density is to estimate

For the kernel estimator, some densities are easier to estimate than others. This is mainly because the kernel estimator uses just one global single smoothing parameter all over the entire real line. Difficulties appear when there are some noticeably high density zones, for which a relatively small bandwidth h may be adequate for having good estimates, but it may give very wiggly estimates in zones where the data are more sparse. On the contrary, a relatively large value of h could give good estimates in low density zones, but it will give oversmoothed estimates in zones with high density data.

Density	Inefficiency
Triweight	1
Normal	1.101
Bimodal 1	1.761
Gamma(3)	3.058
Kurtotic Unimodal	8.772

Table 2.2: Inefficiency $D\left(f\right)/D\left(f_{TRW}\right)$ for several densities. Bimodal 1 is the density $\frac{3}{4}N\left(0,1\right)+\frac{1}{4}N\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)$. Kurtotic unimodal is the density $\frac{2}{3}N\left(0,1\right)+\frac{1}{3}N\left(0,100\right)$.

More formally, how well a particular density can be estimated is related with its curvature. As stated in Subsection 2.1.2, the functional $A\left(f''\right)$ is a measure of the total curvature of f, so its magnitude tells how well a density f can be estimated. High curvatures, i.e., large values of $A\left(f''\right)$, are obtained when f has features like high skewness or several modes. In this cases, kernel estimation becomes more difficult than in cases in which these features are not present.

It should be noted that $A\left(f''\right)$ is not scale invariant, so any value of $A\left(f''\right)$ can be obtained just by changing the scale. It is easy to see that

$$D\left(f\right)=\left[\sigma\left(f\right)^5 A\left(f''\right)\right]^{\frac{1}{4}}, \tag{2.19}$$

where $\sigma\left(f\right)$ is the population standard deviation, is a scale invariant degree of difficulty measure of kernel estimation of f. It can be shown that (2.19) is minimal for f being the $\beta(4,4)$ function, f_{TRW}, also known as the triweight density (see Table 2.1 and plot (c) in Figure 2.1), and its minimum value is 35/243 (Terrel, 1990).

As it was done in Subsection 2.1.3 for comapring efficiency among different kernel functions, the same can be done to compare the performance of $\hat{f}_h$ when estimating a generic density f, using $D\left(f_{TRW}\right)$ as a reference by means of the ratio $D\left(f\right)/D\left(f_{TRW}\right)$.

It is clear from Table 2.2 that densities close to normality are the easiest to estimate for the kernel estimator, and features such as skewness, kurtosis or the existence of high density zones make the estimation more difficult. For example, the table shows that, in a sense, the kurtotic unimodal density is almost nine times more difficult to estimate than the triweight density.

2.1.5 Some modifications of the kernel density estimator

As seen in Subsection 2.1.4, a disadvantage of the kernel density estimator is that it should give an accurate density estimation by means of a single smoothing parameter. This can be an important problem in cases where there are relatively high density zones, such as in densities with high skewness or high kurtosis values. Thus, a natural idea for overcomming this problem is to consider local bandwidths, giving the right or optimal amount of smoothing at each estimation point. As a result, the kernel density estimator is expressed as

$$\hat{f}_L\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\left(x\right)}K\left[\frac{x-X_i}{h\left(x\right)}\right], \tag{2.20}$$

known as the *local kernel density estimator*. Note that h is expressed as a function of the point of estimation x. This implies that at two different points of estimation, x_1, x_2, the corresponding estimations arise from using the same kernel K (for instance, a Gaussian kernel) but different scale parameters $h\left(x_1\right)$, $h\left(x_2\right)$, at each point. A consequence of the dependency of h on x is that, in general, the whole estimation $\hat{f}_L$ is not a density itself, since it needs not to integrate out to one.

Nevertheless, and provided that $f''\left(x\right)\neq 0$, the analogue of (2.12) at the point x is

$$h_{AMISE}\left(x\right)=\left[\frac{A\left(K\right)f\left(x\right)}{\mu_2\left(K\right)^2 f''\left(x\right)^2 n}\right]^{\frac{1}{5}}. \tag{2.21}$$

Choosing h optimally at each x according to (2.21), it can be shown that

$$AMISE\left[\hat{f}_L\right]=\frac{5}{4}\left[\mu_2\left(K\right)^2 A\left(K\right)^4\right]^{\frac{1}{5}}A\left[\left(f^2 f''\right)^{\frac{1}{5}}\right]n^{-\frac{4}{5}},$$

so the rate of convergence of $\hat{f}_L$ and $\hat{f}_h$ coincides and there is no improvement in this sense. Notwithstanding, it can also be shown that $A\left[\left(f^2 f''\right)^{\frac{1}{5}}\right]\leqslant A\left(f''\right)^{\frac{1}{5}}$ for all f, so, at the end, there is always some improvement when choosing $h(x)$ optimally.

Before smoothing, Eq. (2.20) needs $h\left(x\right)$ to be selected. The preliminary estimation of $h\left(x\right)$ is known as the pilot estimation. Since we would like to smooth less in high density regions (and more in low density ones), a logical assumption is to consider $h\left(x\right)$ to vary inversely with the density. The *nearest neighbour density estimator* takes distances from x to the point of interest to be the k-th nearest to x (for some reasonable value of k) in a pilot estimation, which is essentially $h\left(x\right)\propto 1/f\left(x\right)$ (Loftsgaarden and Quesenberry, 1965). However, there are some situations where this assumption is inadequate (Wand and Jones, 1995).

A better idea is to consider the bandwidth h to depend on X_i, so that the single h is replaced by n values $h\left(X_i\right)$, $i=1,2,...,n$,

$$\hat{f}_V\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\left(X_i\right)}K\left[\frac{x-X_i}{h\left(X_i\right)}\right].$$

This is called the *variable kernel density estimator*. A good assumption is to consider $h\left(x_i\right)=h_\vartheta f\left(x_i\right)^{-\frac{1}{2}}$, since under suitable assumptions, the bias results of order $O\left(h^4\right)$, instead of the typical $O\left(h^2\right)$, leaving the variance $O\left(n^{-1}h^{-1}\right)$. Also, taking $h_\vartheta=O\left(n^{-\frac{1}{9}}\right)$ gives a better convergence rate of $MSE=O\left(n^{-\frac{8}{9}}\right)$(Abramson, 1982).

Another situation in which a modification of the kernel estimator is needed is when we want to estimate particularly complex functions. A smart approach is to transform the data to a more convenient scale, estimate the density with the transformed data and then transform back to the original scale. This approach is particularly useful to avoid spurious bumpiness in the tails and reduce boundary bias.

Let $f_X(x)$ be the density in the original scale and $f_Y(y)$ is the transformed density of the (transformed) random variable $Y = g(X)$, with g being a monotonic increasing function. Then, a change of variable gives

$$f_X(x) = f_Y[g(x)]\, g'(x)\,.$$

Estimating f_Y using the kernel density estimator, the *transformation-based kernel density estimator* is

$$\hat{f}_X(x) = \frac{g'(x)}{nh_Y}\sum_{i=1}^{n} K\left[\frac{g(x) - g(X_i)}{h_Y}\right],$$

where h_Y is obtained based on the Y scale.

The proper choice of the function g depends largely on the data. One possibility is to choose g as a parametric family. For example, the shifted power family

$$g(x) = \begin{cases} (x+\lambda_1)^{\lambda_2}\,\mathrm{sign}(\lambda_2) & \lambda_2 \neq 0 \\ \ln(x+\lambda_1)\,, & \lambda_2 = 0 \end{cases},$$

where $\lambda_1 > -\min(X)$ and $\min(X)$ represents the lower endpoint of the support of f_X, can be useful for heavily skewed data (Wand et al., 1991). An alternative is to estimate g nonparametrically. If F_X and F_Y are the distribution functions of f_X and f_Y, a well known result is that $F_Y^{-1}(F_X(X))$ has density f_Y. Thus, F_Y can be chosen to correspond to a relatively easy to estimate density and take $g = F_Y^{-1} \circ \hat{F}_X$, where $\hat{F}_X$ is a kernel estimate of F_X (Ruppert and Cline, 1994).

Sometimes, the true density f may have substantial mass close to the boundary. As mentioned before, the kernel estimator depends on just one somoothing parameter, so it is expected to have a poor performance on the boundaries of this kind of densities. This boundary bias can be corrected by means of the so-called *boundary kernels,* which, as the name suggests, are kernels that are only used on the boundary region, using the common kernel K in the interior. This bias correction comes with a cost: an increase in the inherent variability in the process of estimating f on the boundaries. Also, occasionally, the density estimate is not a good one since it does not integrate to one. A solution to this problem is to normalize to force a unit integral. The interested reader should take a look at Jones (1993), Jones and Foster (1996), Marron and Ruppert (1994). More complicated corrective methods are also possible (Chen, 2000; Scaillet, 2004).

2.1.6 Bandwidth selection

As seen in previous sections, the kernel estimator depends on two choices: the kernel function K and the bandwidth h. In Subsection 2.1.3, it was shown that the choice of the kernel is not of much importance, as in terms of efficiency, all the kernels perform about the same. However, the problem of bandwidth selection is crucial in obtaining a good estimate of the density, since a too small bandwidth gives very wiggly estimates, and too large selections tend to provide very flat estimates.

Ideas on how to select the bandwidth date back to the late 1970s. Since then, the literature on the topic has increased considerably and nowadays there are numerous proposals. This subsection gives a brief overview of the most iconic methods that later gave rise to new and more sophisticated bandwidth selectors.

Simple bandwidth selectors

Recall from Eq. (2.12) that the optimal $AMISE$ bandwidth depends not only on the kernel K, but on the unknown quantity $A(f'')$. When f is thought to be very smooth, h_{AMISE} is computed as if f were normal with variance σ^2, which yields

$$h_{AMISE} = \left[\frac{8\pi^{\frac{1}{2}} A(K)}{3\mu_2(K)^2 n}\right]^{\frac{1}{5}} \sigma. \tag{2.22}$$

The expression (2.22) still depends on the kernel K. Assuming that also K is the Gaussian kernel and replacing σ by an estimation $\hat{\sigma}$, it gives

$$h_{NR} = 1.06 n^{-\frac{1}{5}} \hat{\sigma}, \tag{2.23}$$

which is called the *normal reference rule.* Commonly, the scale measure σ is estimated by $\min\{s, Q/1.34\}$, where s is the sample standard deviation and Q is the interquartile range. This is done fundamentally for reducing the chances of oversmoothing, as $Q/1.34$ protects against outliers in case f has heavy tails. Some other scale estimates have also been studied (Janssen et al., 1995). Of course, when the data is close to normal, it may be expected (2.23) to be a good bandwidth selector, but as long as the data depart from normality, this selector tend to oversmooth and to cover up some important details in the data.

Another of the so-called simple bandwidth selectors is based on the maximal smoothing principle. The idea is to consider the largest degree of smoothing according to the estimated scale of the density, so that the value of the $AMISE$ optimal bandwidth will always be equal or less than some upper bound. It can be shown that

$$h_{AMISE} \leqslant \left[\frac{243 A(K)}{35\mu_2(K)^2 n}\right]^{\frac{1}{5}} \sigma$$

for all densities having standard deviation σ. This bound is reached by the $\beta(4,4)$ function (the already mentioned triweight density, on the bottom left panel in Figure 2.1) (Terrel, 1990).

Taking an estimation of σ, the *oversmoothed bandwidth selector* is

$$\hat{h}_{OS} = \left[\frac{243A\left(K\right)}{35\mu_2\left(K\right)^2 n}\right]^{\frac{1}{5}} s, \tag{2.24}$$

where, as before, s is the sample standard deviation.

Obviously, $\hat{h}_{OS}$ is a larger bandwidth than the one needed for an optimal estimation. However, $\hat{h}_{OS}$ is a good starting point, since at least it is known that the optimal value is somewhere below $\hat{h}_{OS}$. Thus, a common strategy to decide the optimal bandwidth is to consider fractions of $\hat{h}_{OS}$ and take a visual inspection on each estimation to see the emerged features.

Both selectors, (2.23) and (2.24), are very similar when based on standard deviation, since

$$\frac{\hat{h}_{NR}}{\hat{h}_{OS}} \approx 0.93.$$

Cross-validation methods

The bandwidth selectors shown previously are just simple and quick rules for selecting the bandwidth. In this part, some more elaborate, automatic and consistent selectors based on the idea of cross-validation are shown.

The first selector of this class is the so-called least squares cross-validation (LSCV) (Rudemo, 1982; Bowman, 1984). Expanding the $MISE$ of $\hat{f}_h$,

$$MISE\left[\hat{f}_h\right] = \mathbb{E}\left[\int \hat{f}_h\left(x\right)^2 dx\right] - 2\mathbb{E}\left[\int \hat{f}_h\left(x\right) f\left(x\right) dx\right] + \int f\left(x\right)^2 dx. \tag{2.25}$$

The last term on the right-hand side does not depend on h. So, the proposal is to choose the bandwidth as the value of h that minimizes the estimate of the other two terms. For this, it can be shown that an unbiased estimator is

$$LSCV\left(h\right) = \int \hat{f}_h\left(x\right)^2 dx - 2\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{h(-i)}\left(X_i\right), \tag{2.26}$$

where $\hat{f}_{h(-i)}\left(X_i\right)$ is the kernel estimation based on the data except the observation X_i. The bandwidth h that minimizes (2.26) is denoted as $\hat{h}_{LSCV}$.

Sometimes (2.26) has more than one local minimum (Hall and Marron, 1991). The recomendation is to use the largest local minimizer of $LSCV\left(h\right)$, as it produces better results than the global minimizer (Marron, 1993). However, the main drawback is that $\hat{h}_{LSCV}$ lacks of stability even when the sample size increases. This gave rise to some more stable modified versions (Chiu, 1991*a*,*b*, 1992).

Another classic cross-validation approach is the biased cross-validation selector (BCV). This method is based on choosing the bandwidth that minimizes the asymptotic $MISE$, Eq.(2.11). Since $A\left(f''\right)$ is unknown, replacing an estimation is necessary.

A natural estimator of $A\left(f''\right)$ is $A\left(\hat{f}''\right)$. However, it can be proved that (Scott and Terrell, 1987)

$$\mathbb{E}\left[A\left(\hat{f}''\right)\right] = A\left(f''\right) + \frac{1}{nh^5}A\left(K''\right) + O\left(h^2\right),$$

from which, an improved estimate of $A\left(f''\right)$is

$$\hat{A}\left(f''\right) = A\left(\hat{f}''_h\right) - \frac{1}{nh^5}A\left(K''\right), \tag{2.27}$$

where the subscript h in $\hat{f}''_h$ means that this bandwidth was used for estimating both the density itself and its second derivative. Substituting (2.27) in (2.11) gives the following objective function

$$BCV\left(h\right) = \frac{1}{nh}A\left(K\right) + \frac{1}{4}h^4\mu_2\left(K\right)^2\left[A\left(\hat{f}''_h\right) - \frac{1}{nh^5}A\left(K''\right)\right],$$

which minimizer is $\hat{h}_{BCV}$.

The advantage of $\hat{h}_{BCV}$ over $\hat{h}_{LSCV}$ is that the former's sampling distribution is less variable; i.e., it is more stable. In this sense, it can be shown that

$$n^{\frac{1}{10}}\left(\frac{\hat{h}_{BCV}}{h_{AMISE}} - 1\right) \tag{2.28}$$

asymptotically converges in distribution to a $N\left(0, \sigma^2_{BCV}\right)$(Scott and Terrell, 1987). Also, a similar result is valid for least squares cross-validation,

$$n^{\frac{1}{10}}\left(\frac{\hat{h}_{LSCV}}{h_{AMISE}} - 1\right), \tag{2.29}$$

which has a $N\left(0, \sigma^2_{LSCV}\right)$ asymptotic distribution (Hall and Marron, 1987*b*; Scott and Terrell, 1987). As specified by Wand and Jones (1995), the ratio of both asymptotic variances is

$$\frac{\sigma^2_{LSCV}}{\sigma^2_{BCV}} \approx 15.7,$$

which evidences that $\hat{h}_{LSCV}$ selections are much more unstable than $\hat{h}_{BCV}$. However, this stability comes with a charge attached, which is an increase in bias. This fact makes $\hat{h}_{BCV}$, on average, somewhat larger than the $MISE$ optimal bandwidth.

Just like $LSCV(h)$, $BCV(h)$ usually has more than one minimum. As before, it is suggested to use the largest minimizer of $BCV\left(h\right)$(Jones et al., 1996a), although it is also recommended to take $\hat{h}_{BCV}$ as the largest local minimizer less than or equal to $\hat{h}_{OS}$ (Scott, 1992).

Finally, a common drawback of both cross-validation methods is their slow $n^{-\frac{1}{10}}$ rate of convergence, as it follows from (2.28) and (2.29).

Plug-in bandwidth selection

Plug-in bandwidth selection is a faster converging method than least squares and biased cross-validation. The approach consists in substituting in (2.12) an estimate of the unknown $A\left(f''\right)$. This idea is thought to date back to the 1970s (Woodroofe, 1970). Before continuing, it is important to have a look at how to estimate density functionals of the general form $A\left[f^{(m)}\right]$.

Let us consider a density functional of the form:

$$A\left[f^{(m)}\right] = \int f^{(m)}\left(x\right)^2 dx.$$

Integrating by parts and under some smoothness assumptions,

$$A\left[f^{(m)}\right] = \left(-1\right)^m \int f^{(2m)}\left(x\right) f\left(x\right) dx.$$

So, it is sufficient to study functionals of the form

$$\psi_u = \int f^{(u)}\left(x\right) f\left(x\right) dx, \tag{2.30}$$

for u even. Note that Eq. (2.30) is just

$$\psi_u = \mathbb{E}\left[f^{(u)}\left(X\right)\right],$$

so a natural estimator of ψ_u is

$$\hat{\psi}_u = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{(u)}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_\eta^{(u)}\left(X_i - X_j\right), \tag{2.31}$$

where η is a bandwidth and L is a kernel, both possibly different from h and K (Hall and Marron, 1987a; Jones and Sheather, 1991).

The following assumptions are needed to obtain the asymptotic properties of (2.31).

Assumption 2.4. *L is a symmetric kernel of order s, $s = 2, 4, ...$, possesing u derivatives such that*

$$(-1)^{\frac{1}{2}(u+s)+1} L^{(u)}(0)\,\mu_s(L) > 0$$

Assumption 2.5. *The density f has p continuous derivatives, each ultimately monotone, and $p > s$.*

Assumption 2.6. *$\eta = \eta_n$ is a sequence of (positive) bandwidths such that* $\lim_{n\to\infty}\eta = 0$ *and* $\lim_{n\to\infty} n\eta^{2u+1} = \infty$.

Under Assumptions 2.4 to 2.6, it can be shown that the asymptotic MSE of (2.31) is

$$\begin{aligned} AMSE\left[\hat{\psi}_u\right] &= \left[\frac{1}{n\eta^{u+1}}L^{(u)}(0) + \frac{1}{s!}\eta^s\mu_s(L)\,\psi_{u+s} + O\left(\eta^{s+2}\right)\right]^2 \\ &+\frac{2}{n^2\eta^{2u+1}}A\left[L^{(u)}\right]\psi_0 + \frac{4}{n}\left[\int f^{(u)}(x)^2 f(x)\,dx - \psi_u^2\right] \\ &+o\left(\frac{1}{n^2\eta^{2u+1}} + \frac{1}{n}\right) \end{aligned}$$

where the first term on the right hand side corresponds to the asymptotic squared bias of (2.31) and the second term is its asymptotic variance. More details can be found in Wand and Jones (1995).

Due to Assumption 2.4, the main bias term can be made to vanish by choosing η as

$$\eta_{AMSE} = \left[\frac{s!L^{(u)}(0)}{-\mu_s(L)\,\psi_{s+u}n}\right]^{\frac{1}{u+s+1}}. \tag{2.32}$$

Eq. (2.32) is of major importance for the plug-in bandwidth selection. As stated before, the basic idea is to plug in (2.12) an estimate of the unknown $A(f'')$. Expressing (2.12) in terms of ψ_u functionals,

$$h_{AMISE} = \left[\frac{A(K)}{\mu_2(K)^2\,\psi_4 n}\right]^{\frac{1}{5}}. \tag{2.33}$$

Now, following the strategy of plugging an estimate of ψ_4,

$$\hat{h}_{DPI} = \left[\frac{A(K)}{\mu_2(K)^2\,\hat{\psi}_4 n}\right]^{\frac{1}{5}}. \tag{2.34}$$

The problem with (2.34) is that it is not totally automatic, since for obtaining $\hat{\psi}_4$, an initial bandwidth η is needed. A possibility for choosing this initial bandwidth is by means of equation (2.32). If the same kernel K is used, then, from (2.32) the optimal $AMSE$ bandwidth is

$$\eta_{AMSE} = \left[\frac{2K^{(4)}(0)}{-\mu_2(K)\psi_6 n} \right]^{\frac{1}{7}}.$$

Clearly, the problem still remains, since estimating ψ_6 will depend on an initial bandwidth, which in turn will depend on ψ_8, and so on. A common strategy is to estimate ψ_u with some quick and simple rule, like the normal scale rule. Once $\hat{\psi}_u$ is obtained, it is possible to select a bandwidth for estimating ψ_{u-2}. Then, having estimated $\hat{\psi}_{u-2}$, a bandwidth for estimating ψ_{u-4} can be selected, and so forth. In general, this is refer to as a υ stages rule, where estimating $\hat{\psi}_u$ is neccessary for getting into the first step. This procedure is called the υ-stage direct plug-in bandwidth selector, $\hat{h}_{DPI,\upsilon}$.

The following result is a useful one for obtaining a prior estimation $\hat{\psi}_u$ (see, e.g., Wand and Jones (1995), Appendix C). If f is a normal density with variance σ^2, then, for u even,

$$\psi_u = \frac{(-1)^{\frac{u}{2}}\, u!}{(2\sigma)^{u+1}\left(\frac{u}{2}\right)!\pi^{\frac{1}{2}}}. \tag{2.35}$$

Another problem to face is to decide the number of stages υ. Some studies suggest that υ should be at least equal to 2, being $\upsilon = 2$ the most common choice (Aldershof, 1991; Park and Marron, 1992). Thus, for a version of a $\upsilon = 2$ stage plug-in bandwidth selector and using $L = K$, with K a second order kernel, the following steps are a possibility (Wand and Jones, 1995; Sheather and Jones, 1991).

1. Estimate ψ_8 using (2.35), substituting σ by $\hat{\sigma}$, an estimate of scale. This gives

$$\hat{\psi}_8 = \frac{105}{32}\pi^{-\frac{1}{2}}\hat{\sigma}^{-9}$$

2. Estimate ψ_6 by means of the kernel estimator and using the bandwidth

$$\eta_1 = \left[\frac{-2K^{(6)}(0)}{\mu_2(K)\hat{\psi}_8 n} \right]^{\frac{1}{9}}$$

3. Estimate ψ_4 by means of the kernel estimator and using the bandwidth

$$\eta_2 = \left[\frac{-2K^{(4)}(0)}{\mu_2(K)\hat{\psi}_6 n} \right]^{\frac{1}{7}}$$

4. The selected bandwidth is

$$\hat{h}_{DPI,2} = \left[\frac{A(K)}{\mu_2(K)^2\hat{\psi}_4 n} \right]^{\frac{1}{5}}.$$

For a more comprehensive review on bandwidth selection methods in kernel density estimation, the reader is referred to (Jones et al., 1996*a*; Heidenreich et al., 2013).

Smoothed bootstrap

This approach consists in taking a bandwidth that minimizes a smoothed bootstrap approximation to the $MISE$. The first versions of this approach were those from Faraway and Jhun (1990) and Taylor (1989). What makes this approach special is that, in the "bootstrap world", the $MISE$ can be calculated exactly. So, computationally speaking, it is quite competitive compared with other bandwidth selectors.

Consider a random sample $(X_1, X_2, \ldots, X_n)$ coming from an unknown density f. As mentioned in Subsection 1.3.1, two popular discrepancy measures are the integrated squared error and its average, which for the kernel density estimator are just

$$ISE\left[\hat{f}_h\right] = \int \left[\hat{f}_h(x) - f(x)\right]^2 dx \tag{2.36}$$

and

$$MISE\left[\hat{f}_h\right] = \mathbb{E}\left\{ISE\left[\hat{f}_h\right]\right\}. \tag{2.37}$$

The bandwidths that are of interest to come close to are those minimizing Eqs. (2.36) and (2.37), denoted by h_{ISE} and h_{MISE}, respectively. What bootstrap does is to imitate the random mechanism from which the original sample was obtained. This is done by replacing the density f by an estimation.

A possible bootstrap procedure to approximate (2.36) is the following:

1. Choose a pilot bandwidth g and consider the kernel density estimator $\hat{f}_g$.

2. Obtain a bootstrap sample $(X_1^\star, X_2^\star, \ldots, X_n^\star)$ from $\hat{f}_g$.

3. For each $h > 0$, consider the bootstrap version of the kernel density estimator:

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

4. Define the bootstrap version of the integrated squared error,

$$ISE^\star\left[\hat{f}_h^\star\right] = \int \left[\hat{f}_h^\star(x) - \hat{f}_g(x)\right]^2 dx, \tag{2.38}$$

which clearly depends on h.

5. The minimizer of (2.38), $h_{ISE^\star}$, is the analogous version of h_{ISE}.

If the target bandwidth were h_{MISE}, then, the mean of the bootstrap process in (2.38) should be considered; i.e.,

$$MISE^{\star}\left[\hat{f}_h^{\star}\right]=\mathbb{E}^{\star}\left[\int\left[\hat{f}_h^{\star}(x)-\hat{f}_g(x)\right]^2dx\right]. \tag{2.39}$$

Some direct calculations lead to (e.g., as in Marron (1992))

$$MISE^{\star}\left[\hat{f}_h^{\star}\right]=\frac{1}{n}\left[\frac{1}{h}A(K)+A\left(K_h*\hat{f}_g\right)\right]+A\left(K_h*\hat{f}_g-\hat{f}_g\right), \tag{2.40}$$

where, as before, * stands for convolution, and which minimizer is denoted by $h_{MISE^{\star}}$, the bootstrap version of h_{MISE}.

Note in (2.40) that the pilot estimation $\hat{f}_g$, made with the pilot bandwidth g, plays the role of the true density f, and also note that (2.40) depends on the original sample, but not on resamples anymore. Compared with other approaches, this is the great valued property of the bootstrap approach: instead of working via the $AMISE$, it targets the $MISE$ itself.

For choosing the pilot bandwith g, there are several proposals that involve stages of pilot estimations or relying on using reference distributions (Jones et al., 1996*b*). Also, Cao (1993) studied the pilot bandwidth selection problem in this context and proved asymptotic properties for $h_{MISE^{\star}}$, the minimizer of $MISE^{\star}$.

2.2 Kernel distribution estimation

As seen in Subsection 1.3.3, the empirical distribution function $\hat{F}_n$ is already, in a sense, a good estimator of F, although more smoothness is always appreciated. By its construction, kernel estimation gives that additional smoothing when estimating F.

Using that $F(y)=\int_{-\infty}^{y}f(z)\,dz$, it is immediate to formulate a kernel estimator for the distribution function as

$$\hat{F}_h(x)=\int_{-\infty}^{x}\hat{f}_h(z)\,dz.$$

Integrating (2.1) gives

$$\hat{F}_h(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{K}\left(\frac{x-X_i}{h}\right), \tag{2.41}$$

where $\mathbb{K}(x)=\int_{-\infty}^{x}K(z)\,dz$. The estimator (2.41) was introduced in the 1960s by Nadaraya (1964).

2.2.1 Asymptotic approximations of the MSE, $MISE$ and the weighted $MISE$

The well known expression for the mean squared error of $\hat{F}_h(x)$ is

$$\begin{aligned} MSE\left[\hat{F}_h(x)\right] &= \mathbb{E}\left[\hat{F}_h(x) - F(x)\right]^2 \\ &= \mathbb{V}\left[\hat{F}_h(x)\right] + \mathbb{B}^2\left[\hat{F}_h(x)\right]. \end{aligned} \tag{2.42}$$

Let us consider the usual assumptions concerning the kernel K and $\mathbb{K}$, and assume that f is continuous, F' exists, $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$. Then, applying the expectation operator to (2.41),

$$\begin{aligned} \mathbb{E}\left[\hat{F}_h(x)\right] = \mathbb{E}\left[\mathbb{K}\left(\frac{x - X_1}{h}\right)\right] &= \int \mathbb{K}\left(\frac{x-y}{h}\right) f(y)\,dy \\ &= F(x) + \frac{1}{2}h^2 f'(x)\,\mu_2(K) + O\left(h^4\right). \end{aligned} \tag{2.43}$$

Considering that $\int K(z)\,\mathbb{K}(z)\,dz = \frac{1}{2}$,

$$\begin{aligned} \mathbb{E}\left\{\left[\mathbb{K}\left(\frac{x - X_1}{h}\right)\right]^2\right\} &= \int \left[\mathbb{K}\left(\frac{x-y}{h}\right)\right]^2 f(y)\,dy \\ &= F(x) - hf(x)\,C_0 + O\left(h^2\right), \end{aligned} \tag{2.44}$$

where $C_0 = 2\int zK(z)\,\mathbb{K}(z)\,dz$. By (2.43) and (2.44), the asymptotic bias and variance are

$$\mathbb{B}\left[\hat{F}_h(x)\right] = \frac{1}{2}h^2 f'(x)\,\mu_2(K) + O\left(h^4\right) \tag{2.45}$$

and

$$\mathbb{V}\left[\hat{F}_h(x)\right] = \frac{1}{n}F(x)\left[1 - F(x)\right] - \frac{h}{n}f(x)\,C_0 + O\left(\frac{h^2}{n}\right), \tag{2.46}$$

respectively. Now, substituting (2.45) and (2.46) in (2.42),

$$AMSE\left[\hat{F}_h(x)\right] = \frac{1}{n}F(x)\left[1 - F(x)\right] - \frac{h}{n}f(x)\,C_0 + \frac{1}{4}h^4 f'(x)^2 \mu_2(K)^2. \tag{2.47}$$

As for defining a global measure of discrepancy between $\hat{F}_h$ and F, by following similar arguments as those presented for density estimation (Subsection 1.3.1), it is defined

$$MISE\left[\hat{F}_h\right] = \int \mathbb{E}\left[\hat{F}_h(x) - F(x)\right]^2 dx. \tag{2.48}$$

Provided that F has two bounded and continuous derivatives, each ultimate monotone at both tails, and assuming it has enough finite moments, it follows from (2.47) and (2.48) that

$$AMISE\left[\hat{F}_h\right] = \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \frac{h}{n}C_0 + \frac{1}{4}h^4\mu_2(K)^2 A\left(f'\right). \tag{2.49}$$

Differentiating with respect to h and equating to zero, the optimal $AMISE$ bandwidth is

$$h_{AMISE_F} = \left(\frac{C_0}{\mu_2(K)^2 A(f')}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{2.50}$$

where the subscript F is a reminder that this is the distribution case.

To evaluate the performance of the $AMISE$ optimal bandwidth, let us substitute (2.50) in (2.49), giving

$$\inf_h AMISE\left[\hat{F}_h\right] = \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \left\{\frac{3C_0^{\frac{4}{3}}}{4\left[\mu_2(K)^2 A(f')\right]^{\frac{1}{3}}}\right\} n^{-\frac{4}{3}}. \tag{2.51}$$

Although (2.48) can be thought of as a natural extension of (2.42), a more general measure can be defined by introducing weights depending on each point of estimation, giving rise to

$$WMISE\left[\hat{F}_h\right] = \int \mathbb{E}\left[\hat{F}_h(x) - F(x)\right]^2 W(x)\,dF(x), \tag{2.52}$$

where $W(x) \geqslant 0$ is a bounded weight function. Considering that F is smooth enough, then, from (2.47) and (2.52),

$$WAMISE\left[\hat{F}_h\right] = \frac{1}{n}\int F(x)\left[1-F(x)\right]W(x)\,dF(x) - \frac{h}{n}C_0C_1 + \frac{1}{4}h^4\mu_2(K)^2 C_2, \tag{2.53}$$

where $C_1 = \int f^2(x)\,W(x)\,dx$ and $C_2 = \int f'(x)^2 f(x)\,W(x)\,dx$.

As before, differentiating with respect to h and equating to zero, the optimal $WAMISE$ bandwidth is

$$h_{WAMISE} = \left[\frac{C_0C_1}{\mu_2(K)^2 C_2}\right]^{\frac{1}{3}} n^{-\frac{1}{3}}. \tag{2.54}$$

Substituting (2.54) in (2.53), gives

Kernel	$\mu_2(K)$	$C_0/2$	Inefficiency
Uniform	1/3	1/6	1
Epanechnikov	1/5	9/70	1.0041
Biweight	1/7	25/231	1.0082
Triweight	1/9	245/2574	1.0109
Gaussian	1	$1/2\sqrt{\pi}$	1.0233

Table 2.3: Kernel functions and their inefficiency.

$$\inf_h WAMISE\left[\hat{F}_h\right] = \frac{1}{n}\int F(x)\left[1-F(x)\right]W(x)\,dF(x) - \left\{\frac{3\left(C_0C_1\right)^{\frac{4}{3}}}{4\left[\mu_2(K)^2 C_2\right]^{\frac{1}{3}}}\right\} n^{-\frac{4}{3}}. \tag{2.55}$$

In any case, whether from (2.51) or (2.55), it can be concluded that $\hat{F}_h(x)$ is asymptotically more efficient than $\hat{F}_n(x)$, since the constant C_0 is positive for any symmetric kernel (Swanepoel, 1988).

2.2.2 Choosing the kernel K

As with the density estimator, it is also possible to determine the optimal kernel for the distribution estimator. Based on (2.50) and (2.54), the resulting $AMISE$ and $WAMISE$ (equations (2.51) and (2.55)) depend on the kernel function by means of

$$c(K) = \left[\frac{C_0}{2\mu_2^{\frac{1}{2}}}\right]^{\frac{4}{3}}.$$

The optimal kernel K^* will be the one that maximizes $c(K)$ and, as a consequence, will minimize both (2.51) and (2.55). It can be proved that, when it comes to the kernel distribution estimation, the optimal kernel is the uniform kernel function (Jones, 1990). As before, to compare other kernel functions with the optimal uniform kernel, it can be used the inefficiency measure $\left[c(K^*)/c(K)\right]^{\frac{3}{4}}$.

Like in the density estimation case, the message in Table 2.3 is that when estimating the distribution by means of the kernel estimator, using a kernel different from the optimal one is not that serious, as they all perform very similar. Thus, the kernel can be chosen considering other criteria, such as smoothness or ease of implementation. As before, a very common choice is the Gaussian kernel.

2.2.3 Bandwidth selection

In the context of kernel distribution estimation, basically two types of bandwidth selectors have been investigated: the plug-in and cross-validation. The former was studied both theoretically and practically by Altman and Léger (1995) and Polansky and Baker (2000). The latter was analyzed by Sarda (1993), but as shown in Altman and Léger (1995), it needs of very large sample sizes for giving good results. Thus, the modified cross-validation of Bowman et al. (1998) is of greater interest from an applied point of view.

Plug-in bandwidth selection

The approach of Polansky and Baker (2000) is based on taking Eq. (2.48) as a global discrepancy measure. Having obtained an asymptotic approximation for (2.48), which is Eq. (2.49), and an optimal $AMISE$ bandwidth (Eq. (2.50)), Polansky and Baker's bandwidth selector is

$$\hat{h}_{PB} = \left(\frac{C_0}{-\mu_2 (K)^2 \hat{\psi}_2 (g_2)} \right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{2.56}$$

where $\hat{\psi}_r$ is already given in Eq. (2.31) and

$$g_2 = \left(\frac{2L^{(2)} (0)}{-\mu_2 (L)^2 \psi_4} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

where L and K are not necessarily the same kernel. The problem of estimating ψ_2 in (2.56) is the same as that of estimating ψ_4 in Eq. (2.33). Polansky and Baker suggest to use the same iterative method, emphasizing that $\nu = 2$ is sufficient in most cases.

The plug-in bandwidth selection approach of Altman and Léger (1995) is based on taking Eq. (2.52) as a global discrepancy measure. For the sake of simplicity, let us take $W(x) = 1$, so that the discrepancy measure is the one of Cramér–von Mises. To be specific,

$$MISE_C \left(\hat{F}_h \right) = \int \mathbb{E} \left\{ \left[\hat{F}_h (x) - F (x) \right]^2 \right\} f (x) \, dx. \tag{2.57}$$

The approach consists in selecting a bandwidth that minimizes the asymptotic approximation of (2.57). According to Altman and Léger (1995), and under some adequate conditions, it can be proved that the asymptotic $MISE_C$ is

$$\begin{aligned} AMISE_C \left(\hat{F}_h \right) &= h^4 \frac{1}{4} \mu_2 (K)^2 \int f' (x)^2 f (x) \, dx + \frac{1}{n} \int F (x) [1 - F (x)] f (x) \, dx \\ &\quad - \frac{h}{n} C_0 A (f), \end{aligned}$$

from which it follows that the $AMISE_C$ optimal bandwidth is

$$h_C^* = \left(\frac{C_0 A (f)}{\mu_2 (K)^2 \int f' (x)^2 f (x) \, dx} \right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \tag{2.58}$$

Altman and Léger propose to nonparametrically estimate the unknown terms in (2.58), so their plug-in bandwidth is

$$\hat{h}_{AL} = \left(\frac{C_0 \hat{A}(f)}{\mu_2 (K)^2 \hat{D}} \right)^{\frac{1}{3}} n^{-\frac{1}{3}},$$

where

$$\hat{A}(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\nu} K \left(\frac{X_i - X_j}{\nu} \right)$$

and

$$\hat{D} = \frac{1}{n^3 \alpha_b^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} K_b' \left(\frac{X_i - X_j}{\alpha_b} \right) K_b' \left(\frac{X_i - X_k}{\alpha_b} \right),$$

where $D = \int f'(x)^2 f(x)\, dx$, K_b' is the derivative of a kernel K_b (not necessarily the same as K), and α_b is its associated bandwidth parameter. In practice, it is common to chose $\alpha_b = \alpha$ and $K_b = K$.

Cross-validation

The modified cross-validation of Bowman et al. (1998) consists in selecting the bandwidth that minimizes the function

$$CV_B(h) = \frac{1}{n} \sum_{i=1}^{n} \int \left[I_{(-\infty, x]} (X_i) - \hat{F}_{-i}(x) \right]^2 dx,$$

and

$$\hat{F}_{-i}(x) = \frac{1}{n-1} \sum_{j=1, i \neq j}^{n} \mathbb{K} \left(\frac{x - X_j}{h} \right).$$

Bowman et al. (1998) showed that, generally, better results are obtained with their method compared with that of Altman and Léger. However, the main disadvantage is that it is somewhat heavy from the computational point of view.

On the other hand, Sarda's proposal (Sarda, 1993) consists in selecting the bandwidth that minimizes

$$CV_S(h) = \sum_{i=1}^{n} \left[\hat{F}_n (X_i) - \hat{F}_{-i} (X_i) \right]^2,$$

where $\hat{F}_n$ is the empirical distribution function. Nevertheless, this proposal has shown not to provide good results in practice.

2.3 Summary

In this chapter, a brief overview about kernel density and distribution estimators, and some of their main properties, was given. The theoretical developments show that the kernel estimator, both for the density and distribution estimation, is more efficient than traditional tools like the histogram and the empirical distribution function, respectively. It is this advantage that makes prefer the kernel estimator, with a suitable modification, for its use with grouped data.

It has been shown that the kernel estimator depends on two choices: the kernel function and the bandwidth. Although the kernel function is not really important, the right selection of the bandwidth is truly decisive for the estimator to have a good performance. Given its importance, in this chapter an overview of the most iconic techniques for bandwidth selection has also been given, from the fastest and simplest ones to some others more elaborate procedures like the plug-in or the bootstrap.

Certainly, there is no consensus as to which method is the best, since one method may work better than others depending on the intended density estimate. Leaving aside the quick and basic rules for bandwidth selection, on the one hand, the cross-validation methods shown here tend to give relatively small bandwidths (which indeed can be useful when estimating very rough densities), or also often being unstable, in general. This instability can be corrected, but paying a price on the bias. On the other hand, the plug-in and bootstrap bandwidth selectors seem to be effective in achieving a good compromise between bias and variance, as well as having better rates of convergence. That is why they have been so popular for many years. However, in recent times there have been a number of important contributions on the subject of bandwidth selection (Heidenreich et al., 2013).

In the next chapter, an important step in this research will be given. A modification to the kernel estimator will be proposed, so that it can be applied to grouped data. Its asymptotic properties will be studied, and the practical effects of the sample size, the bandwidth and the degree of data grouping on the performance of the estimator will also be shown.

Chapter 3

Kernel density estimation for grouped data

In this chapter, the problem of estimating the probability density function, when the data are grouped, is fully addressed. As seen in Chapter 2, besides being one of the most popular nonparametric estimation techniques, kernel estimation has good statistical properties and has proved to be very effective in finding structure in data sets where the parametric approach is inappropriate. Therefore, in this work, the kernel density estimator is taken as a reference and will be modified, so that it can be used with grouped data.

In this chapter, this new modified estimator will be defined and some important properties will be derived, such as its asymptotic bias and variance. In addition, by means of simulations studies, its performance will be examined under distinct scenarios, like using different sample sizes, bandwidths or varying the degree of grouping of the data. This will allow to give some preliminary practical guidelines.

3.1 Introduction

Because of the limitations of measuring instruments or the inability to monitor systems continuously, strictly speaking all of the experimental data distributions are discrete. As a consequence, it is expected that the measurement uncertainty may have an impact on the estimation of the density function.

Although from slightly different perspectives, this problem has received some attention. A pioneer study related to the subject is that of Hall (1982), in which the influence of rounding errors in kernel density estimation is analyzed. Another early work on this topic is that of Titterington (1983), where the problem of considering grouped, censored or truncated data in kernel density estimation is addressed, under the condition that some information of the overall density must be available (which very frequently is not the case). Strongly based on Hall (1982), Scott and Sheather (1985) study the effect of using equally spaced binned data in kernel density estimation. Assuming that there is a reasonable initial bandwidth, their results allow to control the increase in the $MISE$ due to binning the data by means of choosing an adequate (constant) binwidth.

A similar problem is that related with reducing the computational time of the kernel estimator. Suppose that the density is to be estimated at a grid of points, so that estimates can be plotted. Let us denote the grid points by $b_1, b_2, ..., b_M$. Then, the approach is to use

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

for $j = 1, 2, ..., M$.

Notice that the number of evaluations involved for obtaining $\hat{f}_h\left(b_j\right)$ is nM. Certainly, nM can be a realizable number of operations depending on n and M, but it can rapidly turn worse for other kernel estimators, as in (2.31), which requires $O\left(n^2\right)$ operations. For saving computation time, the idea is to associate the data to grid counts, $c_1, c_2, ..., c_M$, where c_j represents the amount of data on the surroundings of b_j. By doing so, the number of kernel evaluations reduces to only $O\left(M\right)$, which may represent an important computational time reduction, mainly when working with large sample sizes. In this sense, considering general binning rules, accuracy related issues of kernel density estimation based on binned data were studied in Hall and Wand (1996). Their results allow to choose a reasonable number of grid points M to reduce the computational time, while making the error due to binning negligible to some extent.

Clearly, the last problem described is somewhat similar to the problem discussed in this thesis, although there are important differences. On the one hand, in the last problem continuous data are known, and these are grouped on purpose to improve computational speed. Also, one can conveniently choose the length of the intervals to minimize the impact on the $MISE$ due to binning the data. Moreover, the data are binned by means of a constant interval length. On the other hand, in the problem studied in this dissertation, continuous data are unknown and come in a grouped fashion from the very beginning. Furthermore, the length of the intervals is given beforehand and they all are not necessarily of the same length, depending on the experimental conditions on which data were obtained. Thus, the purpose of this study is quite different: given a sample size and an arbitrary set of intervals, not necessarily of the same length, the objective is to obtain the best possible kernel density estimate with some already grouped (or binned) data, under those restrictions over which the data analyst has no control.

Finally, one more clarification is necessary. In survival analysis is common that survival times cannot be observed exactly. For example, in clinical trials, patients can only be examined at intervals whose duration can be one or more months, which may first appear as grouped data. However, for each individual, such follow-up times can be random. This type of data is known as interval censored, in clear connection with the common problem of right censoring. Although there is some resamblence to grouped data, they should not be confused, as in grouped data, monitoring times (i.e., the interval breaks) are the same for each individual.

3.2 Binned kernel density estimator and general grouped data

Suppose that X is the random variable of interest and let $(X_1, X_2, ..., X_n)$ be a random sample from a density f, with distribution function F, and consider a set of intervals $[y_{j-1}, y_j)$, $j = 1, 2, ..., k$. The j-th interval length is $l_j = y_j - y_{j-1}$, its midpoint is $t_j = \frac{1}{2}(y_{j-1} + y_j)$, and denote the number of observations within each interval by $(n_1, n_2, ..., n_k)$. In the spirit of Scott and Sheather (1985), when the interval length l is constant, the binned kernel density estimator may be written as

$$\hat{f}_h^s(x) = \frac{1}{n} \sum_{i=1}^{k} n_i K_h(x - t_i). \tag{3.1}$$

The problem here considered is more general: it is a non-equally spaced grouped data case, where the interval length is not constant. The number of intervals k, the interval lengths $(l_1, ..., l_k)$ and the breaks $(y_0, y_1, ..., y_k)$ typically depend on the sample size n. Also, in more restrictive situations, only the sample proportions $(w_1, w_2, ..., w_k)$ in each interval are available, where $w_j = n_j/n = F_n(y_j-) - F_n(y_{j-1}-)$ is the actual observed random quantity, with $F_n(y-)$ denoting the left-side limit of the empirical distribution function F_n. This is called a general grouped data case.

As mentioned in Chapter 1, some motivation for the present study comes from weed science, where weed emergence observations are often a set of non-equally spaced grouped data. Hydrothermal times (an index combining, for each day, values of temperature and water potential) are typically available in k inspection times and the number of emerged seedlings at each consecutive inspection are observed. Moreover, sometimes only $F_n(y_j)$, $j = 0, 1, ..., k$, are reported. Considering the sample proportions $(w_1, w_2, ..., w_k)$, statistical indices based on a more general version of the binned kernel density estimator have been proposed, and its good performance was experimentally proved using real *Bromus Diandrus* emergence data (Cao et al., 2011). However, a deeper statistical study requires the derivation of asymptotic properties of the binned kernel density estimator in the general grouped data case. The present study theoretically complements the good properties shown in practice by this estimator and opens the possibility to use it in other research areas.

3.3 Asymptotic results

In this section, the general binned kernel density estimator is presented and its asymptotic bias and variance are obtained and compared with those of the standard kernel density estimator, (Eq. (2.1)). For this, the following assumptions are required:

Assumption 3.1. *The kernel K is symmetric probability density function with support in $[-1, 1]$, 6 times differentiable and such that $K^{(6)}$ is bounded.*

Assumption 3.2. *The distribution F has compact support $[\mathcal{L}, \mathcal{U}]$, is 7 times differentiable and its j-th derivative $F^{(j)}$ is bounded for $1 \leqslant j \leqslant 7$.*

Assumption 3.3. *The bandwidth $h = h_n$ is a non random sequence of positive numbers such that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.*

Assumption 3.4. *Given a set of $k = k_n$ intervals $[y_{j-1}, y_j)$, $j = 1, 2, ..., k$, $y_0 \leqslant \mathcal{L}$ and $y_k \geqslant \mathcal{U}$, the average interval length is $\bar{l} = \bar{l}_n = \frac{1}{k}\sum_{i=1}^{k} l_i$, where l_i is the abbreviated notation of the i-th interval length $l_{i,n}$. It is assumed that $\lim_{n\to\infty} \bar{l} = 0$, $\lim_{n\to\infty} n\bar{l} = \infty$ and $\bar{l} = o\left(h^2\right)$. Finally, it is supposed that $\max_i \left|l_i - \bar{l}\right| = \max_{1\leqslant i\leqslant k} \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.*

Assumptions 3.1 and 3.2 are just smoothness and differentiability conditions about the kernel K and the distribution function F, and Assumption 3.3 is the typical one used in kernel density estimation concerning the sample size n and the bandwidth h. However, Assumption 3.4 is of special importance and deserves some comments, since it introduces the necessary conditions to be met between the parameters that define the set of intervals along with the sample size n and the bandwidth h.

Condition $\lim_{n\to\infty} \bar{l} = 0$ simply states that, as the sample size increases, the average interval length shrinks, which means that the whole set of intervals are shrinking as well. However, $\lim_{n\to\infty} n\bar{l} = \infty$ states that n should increase faster than $\bar{l}$ decreases. This is an important condition, as if the intervals shrink faster than n increases, at some point there would be more intervals than data points, and some of the intervals would be empty or there would be not enough density points in each interval.

Condition $\bar{l} = o\left(h^2\right)$ states an intuitive idea: as the sample size n increases, the average length $\bar{l}$ must vanish faster than, at least, h (concretely, faster than h^2). This condition has a practical basis. Since the average distance between points is $\bar{l}$, in order to obtain information of the surroundings at a certain point x, the bandwidth must be greater than $\bar{l}$ at all times. In other words, as n increases, h must vanish, but always behind $\bar{l}$.

Regarding the condition about $\max_i \left|l_i - \bar{l}\right|$, at first this is necessary from the strictly mathematical viewpoint, but in practice it is a way for controlling the variability of the intervals. In other words, in our assumptions we unquestionably accept different interval lengths in order to generalize the binned estimator, but within certain limits, and these limits of maximum variability are controlled by $\bar{l}$ via $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.

Going back to the general grouped data case, and assuming that $(w_1, w_2, ..., w_n)$ are the observed random quantities, the general binned kernel density estimator is defined as

$$\hat{f}_h^g(x) = \sum_{i=1}^{k} w_i K_h(x - t_i), \tag{3.2}$$

where the superscript g stands for the grouped case.

The asymptotic bias and variance of (3.2) are stated in the following theorem. Its proof is included in Appendix D.

Theorem 3.1. *Under assumptions 3.1 to 3.4,*

$$MSE_g = MSE\left[\hat{f}_h^g(x)\right] = \frac{1}{4}h^4\mu_2(K)^2 f''^2(x) + \frac{1}{nh}f(x)A(K) + o\left(h^4\right) + o\left(\frac{1}{nh}\right) \tag{3.3}$$

and

$$MISE_g = MISE\left[\hat{f}_h^g\right] = AMISE\left[\hat{f}_h^g\right] + o\left(h^4\right) + o\left(\frac{1}{nh}\right), \tag{3.4}$$

where

$$AMISE_g = AMISE\left[\hat{f}_h^g\right] = \frac{1}{4}\mu_2(K)^2 h^4 A\left(f''\right) + \frac{1}{nh}A(K). \tag{3.5}$$

Following similar arguments as for the standard kernel density estimator (Eq. (2.1)), it is possible to obtain optimal (global or local) bandwidths. In this case, the asymptotically optimal global bandwidth is obtained from (3.5), which yields

$$h_{AMISE_g} = \left[\frac{A(K)}{\mu_2(K)^2 A(f'')\,n}\right]^{\frac{1}{5}}. \tag{3.6}$$

Two aspects are worth mentioning. Firstly, as long as assumptions 3.1 to 3.4 hold, (2.1) and (3.2) have the same asymptotic expressions for the MSE and the $MISE$. Consequently, (2.12) and (3.6) appear to be the same except for one important difference: in (3.6), $A(f'')$ has to be estimated considering grouped data. In this situation, the estimation of $A(f'')$ requires of adapted density functionals estimators. Based on (2.31), a possible choice is the nonparametric estimator proposed in Cao et al. (2011), given by

$$\hat{A}_g(f'') = \frac{1}{\eta^5}\sum_{i=1}^{k}\sum_{j=1}^{k} L^{(4)}\left(\frac{t_i - t_j}{\eta}\right) w_i w_j, \tag{3.7}$$

where $L^{(4)}$ is the fourth derivative of the kernel L, and η is the auxiliary smoothing parameter. Thus, substituting (3.7) in (3.6), the asymptotically optimal global plug-in bandwidth selector is

$$\hat{h}_g = \left[\frac{A(K)}{\mu_2(K)^2 \hat{A}_g(f'')\, n} \right]^{\frac{1}{5}}. \tag{3.8}$$

Secondly, for a given sample size and an interval partition, the expectation of (3.2) is increased by a term depending on $\bar{l}$, as can be seen in (D.27). Hall (1982) found the same for (3.1), except that he referred to rounding errors in terms expressed as the multiplicative inverse of the number of intervals. Here, it was proved that in the context of different rounding errors (i.e., different interval lengths), its effect can be noticed via $\bar{l}$, although by the assumptions made, it is asymptotically negligible.

3.4 Simulations

To have an idea of the potential of the proposed density estimator, some simulation studies under different scenarios of sample sizes, bandwidths and degree of grouping were performed. For this, the free statistical software R and the nor1mix package have been used to implement the different procedures (R Core Team, 2015; Mächler, 2013).

In the first place, it was of interest to confirm the consistency of (3.2) by the behavior of its $MISE_g$ as the sample size increases, considering two different grouping scenarios. In second place, the $MISE_g$ was studied as a function of h and $\bar{l}$ by means of a fixed sample size. As a result, a 3-D map of the $MISE_g$ was obtained, which enabled to observe its behavior under different grouping situations (by means of $\bar{l}$), as well as detecting minimal zones of the $MISE_g$ for different bandwidths. Additionally, in both studies, the performance of the optimal bandwidth minimizing the $MISE_g$ (denoted by h_{MISE_g}), and the plug-in bandwidth selector for grouped data, $\hat{h}_g$, given in (3.8), was analyzed.

For doing the simulations, a reference density was needed. A normal mixture $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}$ was used, where $\phi_{\mu,\sigma}$ is a $N(\mu, \sigma^2)$ density, $\alpha = (0.70, 0.22, 0.06, 0.02)$, $\mu = (207, 237, 277, 427)$ and $\sigma = (25, 20, 35, 50)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively. This normal mixture was used in weed science to model the relationship between weed emergence of *Bromus diandrus* and hydrothermal time (Cao et al., 2011).

Regarding the kernel function, considering the practical information about these functions given in Subsection 2.1.3, it was decided to use the Gaussian kernel throughout the simulations.

3.4.1 Simulation study 1

In this first simulation experiment, the behavior of the $MISE_g$ is studied as a function of h considering different sample sizes. For this, and regarding Assumption 3.4 and Eq. (3.6), two scenarios that may impact on the behavior of the $MISE_g$ are considered.

As mentioned before, Assumption 3.4 is particularly important for the validity of Theorem 3.1, and thus for the validity of Eq. (3.6). Among all the conditions stated in that assumption, it is specifically important that $\bar{l} = o\left(h^2\right)$ holds. On the one hand, the former condition is necessary from the mathematical point of view, as can be seen in Appendix D. On the other hand, its importance is related with the fact that, when it does not hold, $\bar{l}$ is not shrinking at the right pace with respect to n, which is (theoretically) increasing. In practice, this condition can be interpreted as a heavy grouping case; i.e., a case in which $\bar{l}$ is relatively large compared to the sample size in turn.

To see this, recall from (3.6) that the $AMISE_g$ optimal bandwith is of precise order $n^{-\frac{1}{5}}$. So, since it was assumed that $\bar{l} = o\left(h^2\right)$ holds, then $\bar{l} = o\left(n^{-\frac{2}{5}}\right)$. This should be the right pace at which $\bar{l}$ shrinks with respect to n, and this will be the first scenario. The second scenario is just the opposite, in which $\bar{l}$ does not shrink at the right pace with respect to n. Naturally, it is expected that the second scenario, i.e., a practical heavy grouping case, will have a negative impact on the performance of the estimator (3.2).

Both scenarios can also be succinctly written as follows

- Scenario 1 (S1): $n^{\frac{2}{5}}\bar{l} \to 0$
- Scenario 2 (S2): $n^{\frac{2}{5}}\bar{l} \to \infty$

As to how to calculate the $MISE_g$ in our simulations, let us bring to mind that, for a generic estimator $\hat{\mathfrak{f}}_n$, the $MISE$ can be written as

$$MISE\left(\hat{\mathfrak{f}}_n\right) = \int \mathbb{B}^2\left[\hat{\mathfrak{f}}_n(x)\right]dx + \int \mathbb{V}\left[\hat{\mathfrak{f}}_n(x)\right]dx.$$

Hence, each of the integrals in the last expression can be approximated via Monte Carlo by

$$\int_a^b \mathbb{B}^2\left[\hat{\mathfrak{f}}_n(x)\right]dx \approx (b-a)\frac{1}{B_1}\sum_{i=1}^{B_1}\left[\frac{1}{B}\sum_{j=1}^{B}\hat{\mathfrak{f}}_{n,(j)}(x_i) - \mathfrak{f}(x_i)\right]^2$$

and

$$\int_a^b \mathbb{V}\left[\hat{\mathfrak{f}}_n(x)\right]dx \approx (b-a)\frac{1}{B_1}\sum_{i=1}^{B_1}\left\{\frac{1}{B}\sum_{j=1}^{B}\hat{\mathfrak{f}}^2_{n,(j)}(x_i) - \left[\frac{1}{B}\sum_{j=1}^{B}\hat{\mathfrak{f}}_{n,(j)}(x_i)\right]^2\right\},$$

where $[a,b]$ is the support interval; x_i, $i = 1,2,...,B_1$ is a set of B_1 equally spaced grid points in $[a,b]$, and B is the number of replications, $\hat{\mathfrak{f}}_{n,(j)}(x)$, $j = 1,2,...,B$. The interval considered was $[a,b] = [0, 509.25]$. With this election, the probability under the reference normal mixture is 0.999.

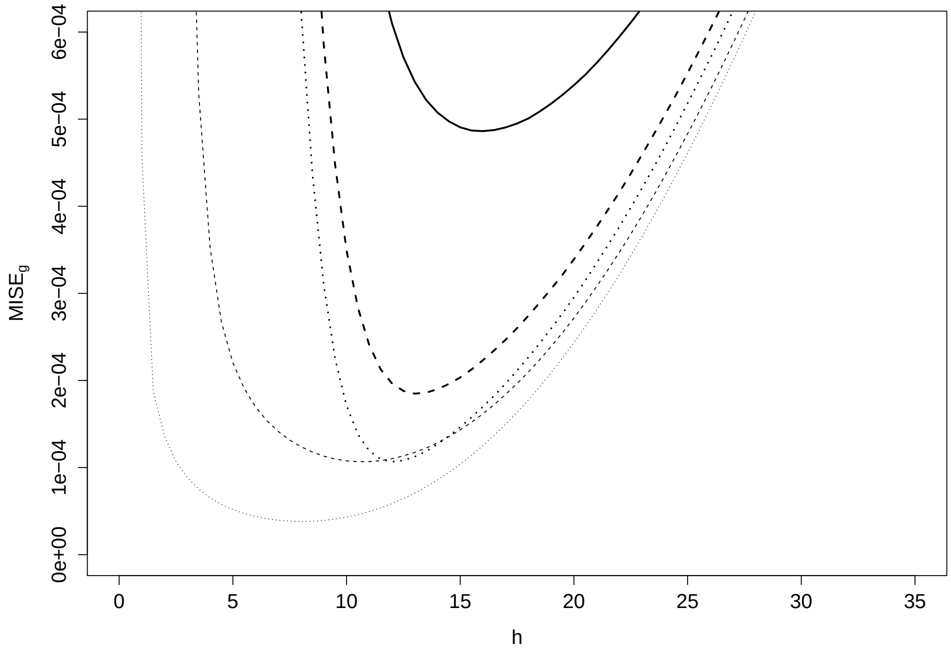

Figure 3.1: $MISE_g$ curves by scenario and sample size. Solid lines are for $n = 60$, dashed lines for $n = 240$, and dotted lines for $n = 960$. Thin lines represent the $MISE_g$ curves in S1, while thick lines represent the $MISE_g$ curves in S2 (note that the $MISE_g$ curves for $n = 60$ are practically identical in both scenarios).

In regard to how to simulate the set of intervals as n increases, three sample sizes were considered: $(n_1, n_2, n_3) = (60, 240, 960)$. Then, the next steps were followed:

Step 3.1. *Consider $\bar{l} = En^{-\alpha}$ and $a_n = Fn^{-\beta}$, where E, α, F and β are positive constants.*

Step 3.2. *Take a small initial set of intervals $\{l_i\}$. For instance, $i = 1, 2, ..., 5$ and $l_1 = \bar{l} - 4a_n$, $l_2 = \bar{l} + 0.5a_n$, $l_3 = \bar{l} - 1.5a_n$, $l_4 = \bar{l} + 3a_n$, $l_5 = \bar{l} + 2a_n$ were considered.*

Step 3.3. *For $i > 5$, $l_i = l_{(i-1)\text{mod}5+1}$, where* mod *stands for the modulo operation. That is to say, the initial set of intervals is repeated one after another, as many times as necessary.*

Note in Step 3.2 that with this initial selection of intervals, $\bar{l}$ remains the same. Although in Step 3.3 this initial set of intervals is repeated as many times as necessary, by the previous selection of intervals, its variability is already kept under control.

Constants E and F are just fitted according to the support interval. For choosing the positive constants α and β, let us consider the following. According to the initial set of intervals in Step 3.2, it follows that

$$\max_i \left| l_i - \bar{l} \right| = 4a_n = 4Fn^{-\beta}.$$

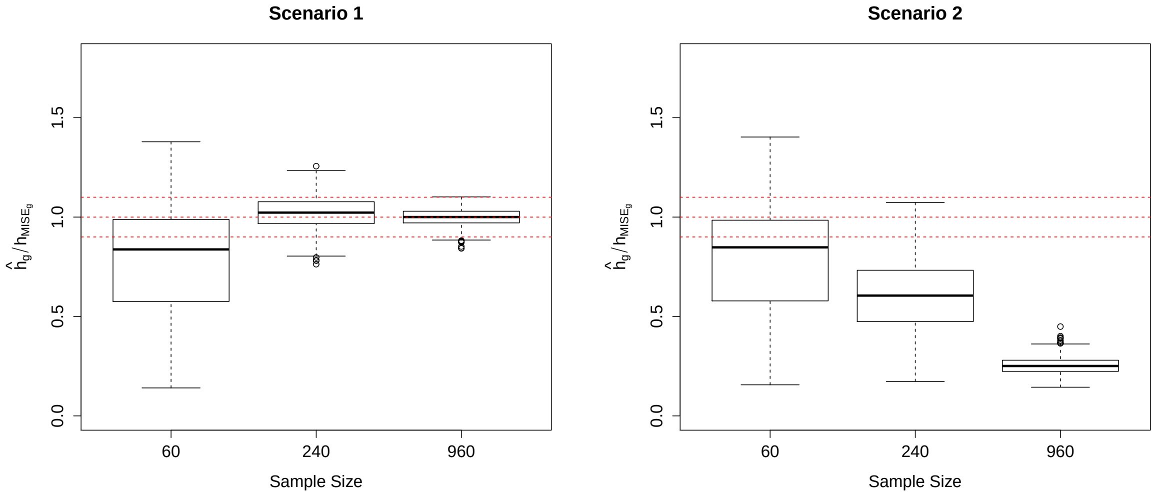

Figure 3.2: Boxplots for $\hat{h}_g/h_{MISE_g}$ for both scenarios.

Assumption 3.4 and Step 3.1 imply that

$$4Fn^{-\beta} = o\left(\bar{l}\right) = o\left(En^{-\alpha}\right),$$

which basically is

$$n^{-\beta} = o\left(n^{-\alpha}\right). \tag{3.9}$$

So, for (3.9) to hold, $n^{\alpha-\beta} \to 0$, which only occurs when $\alpha - \beta < 0$, i.e., when $\beta > \alpha$. Now, recall that for S1, $\bar{l} = o\left(h^2\right) = o\left(n^{-\frac{2}{5}}\right)$ must hold. Thus, according to Step 3.1,

$$\bar{l} = En^{-\alpha} = o\left(n^{-\frac{2}{5}}\right),$$

which basically is

$$n^{-\alpha} = o\left(n^{-\frac{2}{5}}\right), \tag{3.10}$$

which only occurs when $\frac{2}{5} - \alpha < 0$; i.e., when $\alpha > 2/5$.

In brief, for simulating S1, (3.9) and (3.10) must hold, i.e., $\beta > \alpha > 2/5$ must be true. On the other hand, for simulating S2, (3.9) must hold but (3.10) must not hold. It is required that $n^{-\frac{2}{5}}\bar{l} \to \infty$, so both $\beta > \alpha$ and $\alpha < 2/5$ must be true. Particularly, for doing our simulations we chose for S1 $(E, \alpha, F, \beta) = (800, {}^4\!/_5, 150, 1)$, and for S2, $(E, \alpha, F, \beta) = (37.1, {}^1\!/_{20}, 150, 1)$.

For each sample size and each scenario, the simulation experiment was done following these steps:

Step 3.4. *An n-size sample is simulated from the normal mixture reference density f.*

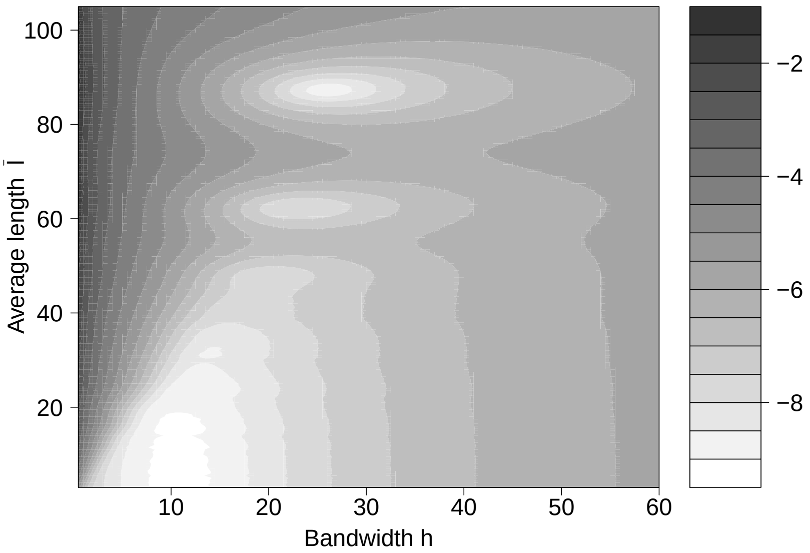

Figure 3.3: Natural logarithm of $MISE_g$ by average length $\bar{l}$ and bandwidth h for a fixed sample size $n = 240$.

Step 3.5. *The data range is divided into intervals* $[y_{i-1}, y_i)$ *of length* l_i*. The basic set of five intervals is repeated until covering the range. For each interval, its midpoint* t_i *and its relative frequency* w_i *are considered.*

Step 3.6. *For a grid of values of* h*, the density estimation* $\hat{f}_h^g(x)$ *is obtained in each of the* $B_1 = 512$ *points.*

Step 3.7. *The process is repeated* $B = 1000$ *times and the* $MISE_g$ *is calculated.*

Figure 3.1 presents the $MISE_g$ curves in both scenarios for the three different sample sizes. The $MISE_g$ curves have the same typical U shape as the $MISE$ curves for the standard kernel estimator, with one global minimum. In both scenarios the $MISE_g$ decreases as the sample size increases, which seems to confirm the consistency for the estimator (3.2). However, the different shape of the $MISE_g$ curves in both scenarios, for a large sample size of $n = 960$, reveals the importance of condition $\bar{l} = o\left(h^2\right)$ to get expressions (3.4) and (3.5). If this condition is not fulfilled, although the estimator (3.2) still seems to be consistent, the expression for the $MISE_g$ will be different to that obtained in Theorem 3.1, as some other additional terms depending on $\bar{l}$ remain important. Therefore, the bandwidth selector given in (3.6) will not be a good approximation for the optimal bandwidth h_{MISE_g}.

It should also be noted in Figure 3.1 that the values of h that minimize the $MISE_g$, by sample size, are not the same in both scenarios. This is an important consideration, since one would expect that the bandwidth selector (3.8) may give good approximations in just one of the scenarios, but not in both simultaneously.

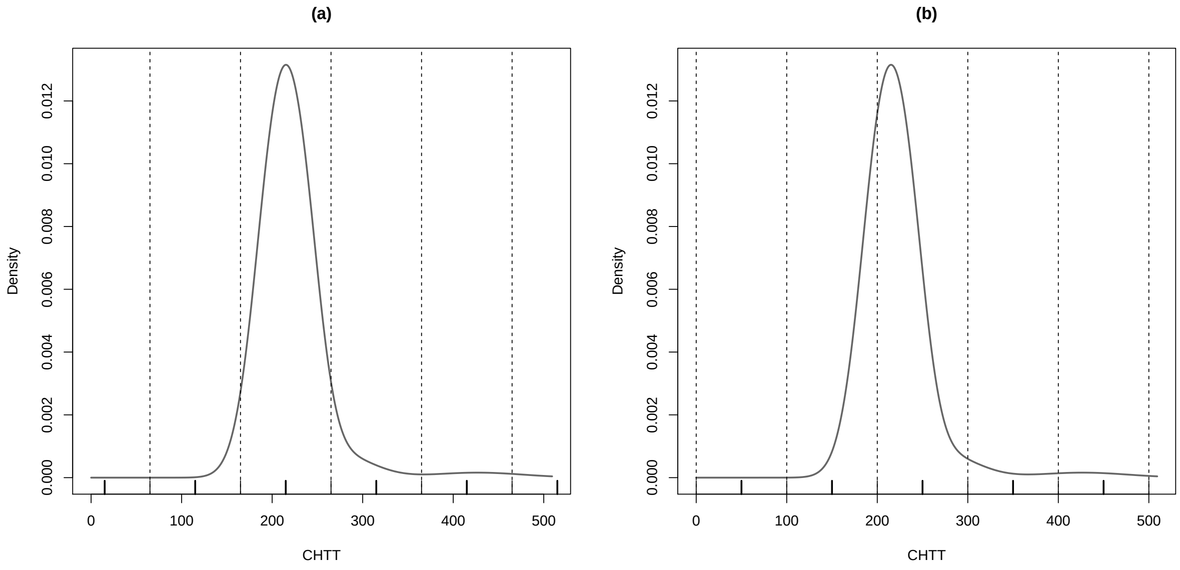

Figure 3.4: Normal mixture reference density for the simulation studies. As well as in (a), in (b) the density's support is roughly divided into five intervals (vertical dashed lines), but slightly shifted. In each case, the intervals capture different zones of the density.

To see this, the practical behavior of the bandwidth selector (3.8) was deeply studied. A second simulation experiment for each sample size and for each scenario went as follows:

Step 3.8. *Simulate an n-size sample from the reference normal mixture density f.*

Step 3.9. *Divide the data range into intervals $[y_{i-1}, y_i)$ of length l_i (according to the previous guidelines in Steps 3.1 to 3.3).*

Step 3.10. *Estimate $A\left(f''\right)$ using equation (3.7) and calculate $\hat{h}_g$ by (3.8).*

Step 3.11. *Compute $\hat{h}_g/h_{MISE_g}$.*

Step 3.12. *Repeat Steps 3.8 to 3.11 $B = 1000$ times.*

Figure 3.2 shows boxplots for $\hat{h}_g/h_{MISE_g}$ in both scenarios for the three different sample sizes. Both scenarios start from the same conditions: a sample size 60 and with relatively heavy grouping, so that the first boxplot in both scenarios reveals that the sampling distribution of $\hat{h}_g$ is not that accurate nor precise. However, as the sample size increases, it is observed that in S1, the sampling distribution of $\hat{h}_g$ becomes more accurate and precise, while in S2, it does get more precise, but is quite far from being accurate. This confirms $\hat{h}_g$ as a good bandwidth selector under scenario S1 conditions; i.e., under the assumption that $\bar{l} = o\left(h^2\right)$ holds.

It is worth emphasizing that this phenomenon is due to the conditions that are met in each scenario. Under S1 conditions, some other terms in the bias of (3.2), which depend

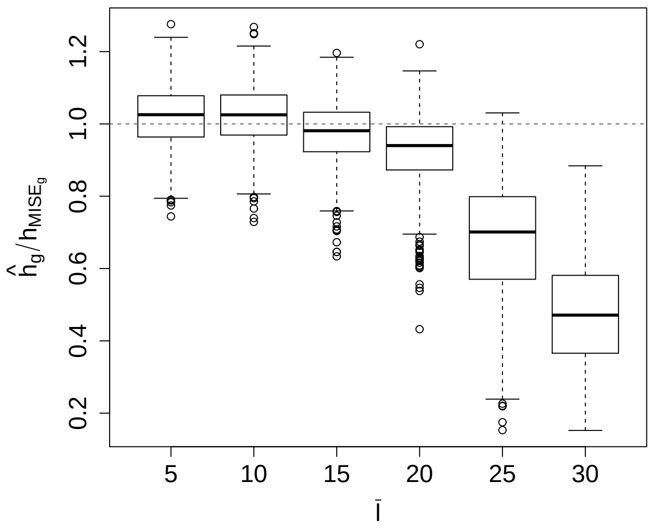

Figure 3.5: Sampling distribution of $\hat{h}_g/h_{MISE_g}$ for different average lengths for sample size $n = 240$.

on $\bar{l}$, becomes quickly negligible. Since (3.6) was obtained from (3.5), and (3.5) is a good approximation of (3.4), $\hat{h}_g$ is then a good approximation of h_{MISE_g}. In contrast, S2 conditions make $\hat{h}_g$ a bad bandwidth selector, since (3.5) is not a good approximation of (3.4) due to the fact that some other terms in the bias, depending on $\bar{l}$, are not vanishing yet.

3.4.2 Simulation study 2

To study situations in which sample size increases and intervals shrink at different paces may be of great theoretical interest, but in practice, that does not occur. What in fact occurs is that there is usually just a single sample with a fixed sample size and a fixed given set of intervals. Thus, the present simulation is of interest because it may give some practical ideas for implementation.

For doing this simulation, a fixed sample size $n = 240$ was considered along with 110 sets of intervals, each set with an average length $\{\bar{l}_1, \bar{l}_2, \bar{l}_3..., \bar{l}_{110}\} = \{1, 2, 3, ..., 110\}$. Also, a grid of 130 values of h $\{h_1, h_2, ..., h_{130}\}$ was considered, with $h_1 = 0.5$ and $h_{i+1} = h_i + 0.5$, for $i = 1, ..., 129$. With the previous specifications, the simulation was carried out as follows:

Step 3.13. *Simulate an n-size sample from the reference normal mixture density f.*

Step 3.14. *Divide the data range into intervals such that its average length is $\bar{l}_i$.*

Step 3.15. *Select h_j and compute $MISE_g$.*

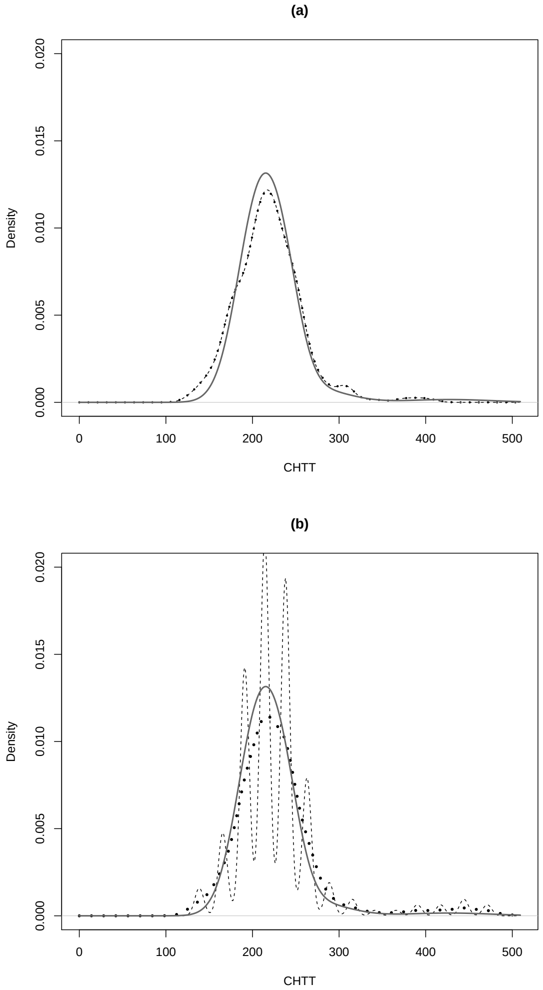

Figure 3.6: Kernel estimation using estimator (3.2) with a sample of size 240. In (a) it was used $\bar{l} = 15$, and bandwidths $h_{MISE_g} = 10.5$ (dotted line) and $\hat{h}_g = 10.2$ (dashed line). In (b) it was used $\bar{l} = 25$ and $h_{MISE_g} = 12.5$ (dotted line) and $\hat{h}_g = 5.2$ (dashed line). In both, solid lines represent the reference mixture density.

Step 3.16. *Repeat the previous steps considering the grid of possible pairs* $\left(\bar{l}_i, h_j\right)$.

Figure 3.3 shows the natural logarithm of the $MISE_g$ as a function of the average length $\bar{l}$ and the bandwidth h for the medium sample size, $n = 240$. The feature that draws most attention at first is the wave-like behavior of the $MISE_g$ natural logarithm, especially at some minima regions. Minima occurring at $\left(h, \bar{l}\right) = (23, 62)$ and $\left(h, \bar{l}\right) = (26, 87)$ are actually abnormal cases where intervals midpoints are, by chance, a good guess of the average location of the data therein. But to better understand this, it is helpfull to visually rely on the graph of the reference density.

Figure 3.4 shows the normal mixture reference density used in this simulations. In Figure 3.4 (a), the density's support has been divided in roughly five intervals (for the sake of simplicity, all of the same length). In Figure 3.4 (b), these same intervals has been slightly shifted, so in each case the intervals capture differents zones of the density function. In Figure 3.4 (a), one of the central intervals captures an area where practically all of the mass is located. Moreover, this interval captures a quite symmetric part of the density. This is particularly advantageous, since the average position of the sample points within that interval is very close to the midpoint, so that in this case, the midpoint is truly representative of the sample points therein. In contrast, in Figure 3.4 (b), this same region of the density has been captured by two central intervals, not one. Each of them capture a zone of the density that is not constant nor symmetric. Thus, the average position of the sample points in each interval would not be that close to the mid interval point; hence, the midpoint would not be a good representative point of the sample points therein.

From the above, the lesson is that even though both cases represent heavy grouping (just a few intervals), it may coincidentally happen that one (or even more) intervals capture areas of density that are constant or symmetrical, and midpoints are really representative, as in Figure 3.4 (a). If this were the case, one would expect a very good performance of the estimator by means of a relatively low $MISE_g$. On the contrary, in cases like Figure 3.4 (b), where midpoints are not representative, one would expect a poor performance, like in Figure 3.4 (b), reflected by a relatively high $MISE_g$. That is what is happening in Figure 3.3, in those minimal white zones associated with large average lengths, and hence the observed wave-like behavior as the average length decreases. Nevertheless, and in general, the estimator should not be expected to perform well in such (or worst) conditions of heavy grouping. Rather, it should be expected to better perform as the grouping becomes lighter; i.e., as the average length decreases.

Applications in mind, the important zone in Figure 3.3 is the minimum located at the bottom left, at $\bar{l}$ about 20 units or less, where the estimator performs the best. Also, in this zone the bandwidth varies from around 8 to 14 units. The average of this range is close to the minimum observed in Figure 3.1 for S1, which is 10.5. In other words, and considering this reference density in particular, given a fixed sample size and a set of intervals, the estimator can be expected to perform well if the average length is less than twenty units.

To better understand this, let us take into account the average sample range $\bar{r}$, which for this simulation happend to be around 340 units. This means that for this sample size, the estimator can be expected to perform well whenever the ratio $\bar{l}/r$ is around 0.06 or less, since in this case, $\bar{l}/r = 20/340 \approx 0.06$.

To reinforce the latter idea, Figure 3.5 shows boxplots with the sampling distribution of $\hat{h}_g/h_{MISE_g}$ for sample size $n = 240$. Note that when $\bar{l}$ is reaching 20 units (i.e., 6% of $\bar{r}$), the distribution of $\hat{h}_g$ starts to behave a bit biased, and as soon as it surpasses this limit, its distribution becomes really biased. This means that when $\bar{l}$ is into the zone of $0.06\bar{r}$, $\hat{h}_g$ gives quite good approximations of h_{MISE_g}, thus making $\hat{f}_h^g$ a good estimator of the true density. Out of that zone of $0.06\bar{r}$, $\hat{h}_g$ is unable to estimate well h_{MISE_g}, and

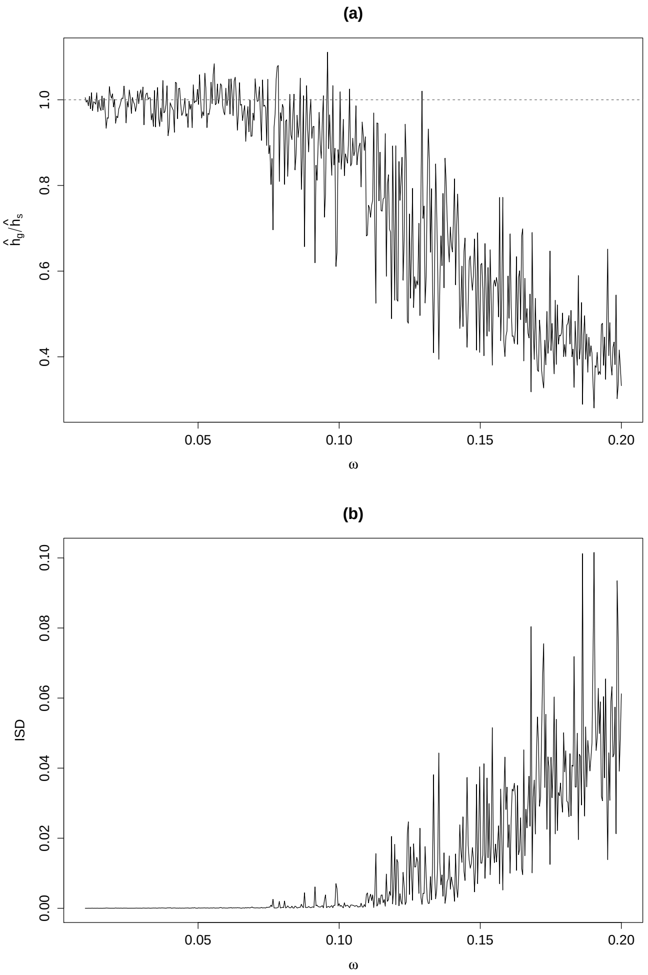

Figure 3.7: (a) $\hat{h}_g/\hat{h}_s$ versus $\omega = \bar{l}/r$. (b) Integrated squared distance (ISD) $\int \left[\hat{f}^g_{\hat{h}_g}(u) - \hat{f}_{\hat{h}_s}(u)\right]^2 du$ versus ω.

then, the estimator is expected to behave badly. This can be visually confirmed in Figure 3.6. In (a), the estimator (3.2) was used over a sample of size 240 and $\bar{l} = 15$, using both h_{MISE_g} and its estimation $\hat{h}_g$. As in this case $\hat{h}_g/h_{MISE_g} \approx 1$, both estimations are practically the same, and they resamble the true density. As opposed, in (b) the estimator was used over the same sample with $\bar{l} = 25$ and $h_{MISE_g}$and $\hat{h}_g$ as well. Notice that as $\hat{h}_g/h_{MISE_g} \approx 0.4$, which is far from the target, the estimator performs poorly, giving a very wiggly estimation.

3.5 Applications

To test the proposed estimator with some real data, it was used the time between eruptions set for the Old Faithful geyser in Yellowstone National Park, Wyoming, United States, available in the R environment for statistical computing datasets.

The 272 sample data was grouped using intervals of different average lengths. Notice that the geyser data is similar to the sample size used in the last simulation, which was 240. This favors for results comparisons. Now there is just one sample and there is no average range $\bar{r}$; instead, the sample range r can be used to express all those different average lengths as a proportion of it, which for these data set is $r = 53$ units. This ratio is called $\omega = \bar{l}/r$.

Since the data come from an unknown density, it is considered as a reference the estimation provided by the standard estimator (2.1) using the complete data and the plug-in bandwidth $\hat{h}_s$ (where s stands for 'standard'), the equivalent version of $\hat{h}_g$ when using complete data.

Figure 3.7 shows the ratio $\hat{h}_g/\hat{h}_s$ and $\int \left[\hat{f}^g_{\hat{h}_g}(u) - \hat{f}_{\hat{h}_s}(u)\right]^2 du$, the integrated squared distance (ISD), versus ω. As it can be seen in Figure 3.7 (a), the bandwidth selector $\hat{h}_g$ works well up to some middle point ω between $[0.05, 0.10]$. Although it is difficult to precise this value, it would not be unreasonable to say that this middle point is around 0.075, which is very close to the previous mark of 0.06. This suggests that the proposed density estimator $\hat{f}^g_h$ will perform well up to approximately $\omega = 0.075$, as can be confirmed in Figure 3.7 (b), where the integrated squared distance erratically begins to increase around this value. This seems to empirically confirm what was found in the simulation experiment.

The last assertion can also be verified in Figure 3.8. In (a), the standard estimator was used as a visual reference. In (b), the general estimator $\hat{f}^g_h$ was used with lightly grouped data, using $\omega = 0.04$, with an estimated bandwidth $\hat{h}_g$ that is almost the same as $\hat{h}_s$. This gives a very acceptable estimation. In (c), the general estimator $\hat{f}^g_h$ was used with somewhat heavy grouped data, using $\omega = 0.08$, with an estimated bandwidth $\hat{h}_g$ that is relatively far from $\hat{h}_s$; thus, the estimator does not perform well in this situation.

3.6 Summary

In this chapter, a generalization of the standard kernel density estimator was proposed and studied. When working with non-equally spaced grouped data, it was found that the bias of the general binned kernel estimator is increased by a term depending on $\bar{l}$, which by the assumptions made, it asymptotically vanishes, making this general estimator an asymptotically unbiased one. This general estimator and the results obtained generalize Scott and Sheather (1985) and Hall and Wand (1996) results, who considered a constant interval length and a constant rounding error, respectively.

By means of simulation studies, it was also investigated the consistency of the general density estimator. It was found that, although consistency is not affected in any case, the importance of condition $\bar{l} = o\left(h^2\right)$ in Assumption 3.4 relies in that it let us determine

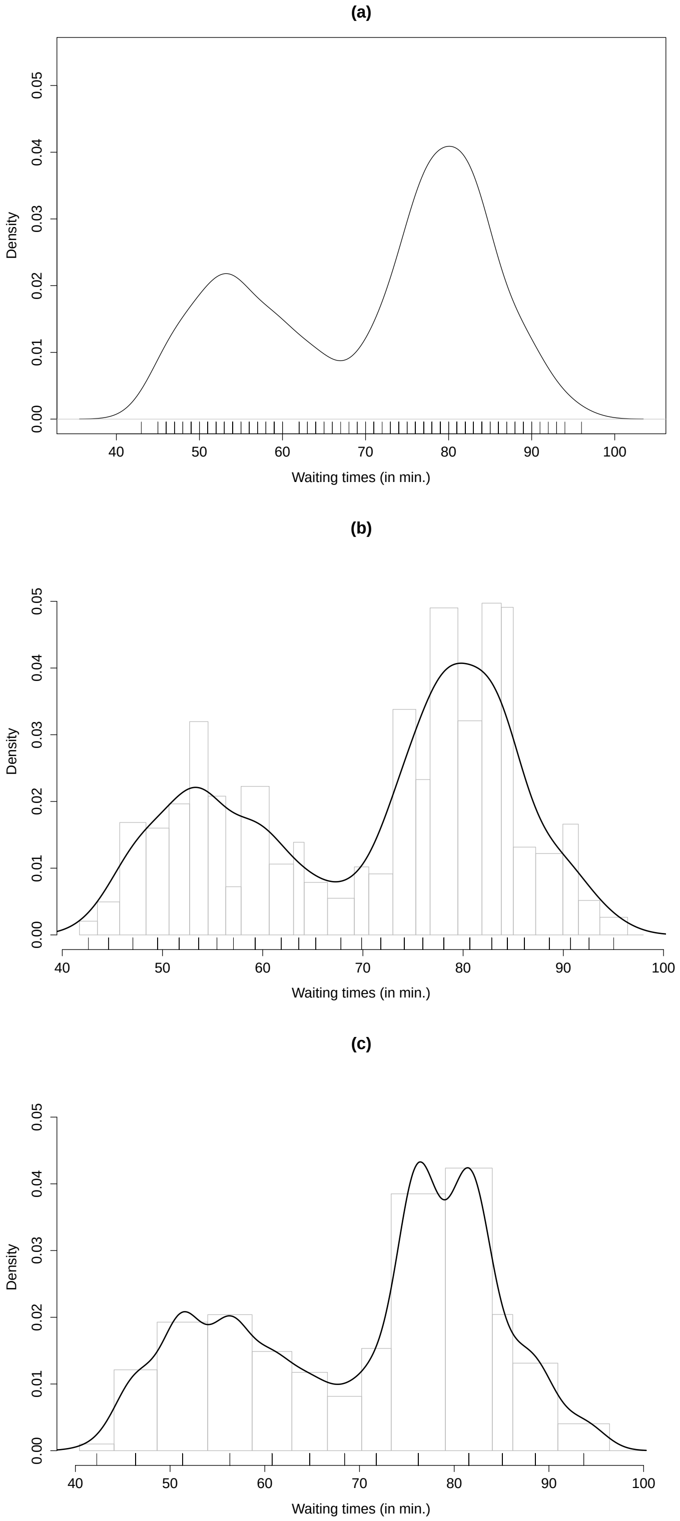

Figure 3.8: Kernel density estimation: (a) Using the standard estimator $\hat{f}_h$ with ungrouped data, and $\hat{h}_s = 2.481$; (b) using $\hat{f}_h^g$ with $\omega = 0.04$ and $\hat{h}_g = 2.428$; (c) using $\hat{f}_h^g$ with $\omega = 0.08$ and $\hat{h}_g = 2.199$.

under what grouping conditions the bandwidth selector $\hat{h}_g$ can be expected to perform succesfully, as well as the estimator $\hat{f}_h^g$, in consequence.

It was found that there is an important relationship between the degree of grouping and the sample size. From the theoretical point of view, as sample size increases, the number of intervals should increase as well, or equivalently, the average length $\bar{l}$ should decrease. In practice, where there is usually just one sample and a given intervals set over which the data analyst has no control, the potential application of the general kernel estimator looks promising whenever data are not heavily grouped, meaning that the average interval length should be not greater than around 6% of the sample data range. Although the application to the real data set of Old Faithful Geyser seems to empirically confirm this, it is of course just a preliminary rule of thumb that should be taken with caution, as it is necessary to perform more simulations and theoretical developments in order to give a general rule of application to different probability density functions and sample sizes.

Finally, more studies are needed regarding the bandwidth selector. On the one hand, it is necessary to theoretically study the functional estimator $\hat{A}_g$, (Eq. (3.7)) as it was used for getting $\hat{h}_g$ without prior information about its statistical properties in this case of grouped data. On the other hand, bear in mind that what makes the general estimator (and whatever other kernel estimator) to work the best is a right bandwidth selection. Thus, it may be thought that even in cases of heavy grouping, by rightly choosing the bandwidth it is possible to correct to some extent the undesirable effect of relatively large interval average lengths on the estimations. Perhaps, it may not be achieved the estimator to work as well as in the case of light grouping, but at least, it may bring some improvement when estimating the density in those cases. This is what the next chapter is about.

Chapter 4

Bandwidth selection in kernel density estimation for grouped data

In the previous chapter, the asymptotic properties of the estimator (3.2) were obtained and, under certain assumptions, it was also obtained an $AMISE$-based bandwidth selector. It was found that the plug-in bandwidth selector for grouped data fails when assumptions are not met, which in practice means the presence of heavy grouping. Also, this bandwidth selector for grouped data happend to coincide with that for continuous data, with the subtle difference that the functional $A\left(f''\right)$ is to be estimated not with continuous, but grouped data.

Trying to cover those cases of heavy grouping, in this chapter, an alternative bootstrap method for bandwidth selection is proposed. Through a comprehensive simulation study, the smoothing parameters obtained by both methods are compared considering different scenarios, including light and heavy grouping depending on the sample size. It is also analyzed the impact of these parameters on the performance of the estimator (3.2). Besides, it is studied the consistency of the estimator of the functional $A\left(f''\right)$ and, at the same time, the consistency for the plug-in bandwidth selector, both considering grouped data.

4.1 Bandwidth selectors

4.1.1 Plug-in bandwidth selector

Remember that under Assumptions 3.1 to 3.4, the asymptotic properties of (3.2) were obtained, from which it follows that (Eq. (3.5))

$$AMISE_g\left(\hat{f}_h^g\right) = \frac{1}{4}\mu_2\left(K\right)^2 h^4 A\left(f''\right) + \frac{1}{nh}A\left(K\right),$$

and

$$h_{AMISE_g} = \left(\frac{A(K)}{\mu_2 (K)^2 A(f'') n} \right)^{\frac{1}{5}}.$$

Unlike h_{AMISE} for continuous data (Eq. (2.12)), in this context, $A(f'')$ has to be estimated using a sample of grouped data. The following estimator was proposed in Cao et al. (2011), (Eq.(3.7))

$$\hat{A}_g (f'') = \frac{1}{\eta^5} \sum_{i=1}^{k} \sum_{j=1}^{k} L^{(4)} \left(\frac{t_i - t_j}{\eta} \right) w_i w_j,$$

where $L^{(4)}$ is the fourth derivative of a possibly different kernel L, and η is an auxiliary smoothing parameter. Plugging $\hat{A}_g (f'')$ in h_{AMISE_g}, a plug-in bandwidth is obtained (Eq.(3.8)),

$$\hat{h}_g = \left(\frac{A(K)}{\mu_2 (K)^2 \hat{A}_g (f'') n} \right)^{\frac{1}{5}}.$$

Next, the consistency of $\hat{h}_g$ as an estimator of the bandwidth minimizing the $MISE_g$ is shown. For this, since h_{AMISE} is asymptotically equivalent to the bandwidth that minimizes the $MISE$, h_{MISE} (Cao, 1990), it is sufficient to prove that (3.7) is a consistent estimator of $A(f'')$.

Generalizing Eq. (2.31) for grouped data, it follows that

$$\hat{\psi}_u^g = \frac{1}{\eta^{u+1}} \sum_{i=1}^{k} \sum_{j=1}^{k} L^{(u)} \left(\frac{t_i - t_j}{\eta} \right) w_i w_j, \tag{4.1}$$

where, as in the case of continuous data, it is sufficient to study functionals for u even. Note that Eq. (3.7) is just a particular case, for $u = 4$.

For obtaining the asymptotic properties of (4.1), the following assumptions are needed.

Assumption 4.1. *The s-th order kernel L (with $s > 0$ and even) is a Lipschitz symmetric density function with support in $[-1, 1]$, $u + 1$ times differentiable and $L^{(u+1)}$continuous. The notation $\mu_s (L) = \int x^s L(x) \, dx$ is used.*

Assumption 4.2. *The distribution function F is a $p + 1$ times differentiable function with compact support $[\mathcal{L}, \mathcal{U}]$, such that $F^{(p+1)}$ is continuous, $p \geqslant \max\{u, s + 1\}$, and $A\left(F^{(u+1)}\right) = A\left(f^{(u)}\right) < \infty$, where f is the density function.*

Assumption 4.3. *The bandwidth $\eta = \eta_n$ is a non random sequence of positive numbers such that $\lim_{n\to\infty} \eta = 0$ and $\lim_{n\to\infty} n\eta^{2u} = \infty$.*

Assumption 4.4. *Given a set of $k = k_n$ intervals $[y_{j-1}, y_j)$, $j = 1, 2, ..., k$, with $y_0 \leqslant \mathcal{L}$ and $y_k \geqslant \mathcal{U}$, the average interval length is $\bar{l} = \bar{l}_n = \frac{1}{k} \sum_{i=1}^{k} l_i$, where $l_i = l_{i,n}$ is the length* *of the i-th interval. It is assumed that* $\lim_{n\to\infty} \bar{l} = 0$*,* $\lim_{n\to\infty} n\bar{l} = \infty$ *and* $\bar{l} = o\left(\eta^{2u+1}\right)$*. Finally, we suppose that* $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$*.*

The consistency of (4.1) is stated in the following theorem. Its proof is included in Appendix E.1.

Theorem 4.1. *Under Assumptions 4.1 to 4.4,* $\hat{\psi}_u^g \to \psi_u$*, in probability.*

As mentioned above, it is observed that for light grouping frameworks, the behavior of $\hat{h}_g$ is satisfactory. However, when grouping is heavy, the results obtained for this bandwidth selector are rather deficient. In the following subsection, an alternative bootstrap procedure for bandwidth selection will be proposed.

4.1.2 Bootstrap bandwidth selector

Recall that w_j is the proportion of observations in the j-th interval, for $j = 1, 2, ..., k$. Using standard calculations (Appendix E.2), it is straightforward to obtain a closed expression for the $MISE_g$.

Theorem 4.2. *Let* F *be a distribution with probability density* $F' = f$*, and* K *a kernel function. Let* $(X_1, \ldots, X_n)$ *be a random sample from* f*. Consider a set of intervals* $[y_{j-1}, y_j)$*,* $j = 1, 2, \ldots, k$*, whose j-th midpoint is given by* $t_j = (y_{j-1} + y_j)/2$*. Let* $(n_1, \ldots, n_k)$ *be the number of observations within each interval, and let* $(w_1, \ldots, w_k)$*, be the sample proportions, where* $w_j = n_j/n$*. Assume that* $F(y_k) = 1$ *and* $F(y_0) = 0$*. Then,*

$$\begin{aligned} MISE\left(\hat{f}_h^g\right) &= \mathbb{E}\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right] \\ &= \int \left[\sum_{i=1}^{k} p_i K_h(x - t_i) - f(x)\right]^2 dx \\ &\quad + \frac{A(K)}{nh} - \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j (K * K)_h (t_i - t_j), \end{aligned} \tag{4.2}$$

where $p_j = F(y_j) - F(y_{j-1})$, for $j = 1, 2, ..., k$, and the symbol $*$ stands for convolution.

Eq. (4.2), considered as a function of h (and denoted by $MISE_g(h)$), can be used in simulations to approximate h_{MISE_g}, the optimal global bandwidth minimizing the mean integrated squared error.

Based on bootstrap techniques, in this subsection, an estimator of the optimal bandwidth minimizing the $MISE_g$ is defined. Bootstrap procedures for bandwidth selection in kernel density estimation for continuous data have been studied since the works of Taylor (1989); Faraway and Jhun (1990); Cao (1993); Marron (1992). It is now proposed a bootstrap bandwidth selection method for grouped data.

Let

$$\hat{f}_{\zeta}^{g}(x) = \sum_{i=1}^{k} w_i K_{\zeta}(x - t_i)$$

be the estimator (3.2) based on a pilot bandwidth ζ. Draw a bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{f}_{\zeta}^{g}$ and, given a bandwidth h, consider the analogue of the kernel density estimator, $\hat{f}_{h}^{g*}(x) = \sum_{i=1}^{k} w_i^* K_h(x - t_i)$, where $w_i^* = F_n^*(y_j-) - F_n^*(y_{j-1}-)$, with $F_n^*(y) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(X_i^*)$. The bootstrap version of the mean integrated squared error, $MISE^*$, is defined as

$$MISE^*\left(\hat{f}_h^{g*}\right) = \mathbb{E}^*\left[\int \left(\hat{f}_h^{g*}(x) - \hat{f}_{\zeta}^{g}(x)\right)^2 dx\right]. \tag{4.3}$$

Using a parallel process to that followed to obtain (4.2), it is possible to derive a closed representation for (4.3) (Appendix E.2.1). This expression is given by

$$\begin{aligned} MISE^*\left(\hat{f}_h^{g*}\right) &= \int \left[\sum_{i=1}^{k} w_i^{\zeta} K_h(x - t_i) - \hat{f}_{\zeta}^{g}(x)\right]^2 dx \\ &+ \frac{A(K)}{nh} - \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} w_i^{\zeta} w_j^{\zeta} (K * K)_h (t_i - t_j), \end{aligned} \tag{4.4}$$

where $w_i^{\zeta} = \mathbb{E}^*[w_i^*] = \hat{F}_{\zeta}(y_i) - \hat{F}_{\zeta}(y_{i-1})$, being

$$\hat{F}_{\zeta}(y) = \int_{-\infty}^{y} \hat{f}_{\zeta}^{g}(u)du = \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{y - t_i}{\zeta}\right),$$

with $\mathbb{K}(u) = \int_{-\infty}^{u} K(v)\,dv$.

Doing some elaborations, it is possible to show that

$$\begin{aligned} MISE^*\left(\hat{f}_h^{g*}\right) &= \frac{n-1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k} w_i^{\zeta} w_j^{\zeta} (K * K)_h (t_i - t_j) \\ &\quad -2\sum_{i=1}^{k}\sum_{j=1}^{k} w_i^{\zeta} w_j (K_h * K_{\zeta})(t_i - t_j) \\ &\quad +\sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j (K * K)_{\zeta}(t_i - t_j) + \frac{A(K)}{nh}. \end{aligned} \tag{4.5}$$

Note that expression (4.5) directly allows to evaluate the $MISE^*$ over a grid of values of h without using Monte Carlo. In other words, this bootstrap bandwidth selection, unlike many other bootstrap procedures, does not require the generation of any bootstrap resample in practice. This feature is also fulfilled in the bootstrap method proposed by Cao (1993) for the continuous data case.

The bootstrap bandwidth h^*_{MISE} is obtained minimizing (4.5), i.e.,

$$h^*_{MISE} = \arg\min_{h>0} MISE^* \left(\hat{f}_h^{g*}\right),$$

and if K is a Gaussian kernel, it is straightforward to see that

$$\begin{aligned} MISE^* \left(\hat{f}_h^{g*}\right) &= \frac{n-1}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} w_i^\zeta w_j^\zeta K_{\sqrt{2}h} \left(t_i - t_j\right) \\ & -2 \sum_{i=1}^{k} \sum_{j=1}^{k} w_i^\zeta w_j K_{\sqrt{h^2+\zeta^2}} \left(t_i - t_j\right) \\ & + \sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j K_{\sqrt{2}\zeta} \left(t_i - t_j\right) + \frac{A(K)}{nh}. \end{aligned} \tag{4.6}$$

where K_δ stands for a Gaussian density function with mean 0 and standard deviation δ.

An important issue in the previous bootstrap method is the choice of the pilot bandwidth ζ. In the bootstrap procedures for bandwidth selection in kernel density estimation, it is well studied that the pilot bandwidth ζ should be the optimal value ζ_{opt} that minimizes $\mathbb{E}\left\{\left[\hat{A}_\zeta(f'') - A(f'')\right]^2\right\}$(Cao, 1993). Assuming that f is $N\left(\mu, \sigma^2\right)$ and K is a Gaussian kernel, it is easy to see that for continuous data (see Appendix E.3),

$$\zeta_{opt} \approx 0.78\sigma n^{-\frac{2}{13}}. \tag{4.7}$$

Intuitively, it seems clear that for grouped data, ζ_{opt} will tend to increase as average length increases. For obtaining a relationship between the optimal ζ for grouped data, ζ_{opt_g}, and ζ_{opt}, some simulations for different average lengths and sample sizes were performed. Results suggest that for sample sizes $n < 150$, $\zeta_{opt_g} \approx 0.8\zeta_{opt}$ for $\omega = \frac{\bar{l}}{r} \leqslant 0.10$, where r is the sample range, and $\zeta_{opt_g} \approx \zeta_{opt}(4\omega + 0.4)$ otherwise. For sample sizes $n \geqslant 150$, $\zeta_{opt_g} \approx \zeta_{opt}$ for $\omega \leqslant 0.075$; otherwise, $\zeta_g \approx \zeta_{opt}(7\omega + 0.5)$ (see Appendix E.4). Once obtained ζ_{opt_g}, h^*_{MISE} can be numerically approximated.

4.2 Simulation studies

In Section 3.4, it was already studied the practical behavior of the plug-in bandwidth selector for light and heavy grouping conditions. Figure 3.1 showed that althought consistency seems to be confirmed in both scenarios, the values of h that minimize the $MISE_g$, h_{MISE_g}, are different from one scenario to another. Figure 3.2 confirms that only under the conditions of S1, the plug-in selector (3.8) performs well and gives good approximations for h_{MISE_g}.

The present simulation study compares not only the practical behavior of both the

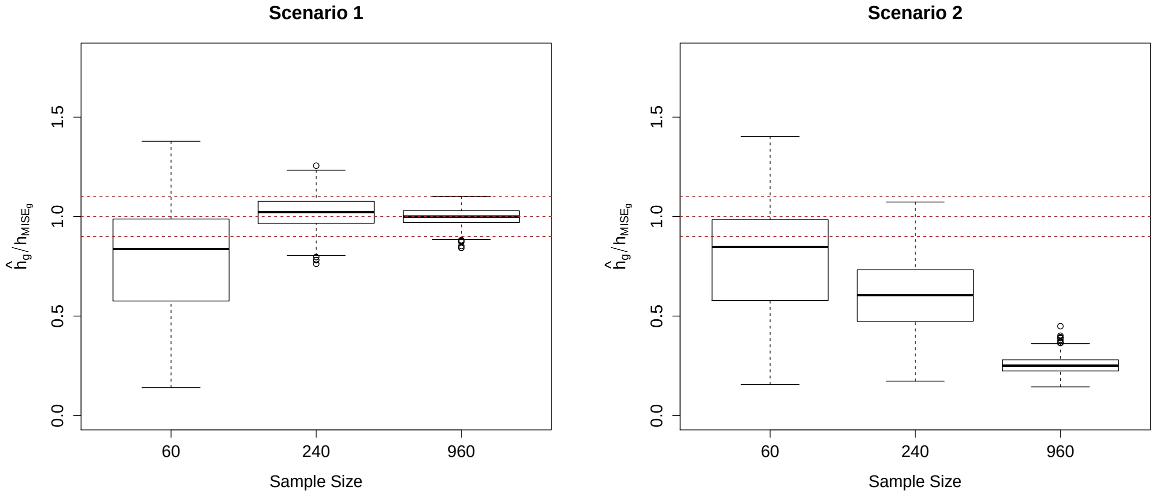

Figure 4.1: Boxplots for $\hat{h}_g/h_{MISE_g}$ for both scenarios.

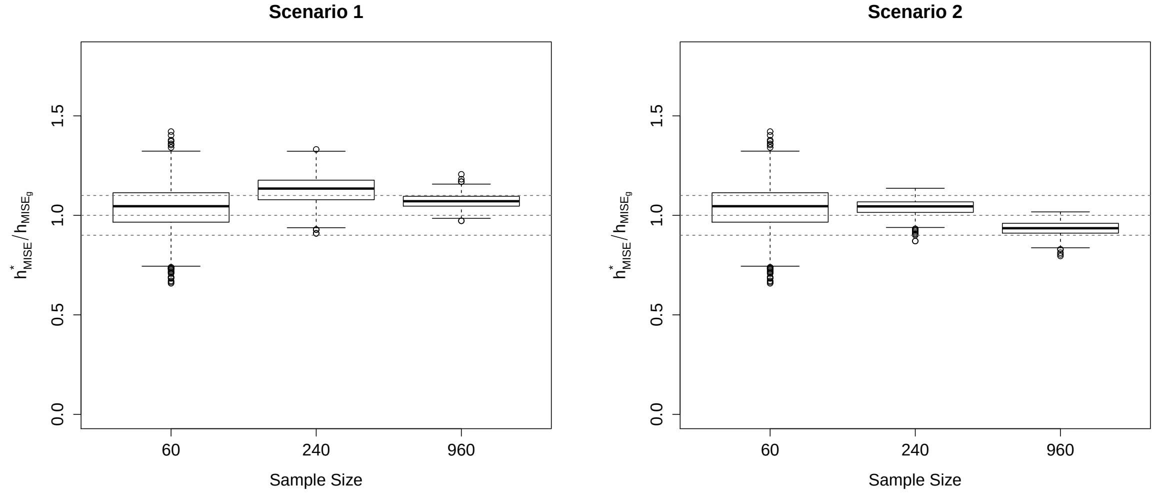

Figure 4.2: Boxplots for h^*_{MISE}/h_{MISE_g} for both scenarios.

plug-in and the bootstrap selector in giving good approximations for h_{MISE_g}, but also the impact of these bandwidth selectors, $\hat{h}_g$ and h^*_{MISE}, on the quality of the density estimator $\hat{f}^g_h$ via the $MISE_g$. For these purposes, the free statistical software R and some particular packages were used (R Core Team, 2015; Mächler, 2013; Wand, 2013).

As in Section 3.4, it is considered as a reference density the normal mixture $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}(x)$, where $\phi_{\mu,\sigma}$ is a $N\left(\mu, \sigma^2\right)$ density, with weights $\alpha = (0.70, 0.22, 0.06, 0.02)$, means $\mu = (207, 237, 277, 427)$, and standard deviations $\sigma = (25, 20, 35, 50)$.

For comparing the accuracy of bandwidth selectors $\hat{h}_g$ and h^*_{MISE}, the following steps were followed:

1. Simulate an n-size sample from f.

2. Divide the data range into intervals $[y_{i-1}, y_i)$ of length l_i according to the previous guidelines (see Subsection 3.4.1).

3. Using a Gaussian kernel, L, and a similar iterative process to that described in Wand and Jones (1995), but adapted to grouped data, select a pilot bandwidth η and estimate $A(f'')$ according to equation (3.7). Then, compute $\hat{h}_g$.

4. Selecting a pilot bandwidth ζ_{opt_g} as described in Subsection 4.1.2, approximate h^*_{MISE}.

5. Compute the ratios $\frac{\hat{h}_g}{h_{MISEg}}$ and $\frac{h^*_{MISE}}{h_{MISE_g}}$.

6. Repeat the previous steps 1000 times.

For making readability easier, Figure 3.2 is included again, which now is presented as Figure 4.1.

Regarding the plug-in selector (Figure 4.1), it is clear that starting from $n = 60$ with heavy grouping, under conditions in S1, $\hat{h}_g$ improves its performance in approximating h_{MISE_g} as sample size increases, since heavy grouping fades at the right pace. The opposite occurs in S2, where starting in the same conditions of sample size and heavy grouping, $\hat{h}_g$ performs worse at each stage.

On the other hand, it is evident from Figure 4.2 that the bootstrap bandwidth selector outperforms the plug-in selector in approaching h_{MISE_g}. Despite a slight bias, in general, the bootstrap selector shows more stability under any sample size and scenario, which means that it can be used both in cases of light or heavy grouping. On the contrary, the plug-in selector outperforms the bootstrap selector only in medium to large sample sizes, and only in cases of light grouping, as can be seen in Figure 4.1.

To see the effect of using $\hat{h}_g$ or h^*_{MISE} in both scenarios on the performance of the estimator $\hat{f}^g_h(x)$, the $MISE_g(h)$ was evaluated at every single of the 1000 bandwidths $\hat{h}_g$ and h^*_{MISE}, and compared with the $MISE_g(h)$ evaluated at h_{MISE_g}. The results can be

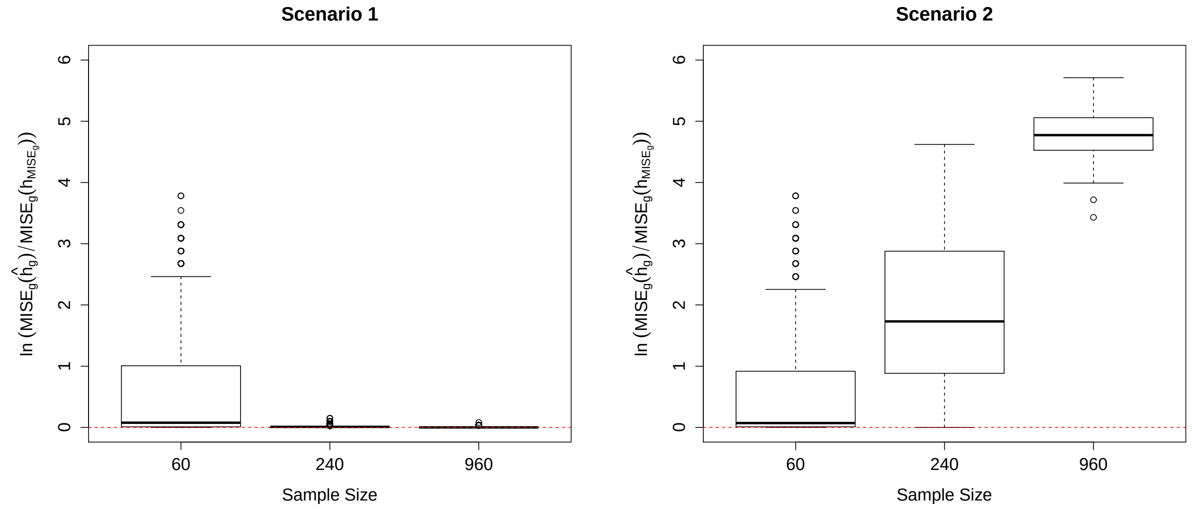

Figure 4.3: Box-plots for $\ln\left[\frac{MISE_g(\hat{h}_g)}{MISE_g(h_{MISE_g})}\right]$ for both scenarios.

seen in Figure 4.3 and Figure 4.4 for the plug-in and bootstrap bandwidths, respectively. They are presented as natural logarithms for comparison purposes.

The consequences are clear: when using $\hat{h}_g$, while in S1 the quality of $\hat{f}_h^g(x)$ becomes better as n increases, S2 shows on average increasingly disastrous density estimations that go as far as five orders of magnitude compared with S1 (Figure 4.3). On the other hand, as happened with the plug-in selector, when using the bootstrap bandwidth in S1 (Figure 4.4), the quality of the density estimations improves as sample size increases. Morover, the bootstrap selector performs better than the plug-in for sample size 60. In Scenario S2, similar results are shown, although it looks like the effect of heavy grouping makes the estimator more sensitive to slight changes in the bandwidth, as can be seen for sample size 960.

A visual support can be useful to better understand the latest ideas. Figure 4.5 shows the application of the estimator (3.1) in the case of heavy grouping; that is, S2. To estimate the density it was used the average bandwidth returned by the plug-in and bootstrap selectors for this scenario at each sample size. The first row shows that both selectors provide acceptable bandwidths for sample size 60, although the estimation using the bootstrap bandwidth is somewhat better. The second row shows how the estimation using the plug-in bandwidth begins to deteriorate for sample size 240, while the estimation using the bootstrap bandwidth still holds acceptable. The third row, for sample size 960, shows that the estimate obtained with the plug-in bandwidth is very wiggly, while the estimate obtained with the bootstrap bandwidth, although it could be improved, it still remains closer to the true density.

The above suggests that no matter the sample size, it should be avoided using $\hat{h}_g$ in

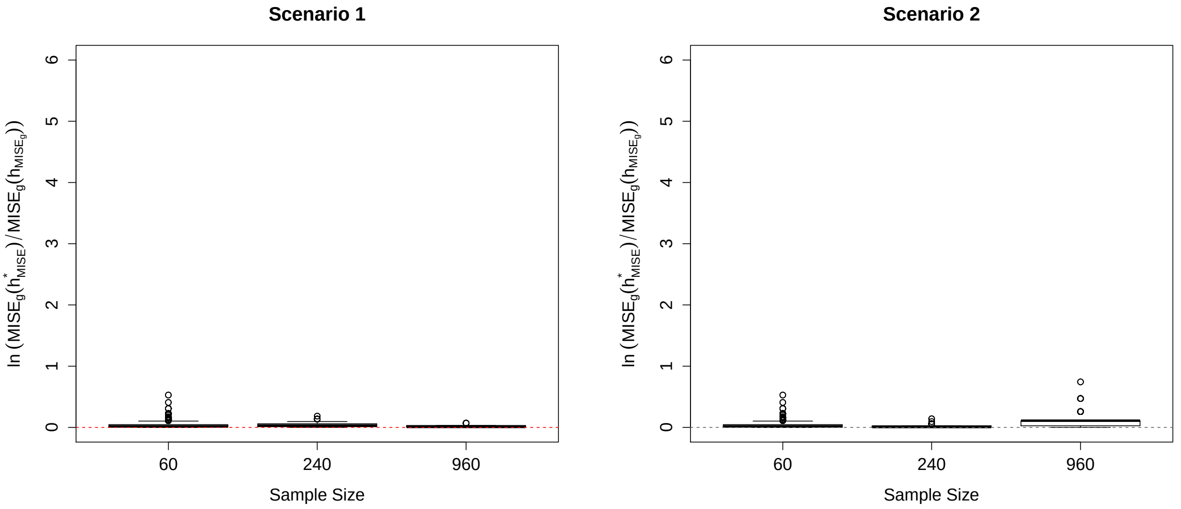

Figure 4.4: Box-plots for $\ln\left[\frac{MISE_g\left(h^*_{MISE}\right)}{MISE_g\left(h_{MISE_g}\right)}\right]$ for both scenarios.

cases of heavy grouping. In that instance, it would be preferable to use the bootstrap bandwidth selector, h^*_{MISE}.

4.3 Summary

It has been shown that under suitable assumptions, $\hat{\psi}^g_u$ is asymptotically consistent, so that $\hat{h}_g$ approaches to h_{MISE_g} as the sample sizes increases. However, in practice, there are some limitations that need to be considered in order to get the best performance of $\hat{f}^g_h$ in each situation. In previous analyses, it was found that the plug-in bandwidth selector, $\hat{h}_g$, has some limitations when performing in heavy grouping conditions. Thus, for overcoming these inconveniences, in this chapter, it was proposed and studied the performance of an alternative bootstrap bandwidth selector.

Given a grouped data sample, simulation studies showed that the plug-in bandwidth selector $\hat{h}_g$ should be the first option only when sample size is medium or large, and grouping is not heavy. Under other conditions, whether sample size is small or grouping is heavy, other bandwidth selectors should be considered.

Bootstrap bandwidth selector appears to be an option that in general outperforms $\hat{h}_g$. Although slightly biased, stability under any scenario or sample size is its best feature, giving quite acceptable density estimations. Results also show that despite the slight bias of bootstrap bandwidth selectors, they exhibit a good performance in nonparametric density estimation.

It is important to stress that since $\hat{h}_g$ is focused on minimizing the $AMISE_g$, there are some other terms in the $MISE_g$ series expansion that depend on the average length and

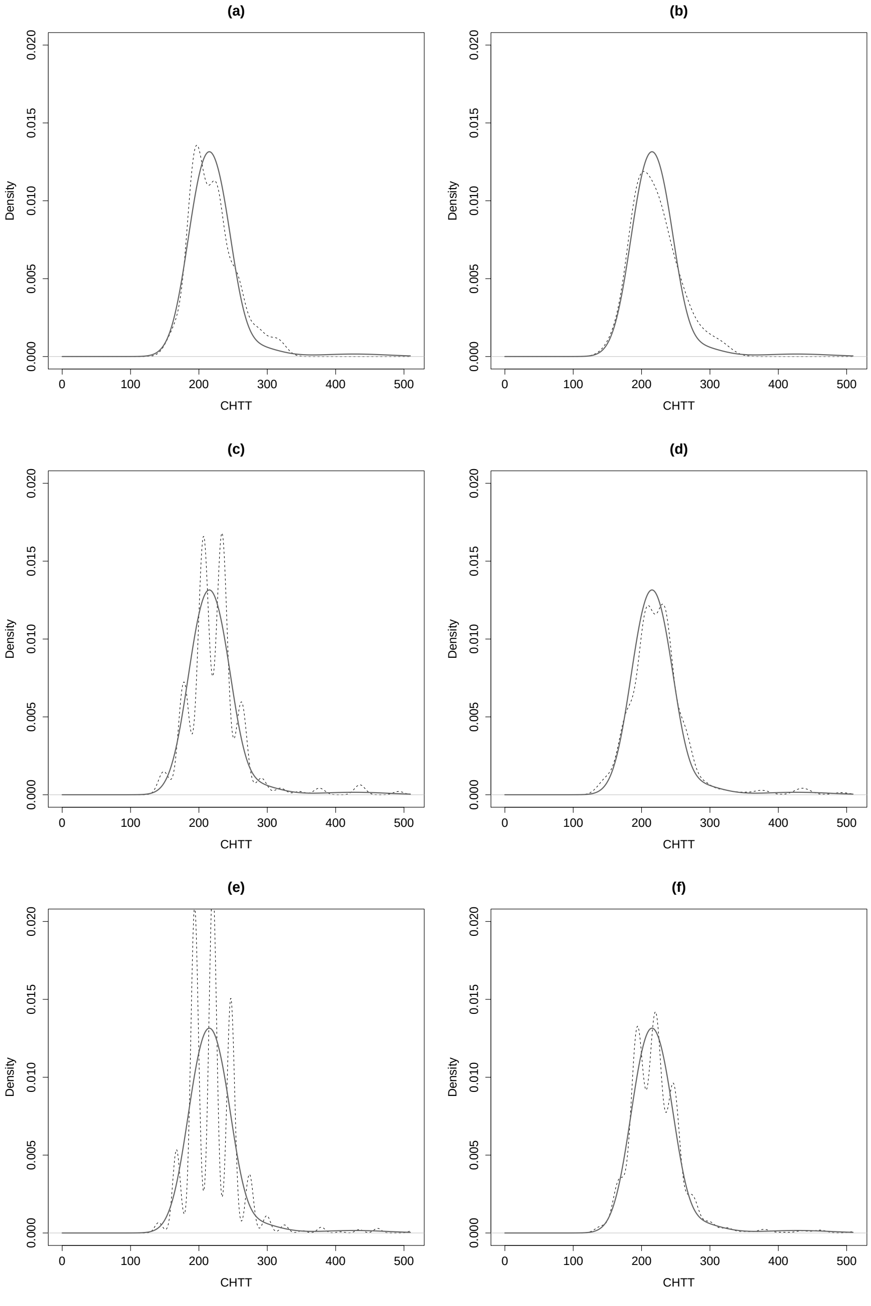

Figure 4.5: Kernel density estimation for heavy grouped data (S2). On the left, density estimations using the plug-in bandwidth selector (pbw) ; on the right, using the bootstrap selector (bbw). The first row corresponds to sample size 60; the second row, 240, and the third one, 960. Bandwidths used were: (a) pbw=12.4; (b) bbw=16.6; (c) pbw=7.8; (d) bbw=11.9; (e) pbw=3.1; (f) bbw=8.6.

that, under heavy grouping, they are not negligible at all. The result is a bad performance of $\hat{h}_g$ in all cases of heavy grouping, whether sample size is large or small. On the contrary, the bootstrap bandwidth selector obtains good pilot information about the distribution via $\hat{f}_\zeta$, which allows it to reproduce important features of the distribution under any scenario or sample size.

Finally, concepts like heavy or light grouping deserve some handy reference. Simulation studies suggest that, in general, light grouping can be considered when $\omega < 0.075$, approximately. The transition between light and heavy grouping (medium grouping, so to speak) could be considered when ω is somewhere between 0.075 and 0.10. Typically, heavy grouping cases can be considered when $\omega > 0.10$.

Chapter 5

Kernel distribution estimation for grouped data

In this chapter, the problem of estimating the distribution function F with grouped data is addressed. Based on the density estimator defined in Chapter 3, Eq. (3.2), in this chapter, an appropriate estimator of the cumulative distribution function F is derived, which by construction, is already adapted for grouped data. Its asymptotic properties are derived, and its performance in different grouping scenarios is analyzed through simulation studies. Also, a brief study on bandwidth selection in this context is included.

5.1 Introduction

As mentioned in Subsection 1.3.3, just as the density function f does, the distribution function F also describes the structure of a data set, but from another point of view. In many applications, including some problems of weed science, data are given not only in an aggregated fashion, but also cumulative. In these cases, the most appropiate approach is to estimate not the density f, but the distribution function F.

It is straightforward to construct a kernel estimator for the distribution F based on Eq. (2.1), as was explained in Section 2.2. Nevertheless, the topic of kernel distribution estimation has not been as popular as kernel density estimation, and the same is valid for the case of grouped data.

A seminal paper on this topic is the one of Turnbull (1976), which is concerned with the nonparametric estimation of F when data are grouped, censored or truncated, by means of an algorithm based on the idea of self-consistency. This work is very related to previous and later works about survival curves, hazard models and censored data.

As explained in Section 3.1, the problem studied in this thesis is of a different kind. Given a set of intervals, not necessarily of the same length, and given the number (or the proportion) of data in each interval, the objective is to use kernel estimation to estimate

f or F in the best possible way by rightly choosing the bandwidth under those particular grouping conditions, whether light or heavy.

5.2 Asymptotic results

Integrating Eq. (3.1), the kernel distribution estimator for binned or grouped data is

$$\hat{F}_h^g(x) = \int_{-\infty}^{x} \hat{f}_h^g(u)\, du = \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x-t_i}{h}\right), \tag{5.1}$$

where $\mathbb{K}(x) = \int_{-\infty}^{x} K(z)\, dz$. The asymptotic bias and variance of (5.1) are stated in the next theorem. Its proof is included in the Appendix F.

Assumption 5.1. *The kernel K is a symmetric probability density function with support in $[-1,1]$, at least 5-times differentiable and such that $K^{(5)}$ is bounded.*

Assumption 5.2. *The distribution F has compact support $[\mathcal{L},\mathcal{U}]$, it is 7-times differentiable and $F^{(7)}$ is bounded.*

Assumption 5.3. *The bandwidth $h = h_n$ is a non random sequence of positive numbers such that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.*

Assumption 5.4. *Given a set of $k = k_n$ intervals $[y_{j-1}, y_j)$, $j = 1, 2, ..., k$, $y_0 \leqslant \mathcal{L}$ and $y_k \geqslant \mathcal{U}$, the average interval length is $\bar{l} = \bar{l}_n = \frac{1}{k}\sum_{i=1}^{k} l_i$, where l_i is the abbreviated notation of the i-th interval length $l_{i,n}$. It is asumed that $\lim_{n\to\infty} n\bar{l} = \infty$ and $\bar{l} = o\left(h^2\right)$. Finally, we suppose that $\max_i \left|l_i - \bar{l}\right| = \max_{1\leqslant i\leqslant k} \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.*

Note that Assumptions 5.1 to 5.4 are basically the same as in the case of density estimaton for grouped data (Section 3.3). The only slight difference is the hypothesis about the differentiability of the kernel K. This makes sense: the estimator (3.2) requires the kernel K to be at least 6 times differentiable. Since the estimator (5.1) contains the integral of K, it only needs the kernel K to be at least 5 times differentiable. Besides that, the assumptions regarding the distribution F and the elements that characterize a grouped data set (the number of intervals k, the set of breaks $(y_0, y_1, \ldots, y_k)$, the average length $\bar{l}$, the maximum absolute difference $\max_i \left|l_i - \bar{l}\right|$ and their relationship with n and h) are the same.

Theorem 5.1. *Under Assumptions 5.1 to 5.4,*

$$MSE\left[\hat{F}_h^g(x)\right] = \frac{h^4}{4} F''(x)^2 + \frac{1}{n} F(x)\left[1 - F(x)\right] - \frac{h}{n} F'(x)\, C_0 + O\left(\frac{h^2}{n}\right) + o\left(h^4\right)$$

and

$$MISE\left[\hat{F}_h^g\right] = AMISE\left[\hat{F}_h^g\right] + O\left(\frac{h^2}{n}\right) + o\left(h^4\right),$$

where

$$AMISE\left[\hat{F}_h^g\right] = \frac{h^4}{4}\mu_2\left(K\right)^2 A\left(f'\right) + \frac{1}{n}\int F\left(x\right)\left[1 - F\left(x\right)\right]dx - \frac{h}{n}C_0 \tag{5.2}$$

and

$$C_0 = 2\int zK\left(z\right)\mathbb{K}\left(z\right)dz.$$

From Eq. (5.2), it is immediate to get an asymptotically optimal global bandwidth. Taking the first derivative of (5.2), equating to zero and solving for h, it is obtained

$$\mathfrak{h}_{AMISE} = \left[\frac{C_0}{n\mu_2\left(K\right)^2 A\left(f'\right)}\right]^{\frac{1}{3}}, \tag{5.3}$$

Note that Eq. (5.3) coincides with Eq. (2.50), since both are obtained via the $MISE$ considering no weights $w(x)$, as in (2.48), in contrast with (2.52). Regarding C_0, recall from Subsection 2.2.1 that it is a key constant that let $\hat{F}_h$, and hence, $\hat{F}_h^g$, be asymptotically more efficient than the empirical distribution function $\hat{F}_n$, since this constant is always positive.

For Eq. (5.3) to be a practical expression, an estimate of $A\left(f'\right)$ is required, since the remaining factors depend on known quantities. As it was done in Section 3.3 regarding the asymptotically optimal global bandwidth for kernel density estimation for grouped data, in this case it will also be used a nonparametric estimate of $A\left(f'\right)$. For estimating $A\left(f'\right)$, as proposed by Polansky and Baker (2000) (see Subsection 2.2.3, Eq.(2.56)) in the context of grouped data, the sample of mid points will be used instead of the complete sample (which in practice is unknown). Let us call $\hat{A}_{PB_g}\left(f'\right)$ the estimate of $A\left(f'\right)$ using the Polansky and Baker method in the grouped data case. Pluging $\hat{A}_{PB_g}\left(f'\right)$ into (5.3) gives a practical plug-in expression,

$$\hat{h}_{PB_g} = \left[\frac{C_0}{n\mu_2\left(K\right)^2 \hat{A}_{PB_g}\left(f'\right)}\right]^{\frac{1}{3}}. \tag{5.4}$$

5.3 Simulations

In this section, the effectiveness of the estimator (5.1) will be tested by means of a simulation study. The procedure will be parallel to the simulation study in Section 3.4, and the same normal mixture and different scenarios of sample sizes, bandwidths and degree of grouping will be considered. Besides the free statistical software `R` and the already

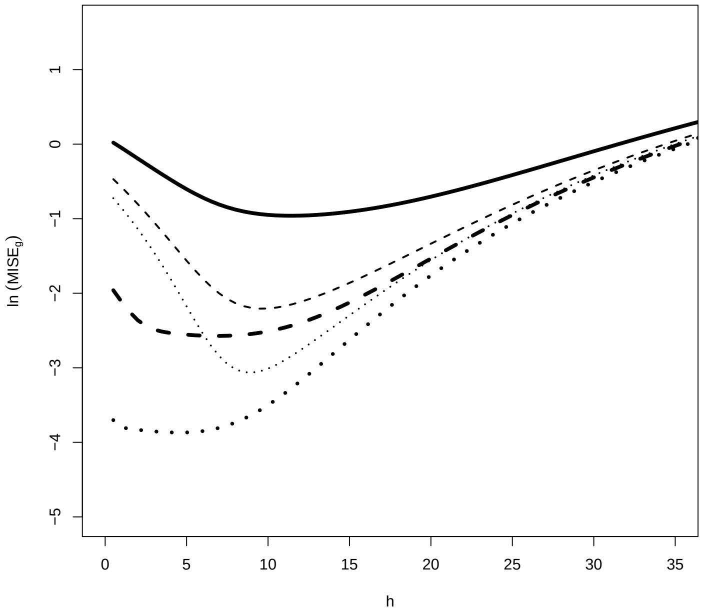

Figure 5.1: $\ln(MISE_g)$ curves by scenario and sample size. Solid lines are for $n = 60$, dashed lines for $n = 240$ and dotted lines for $n = 960$. Thick lines represent curves in S1, while thin lines represent curves in S2 (note that curves for $n = 60$ are practically the same in both scenarios).

used package `nor1mix`, it will also be used the package `kerdiest` (Quintela-del-Río and Estévez-Pérez, 2012).

5.3.1 Simulation study 1

In this first simulation study, the behavior of the $MISE$ in the case of grouped data ($MISE_g$) is studied depending on the bandwidth h for the three different sample sizes considered. As it was done in the case of the density, in the actual case of the distribution, two different scenarios are considered based on Assumption 5.4 and Eq. (5.3). These two scenarios may impact the behavior of the bandwidth selector (5.4) and, therefore, the performance of the estimator (5.1).

From Eq. (5.3), the asymptotically optimal global bandwidth is $O\left(n^{-1/3}\right)$. Since it is assumed that $\bar{l} = o\left(h^2\right)$, then $\bar{l} = o\left(n^{-2/3}\right)$ should be the right pace at which the average length decreases as the sample size increases. In other words, under this scenario (S1), it should be expected that the bandwidth selector (5.4) gives good approximations to the values of h that minimize the $MISE_g$, $h_{MISE_{\hat{F}}}$. The second scenario (S2) is just the opposite, in which the statement $\bar{l} = o\left(n^{-2/3}\right)$ does not hold.

Those two scenarios can be expressed as follows

- S1: $n^{-2/3}\bar{l} \to 0$

- S2: $n^{-2/3}\bar{l} \to \infty$

The procedure for calculating the $MISE_g$ is the same as the one explained in Subsection 3.4.1. Also, the same Steps 3.1 to 3.3 are to be followed to simulate the intervals set as n increases. As to how to choose the constants α and β, based on Eqs. (3.9) and (3.10), it is inferred that $\beta > \alpha > 2/3$ must be true for simulating S1, and both $\beta > \alpha$ and $\alpha < 2/3$ must hold for simulating S2. Thus, the same values (E, α, F, β) used for each scenario in the case of the density, are still valid for the distribution case. To carry out the simulation experiment, the same Steps 3.4 to 3.7 will be followed.

Figure 5.1 shows the $MISE_g$ curves for the three different sample sizes in both scenarios. Note that a semilogarithmic scale was used in order to better appreciate the minima values, which was not necessary in the case of the density function (see Figure 3.1). This is because in the case of the distribution, very little differences are found in the $MISE_g$ curves for small values of h, particularly for the largest sample size. This suggests that even in the case of grouped data, little deviations from the optimal bandwidth may still give quite good distribution estimates (particularly for large sample sizes), making the distribution estimation a relatively more robust procedure than the density estimation.

As happend with the density estimation for grouped data for both scenarios, the $MISE_g$ decreases as the sample size increases, which seems to confirm consistency of the estimator (5.1). However, it is expected that the bandwidth selector (5.4) will give good aproximations to $h_{MISE_{\hat{F}}}$ whenever $\bar{l} = o\left(n^{-2/3}\right)$ holds (i.e., under S1 conditions).

To confirm the latter, a second simulation experiment was conducted. This simulation follows the same Steps 3.8 to 3.11, and consists in comparing the sample distribution of (5.4) with the target values h_{MISE_g}, the ones that minimize the $MISE_g$ in each sample size and scenario.

Figure 5.2 resembles to the behavior of the sampling distribution of (3.8) in the case of density estimation. Starting from the same grouping conditions and sample size, in S1, the sampling distribution gets narrower and accurate as the sample size increases, while in S2 it gets precise but far from the target value. This confirms that (5.4) is a good bandwidth selector as long as S1 conditions hold. The explanation is the same as before: under S1 conditions, $\bar{l}$ decreases faster enough as the sample size increases, so that the remaining terms of the bias of (5.1) quickly become negligible. On the contrary, under S2 conditions, those remaining terms depending on $\bar{l}$ do not vanish as fast as required for (5.4) to be a good bandwidth selector.

Figure 5.3 shows the impact of the bandwidth selector (5.4) on the distribution estimator (5.1). Clearly, in S2, the impact of poor bandwidth selections are evident in the quality of the estimation of the distribution, wich negatively increases by up to two orders of magnitude. However, it should be noted that in the present case of the distribution, a poor bandwidth choice does not impact so negatively in the corresponding estimates as in the case of the density (see Figure 4.3).

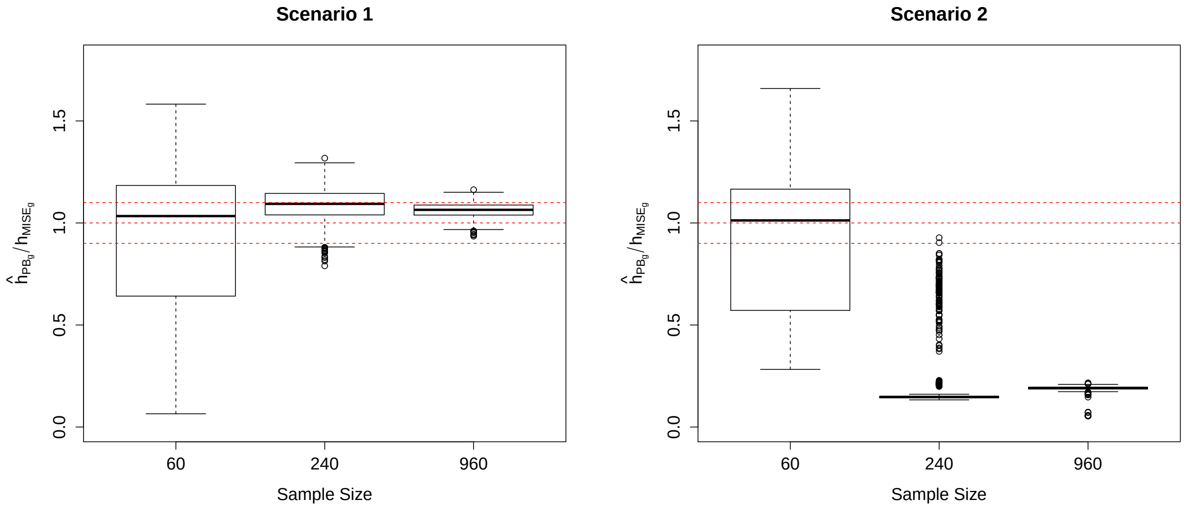

Figure 5.2: Boxplots for $\hat{h}_{PB_g}/h_{MISE_g}$ for both scenarios.

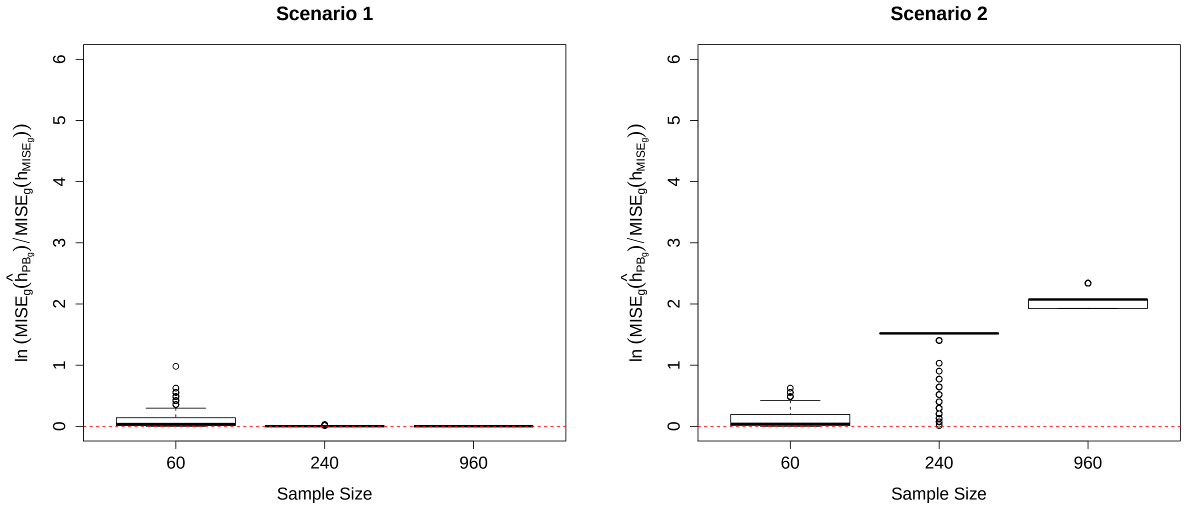

Figure 5.3: Box-plots for $\ln\left[\frac{MISE_g(\hat{h}_{PB_g})}{MISE_g(h_{MISE_g})}\right]$ for both scenarios.

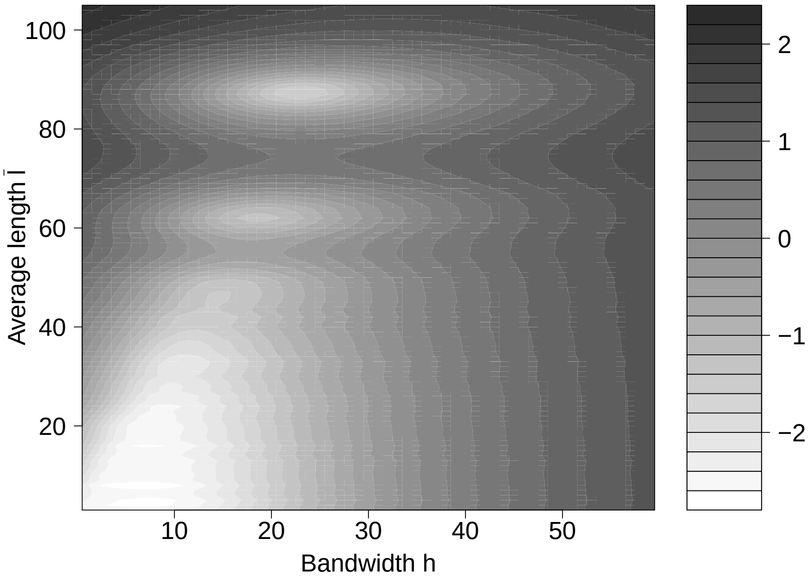

Figure 5.4: Natural logarithm of $MISE_g$ by average length $\bar{l}$ and bandwidth h for a fixed sample size $n = 240$.

5.3.2 Simulation study 2

It is of interest to study situations in which it is ideally observed the sample size increasing and the average length decreasing at different rates, but in practice this seldom really occurs. Thus, this simulation deals with a more factual situation in which there is a given sample size and a given set of fixed intervals.

For this simulation, it is considered a sample size $n = 240$, a set of average lengths and a grid of values of h, just as did in Subsection 3.4.2. The same Steps 3.13 to 3.16 are followed.

Natural logarithms of $MISE_g$ are shown in Figure 5.4. As in the case of the density, it is striking the wavelike behavior and the three minima regions. As before, the explanation is the same: sometimes, by chance, midpoints are really representative of the average location of datapoints into the intervals, and the estimator performs well, even though the two minima closest to the top correspond to heavy grouping; however, in general, it should not be expected the estimator (5.1) to perform well in such instances.

What seems really interesting from the practical viewpoint is the minimum closest to the bottom left, where the estimator reaches its better performance and clearly corresponds to cases of light grouping. This zone is caracterized by an average length $\bar{l} \approx 23$ units or less, or dividing by the average range $\bar{r}$, by a ratio $\bar{l}/\bar{r} \approx 0.08$ or less, which gives a clue about when to expect a good performance of the estimator (5.1) in a practical situation.

Figure 5.5 supports the last statement. When the average length $\bar{l}$ is around 20 units

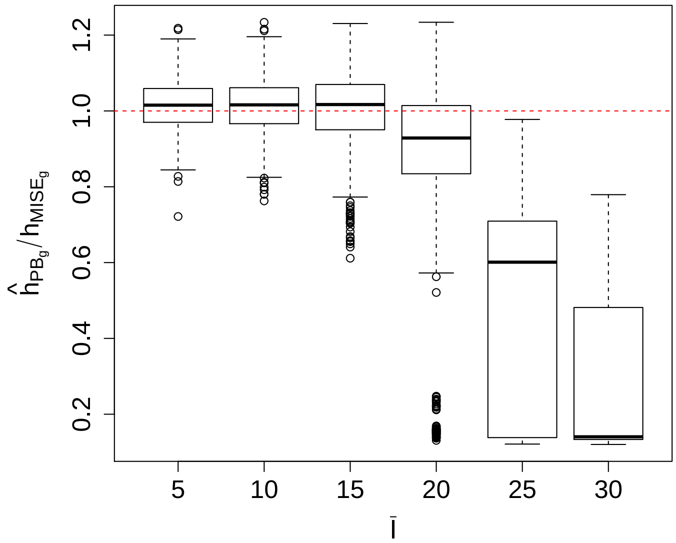

Figure 5.5: Sampling distribution of $\hat{h}_{PB_g}/h_{MISE_g}$ for different average lengths for sample size n=240.

or less, the sampling distribution of (5.4) is centered somewhere around the target value and its variability is more or less constant. As soon as $\bar{l}$ reaches the value of 20, it seems to perform more unstable: a bit biased and more dispersed. When $\bar{l}$ surpasses the value of 20, the bandwidth selector behaves poorly.

Figure 5.6 is a visual example of what may happen when estimating the distribution for values greater or lesser than around 20 units. In the case of $\bar{l} = 15$, using the optimal h_{MISE_g} and the estimated bandwidth $\hat{h}_{PB_g}$, both estimations are indistinguishable since both bandwidths are practically the same (i.e., the bandwidth selector performs well). On the contrary, when $\bar{l} = 25$, $\hat{h}_{PB_g}$ is quite far from its target, h_{MISE_g}. Then, the corresponding estimates are notoriously different.

5.4 Applications

As in the case of kernel density estimation for grouped data, to verify the performance of the estimator (5.1) via the bandwidth selector (5.4), the Old Faithful geyser data will be used. It contains the time between eruptions for the Old Faithful geyser in the Yellowstone National Park, Wyoming, United States. The dataset is available in the R environment for statistical computing. It is worth to note that the sample size of this data set is 272, similar to 240, one of the sample sizes considered in the previous simulation studies, which

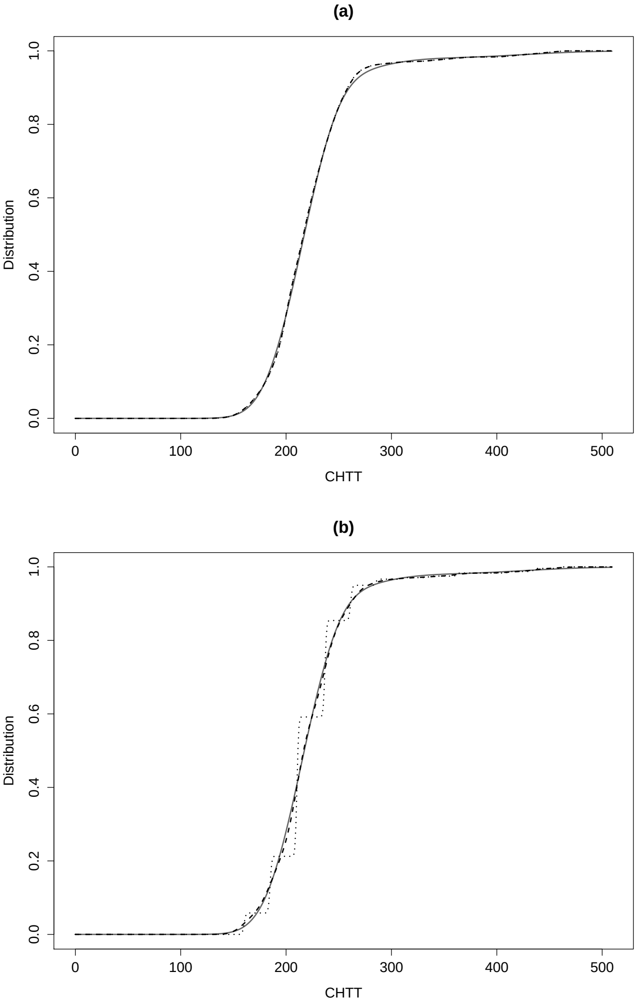

Figure 5.6: Kernel estimation using estimator (5.1) with a sample of size 240. In (a) $\bar{l} = 15$ and bandwidths $h_{MISEg} = 7$ (dashed line) and $\hat{h}_{PB_g} = 6.4$ (dotted line). In (b), $\bar{l} = 25$ and $h_{MISEg} = 9.5$ (dashed line) and $\hat{h}_{PB_g} = 1.3$ (dotted line). In both, solid lines represent the reference mixture distribution.

favors for comparison.

Instead of the average range $\bar{r}$, it will be used the data range r and the ratio $\omega = \bar{l}/r$. The reference distribution will be the one provided by the standard estimator (2.41) for complete data using the plug-in bandwidth $\hat{h}_{PB}$.

Figure 5.7 (a) shows similar results to that obtained by the simulation study in Subsection 5.3.2: the ratio $\hat{h}_{PB_g}/\hat{h}_{PB}$ appears to have an average value of 1 up to approximately $\omega = 0.075$. From this value onwards, the bandwidth selections begin to fall short. This is also verified in 5.7 (b), where the ISD between $\hat{F}_h^g$ and $\hat{F}_h$ begin to markedly increase starting from $\omega \approx 0.075$. However, in this case of distribution estimation, after $\omega \approx 0.075$, bandwidth selections begin to fall short not so quickly as in the case of the density (Figure

3.7 (a)). This means that the bad performance of the estimator begins to be notorious for a bit more heavy grouped data; i.e., this suggests again that kernel distribution estimation for grouped data is somewhat more robust than kernel density estimation in the same case.

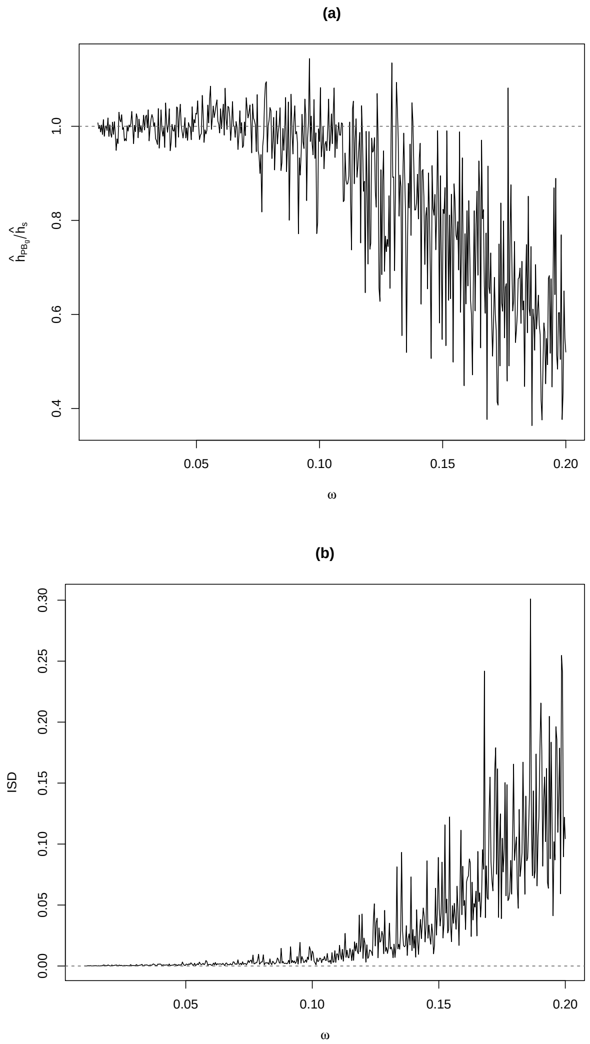

Figure 5.7: (a) $\hat{h}_{PB_g}/\hat{h}_{PB}$ versus $\omega = \bar{l}/r$. (b) Integrated squared distance (ISD) $\int \left[\hat{F}^g_{\hat{h}_{PB_g}}(u) - \hat{F}_{\hat{h}_{PB}}(u) \right]^2 du$ versus ω.

The latter can be seen in Figure 5.8. Note that unlike the case of density estimation (Figure 3.8), in this case it was neccesary to consider a more heavy grouped data case to really start noticing the bad performance of the bandwidth selector $\hat{h}_{PB_g}$ as well as the estimator $\hat{F}^g_h$.

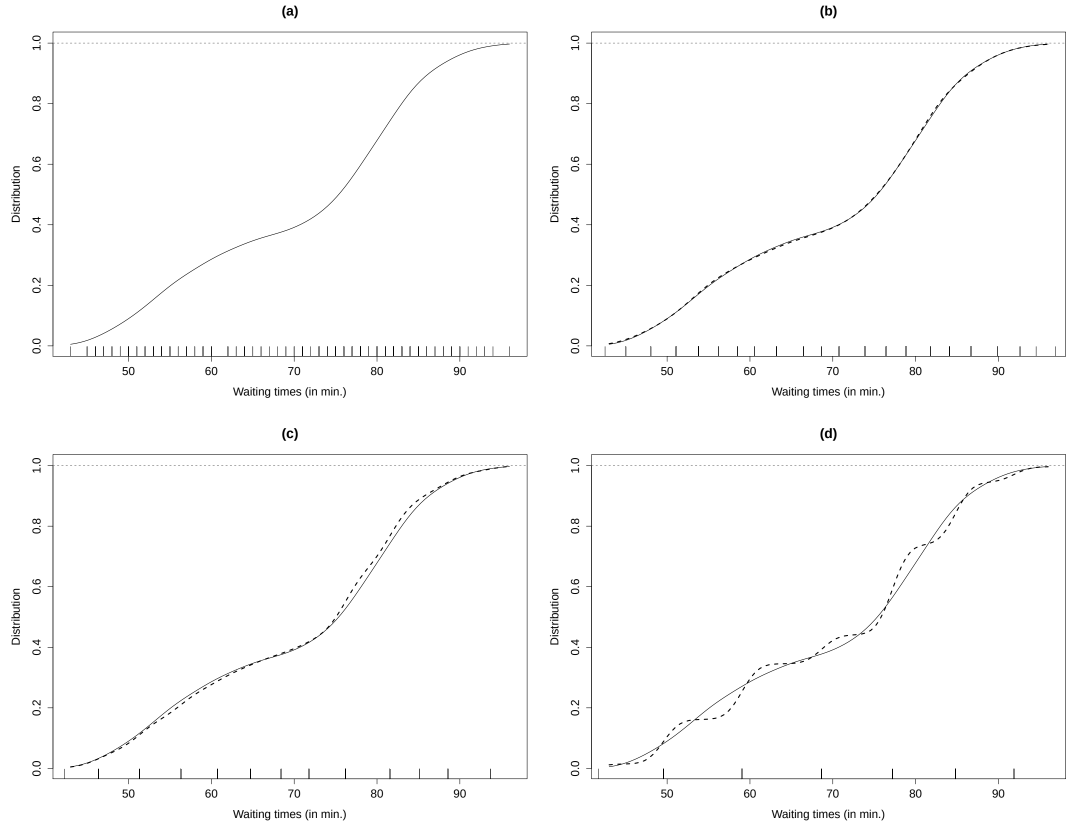

Figure 5.8: Kernel distribution estimation: (a) using the standard estimator $\hat{F}_h$ with ungrouped data and $\hat{h}_s = 2.012$; (b) using $\hat{F}_h^g$ with $\omega = 0.05$ and $\hat{h}_s = 2.013$; (c) using $\hat{F}_h^g$ with $\omega = 0.08$ and $\hat{h}_s = 1.937$; (d) using $\hat{F}_h^g$ with $\omega = 0.15$ and $\hat{h}_s = 1.571$. In all four cases, the solid line represents the kernel distribution estimation using $\hat{F}_h$(ungrouped data), while dashed lines are the kernel distribution estimations using $\hat{F}_h^g$ (grouped data).

5.5 Summary

In short, it has been shown that under the right assumptions, the kernel distribution estimator is an effective tool due to the good performance of the corresponding plug-in bandwidth selector. In practice, when there is a fixed sample size and a given set of intervals, the good performance of the plug-in bandwidth selector is limited to a certain degree of grouping, which in this context may be referred to as light grouping.

The different simulations performed in this chapter show that the kernel distribution estimator is somewhat more robust than the kernel density estimator, in the sense that bandwidth selections slightly different from the optimal bandwidth do not greatly influence the distribution estimation, as it does in the case of kernel density estimation.

The latter is reinforced by studying the behavior of the plug-in bandwidth selector considering different grouping levels. Our simulations show that, in the case of the distribution, the plug-in bandwidth selector can give good results in grouping conditions for which, in the case of density estimation, it falls short. This property gives the kernel distribution estimator certain advantage over the kernel density estimator, since its plug-in bandwidth resists more in cases of relatively heavy grouping; thus, letting the kernel distribution estimator perform well at grouping levels in which the kernel density estimator needs the use of somewhat more elaborated bandwidth selectors. Because of this, in the case of distribution estimation, an alternative bandwidth selector is not that necessary as in the case of density estimation. Nevertheless, it would be of great advantage to propose a more accurate bandwidth selector. This is an interesting issue for future research.

Chapter 6

Applications and empirical studies

A key part of any research in statistical techniques is its application to real data. This chapter considers three real grouped data sets on seedling emergence obtained in weed science studies[1]. The aim is to estimate the density and distribution functions on each data set by means of the estimators studied in the previous chapters, and to compare their performance with parametric techniques commonly used by weed scientists. Also, based on those grouped data sets, the bandwidth selectors studied in Chapter 4 are tested over different grouping conditions through a simulation study. The results suggest that, in general, the nonparametric techniques proposed in this work perform acceptably and, in some cases, they would be more suitable than parametric methods for studying emergence curves in weed science.

6.1 About the grouped data sets

The three grouped data sets considered refers to *Phalaris paradoxa L.* (hood canary grass) seedling emergence, which is one of the most problematic weeds of winter cereals in Mediterranean climates (Alemseged et al., 2001; Jiménez-Hidalgo et al., 1997). It is very abundant in cereal fields in southern Spain, where it represents a major problem (González-Andújar and Saavedra, 2003). *Phalaris paradoxa* is an agressive crop competitor and, when unmanaged, it may reduce wheat yields up to 40% (Delow and Milne, 1986). The way of controlling this weed is mainly by herbicides, implying a realtively large investment in this kind of products and the possibility of creating herbicide resistant populations. Some changes in farming practices have exacerbated the incidence of this weed, especially the adoption of conservation tillage.

Seedling emergence experiments were conducted from fall to spring during three consecutive seasons in an area with no previous history of *P. paradoxa* infestation in the ETSIA experimental field of the University of Seville (37.35 N, 5.93 W; 21 m a.s.l., Seville,

[1] Very special thanks to José María Urbano, from the University of Seville, who kindly provided these data.

Experiment 1

CHTT	*A. Counts*	*C. counts*	w_i	*C. proportion*
41.08	9.5	9.5	0.057	0.057
82.16	14.5	24.0	0.087	0.144
103.42	27.0	51.0	0.162	0.306
124.68	36.5	87.5	0.219	0.526
171.79	38.8	126.3	0.233	0.759
231.32	16.5	142.8	0.099	0.858
243.74	6.5	149.3	0.039	0.897
269.13	5.8	155.1	0.035	0.932
346.03	5.3	160.4	0.032	0.963
422.16	2.0	162.4	0.012	0.975
471.02	1.3	163.7	0.008	0.983
519.68	1.0	164.7	0.006	0.989
593.37	0.5	165.2	0.003	0.992
642.64	1.3	166.5	0.008	1.000

Table 6.1: Average counts, cumulative average counts, weights and cumulative proportions of *P. paradoxa* seeds emerged at each CHTT.

Andalusia, Southern Spain). For the experiments, mature caryopses of *P. paradoxa* (from now on, seeds) were collected in June 2005 from a wheat field near Jerez, about 90 km Southwest of Seville, and stored in airtight containers at 4°C until ready for use.

In each study season, four 25 x 25 cm plots were randomly established and the soil up to 5 cm deep was replaced by a substrate. The substrate was a mixture of 50% Kekkilä garden peat (Kekkilä Oy, Finland), 25% sand, and 25% local silt loam soil. After sterilization by steam under pressure, the amount of substrate for each plot was mixed with 500 *P. paradoxa* seeds and incorporated to plots on 11, 22, and 29 November 2005, 2006, and 2007, respectively, within the local range of cereal sowing dates. Seed losses to surface-foraging predators were prevented by placing 2-mm mesh cages over the plots.

In each season, numbers of emerged seedlings were recorded at weekly intervals from sowing until seedling emergence ceased (approx. mid April). Censed seedlings were immediately removed with minimum disturbance of the substrate.

Climatic variables were obtained from a meteorological station located 15 km away from the experimental field. Soil temperature and water potential (ψ) at 5 cm depth were estimated using the STM^2 software (Spokas and Forcella, 2009). STM^2 requires inputs of daily weather data, along with information on the geographical location, soil texture and organic matter content.

Soil temperature and water potential were used to calculate HTT for day t, $\theta_{HTT}(t)$, by means of the following equation (Schutte et al., 2008):

$$\theta_{HTT}(t) = \theta_H(t) \cdot \theta_T(t),$$

Experiment 2

CHTT	*A. counts*	*C. counts*	w_i	*C. proportion*
88.91	2.0	2.0	0.012	0.012
102.20	40.8	42.8	0.242	0.254
128.94	7.3	50.1	0.043	0.297
183.71	6.8	56.9	0.040	0.337
265.37	8.0	64.9	0.047	0.385
311.71	8.5	73.4	0.050	0.435
317.70	27.0	100.4	0.160	0.595
323.13	14.0	114.4	0.083	0.679
348.51	30.8	145.2	0.183	0.861
402.07	5.0	150.2	0.030	0.891
455.87	12.3	162.5	0.073	0.964
504.58	4.3	166.8	0.026	0.989
559.84	1.8	168.6	0.011	1.000

Table 6.2: Average counts, cumulative average counts, weights and cumulative proportions of *P. paradoxa* seeds emerged at each CHTT.

Experiment 3

CHTT	*A. counts*	*C. counts*	w_i	*C. proportion*
88.67	138.3	138.3	0.813	0.813
112.32	4.8	143.1	0.028	0.841
176.92	13.8	156.9	0.081	0.922
235.57	6.5	163.4	0.038	0.961
253.25	2.5	165.9	0.015	0.975
259.55	0.3	166.2	0.002	0.977
280.57	1.3	167.5	0.008	0.985
295.29	1.8	169.3	0.011	0.995
316.14	0.5	169.8	0.003	0.998
336.98	0.3	170.1	0.002	1.000

Table 6.3: Average counts, cumulative average counts, weights and cumulative proportions of *P. paradoxa* seeds emerged at each CHTT.

where $\theta_H(t) = I_{[\psi(b),\infty)}(\psi_t)$. Therefore, $\theta_H(t) = 1$ when the actual water potential at day t, $\psi(t)$, is larger than or equal to the base water potential for seed germination, ψ_b; otherwise, $\theta_H(t) = 0$ and

$$\theta_T(t) = \max\{T(t) - T_b, 0\},$$

where $T(t)$ is the daily average soil temperature at day t and T_b is the base temperature for seed germination. Cumulative hydrothermal time (CHTT) starting at weed sowing up to day s is defined as

$$\Theta_{CHTT}(s) = \sum_{t=1}^{s} \theta_{HTT}(t).$$

Base temperature (T_b) and water potential (ψ_b) for *P. paradoxa* seedling emergence were considered at 0.8°C and -1.50 MPa, respectively.

Concerning the data, it should be noted that time (in seconds, hours, or days) and cumulative hydrothermal time are not changing synchronously. At a given time t_{i-1}, there is an observed cumulative hydrothermal time CHTT_{i-1}. For a next inspection at time t_i, there is an associated CHTT_i. There are two possibilities: 1) $\text{CHTT}_i > \text{CHTT}_{i-1}$, or 2) $\text{CHTT}_i = \text{CHTT}_{i-1}$. In the first case, the number of emerged seeds between CHTT_{i-1} and CHTT_i, n_i, is associated with the midpoint of the interval. In the second case, n_i is just associated to the still observed value $\text{CHTT}_i = \text{CHTT}_{i-1}$ (which can be thought of as a "midpoint" itself). Under these considerations, the grouped data sets are shown in Tables 6.1, 6.2 and 6.3. The first column, CHTT, refers to the values of the so called "midpoints". The second column stands for "average counts" observed at each CHTT, since in each experiment there were four repetitions. The elements in the third column are just the cumulative counts; the fourth column contains the "weight" associated to each CHTT, by means of $w_i = n_i/n$, where n is the total number of emerged seedlings. The fifth column contains the cumulative proportion of emerged seedlings at each CHTT.

6.2 Density estimation and simulation study

A first goal is to determine the structure of the data by means of estimating the density. For this, a suitable bandwidth selector, plug-in or bootstrap, has to be picked. A quick first analysis of the data reveals that the average distance between CHTT is around 48 units for the first two experiments and 21 units for the third. Also, for each experiment, the range of the data is roughly 602, 471 and 248 units, respectively. Dividing the former by the latter in each experiment gives $\omega_1 \approx 0.081$, $\omega_2 \approx 0.101$ and $\omega_3 \approx 0.084$. According to our previous guidelines, the advise is to choose the plug-in selector whenever the data is lightly grouped, and to choose the bootstrap selector otherwise. The values ω_1, ω_2 and ω_3 suggest that data are somewhat heavily grouped, so, in principle, the bootstrap selector

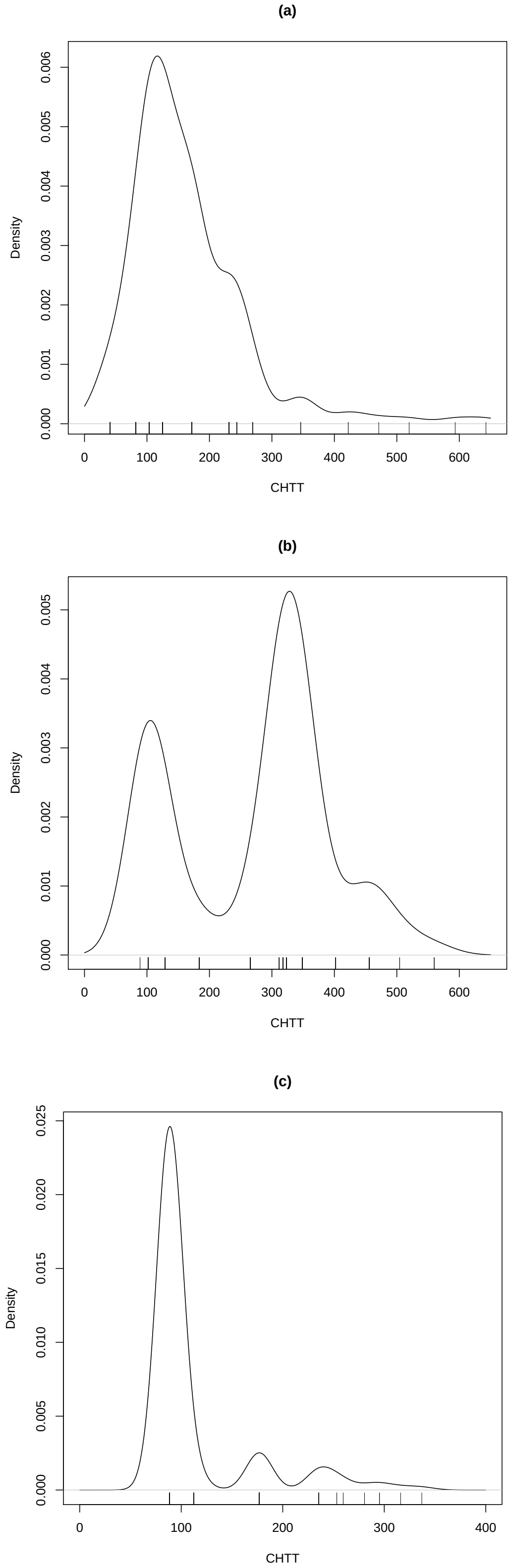

Figure 6.1: Kernel density estimation of *P. paradoxa* seedling emergence: (a) experiment 1 (Table 6.1), using bandwidth $h^*_{MISE} = 27.2$; (b) experiment 2 (Table 6.2), with bandwidth $h^*_{MISE} = 33.7$; (c) experiment 3 (Table 6.3), using bandwidth $h^*_{MISE} = 13.0$. The Gaussian kernel was used in all three cases.

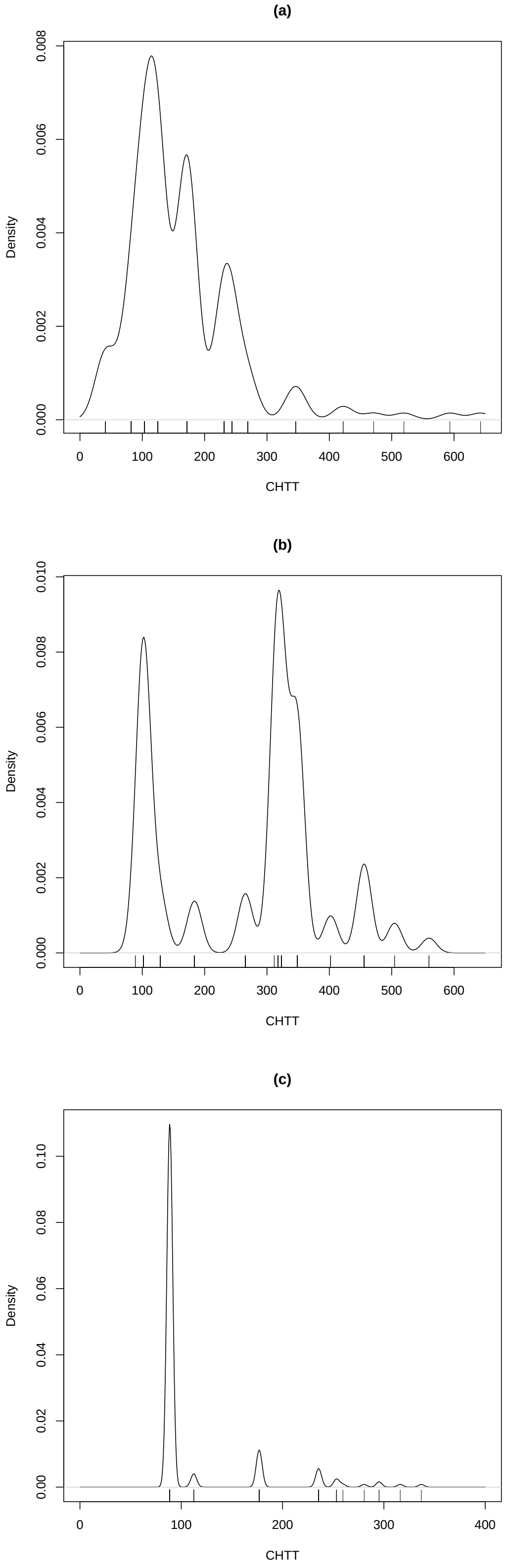

Figure 6.2: Kernel density estimation of *P. paradoxa* seedling emergence: (a) experiment 1 (Table 6.1), using bandwidth $\hat{h}_g = 16.5$; (b) experiment 2 (Table 6.2), with bandwidth $\hat{h}_g = 12.0$; (c) experiment 3 (Table 6.3), using bandwidth $\hat{h}_g = 2.9$. The Gaussian kernel was used in all three cases.

should be chosen.

The possible structure of the data in each experiment is shown in Figures 6.1 and 6.2, using the bootstrap and plug-in selectors, repectively. As can be seen, the density estimations in Figure 6.2 are too wiggly due to the relatively small bandwidths given by the plug-in selector. On the other hand, softer and more reasonable structures are obtained using the bootstrap selector.

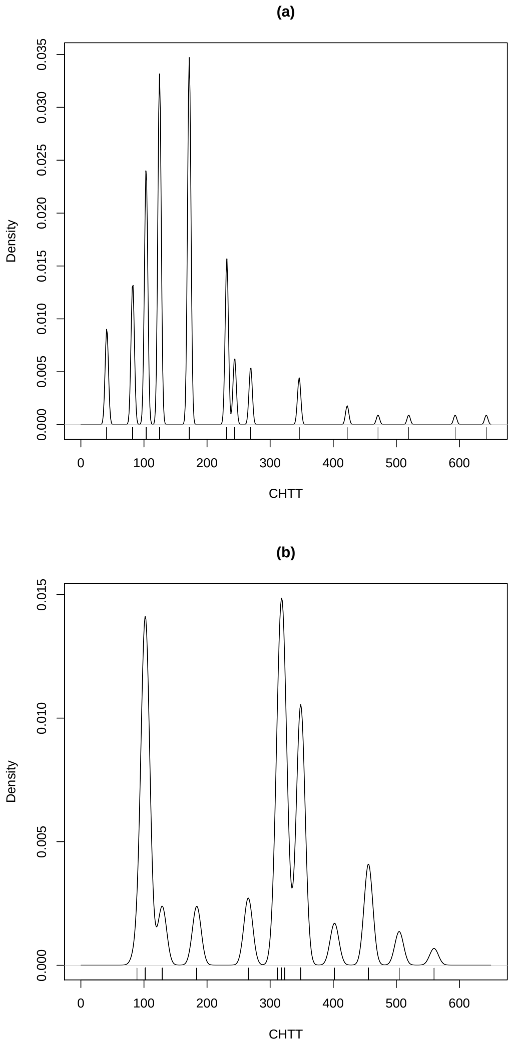

Figure 6.3: Kernel density estimation of *P. paradoxa* seedling emergence: (a) experiment 1 (Table 6.1), using the Sheather and Jones bandwidth $\hat{h}_{SJ} = 2.5$; (b) experiment 2 (Table 6.2), with bandwidth $\hat{h}_{SJ} = 6.9$. The Gaussian kernel was used in both cases. It was not possible to obtain the Sheather and Jones bandwidth using the data from experiment 3 (Table 6.3), due to the data sparseness.

Besides, Figure 6.3 shows the density estimations when using the method of Sheather and Jones, one of the most popular data-driven bandwidth selectors over the past years

(Sheather and Jones, 1991). Clearly, the Sheather and Jones bandwidth selector is not adequate at all to be used with grouped data, since due to the lack of information, it seems to be even more sensitive than the plug-in bandwidth selector $\hat{h}_g$. Note that the Sheather and Jones bandwidths are still lower than those given by the plug-in selector when using data from experiments 1 and 2, and when using those from experiment 3, the method is unable to deal with the data sparseness and hence giving no results. Thus, from now on, the structures shown in Figure 6.1 will be considered.

At this point, the density estimations obtained for each experiment may be used by weed scientists to determine probabilities of *P. paradoxa* seedling emergence. However, one more step further will be given. Trusting on the visual impression and assuming that the density estimations are valid, let us consider these estimates as a sort of "pilot" density estimations and then propose acceptable models for the seedling emergence in each experiment. In doing so, it is possible to test the recommendations for chosing the bandwidth selector by means of a simulation study, considering samples from this models and evaluating the density estimates based on plug-in or bootstrap selectors at different grouping conditions.

Reasonable normal mixtures for the seedling emergence patterns in the three experiments considered are

$$f_1\left(x\right) = \sum_{i=1}^{3} \alpha_{1i} \phi_{\mu_{1i},\sigma_{1i}}\left(x\right),$$

$$f_2\left(x\right) = \sum_{i=1}^{3} \alpha_{2i} \phi_{\mu_{2i},\sigma_{2i}}\left(x\right)$$

and

$$f_3\left(x\right) = \sum_{i=1}^{3} \alpha_{3i} \phi_{\mu_{3i},\sigma_{3i}}\left(x\right),$$

where $\phi_{\mu,\sigma}$ is a $N\left(\mu,\sigma\right)$, μ_{ji} is the i-th component of the j-th vector of means, σ_{ji} is the i-th component of the j-th vector of standard deviations, and α_{ji} is the i-th component of the j-th vector of weights. Specifically, $\vec{\mu}_1 = (150, 230, 340)$, $\vec{\sigma}_1 = (38, 25, 30)$, $\vec{\alpha}_1 = (0.79, 0.16, 0.05)$; $\vec{\mu}_2 = (110, 340, 490)$, $\vec{\sigma}_2 = (30, 48, 40)$, $\vec{\alpha}_2 = (0.30, 0.61, 0.09)$; $\vec{\mu}_3 = (90, 175, 240)$, $\vec{\sigma}_3 = (38, 25, 30)$, $\vec{\alpha}_3 = (0.83, 0.10, 0.07)$.

To assess the accuracy of the estimations at different degrees of grouping, the mean integrated squared error using the kernel density estimator with complete data, $MISE\left(\hat{f}_h\right)$, considering (2.34) as the bandwidth selector , was considered as a reference. At each grouping level, once the density was estimated, the integrated squared error $ISE\left(\hat{f}_h^g\right) = \int \left[\hat{f}_h^g\left(u\right) - f_j\left(u\right)\right]^2 du$, for $j = 1, 2, 3$, was calculated. Then, to assess the accuracy, it was used the ratio $\rho = ISE\left(\hat{f}_h^g\right) / MISE\left(\hat{f}_h\right)$.

This process is summarized in the following steps:

1. Consider a set of average interval lengths values $\left\{\bar{l}_i\right\}$, $i = 1, 2, 3, \cdots, m$.

2. Consider the model f_j, $j = 1, 2, 3$, and simulate a sample of size $n_0 = 170$ (this sample size is similar to those used in the three experiments).

3. At the i-th trial, group the data according to $\bar{l}_i$.

4. Consider $\omega_i = \bar{l}_i / r_i$, where r_i is the data range. The grouped data consists of the midpoints of the intervals repeated as many times as the number of data in each interval. Using this censored sample, obtain the bandwidth with both the plug-in and bootstrap selectors.

5. Use each bandwidth to estimate the density and for $j = 1, 2, 3$, calculate $ISE_i\left(\hat{f}_h^g\right) = \int \left[\hat{f}_h^g(u) - f_j(u)\right]^2 du$.

6. Compute $\rho_i = ISE_i\left(\hat{f}_h^g\right) / MISE\left(\hat{f}_h\right)$.

7. In order to obtain the average trend, for each i, repeat the previous steps 1000 times and obtain the average $ISE_i\left(\hat{f}_h^g\right)$ (i.e., the $MISE_i\left(\hat{f}_h^g\right)$). Then, compute $\overline{\rho_i} = MISE_i\left(\hat{f}_h^g\right) / MISE\left(\hat{f}_h\right)$.

As before in this work, the simulation was done using the environment for statistical computing `R` (R Core Team, 2015).

Figure 6.4 shows the common logarithm of ρ_i and $\overline{\rho_i}$ at each value ω_i for both the plug-in and the bootstrap bandwidth selectors presented in Section 4.1, when simulating from each of the three models proposed. In general, the first thing to note is that for small values of ω, the average trend of $\log_{10}(\rho)$ is close to zero, which means that the $MISE_i\left(\hat{f}_h^g\right)$ is quite close to the $MISE\left(\hat{f}_h\right)$. In other words, for ω up to around 0.075 (or a little bit more in some case), both selectors perform well in general, and so, the kernel density estimator for grouped data seems to perform as good as in the case of continuous data. Nevertheless, it has to be mentioned that in the cases (b) and (c), the plug-in bandwidth performs a bit better on average than the bootstrap one. Based on this, it could be said that in cases of light grouping, although bootstrap selectors perform well, the first choice should be the plug-in selector.

The situation is quite different for large values of ω. Of course, since there is more uncertainty in the data, it is naturally expected that density estimations get worse as the degree of grouping increases. However, and because of that, it is remarkable to observe the smaller error in density estimations when using the bootstrap with respect to the plug-in selector.

		(a)		(b)		(c)	
		$p.i.$	b	$p.i.$	b	$p.i$	b
ω	0.05	0.0095	0.0091	0.0552	0.2358	0.0512	0.1282
	0.10	0.2405	0.0885	0.3315	0.4292	0.6929	0.3384
	0.15	1.6945	0.2679	1.2615	0.6879	1.7774	0.6266
	0.20	2.1798	0.5104	1.6216	0.8980	2.0585	0.9721

Table 6.4: $\log_{10} \overline{\rho}$ for different values of ω, considering the plug-in ($p.i$) and the bootstrap (b) selectors in each of the three models proposed: (a) f_1, (b) f_2, (c) f_3.

For example, when $\omega = 0.15$ (which means that data is contained in around 7 intervals), and considering the three models jointly, $\log_{10}(\overline{\rho})$ ranged from around 0.3 to 0.7, meaning that the $MISE_i\left(\hat{f}_h^g\right)$ ranged from approximately 2 to 5 times $MISE\left(\hat{f}_h\right)$. In contrast, when using the plug-in selector, $\log_{10}(\overline{\rho})$ happend to range from around 1.3 to 1.8, so that, on average, the $ISE_i\left(\hat{f}_h^g\right)$ roughly went from 20 to 60 times the $MISE\left(\hat{f}_h\right)$. Moreover, considering the most extreme grouping case, $\omega = 0.20$ (data contained in just around five intervals), note that $\log_{10}(\overline{\rho})$ did not surpass 1 when using the bootstrap selector; i.e., the $MISE_i\left(\hat{f}_h^g\right)$ was, overall the three models, at most 10 times $MISE\left(\hat{f}_h\right)$. On the other hand, when using the plug-in selector, $\log_{10}(\rho) \approx 2$ in all three cases, meaning that the $MISE_i\left(\hat{f}_h^g\right)$ was around a hundred times $MISE\left(\hat{f}_h\right)$. Table 6.4 helps to clarify this.

To have an idea of how the density estimates look at different grouping conditions, three representative values for ω were considered: 0.05, 0.10, 0.15. The first still represents light grouping; the second one may represent a degree of grouping somewhat in the border between light and heavy grouping. The third one is clearly a case of heavy grouping. In each case, the models f_1, f_2 and f_3 were estimated considering the plug-in for the first value of ω, and the bootstrap selector for the other two values.

Figure 6.5 shows how the estimates naturally become more and more deficient as ω increases. In the lightest case, the estimator is capable to approximate the structure of the data in all cases. As the grouping effect increases, the quality of the estimations diminishes, but it is worthy of attention that even when there are as few as 10 to 7 intervals, the estimator is able to still reveal some of the structure of the data by choosing the right bandwidth selector.

6.3 A comparison between kernel distribution estimation and parametric approaches

At the beginning of this thesis, it was mentioned that in trying to assess the relationship between seedling emergence and cumulative hydrothermal time, weed scientists have typically used parametric regression models. However, the main drawback of this approach is perhaps its rigidity to capture complex details in the distribution, like thin spikes, heavy

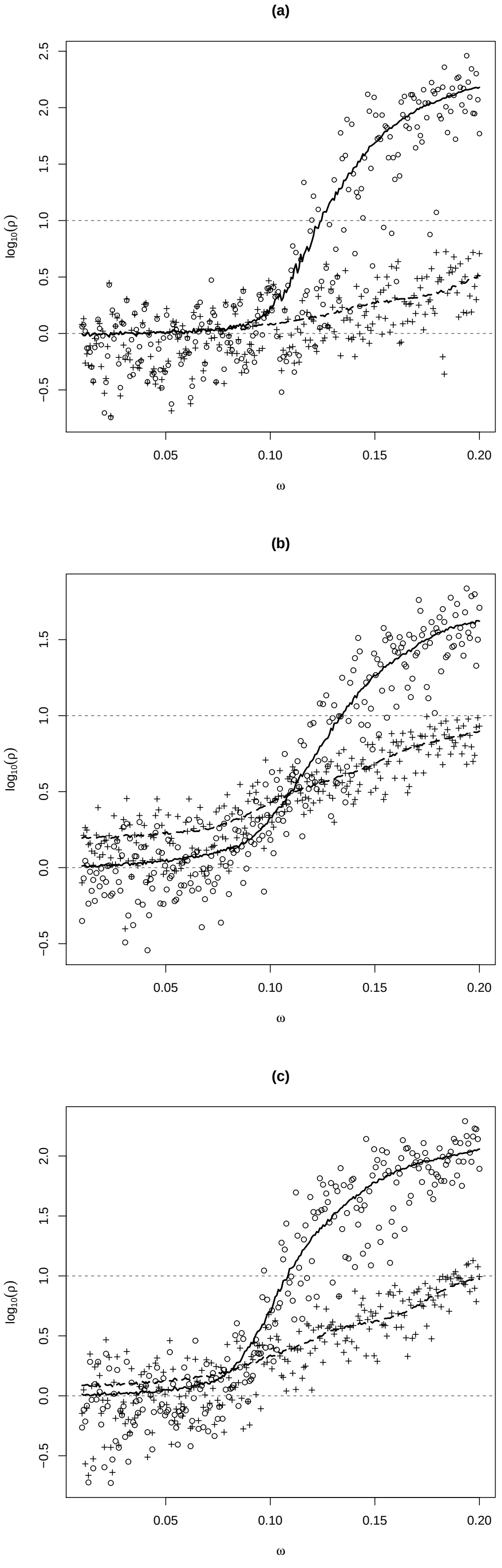

Figure 6.4: $\log_{10}\rho$ and $\log_{10}\overline{\rho}$ versus ω when simulating from (a) model f_1; (b) model f_2; (c) model f_3. Empty circles $\circ$ and $+$ signs correspond to $\log_{10}\rho$ when using the plug-in and the boostrap selectors, respectively. The solid and dotted lines are $\log_{10}\overline{\rho}$, when usign the plug-in and bootstrap selector, respectively.

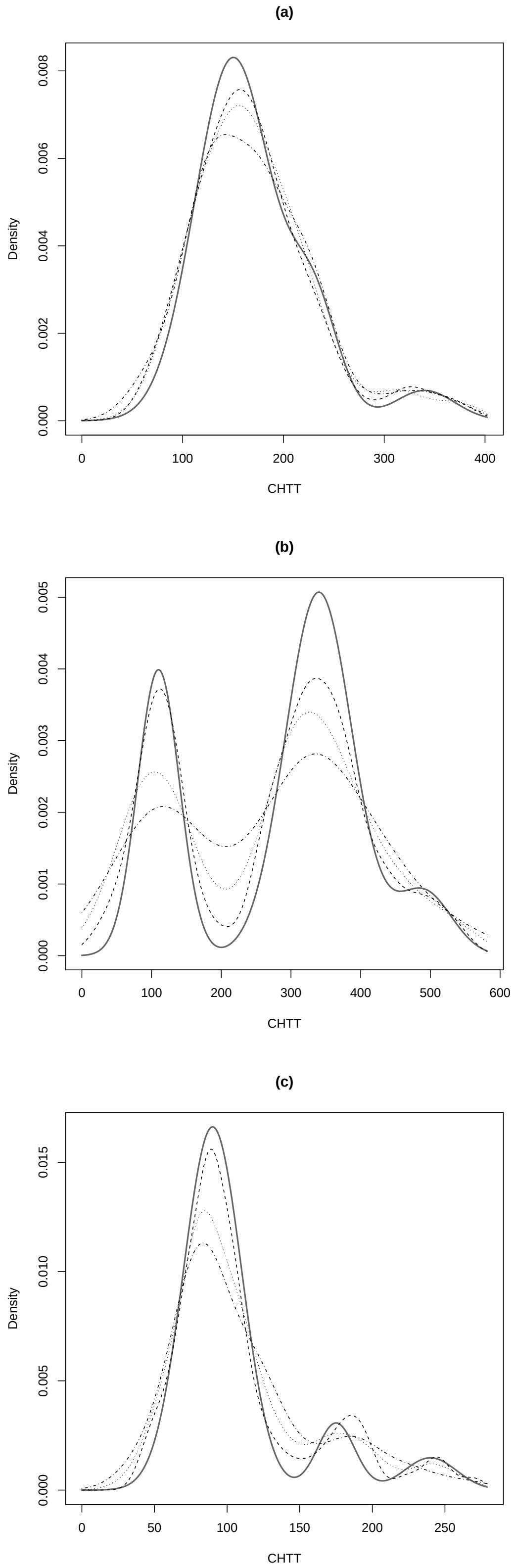

Figure 6.5: Kernel density estimation for: (a) model f_1; (b) model f_2; (c) model f_3, all in solid line. The dashed line corresponds to $\omega = 0.05$, using the plug-in selector. The bootstrap selector was used for $\omega = 0.10$ (dotted line) and $\omega = 0.15$ (dot-dashed line).

tails or subtle details in certain zones. Nonparametric techniques are characterized by its flexibility, which suggests that kernel distribution estimation could be a good option to describe that relationship.

In this section, the performance of some nonlinear parametric regression models is compared to the performance of the nonparametric kernel distribution estimator for grouped data proposed in (5.1). Typical nonlinear regression models used in weed science are the Logistic, Gompertz and Weibull models.

The `R` function `nls` can be used for fitting nonlinear regression. A common problem when using nonlinear least-squares algorithms is that most of them require to specify starting values for the parameters. For getting them automatically, the self-starting nonlinear models can be used (see, for instance, Bates and Watts (1988) for a description of some techniques for finding starting values, and Pinheiro and Bates (2000) for more information about the self-starting models available in `R`).

In `R`, Logistic, Gompertz and Weibull self-starting models are defined as follows:

- Logistic

$$m_{L_\Phi}(x) = \frac{\phi_1}{1 + \exp\left[\left(\phi_2 - x\right)/\phi_3\right]}$$

- Gompertz

$$m_{G_\Phi}(x) = \phi_1 \exp\left[-\phi_2 \phi_3^x\right]$$

- Weibull

$$m_{W_\Phi}(x) = \phi_1 - \phi_2 \exp\left[-\exp\left(\phi_3\right) x^{\phi_4}\right],$$

where Φ is the vector of parameters in each model, and x may refer to the cumulative hydrothermal time in the weed science context.

6.3.1 Comparison of goodness of fit to real data

In this subsection, the three real data sets (Tables 6.1, 6.2 and 6.3) will be used to compare the goodness of fit of the above parametric models and the kernel distribution estimator for grouped data. For this, the standard error of the estimate will be used, defined in general as

$$S = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(Y_i - Y_i'\right)^2},$$

where Y_i is the i-th observed value and Y_i' is the i-th model predicted value. In the present context, Y_i represents the observed cumulative emergence fraction at midpoint t_i, $\hat{F}_n\left(t_i\right)$,

	(a)	(b)	(c)
K	5.2	5.8	12.9
L	3.3	9.3	4.2
G	2.5	9.9	3.9
W	3.2	8.4	2.4

Table 6.5: Standard error of the estimate (multiplied by 100, for readability) when using kernel distribution estimation (K), Logistic (L), Gompertz (G) and Weibull (W) regressions considering data from: (a) Table 6.1, (b) Table 6.2, (c) Table 6.3.

and Y_i' may represent whether the kernel distribution estimation at t_i, $\hat{F}_h^g(t_i)$, or the parametric regression estimation at t_i, $\hat{m}_\Phi(t_i)$.

Figure 6.6 shows the three parametric models considered fitted to the data, as well as the kernel distribution estimation in each case. Regarding the first experiment data (Figure 6.6 (a)), it seems that all the parametric models fit reasonably well. All three, Logistic, Gompertz and Weibull regressions seem to fit very natural to the data, although on the region near to zero, the latter does not satisfy the condition of being greater or equal to zero, as a distribution function should be. Of course, this could be achieved by setting some parameters beforehand, but this could be a disadvantage, because the model would fit the data forcedly. In turn, the flexibility of the kernel distribution estimation allows to better describe the data structure in the top right region, close to one, where all the parametric models fall short due to its rigidity.

Figure 6.6 (b) shows data whose structure is somewhat more complicated. The location of the points suggests that the parametric models will have serious problems to smoothly adjust to the data, as can be confirmed by the graphs. Again, some parameters could be set in advance to try the models to fit the data while satisfying the characteristics of a distribution function, but the result could be quite forced, or there may be errors due the impossibility of convergence in the remaining parameters. It is in this situation where the flexibility of the kernel distribution estimation method greatly exceeds the possibilities of parametric methods. As can be seen in the figure, it smoothly follows through the empirical distribution function.

Figure 6.6 (c) shows a challenging situation, especially for parametric models. There is an important lack of information regarding the cumulative emergence of seeds for small values of the cumulative hydrothermal time. Since the data available corresponds only to high values of the cumulative emergence, when leaving parameters to freely vary, the parametric methods “assume” that the available points correspond to the top of a sigmoid, whose bottom is located in a negative region for the cumulative hydrothermal time. This is a nonsense region for a distribution function of seedling emergence. Hence, under these circumstances, it is necessary to force parametric models to fit to the data fixing beforehand some parameters, which in some cases may not lead to wise results.

An example of this is the Weibull model fit. As can be seen in Figure 6.6 (c), although

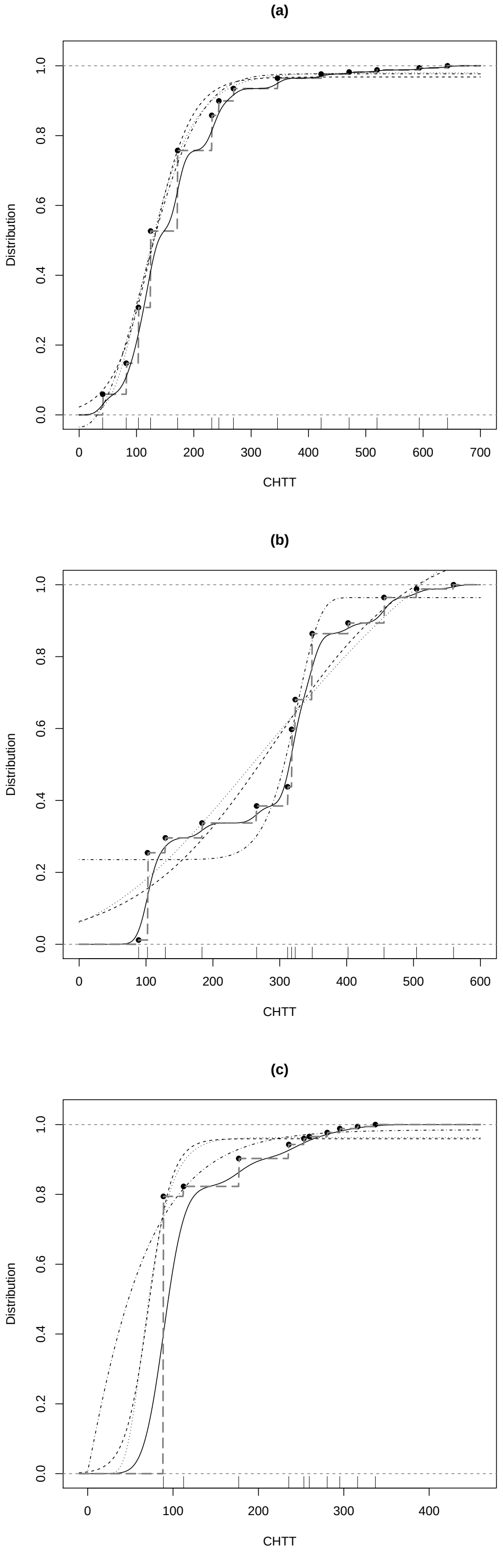

Figure 6.6: Cumulative emergence data (solid dots) from: (a) Table 6.1, (b) Table 6.2, (c) Table 6.3. The models fitted are the kernel distribution estimation (solid line), Logistic (dashed), Gompertz (dotted) and Weibull (dotdashed). The empirical distribution function has also been added (longdash).

they tend to oversimplify the structure, somewhat reasonable functions were obtained when using Logistic or Gompertz regressions, but the Weibull one is unacceptable in the sense that can generate discontinuities near zero. In contrast, the kernel distribution estimation does not have any of the difficulties mentioned above, and just by appropriately choosing the bandwidth may give quite reasonable results. Again, it can be seen how the kernel distribution estimation smoothly passes through the empirical distribution function, being able to better describe the upper part of the distribution.

Some information regarding the goodness of fit of the three parametric models and the kernel distribution estimation can be seen in Table 6.5. In case (a), it is clear that the location of the points favors the three parametric models to fit well to the data, as confirmed by the values of the standard error of the estimate. Besides, without necessarily being a bad choice, the kernel distribution estimation has the worst punctuation.

Case (b) numerically confirms what has been explained: the rigidity of the parametric models makes it difficult to adequately describe the distribution of such data. Clearly, the kernel distribution estimation outperforms all of them, obtaining by far the best score.

Case (c) is worth to analyze. Apparently, just based on the standard error of the estimate, any of the three parametric models could be considered as a better choice than the kernel distribution estimation. However, some aspects should be taken into account before deciding. On the one hand, the standard error of the estimate certainly supports that the parametric models are, by far, closer to the observed data points than the kernel distribution estimation, but, since they have been forced to meet some conditions, the price to pay is that they oversimplify the structure. So, paradoxically, on average they are closer to the data, but barely describe the distribution. On the other hand, the kernel distribution estimation appears to be more distant to the data points, but it seems to better describe the distribution. Indeed, its relatively high score regarding the standard error of the estimate is mainly due to its vertical distance to the most left-handed data point. This fact is, by the way, consistent with the lack of information on that region.

6.3.2 Comparison of goodness of fit to model distributions

Next, the goodness of fit of the parametric models and the kernel distribution estimator will be tested by means of a simulation. It will be carried out obtaining samples from the models f_1, f_2, f_3, and reproducing the grouping conditions in each experiment. To evaluate the closeness of the parametric models and the kernel density estimator for grouped data to the model distributions F_1, F_2, F_3, the $MISE$ will be used.

To carry out the simulation, the grouping conditions of the data in each experiment (see Tables 6.1, 6.2 and 6.3) were reproduced; i.e., based on the normal mixtures models associated to each experiment, f_1, f_2 and f_3 (see Section 6.2), samples were simulated and grouped in such a way that the grouped samples had $\omega_1 \approx 0.081$ when using f_1, $\omega_2 \approx 0.101$ and $\omega_3 \approx 0.084$ when using f_2 and f_3, respectively.

As for the kernel distribution estimation, the simulation went as follows:

1. Consider the model f_j, $j = 1, 2, 3$, and simulate a sample of size $n_0 = 170$.
2. Group the sample in such a way that the ratio $\bar{l}/r \approx \omega_j$. The grouped sample consists of the midpoints of the intervals repeated as many times as the number of data in each interval.
3. Considering the grouped sample, obtain the bandwidth using the Polansky & Baker selector (2.56) and estimate the distribution over a suitable grid of values $\{x_i\}$, $i = 1, 2, 3, \cdots, m_1$, using $\hat{F}_h^g$ (Eq. 5.1).
4. Calculate the integrated squared error $ISE\left(\hat{F}^g_{\hat{h}_{PB_g}}\right) = \int \left[\hat{F}^g_{\hat{h}_{PB_g}}(u) - F_j(u)\right]^2 du$, where F_j is the distribution function related to the model f_j; i.e., $F_j' = f_j$.
5. Repeat the process 1000 times and obtain the average and the standard deviation of those thousand $ISE\left(\hat{F}^g_{\hat{h}_{PB_g}}\right)$.

As for the parametric models, the simulation went as follows:

1. Consider the model f_j, $j = 1, 2, 3$, and simulate a sample of size $n_0 = 170$.
2. Group the sample in such a way that the ratio $\bar{l}/r \approx \omega_j$. The grouped sample consists of the midpoints of the intervals repeated as many times as the number of data in each interval.
3. Using the grouped sample, estimate the empirical distribution function (Eq. 1.5) over a suitable grid of points $\{x_i\}$, $i = 1, 2, 3, \cdots, m_2$.
4. With the set of pairs $\left[x_i, \hat{F}_n(x_i)\right]$, adjust the models m_{L_Φ}, m_{G_Φ}, m_{W_Φ} and estimate the parameter vector Φ by $\hat{\Phi}$.
5. For each model, obtain the $ISE\left[m_{L_{\hat{\Phi}}}\right]$, $ISE\left[m_{G_{\hat{\Phi}}}\right]$, $ISE\left[m_{W_{\hat{\Phi}}}\right]$.
6. Repeat the process 1000 times and obtain the average and standard deviation of those thousand ISE for each model.

For the kernel distribution estimation and for the Polansky & Baker bandwidth selector, the Gaussian kernel was used along with the `R` package `kerdiest` (Quintela-del-Río and Estévez-Pérez, 2012).

Table 6.6 shows results about how well the set of estimated distribution and regression curves approximate the actual underlying cumulative emergence curves F_j, $j = 1, 2, 3$. Regarding the first model, f_1, it can be noted that among the three parametric models used, one of them, the Gompertz model, is a solid candidate to model the distribution F_1

	(a)		(b)		(c)	
	Mean	SD	Mean	SD	Mean	SD
K	0.182	0.148	0.450	0.398	0.207	0.148
L	0.319	0.180	2.592	1.168	0.793	0.347
G	0.187	0.154	3.124	1.766	0.464	0.196
W	4.532	7.239	3.995	1.032	NA	NA

Table 6.6: Mean and standard deviation of the ISE using kernel distribution estimation (K), Logistic (L), Gompertz (G) and Weibull (W) regressions when simulating samples from: (a) f_1, (b) f_2, (c) f_3.

of seedling emergence. But also, note that kernel distribution estimation is a very good option for modelling seedling emergence under this model. Both Gompertz and kernel distribution estimation have very similar performances concerning closeness to F_1 (mean ISE) and precision (SD). Knowing the rigidity of parametric models, F_1 should be an easy curve to describe, with no complex features nor very special variations. This can be confirmed in a sense by looking at Figure 6.1, plot (a). Despite some specific details of the density on the right tail, its main feature is given by the big bell-shaped part on the left.

Concerning the second model, f_2, the results show that its distribution F_2 is a bit more difficult to approximate. Among the parametric models, the one that seems to better perform is the Logistic, but the kernel distribution estimator clearly outperforms it. Certainly, based on Figure 6.1, plot (b), it can be inferred that the combination of notably high and low density zones gives the distribution some details that are difficult to capture by the rigid parametric models.

With respect to model f_3, the Logistic and Gompertz models had better results than those obtained in model f_2, and regarding the self-starting Weibull algortihm, it was striking that for the majority of samples it was not capable of finding the optimal parameters for adjusting the model. This is, in fact, an added problem that may appear in practice, and the ease of use of the automatic procedures should be replaced by manual procedures in which the user has to figure out either the starting values of the parameters or, plainly, to select the optimal parameters “by eye”. In contrast, the nonparametric kernel distirbution estimator performed very decently, obtaining similar results to those obtained in model f_1.

Based on Table 6.6, Figure 6.7 shows examples of kernel distribution estimation compared to the parametric regression fits in each model. In (a), except the Weibull one, Logistic and Gompertz regressions as well as the kernel distribution estimation fit very well and very similarly to each other, although the nonparametric method seems to adjust better at subtle features of the distribution, like at the most upper right or bottom left part of the plot. These little differences may give this method a slight advantage over the parametric regression, as numbers confirm in Table 6.6.

Case (b) clearly shows that sometimes, parametric models may be innapropiate, as they may simplify too much. Even though it is not a very complex one, the structure of

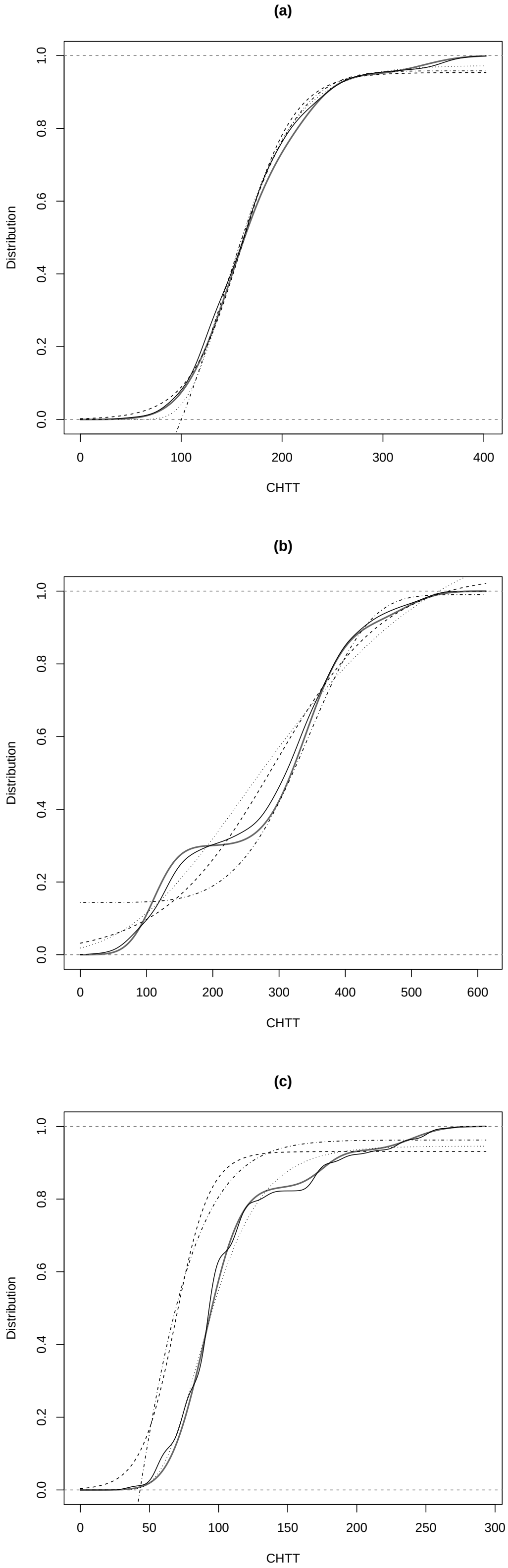

Figure 6.7: Comparison between kernel distribution estimation (solid, black line) and parametric regressions: Logistic (dashed), Gompertz (dotted), Weibull (dotdashed). The actual distributions (a) F_1, (b) F_2, (c) F_3, are in grey, thick solid line.

the data gives the distribution function special characteristics that are impossible for the parametric model to reproduce. It is in cases like this where the flexibility of nonparametric methods is very suitable.

Case (c) is somewhat similar to case (b), in the sense that the distribution shows specific features that the parametric model cannot reproduce, tending to oversimplify the structure. Although the Gompertz regression model roughly gives a good fit, the nonparametric approach is able to more finely describe those subtle details, mainly at the most upper right part of the plot. The other two regression models are clearly out of consideration.

6.4 Summary

In this chapter, kernel density estimation has been used for estimating the structure of three grouped data sets coming from real experiments performed in weed science. These data sets consist of the number of seedlings (*P. paradoxa*) emerged at certain cumulative hydrothermal times.

By means of rightly choosing the bandwidth, kernel density estimation proved to be an effective tool for finding structure in the data, even though the data sets were heavy grouped. Taking those pilot density estimations as a reference, suitable normal mixture models were proposed for describing the emergence of seedlings, allowing subsequent simulation studies where the plug-in and bootstrap bandwidth selectors were tested under different grouping conditions.

The results confirmed what was seen in previous chapters; namely,

1. If data grouping is light ($\omega < 0.075$), both plug-in or bootstrap bandwidth selectors may be used, although it is slightly preferable to use the plug-in selector.

2. In case of heavy grouping ($\omega > 0.075$), the plug-in selector is not recommended at all, and the bootstrap selector should be prefered in any case.

Proceeding in this way, the density estimation error remained under control in both cases, but most importantly, it remained fairly bounded in cases of heavy grouping, allowing kernel density estimator to detect most or some of the data structure, even in cases of very heavy grouping.

The first sumulation study showed that some densities are more difficult to estimate than others. Those having multiple modes or alternated areas of high and low density (spiky modes) are more complicated for the kernel density estimator. This was to be expected: if grouping itself hides valuable information about the density of the data, the loss of information becomes more pronounced in cases of greater curvature. The latter, and considering that only a single bandwidth all over the support is used for capturing information, make these type of densities more challenging to estimate. Nevertheless, using the adequate bandwidth selector according to the degree of grouping, the estimator showed to overcome that problem to some extent.

On the other hand, a comparison of the goodness of fit of nonlinear parametric regression methods and the kernel distribution estimator was made considering the three real data sets available. It showed that, unless the distribution is relatively smooth and sigmoidal shape, the former may have serious problems to describe some specifics of the data distribution. Instead, the nonparametric method proved to be a good choice overall, giving quite competitive results both with sigmoidal or more curvy distributions functions.

Moreover, the comparison of goodness of fit between the already mentioned methods was also made by means of a simulation study. By simulating the same grouping conditions found in the three real grouped data sets, this study corroborated that, on average, the kernel distribution estimator performed quite competitive or better than those traditional parametric approaches used in weed science. Thus, kernel distribution estimation is a valid option for describing the relationship between seedling emergence and cumulative hydrothermal time. Furthermore, its flexibility was found to be really helpful in describing structures that are not that simple, as those having more than one mode or having subtle features or variations in the density. Parametric models, due to its rigidity, showed to be limited for describing those cases.

Discussion and conclusions

After all the theoretical and computational work done in this dissertation, this chapter presents a conceptual discussion of both the statistical tools that have been proposed and the results obtained. Also, the main conclusions of this work are set and some research lines for the future are identified, which could help to get a deeper understanding about the problems posed by the limited information provided by grouped data compared to complete data.

Discussion

Samples of grouped data are very useful in descriptive statistics, since in just a glance, they give an idea of how the data at hand are distributed. This simplicity makes tools like the histogram so popular in various areas of knowledge. However, while an advantage in some sense, grouped data can also be a disadvantage if inferences are desired within a high degree of accuracy.

An example of the latter is the problem that gave rise to this thesis: in weed science, it is essential to estimate probabilities of seedling emergence as accurately as possible, since implementing efficient programs to eradicate weeds depends on them. When from experimental reasons data are obtained in grouped fashion (i.e., data cannot be ungrouped), making clear-cut inferences is an issue that can be challenging. This is essentially the problem that this thesis has tried to solve.

The problem at hand has been tackled trying to be simple, but formal. Thus, since the theory shows that the kernel estimator (whether for estimating the density or the distribution) is asymptotically more efficient than very basic tools like the histogram or the empirical distribution function, the natural choice has been to choose the kernel estimator and to propose a suitable modification, so it can be used with grouped data. This modification is, essentially, a way to disaggregate the data. Since it is not even known the way in which the data are distributed into any interval, the easiest way of disaggregation has been to propose the midpoint of every interval as a representative of the data whithin, and to consider it as many times as data therein. In other words, it has been implicitly assumed that the distribution of the data within the intervals is symmetric, which occurs when the probability density function, in a given interval, is symmetric. The simplest case of this situation is that of a constant density within every interval.

The above is, perhaps, the first of the limitations of the estimator that can be pointed out. This lets us understand one of the reasons why, in general, kernel density (or distribution) estimation for grouped data gets better as the interval lengths decreases. Assume there is a nontrivial density function (i.e., one with a certain degree of curvature). There could be some cases in which, by chance, one or more intervals capture a symmetric region of the density, so that, in these cases, the midpoint choice is adequate. Nevertheless, it is not reasonable to expect that to happen with all intervals and, therefore, the choice of the midpoint is not suitable, in general, when the intervals are large. On the other hand, when the intervals are small, to consider that the region of the density captured by the intervals is symmetric is not that severe; hence, the distance between the truly representative point of the data whithin the intervals and the midpoints tends to be small. This reasoning leads to think that some improvement in the quality of the estimates can be achieved when considering more complex forms of disaggregating the data.

The choice of representative data points influences the way in which the asymptotic properties of the estimator are obtained. Based on the experience of weed scientists, in this dissertation, it has been considered that the intervals are of different length and that they remain fixed from one experiment to another. From the statistical point of view, this means that from one trial to another, the observed random quantity is the number (or the proportion) of data within the intervals, but not the midpoints. Indeed, considering a scheme in which intervals change from trial to trial (and therefore, the midpoints as well) adds a source of variability. From the mathematical point of view, it represents an additional challenge in obtaining the asymptotic properties of the estimator. Of course, to consider other criteria of data disaggregation could also increase the complexity of the mathematical treatment.

In addition to the suitability of the midpoints as representative of the data within the intervals, another factor that affects the quality of the estimation is the bandwidth selection. In this sense, it has been proved that under the assumptions made about the density and kernel functions, as well as the variability of the intervals and their asymptotic relationship to the bandwidth, the expression for the $AMISE$ optimal bandwidth selector for grouped data fairly coincides with the $AMISE$ optimal bandwidth selector for ungrouped data. It is important to note that this is only true when those assumptions hold. Otherwise, the non leading terms of the Taylor's representation of the $MISE$ are not negligible, and that $AMISE$ expression is not a good approximation of the $MISE$.

Naturally, this leads us to basically distinguish two scenarios: light and heavy grouping. Speaking in these terms, the use of the plug-in selector is only indicated in cases of light grouping (that is why the non-leading terms of the Taylor's representation of the $MISE$ actually vanish). However, it became necessary to define what is meant by light grouping in practice. Moreover, it was also necessary to propose an alternative selector for those cases in which the plug-in is not adequate; i.e., heavy grouping cases.

Given a reasonable sample size and different sets of intervals, the different simulations performed helped to identify what light grouping is in practice. Facing applications, the importance of this is huge, since it allows to establish guidelines on when to conveniently use the plug-in selector. For those cases in which the plug-in selector cannot be used, it was proposed a bootstrap bandwidth selector, which is based on minimizing a closed expression of the bootstrap version of the $MISE$.

The results obtained when using the bootstrap bandwidth selector where highly satisfactory. A preliminary reading of the results showed that the bootstrap bandwidth selector outperformed the plug-in in general. Although the plug-in selector slightly outperformed the bootstrap selector in cases of light grouping, the bootstrap selector clearly had a better performance than the plug-in selector in cases of heavy grouping. It is important to highlight that the bootstrap bandwidth selector has the valuable advantage of getting pilot information about the distribution, since it operates by first obtaining a pilot estimation, for which it is necessary a pilot bandwidth. This way, it reproduces important features of the distribution under any scenario or sample size. The key is to adequately select the pilot bandwidth. To do that, a proper estimation of the curvature is needed.

Another viewpoint to study the structure of the data is through the distribution function. As derived from the kernel density estimator for grouped data, the kernel distribution estimator for grouped data showed to be an effective tool due to the good performance of its own plug-in bandwidth selector. However, compared with the kernel density estimator, it was observed a subtle difference that may be important in practice: the kernel distribution estimator is somewhat more “robust” than the kernel density estimator, since slighlty different bandwidth selections from the optimal bandwidth do not have such a great impact on the distribution estimation, as it occurs in the case of kernel density estimation. This difference gives the kernel distribution estimator certain advantage in practice, as it seems to be more resistant to the presence of heavy grouping; i.e., it may perform acceptably in cases when the kernel density estimator for grouped data fails, hence, needing the use of more elaborated selectors, like the bootstrap. This evidence suggests that it would be very interesting to propose an alternative bandwdith selector for the kernel distribution estimator for grouped data, to be used in cases of heavy or very heavy grouping.

The chapter about applications to real data confirmed that kernel density and distribution estimators for grouped data are worth to study. The real datasets showed different types of structures that allowed to evaluate the performance of these estimators in such contexts. Despite considering more complex densities (i.e., with more curvature), it was positively striking to note that the guidelines obtained in previous chapters, based on relatively smooth densities, were still valid to some extent to identify regions of light and heavy grouping. Of course, the more the curvature, the more difficult for both estimators to estimate the structure of the data, and regions of heavy or light grouping may slightly change. However, it seems that the boundary between light and heavy grouping can be identified within certain ranges, regardless of the complexity of the data structure.

The last of the applications turned out to be quite innovative and successful. When comparing the kernel distribution estimator versus the typical nonlinear regression models used by weed scientists, it was observed that the flexibility of the nonparametric tool can be decisive to adequately describe the data structure without actually oversimplifying, as may occur with some parametric methods. These subtle differences (or sometimes not that subtle) can make the difference between accurate and inaccurate seedling emergence prediction, which is the basis for proper implementation of mechanisms for weed erradication, a matter of great importance from the social and economical standpoint.

Finally, it is worth mentioning that most of the contents of this study have been submitted (or are about to be sent) to specialized journals on the subject (Reyes et al. (2015*a*), Reyes et al. (2015*b*), González-Andujar et al. (2015)).

Conclusions

The findings are encouraging. On the one hand, the objective of smoothly estimating the density or the distribution function when the data at hand are grouped has been met, which in the case of weed science, it means to be able to obtain more accurate probabilities of seedling emergence. On the other hand, a complete and formal theoretical work has been done regarding the asymptotic properties of the kernel density and distribution estimators proposed, as well as a comprehensive study regarding some bandwidth selectors. These bandwidth selectors have proven to be effective in different grouping scenarios, keeping the estimation error relatively under control. Lastly, helpful guidelines of use have been established for identifying light and heavy grouping in practice.

There are some possible future work lines. For example,

1. To explore more complex criteria of data disaggregation, which, in turn, would entail more complex mathematical treatments.

2. To look for an automatic plug-in bandwidth selector for grouped data in both kernel density and distribution estimation cases (perhaps, by a possible modification to the plug-in selector that allows to automatically correct the lack of information when working with grouped data, specially in heavy grouping scenarios).

3. Since the estimation of the curvature is a key for both selectors (the plug-in and the bootstrap), it will be helpful to consider more elaborate and effective ways for estimating the curvature.

4. Besides the plug-in, to propose alternative bandwidth selectors in the case of the

kernel distribution estimation for grouped data. Presumably, this would considerably improve the performance of this tool in cases of heavy or very heavy grouping.

5. To apply the proposed nonparametric estimators considering data coming from other areas of knowledge. Also, to consider more complex data structures.

6. To consider other approaches, like nonparametric isotonic regression, or more complex models like those based on nonhomogeneous Poisson processes.

Appendix A

Notation

This is a summary of some of the main notation used in this dissertation.

Let ϱ be a real valued univariate function.

- $\varrho_h(u) = \frac{1}{h}\varrho\left(\frac{u}{h}\right)$, for $h > 0$.
- $\int \varrho(x)\,dx = \int_{-\infty}^{\infty} \varrho(x)\,dx$.
- $\varrho^{(r)}(x) = \frac{d^r}{dx^r}\varrho(x)$.
- $A(\varrho) = \int \varrho(x)^2\,dx$.
- $\mu_l(\varrho) = \int x^l \varrho(x)\,dx$.
- The convolution of ϱ and ϱ_0, where ϱ_0 is another real valued function, is represented by

$$(\varrho * \varrho_0)(x) = \int \varrho(x-u)\,\varrho_0(u)\,du.$$

- Given a density f and r an even integer,

$$\psi_r = \int f^{(r)}(x)\,f(x)\,dx.$$

Let a_n and b_n be two real valued deterministic sequences.

- $a_n = O(b_n)$ as $n \to \infty$ if and only if $\limsup_{n\to\infty} |a_n/b_n| < \infty$.
- $a_n = o(b_n)$ as $n \to \infty$ if and only if $\lim_{n\to\infty} |a_n/b_n| = 0$.
- $a_n \sim b_n$ if and only if $\lim_{n\to\infty} (a_n/b_n) = 1$.

Some other abbreviations:

- HTT: *hydrothermal time*.

- CHTT: *cumulative hydrothermal time.*

- MSE: *mean squared error.*

- MISE: *mean integrated squared error.*

- AMSE: *asymptotic mean squared error.*

- AMISE: *asymptotic mean integrated squared error.*

Appendix B

Asymptotic notation and Taylor expansion

Order and asymptotic notation

The order notation O ("big oh") and o ("little oh") is commonly used for the large sample analysis of density estimators. Although this notation is defined for general real valued functions (Serfling, 1980), in this context it will be enough to consider it for real valued sequences.

Let a_n and b_n be sequences of real numbers. On the one hand, it is said that a_n is "big oh" b_n (i.e., a_n is of order b_n) as n increases, which is written as $a_n = O\left(b_n\right)$, if and only if

$$\lim_{n\to\infty} \sup \left|\frac{a_n}{b_n}\right| < \infty;$$

that is to say, $a_n = O\left(b_n\right)$ if $\left|\frac{a_n}{b_n}\right|$ remains bounded as n increases.

On the other hand, it is said that a_n is "little oh" b_n (i.e., a_n is of small order b_n) as n increases, written as $a_n = o\left(b_n\right)$, if and only if

$$\lim_{n\to\infty} \left|\frac{a_n}{b_n}\right| = 0.$$

In the case of sequences, it is usually understood that n increases with no limit, so the condition "as $n \to \infty$" will be assumed.

Besides, the notation $a_n = O\left(1\right)$ means that a_n is bounded, and $a_n = o\left(1\right)$ means that a_n approaches to zero as n increases. It is also said that a_n is asymptotically equivalent to b_n, which is expressed as $a_n \sim b_n$, if and only if

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 1.$$

Taylor expansion

Taylor expansion is a very useful tool for getting asymptotic approximations. Let f be an $m+1$ times differentiable function. Then, Taylor's theorem states that

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2!} f''(a)(x-a)^2 + ... + \frac{1}{m!} f^{(m)}(a)(x-a)^m + R_m(x). \quad \text{(B.1)}$$

The polynomial

$$f(a) + f'(a)(x-a) + \frac{1}{2!} f''(a)(x-a)^2 + ... + \frac{1}{m!} f^{(m)}(a)(x-a)^m$$

is called the m-th Taylor expansion of f around a, and $R_m(x)$ is called the remainder term. A possible explicit expression for R_m is the following, due to Lagrange,

$$R_m(x) = \frac{1}{(m+1)!} f^{(m+1)}(c)(x-a)^{m+1}, \quad \text{(B.2)}$$

for some value c between x and a. When x is near a, $R_m(x)$ is small and $R_m(x) = o([x-a]^m)$ when $x \to a$.

In some multivariate applications, Taylor's theorem is also very useful. Let $\vec{x}$ and $\vec{x_0}$ be d-dimensional vectors. Then, the second order Taylor's formula is

$$\begin{aligned} f(\vec{x}) &= f(\vec{x_0}) + \sum_{i=1}^{d} (x_i - x_{0i}) \frac{\partial f}{\partial x_i}(\vec{x_0}) + \frac{1}{2} \sum_{i,j=1}^{d} (x_i - x_{0i})(x_j - x_{0j}) \frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x_0}) + \\ & R_2(\vec{x}, \vec{x_0}), \end{aligned}$$

where

$$R_2(\vec{x}, \vec{x_0}) = \frac{1}{3!} \sum_{i,j,k=1}^{d} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(\vec{c})(x_i - x_{0i})(x_j - x_{0j})(x_k - x_{0k}),$$

and $\vec{c}$ is somewhere on the line joining $\vec{x}$ and $\vec{x_0}$.

Appendix C

Some useful results

Result C.1. *Under Assumption 3.4,*

$$l_{max} = O\left(\bar{l}\right).$$

Proof. By definition,

$$l_{max} = \max_i |l_i|. \tag{C.1}$$

Adequately modifying (C.1), in terms of $\bar{l}$,

$$l_{max} = \max_i \left|l_i - \bar{l} + \bar{l}\right| \leqslant \bar{l} + \max_i \left|l_i - \bar{l}\right|.$$

But, from assumption 3.4, $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$. Hence,

$$l_{max} = O\left(\bar{l}\right). \tag{C.2}$$

□

Result C.2. *Under Assumption 3.4,*

$$\overline{l^2} = \bar{l}^2 + o\left(\bar{l}^2\right).$$

Proof. By definition,

$$\overline{l^2} = \frac{1}{k}\sum_{i=1}^{k} l_i^2.$$

The last expression can be modified as

$$\frac{1}{k}\sum_{i=1}^{k} l_i^2 = \frac{1}{k}\sum_{i=1}^{k} l_i \left(l_i - \bar{l}\right) + \frac{1}{k}\bar{l}\sum_{i=1}^{k} l_i,$$

from which,

$$\overline{l^2} = \bar{l}^2 + \frac{1}{k}\sum_{i=1}^{k} l_i \left(l_i - \bar{l}\right). \tag{C.3}$$

Bounding the second term on the right hand side of (C.3),

$$\begin{aligned}\left|\frac{1}{k}\sum_{i=1}^{k} l_i \left(l_i - \bar{l}\right)\right| &\leqslant \frac{1}{k}\max_i \left|l_i - \bar{l}\right| \sum_{i}^{k} l_i \\ &\leqslant o\left(\bar{l}^2\right),\end{aligned}$$

since from assumption 3.4, $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$. Finally, going back to (C.3),

$$\overline{l^2} = \bar{l}^2 + o\left(\bar{l}^2\right). \tag{C.4}$$

□

Result C.3. *Under Assumption 3.4 and Eq. (C.2),*

$$\max_i \left|l_i^2 - \overline{l^2}\right| = o\left(\bar{l}^2\right).$$

Proof. Adequately modifying $\left|l_i^2 - \overline{l^2}\right|$,

$$\begin{aligned}\left|l_i^2 - \overline{l^2}\right| &= \left|l_i^2 - l_i\bar{l} + l_i\bar{l} - \overline{l^2}\right| \\ &\leqslant l_i \left|l_i - \bar{l}\right| + \left|l_i\bar{l} - \overline{l^2}\right| \\ &\leqslant l_{max} \max_i \left|l_i - \bar{l}\right| + \left|l_i\bar{l} - \overline{l^2}\right|. \end{aligned}\tag{C.5}$$

Let us find an upper bound for $\left|l_i\bar{l} - \overline{l^2}\right|$.

$$\begin{aligned}\left|l_i\bar{l}-\overline{l^2}\right| &= \left|l_i\frac{1}{k}\sum_{j=1}^{k}l_j-\frac{1}{k}\sum_{j=1}^{k}l_j^2\right| \\ &= \left|\frac{1}{k}\sum_{j=1}^{k}\left(l_il_j-l_j^2\right)\right| \\ &\leqslant \frac{1}{k}\sum_{j=1}^{k}l_j\left|l_i-l_j\right| \\ &\leqslant \frac{1}{k}\sum_{j=1}^{k}l_j\left(\left|l_i-\bar{l}\right|+\left|\bar{l}-l_j\right|\right),\end{aligned}$$

but $\frac{1}{k}\sum_{j=1}^{k}l_j\left(\left|l_i-\bar{l}\right|+\left|\bar{l}-l_j\right|\right)=\bar{l}\left|l_i-\bar{l}\right|+\frac{1}{k}\sum_{j=1}^{k}l_j\left|\bar{l}-l_j\right|$. Then,

$$\begin{aligned}\left|l_i\bar{l}-\overline{l^2}\right| &\leqslant \bar{l}\left|l_i-\bar{l}\right|+\frac{1}{k}\sum_{j=1}^{k}l_j\left|\bar{l}-l_j\right| \\ &\leqslant \bar{l}\max_i\left|l_i-\bar{l}\right|+\bar{l}\max_i\left|\bar{l}-l_i\right| \\ &\leqslant 2\bar{l}\max_i\left|\bar{l}-l_i\right|.\end{aligned}$$

Now, going back to Eq. (C.5),

$$\begin{aligned}\left|l_i^2-\overline{l^2}\right| &\leqslant l_{max}\max_i\left|l_i-\bar{l}\right|+2\bar{l}\max_i\left|\bar{l}-l_i\right| \\ &\leqslant \left[l_{max}+2\bar{l}\right]\max_i\left|l_i-\bar{l}\right|. \qquad \text{(C.6)}\end{aligned}$$

Eq. (C.6) is valid for all i, particularly for the maximum. So,

$$\begin{aligned}\max_i\left|l_i^2-\overline{l^2}\right| &\leqslant \left[l_{max}+2\bar{l}\right]\max_i\left|l_i-\bar{l}\right| \\ &\leqslant O\left(\bar{l}\right)o\left(\bar{l}\right) \\ &\leqslant o\left(\bar{l}^2\right), \qquad \text{(C.7)}\end{aligned}$$

since by Assumption 3.4 and the Eq. (C.2), $\max_i\left|l_i-\bar{l}\right|=o\left(\bar{l}\right)$ and $l_{max}=O\left(\bar{l}\right)$.

□

Result C.4. *Consider the distribution function F at the endpoints y_i and y_{i-1}. By Taylor's theorem, there exist some $\tau_i\in[t_i,y_i]$ and $\tau_{i-1}\in[y_{i-1},t_i]$ such that*

$$\begin{aligned} F\left(y_i\right) &= F\left(t_i\right)+F'\left(t_i\right)\left(y_i-t_i\right)+\frac{1}{2}F''\left(t_i\right)\left(y_i-t_i\right)^2+\cdots \\ &\quad +\frac{1}{m!}F^{(m)}\left(t_i\right)\left(y_i-t_i\right)^m+\frac{1}{(m+1)!}F^{(m+1)}\left(\tau_i\right)\left(y_i-t_i\right)^{m+1}, \end{aligned} \tag{C.8}$$

and

$$\begin{aligned} F\left(y_{i-1}\right) &= F\left(t_i\right)+F'\left(t_i\right)\left(y_{i-1}-t_i\right)+\frac{1}{2}F''\left(t_i\right)\left(y_{i-1}-t_i\right)^2+\cdots \\ &\quad +\frac{1}{m!}F^{(m)}\left(t_i\right)\left(y_{i-1}-t_i\right)^m+\frac{1}{(m+1)!}F^{(m+1)}\left(\tau_{i-1}\right)\left(y_{i-1}-t_i\right)^{m+1}. \end{aligned} \tag{C.9}$$

Substracting (C.9) from (C.8), we have

$$F\left(y_i\right)-F\left(y_{i-1}\right)=\sum_{j=1}^{m}\frac{1}{j!}F^{(j)}\left(t_i\right)\alpha_{ji}+R_\tau, \tag{C.10}$$

where

$$R_\tau=\frac{1}{(m+1)!}\left[F^{(m+1)}\left(\tau_i\right)\left(y_i-t_i\right)^{m+1}-F^{(m+1)}\left(\tau_{i-1}\right)\left(y_{i-1}-t_i\right)^{m+1}\right] \tag{C.11}$$

and

$$\alpha_{ji}=\left(\frac{l_i}{2}\right)^j-\left(-\frac{l_i}{2}\right)^j=\begin{cases}0 & \textit{for } j \textit{ even} \\ 2\left(\frac{l_i}{2}\right)^j & \textit{else}\end{cases}. \tag{C.12}$$

Next, it will be proved that under Assumption 3.2,

$$\left|(m+1)!R_\tau\right|\leqslant 2\mathfrak{L}_{F^{(m+1)}}\left(\frac{l_i}{2}\right)^{m+2}+\parallel F^{(m+1)}\parallel_\infty \alpha_{m+1,i},$$

where $\mathfrak{L}_{F^{(m+1)}}$ is the Lipschitz constant of F^{m+1}.

Proof. Let us add $F^{(m+1)}\left(t_i\right)-F^{(m+1)}\left(t_i\right)$ to both $F^{(m+1)}\left(\tau_i\right)$ and $F^{(m+1)}\left(\tau_i\right)$. Through associative operations we have

$$\begin{aligned} \left|F^{(m+1)}\left(\tau_i\right)\mathfrak{o}_1-F^{(m+1)}\left(\tau_{i-1}\right)\mathfrak{o}_2\right| &\leqslant \left|F^{(m+1)}\left(\tau_i\right)-F^{(m+1)}\left(t_i\right)\right|\left|\mathfrak{o}_1\right| \\ &\quad +\left|F^{(m+1)}\left(\tau_{i-1}\right)-F^{(m+1)}\left(t_i\right)\right|\left|\mathfrak{o}_2\right| \\ &\quad +\left|F^{(m+1)}\left(t_i\right)\right|\left|\alpha_{m+1,i}\right|, \end{aligned}$$

where $\mathfrak{o}_1 = \left(\frac{l_i}{2}\right)^{m+1}$ and $\mathfrak{o}_2 = \left(-\frac{l_i}{2}\right)^{m+1}$.

As long as $F^{(m+1)}$ is Lipschitz,

$$\left|F^{(m+1)}\left(\tau_i\right) - F^{(m+1)}\left(t_i\right)\right| \leqslant \mathfrak{L}_{F^{(m+1)}}\left|\tau_i - t_i\right|,$$

and

$$\left|F^{(m+1)}\left(\tau_{i-1}\right) - F^{(m+1)}\left(t_i\right)\right| \leqslant \mathfrak{L}_{F^{(m+1)}}\left|\tau_{i-1} - t_i\right|.$$

Since $|\tau_i - t_i| \leqslant \frac{1}{2}l_i$ and $|\tau_{i-1} - t_i| \leqslant \frac{1}{2}l_i$, then,

$$\left|F^{(m+1)}\left(\tau_i\right)\mathfrak{o}_1 - F^{(m+1)}\left(\tau_{i-1}\right)\mathfrak{o}_2\right| \leqslant \mathfrak{L}_{F^{(m+1)}} l_i \mathfrak{o}_1 + \parallel F^{(m+1)} \parallel_\infty \alpha_{m+1,i}, \qquad \text{(C.13)}$$

or, equivalently, recalling (C.11) ,

$$\left|(m+1)!R_\tau\right| \leqslant 2\mathfrak{L}_{F^{(m+1)}}\left(\frac{l_i}{2}\right)^{m+2} + \parallel F^{(m+1)} \parallel_\infty \alpha_{m+1,i}. \qquad \text{(C.14)}$$

□

Result C.5. *Under Assumptions 3.1 and 3.3, for a fixed $x \in (y_0, y_k)$ and a sufficiently large sample size n,*

$$\int_{y_0}^{y_k} \phi_1''(t)\,dt = \phi_1'(y_k) - \phi_1'(y_0) = 0,$$

where $\phi_1(t) \equiv F'(t)\,K\left(\frac{x-t}{h}\right)$.

Proof. By definition,

$$\phi_1'(t) = F'(t)\,K'\left(\frac{x-t}{h}\right)\left(-\frac{1}{h}\right) + K\left(\frac{x-t}{h}\right)F''(t).$$

Consider a fixed point $x \in (y_0, y_k)$. Then, $x - y_0 = d_0 > 0$, so,

$$\phi_1'(y_0) = F'(y_0)\,K'\left(\frac{d_0}{h}\right)\left(-\frac{1}{h}\right) + K\left(\frac{d_0}{h}\right)F''(y_0).$$

Using Assumption 3.3, since h approaches to zero as n increases, there is a sufficiently large sample size n such that $d_0/h > 1$. By Assumption 3.1, the kernel support is $[-1, 1]$. Then, $K\left(\frac{d_0}{h}\right) = K'\left(\frac{d_0}{h}\right) = 0$. Thus,

$$\phi_1'(y_0) = 0.$$

When choosing y_k, $x - y_k = d_k < 0$. As before, there is a sufficiently large sample size n such that $d_0/h < -1$, so that $K\left(\frac{d_k}{h}\right) = K'\left(\frac{d_k}{h}\right) = 0$. Then,

$$\phi_1'(y_k) = 0,$$

and

$$\int_{y_0}^{y_k} \phi_1''(t)\, dt = \phi_1'(y_k) - \phi_1'(y_0) = 0. \tag{C.15}$$

□

Result C.6. *Under assumption 3.1 and Eq. (C.7),*

$$\left| \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2} \right) \int_{y_{i-1}}^{y_i} \phi_1''(t)\, dt \right| \leqslant o\left(\frac{\overline{l}^2}{h} \right).$$

Proof. On the one hand, based on the definition, the first and second derivatives of $\phi_1(t)$ are

$$\phi_1'(t) = F'(t)\, K'\left(\frac{x-t}{h}\right)\left(-\frac{1}{h}\right) + K\left(\frac{x-t}{h}\right) F''(t)$$

and

$$\phi_1''(t) \quad = \quad K\left(\frac{x-t}{h}\right) F'''(t) - \frac{2}{h} F''(t)\, K'\left(\frac{x-t}{h}\right) + \frac{1}{h^2} F'(t)\, K''\left(\frac{x-t}{h}\right).$$

Note that if $t \notin [x-h, x+h]$, $\left|\frac{x-t}{h}\right| > 1$; hence, $\phi_1''(t) = 0$. Otherwise,

$$\left|\phi_1''(t)\right| \leqslant \| \phi_{10} \|_\infty + \frac{1}{h} \| \phi_{11} \|_\infty + \frac{1}{h^2} \| \phi_{12} \|_\infty . \tag{C.16}$$

where $\phi_{10}(t) = K\left(\frac{x-t}{h}\right) F'''(t)$, $\phi_{11}(t) = -2F''(t)\, K'\left(\frac{x-t}{h}\right)$ and $\phi_{12}(t) = F'(t)\, K''\left(\frac{x-t}{h}\right)$. On the other hand, using (C.7)

$$\begin{aligned} \left| \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2} \right) \int_{y_{i-1}}^{y_i} \phi_1''(t)\, dt \right| &\leqslant \max_i \left| l_i^2 - \overline{l^2} \right| \sum_{i=1}^{k} \left| \int_{y_{i-1}}^{y_i} \phi_1''(t)\, dt \right| \\ &\leqslant o\left(\overline{l}^2\right) \int_{y_0}^{y_k} \left|\phi_1''(t)\right| dt, \end{aligned}$$

and since $\phi_1''(t) = 0 \;\forall\; t \in [x-h, x+h]$, then

$$\left| \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2} \right) \int_{y_{i-1}}^{y_i} \phi_1''(t)\, dt \right| \leqslant o\left(\overline{l}^2\right) \int_{x-h}^{x+h} \left|\phi''(t)\right| dt. \tag{C.17}$$

Considering Eq. (C.16),

$$\begin{aligned}\int_{x-h}^{x+h} \left|\phi_1''(t)\right| dt &\leqslant \parallel \phi_{10} \parallel_\infty \int_{x-h}^{x+h} dt + \frac{\parallel \phi_{11} \parallel_\infty}{h} \int_{x-h}^{x+h} dt + \frac{\parallel \phi_{12} \parallel_\infty}{h^2} \int_{x-h}^{x+h} dt \\ &\leqslant 2h \parallel \phi_{10} \parallel_\infty + 2 \parallel \phi_{11} \parallel_\infty + \frac{1}{h} \parallel \phi_{12} \parallel_\infty \\ &= O\left(\frac{1}{h}\right),\end{aligned}$$

and so, Eq. (C.17) becomes

$$\left|\sum_{i=1}^{k} \left(l_i^2 - \overline{l^2}\right) \int_{y_{i-1}}^{y_i} \phi_1''(t)\, dt\right| \leqslant o\left(\frac{\bar{l}^2}{h}\right). \tag{C.18}$$

□

Theorem C.1. ***Ostrowski's inequality.*** *Let $\mathfrak{f}$ be a continuous real function such that $\mathfrak{f} \in C^1([a,b])$, $x \in (a,b)$. Then, for all $x \in (a,b)$,*

$$\left|\mathfrak{f}(x) - \frac{1}{b-a}\int_a^b \mathfrak{f}(t)\, dt\right| \leqslant \left[\frac{1}{4} + \frac{\left(x - \frac{a+b}{2}\right)^2}{(b-a)^2}\right](b-a)\,\mathfrak{L}_{\mathfrak{f}}, \tag{C.19}$$

where $\mathfrak{L}_{\mathfrak{f}} = \parallel \mathfrak{f}' \parallel_\infty$ is the Lipschitz constant for $\mathfrak{f}$. (Ostrowski, 1938; Anastassiou, 1995).

Theorem C.2. ***Multivariate Ostrowski's inequality.*** *Let $\mathfrak{f} \in C^1(\Pi[a_i,b_i])$, where $a_i < b_i$; $a_i, b_i \in \mathbb{R}$, $i = 1,...,k$, and let $\vec{x}_0 = (x_{01},...,x_{0k}) \in \Pi_{i=1}^k [a_i,b_i]$ be fixed. Then,*

$$\left|\mathfrak{f}(\vec{x}_0) - \frac{1}{\Pi_{i=1}^k \mathfrak{c}_\mathfrak{i}}\int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \mathfrak{f}(Z_1,\ldots,Z)\, dZ_1 \ldots dZ_k\right| \leqslant \sum_{i=1}^{k}\left[\frac{\mathfrak{c}_{\mathfrak{aoi}}^2 + \mathfrak{c}_{\mathfrak{boi}}^2}{2\mathfrak{c}_\mathfrak{i}}\right]\left\|\frac{\partial \mathfrak{f}}{\partial Z_i}\right\|_\infty, \tag{C.20}$$

where $\mathfrak{c}_{\mathfrak{aoi}} = (x_{0i} - a_i)$, $\mathfrak{c}_{\mathfrak{boi}} = (b_i - x_{0i})$ and $\mathfrak{c}_\mathfrak{i} = b_i - a_i$. Clearly, Eq. (C.20) generalizes Eq. (C.19) (Anastassiou, 1997).

Appendix D

Proof of Theorem 3.1

Proof. To prove Theorem 3.1, Taylor's theorem (Appendix B) will be used intensively.

Let us first obtain the bias of the estimator $\hat{f}_h^g$. Applying the expectation operator to (3.2), we obtain

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h^g(x)\right] &= \frac{1}{h}\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)\mathbb{E}\left[w_i\right]\\ &= \frac{1}{h}\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)p_i\\ &= \frac{1}{h}\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)\left[F\left(y_i\right)-F\left(y_{i-1}\right)\right], \qquad \text{(D.1)}\end{aligned}$$

where $p_i = \left[F\left(y_i\right)-F\left(y_{i-1}\right)\right]$ is the difference of the distribution function evaluated at the limits of the i-th interval, and all the effect of grouping the data is in p_i. According to our objectives, to get it in a more tractable and intuitive form, let us use a Taylor expansion of p_i around t_i. This was already done in (C.10):

$$F\left(y_i\right)-F\left(y_{i-1}\right) = \sum_{j=1}^{m}\frac{1}{j!}F^{(j)}\left(t_i\right)\alpha_{ji} + R_\tau, \qquad \text{(D.2)}$$

where the values of $\alpha_{ji} = \left(y_i - t_i\right)^j - \left(y_{i-1}-t_i\right)^j$ are given in (C.12) and R_τ is given in Eq. (C.11).

An aspect to consider is how many terms we should take into account in (C.10). For this, consider Eq. (C.14). Note that taking $m+1$ as an even number simplifies the result, since by the parity conditions (C.12), the second term on the right hand side is zero. Then, taking $m+1=2$ would be too little, with just one leading term, and taking more than four would be too much and unnecessary. Then, $m+1=4$ seems an adequate choice, giving:

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h^g(x)\right] &= \frac{1}{h}\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)\left[F'(t_i)\,\alpha_{1i} + \frac{1}{3!}F'''(t_i)\,\alpha_{3i} + R_\tau\right], \\ &= \frac{1}{h}(A+B+C), \end{aligned} \tag{D.3}$$

where

$$A = \sum_{i=1}^{k} l_i F'(t_i)\, K\left(\frac{x-t_i}{h}\right),$$

$$B = \frac{1}{24}\sum_{i=1}^{k} l_i^3 F'''(t_i)\, K\left(\frac{x-t_i}{h}\right)$$

and

$$C = \frac{1}{24}\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)\left[F^{(4)}(\tau_i)\left(\frac{l_i}{2}\right)^4 - F^{(4)}(\tau_{i-1})\left(-\frac{l_i}{2}\right)^4\right].$$

Note that there are two leading terms (since the term containing α_{2i} is zero) plus the remainder, which can be easily bounded by (C.14).

Recall that the number of intervals k increases as the sample size does, so we may then approximate A and B by integrals over the support. Recall $\phi_1(t) = F'(t)\, K\left(\frac{x-t}{h}\right)$. Then,

$$A = \sum_{i=1}^{k} l_i \phi_1(t_i).$$

Using again a Taylor expansion, the integral over the i-th interval can be expressed as

$$\begin{aligned}\int_{y_{i-1}}^{y_i} \phi_1(t)\,dt &= \int_{y_{i-1}}^{y_i}\left[\phi_1(t_i) + (t-t_i)\,\phi_1'(t_i) + \frac{1}{2}(t-t_i)^2\,\phi_1''(t_i)\right. \\ &\quad \left. + \frac{1}{3!}(t-t_i)^3\,\phi_1'''(t_i) + \frac{1}{4!}(t-t_i)^4\,\phi_1^{(4)}(\xi_i)\right]dt,\end{aligned}$$

for some ξ_i between t and t_i. Then, using the change of variable

$$s = t - t_i \tag{D.4}$$

and by the parity properties in (C.12),

$$\int_{y_{i-1}}^{y_i} \phi_1(t)\,dt = l_i\phi_1(t_i) + \frac{1}{24}l_i^3\phi_1''(t_i) + \frac{1}{4!80}l_i^5\phi_1^{(4)}(\xi_i). \tag{D.5}$$

Adding all over the k intervals in both sides of (D.5),

$$\int_{y_0}^{y_k} \phi_1(t)\,dt = \sum_{i=1}^{k} l_i \phi_1(t_i) + \frac{1}{24}\sum_{i=1}^{k} l_i^3 \phi_1''(t_i) + \frac{1}{4!80}\sum_{i=1}^{k} l_i^5 \phi_1^{(4)}(\xi_i), \tag{D.6}$$

i.e., since the first term on the right hand side of (D.6) is A,

$$A = \int_{y_0}^{y_k} \phi_1(t)\,dt - \frac{1}{24}\sum_{i=1}^{k} l_i^3 \phi_1''(t_i) - \frac{1}{4!80}\sum_{i=1}^{k} l_i^5 \phi_1^{(4)}(\xi_i). \tag{D.7}$$

Look that, by the definition of $\phi_1(t)$, every time we differentiate the function, we will get an h in the denominator, so that the n-th derivative of $\phi_1(t)$ has the following expression

$$\phi_1^{(n)}(t) = \phi_{10} + \frac{1}{h}\phi_{11} + \frac{1}{h^2}\phi_{12} + ... + \frac{1}{h^n}\phi_{1n}, \tag{D.8}$$

where $\phi_{10}, \phi_{11}, ..., \phi_{1n}$ are functions of products of derivatives of $F'(t)$ and $K\left(\frac{x-t}{h}\right)$. So,

$$\begin{aligned} \left|\sum_{i=1}^{k} l_i^5 \phi_1^{(4)}(\xi_i)\right| &\leqslant k l_{max}^5 \parallel \phi_1^{(4)} \parallel_\infty \\ &= \frac{y_k - y_0}{\bar{l}} O\left(\bar{l}^5\right) O\left(\frac{1}{h^4}\right) \\ &= O\left(\frac{\bar{l}^4}{h^4}\right), \end{aligned} \tag{D.9}$$

where we used the result (C.2), we also used that $\bar{l} = \frac{y_k - y_0}{k}$, and by Eq. (D.8), $\parallel \phi_1^{(4)} \parallel_\infty = O\left(h^{-4}\right)$.

Let us go back to Eq. (D.7). The second term on the right hand side can be decomposed as

$$\frac{1}{24}\sum_{i=1}^{k} l_i^3 \phi_1''(t_i) = \frac{1}{24}\left[\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i \phi_1''(t_i) + \overline{l^2}\sum_{i=1}^{k} l_i \phi_1''(t_i)\right]. \tag{D.10}$$

Following the same argument as for obtaining Eq. (D.7),

$$\overline{l^2}\sum_{i=1}^{k} l_i \phi_1''(t_i) = \overline{l^2}\int_{y_0}^{y_k} \phi_1''(t)\,dt - \frac{\overline{l^2}}{24}\sum_{i=1}^{k} l_i^3 \phi_1^{(4)}(t_i) - \frac{\overline{l^2}}{4!80}\sum_{i=1}^{k} l_i^5 \phi_1^{(6)}(\xi_i).$$

Note that for a fixed x and for a sufficiently large sample size, $\int_{y_0}^{y_k} \phi_1''(t)\,dt = \phi_1'(y_k) - \phi_1'(y_0) = 0$, as demonstrated in (C.15). As for the other two terms, following the same lines as for (D.9),

$$\begin{aligned}\left|\overline{l^2}\sum_{i=1}^{k} l_i^5 \phi_1^{(6)}(\xi_i)\right| &\leqslant O\left(\bar{l}^2\right) k l_{max}^5 \parallel \phi_1^{(6)} \parallel_\infty \\ &= O\left(\frac{\bar{l}^6}{h^6}\right), \end{aligned} \tag{D.11}$$

where result (C.4) has been used. Similarly,

$$\begin{aligned}\left|\overline{l^2}\sum_{i=1}^{k} l_i^3 \phi_1^{(4)}(\xi_i)\right| &\leqslant O\left(\bar{l}^2\right) k l_{max}^3 \parallel \phi_1^{(4)} \parallel_\infty \\ &= O\left(\frac{\bar{l}^4}{h^4}\right). \end{aligned} \tag{D.12}$$

Updating Eq. (D.10) using (D.12) and (D.11),

$$\frac{1}{24}\sum_{i=1}^{k} l_i^3 \phi_1''(t_i) = \frac{1}{24}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i \phi_1''(t_i) + O\left(\frac{\bar{l}^4}{h^4}\right). \tag{D.13}$$

Proceeding as before, the first term on the right hand side of (D.13) can be expressed as

$$\begin{aligned}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i \phi_1''(t_i) &= \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}\phi_1''(t)\,dt \\ &-\frac{1}{4!}\sum_{i=1}\left(l_i^2-\overline{l^2}\right) l_i^3 \phi_1^{(4)}(t_i) - \frac{1}{4!80}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i^5 \phi_1^{(6)}(\xi_i). \end{aligned} \tag{D.14}$$

Now, using (C.7), it is easy to prove that

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i^5 \phi_1^{(6)}(\xi_i)\right| &\leqslant \max_i \left|l_i^2-\overline{l^2}\right| O\left(\bar{l}^5\right)\frac{y_k-y_o}{\bar{l}} \parallel \phi_1^{(6)} \parallel_\infty \\ &= o\left(\frac{\bar{l}^6}{h^6}\right), \end{aligned} \tag{D.15}$$

and very similarly,

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_i^3\phi_1^{(4)}(t_i)\right| \leqslant \max_i \left|l_i^2-\overline{l^2}\right| O\left(\bar{l}^3\right)\frac{y_k-y_o}{\bar{l}} \parallel \phi_1^{(4)} \parallel_\infty$$
$$= o\left(\frac{\bar{l}^4}{h^4}\right). \qquad \text{(D.16)}$$

Equations (D.16), (D.15), (C.18) and (D.13) imply that

$$\frac{1}{24}\sum_{i=1}^{k} l_i^3\phi_1''(t_i) = o\left(\frac{\bar{l}^2}{h}\right),$$

and so, going back to Eq. (D.7),

$$A = \int_{y_0}^{y_k}\phi_1(t)\,dt + o\left(\frac{\bar{l}^2}{h}\right).$$

Consider the change of variable

$$r = \frac{x-t}{h}. \qquad \text{(D.17)}$$

.

Then,

$$\int_{y_0}^{y_k}\phi_1(t)\,dt = h\int_{\frac{x-y_k}{h}}^{\frac{x-y_0}{h}} F'(x-hr)\,K(r)\,dr.$$

Note that as n increases, the limits of integration tends to $-\infty$ and ∞. By a Taylor expansion of $F'(x-hr)$ around x, and since $F'(x) = f(x)$,

$$\int_{y_0}^{y_k}\phi_1(t)\,dt = h\int K(r)\left[f(x) + f'(x)(-hr) + \frac{1}{2}f''(x)(-hr)^2 + O\left(h^3\right)\right]dr.$$

Due to the properties of the kernel K,

$$\int_{y_0}^{y_k}\phi_1(t)\,dt = h\left[f(x) + \frac{1}{2}h^2f''(x)\,\mu_2(K) + O\left(h^3\right)\right], \qquad \text{(D.18)}$$

so

$$A = h\left[f(x) + \frac{1}{2}h^2f''(x)\,\mu_2(K) + O\left(h^3\right)\right] + o\left(\frac{\bar{l}^2}{h}\right). \qquad \text{(D.19)}$$

Let us now define $\phi_2 \equiv F'''(t)\,K\left(\frac{x-t}{h}\right)$. Then,

$$B_1 = 24B = \sum_{i=1}^{k} l_i^3 \phi_2 (t_i) ,$$

which can be decomposed as

$$B_1 = \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2} \right) l_i \phi_2 (t_i) + \overline{l^2} \sum_{i=1}^{k} l_i \phi_2 (t_i) . \tag{D.20}$$

The first term on the right hand side of Eq. (D.20) can be bounded as

$$\begin{aligned} \left| \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2} \right) l_i \phi_2 (t_i) \right| &\leqslant \sum_{i=1}^{k} \left| \left(l_i^2 - \overline{l^2} \right) l_i \phi_2 (t_i) \right| \\ &\leqslant \max_i \left| l_i^2 - \overline{l^2} \right| l_{max} k \parallel \phi_2 \parallel_\infty \\ &= o \left(\bar{l}^2 \right) , \end{aligned} \tag{D.21}$$

where (C.7), (C.2), $k = \frac{y_k - y_0}{\bar{l}}$ and $\parallel \phi_2 \parallel = O(1)$ were considered.

As to the second term on the right hand side of (D.20), Ostrowski's inequality, given in Eq. (C.19), and (C.7) are to be used. Multiplying Ostrowski's inequality by l_i and evaluating it at the midpoint t_i, we get

$$\left| l_i \phi_2 (t_i) - \int_{y_{i-1}}^{y_i} \phi_2 (t) \, dt \right| \leqslant \frac{1}{4} l_i^2 \parallel \phi_2' \parallel_\infty .$$

Summing all over the k intervals,

$$\sum_{i=1}^{k} \left| l_i \phi_2 (t_i) - \int_{y_{i-1}}^{y_i} \phi_2 (t) \, dt \right| \leqslant \frac{1}{4} \parallel \phi_2' \parallel_\infty k \overline{l^2}.$$

Given the definition of $\phi_2 (t)$, its first derivative is

$$\phi_2' (t) = \phi_{20} (t) + \frac{1}{h} \phi_{21} (t) ,$$

where $\phi_{20} (t) = K \left(\frac{x-t}{h} \right) F^{(4)} (t)$ and $\phi_{21} (t) = -F''' (t) K' \left(\frac{x-t}{h} \right)$. So,

$$\begin{aligned} \sum_{i=1}^{k} \left| l_i \phi_2 (t_i) - \int_{y_{i-1}}^{y_i} \phi_2 (t) \, dt \right| &\leqslant \frac{1}{4} \left[\parallel \phi_{20} \parallel_\infty + \frac{1}{h} \parallel \phi_{21} \parallel_\infty \right] \frac{y_k - y_0}{\bar{l}} O \left(\bar{l}^2 \right) \\ &= O \left(\frac{\bar{l}}{h} \right) , \end{aligned}$$

i.e.,

$$\sum_{i=1}^{k} l_i \phi_2(t) = \int_{y_0}^{y_k} \phi_2(t)\,dt + O\left(\frac{\bar{l}}{h}\right). \tag{D.22}$$

Using (C.4), (D.22) and (D.21), Eq. (D.20) becomes

$$B_1 = \left[\bar{l}^2 + o\left(\bar{l}^2\right)\right]\left[\int_{y_0}^{y_k} \phi_2(t)\,dt + O\left(\frac{\bar{l}}{h}\right)\right] + o\left(\bar{l}^2\right),$$

which simplifying gives

$$B_1 = \bar{l}^2 \int_{y_0}^{y_k} \phi_2(t)\,dt + O\left(\frac{\bar{l}^3}{h}\right). \tag{D.23}$$

Again, considering the change of variable (D.17),

$$\int_{y_0}^{y_k} \phi_2(t)\,dt = h\int_{\frac{x-y_k}{h}}^{\frac{x-y_0}{h}} F'''(x-hr)\,K(r)\,dr.$$

By a Taylor expansion of $F'''(x-hr)$ around x, and since $F'(x) = f(x)$,

$$\int_{y_0}^{y_k} \phi_2(t)\,dt = h\int K(r)\left[f''(x) + f'''(x)(-hr) + \frac{1}{2}f^{(4)}(x)(-hr)^2 + O\left(h^3\right)\right]dr,$$

and due to the properties of the kernel K,

$$\int_{y_0}^{y_k} \phi_2(t)\,dt = h\left[f''(x) + \frac{1}{2}h^2 f^{(4)}(x)\,\mu_2(K) + O\left(h^3\right)\right]. \tag{D.24}$$

Thus, substituting (D.24) in (D.23) and dividing by 24,

$$B = \frac{1}{24}\bar{l}^2 h\left[f''(x) + \frac{1}{2}h^2 f^{(4)}(x)\,\mu_2(K) + O\left(h^3\right)\right] + O\left(\frac{\bar{l}^3}{h}\right). \tag{D.25}$$

Regarding the term C in Eq. (D.3), using (C.13) it is obtained

$$\left|F^{(4)}(\tau_i)\left(\frac{l_i}{2}\right)^4 - F^{(4)}(\tau_{i-1})\left(-\frac{l_i}{2}\right)^4\right| = O\left(\bar{l}^5\right),$$

then,

$$\begin{aligned}\left|\sum_{i=1}^{k} K\left(\frac{x-t_i}{h}\right)\left[F^{(4)}(\tau_i)\left(\frac{l_i}{2}\right)^4 - F^{(4)}(\tau_{i-1})\left(-\frac{l_i}{2}\right)^4\right]\right| &\leqslant k \parallel K \parallel_\infty O\left(\bar{l}^5\right) \\ &= O\left(\bar{l}^4\right),\end{aligned} \tag{D.26}$$

since $k = (y_k - y_0)/\bar{l}$.

By (D.26), (D.25) and (D.19), doing a general update of Eq. (D.3) and using Assumption 3.4, results in

$$\mathbb{E}\left[\hat{f}_h^g(x)\right] = \left[f(x) + \frac{1}{2}h^2 f''(x)\mu_2(K) + o\left(h^2\right)\right] + \left[\frac{\bar{l}^2}{24}f''(x) + o\left(\bar{l}^2\right)\right]. \qquad \text{(D.27)}$$

Using again Assumption 3.4, and substracting $f(x)$, it is obtained

$$\mathbb{B}\left[\hat{f}_h^g(x)\right] = \frac{1}{2}h^2 f''(x)\mu_2(K) + o\left(h^2\right). \qquad \text{(D.28)}$$

Let us now apply the variance operator to the estimator (3.2),

$$\begin{aligned}\mathbb{V}\left[\hat{f}_h^g(x)\right] &= \mathbb{V}\left[\frac{1}{h}\sum_{i=1}^{k} w_i K\left(\frac{x-t_i}{h}\right)\right] \\ &= \frac{1}{h^2}\left\{\sum_{i=1}^{k}\mathbb{V}\left[w_i K\left(\frac{x-t_i}{h}\right)\right] + 2\sum_{i<j}\mathbb{C}\left[w_i K\left(\frac{x-t_i}{h}\right), w_j K\left(\frac{x-t_j}{h}\right)\right]\right\} \\ &= \frac{1}{h^2}\left\{\sum_{i=1}^{k} K^2\left(\frac{x-t_i}{h}\right)\mathbb{V}\left[w_i\right] + 2\sum_{i<j} K\left(\frac{x-t_i}{h}\right) K\left(\frac{x-t_j}{h}\right)\mathbb{C}\left[w_i, w_j\right]\right\}.\end{aligned}$$

Note that the vector $(n_1, n_2, ...n_k)$ follows a multinomial distribution with parameters $n = \sum_{i=1}^{k} n_i$ and $\vec{p} = (p_1, p_2, ..., p_k)$, where $p_i = \mathbb{E}[w_i]$. Since $\mathbb{V}[n_i] = np_iq_i$ and $p_i = [F(y_i) - F(y_{i-1})]$, where $q_i = 1 - p_i$, $\mathbb{C}[n_i, n_j] = -np_ip_j$ for $i \neq j$, we have

$$\mathbb{V}\left[\hat{f}_h^g(x)\right] = \frac{1}{h^2}(D + E), \qquad \text{(D.29)}$$

where

$$D = \frac{1}{n}\sum_{i=1}^{k} K^2\left(\frac{x-t_i}{h}\right) p_i q_i$$

and

$$E = -\frac{2}{n}\sum_{i<j} K\left(\frac{x-t_i}{h}\right) K\left(\frac{x-t_j}{h}\right) p_i p_j.$$

By Eq. (C.10), it can be written

$$p_i = F'(t_i)\alpha_{1i} + \frac{1}{3!}F'''(t_i)\alpha_{3i} + R_\tau \qquad \text{(D.30)}$$

and

$$q_i = 1 - F'(t_i)\,\alpha_{1i} - \frac{1}{3!}F'''(t_i)\,\alpha_{3i} - R_\tau.$$

Multipliying p_i by q_i and according to Eq. (C.12),

$$p_i q_i = F'(t_i)\, l_i + O\left(\bar{l}^2\right).$$

Define

$$D_1 = nD = \sum_{i=1}^{k} K^2\left(\frac{x-t_i}{h}\right)\left[F'(t_i)\, l_i + O\left(\bar{l}^2\right)\right],$$

and define $\phi_3(t) \equiv K^2\left(\frac{x-t}{h}\right)F'(t)$, so that

$$\begin{aligned} D_1 &= \sum_{i=1}^{k} K^2\left(\frac{x-t_i}{h}\right)F'(t_i)\, l_i + O\left(\bar{l}^2\right) k \parallel K^2 \parallel_\infty \\ &= \sum_{i=1}^{k} l_i\phi_3(t_i) + O\left(\bar{l}\right). \end{aligned} \tag{D.31}$$

Using a Taylor series expansion, the integral over the i-th interval is

$$\begin{aligned} \int_{y_{i-1}}^{y_i} \phi_3(t)\,dt &= \int_{y_{i-1}}^{y_i} \Bigg[\phi_3(t_i) + (t-t_i)\,\phi_3'(t_i) + \frac{(t-t_i)^2}{2}\phi_3''(t_i) + \frac{(t-t_i)^3}{3!}\phi_3'''(t_i) \\ &\quad + \frac{(t-t_i)^4}{4!}\phi_3^{(4)}(\xi_i)\Bigg]\,dt \\ &= l_i\phi_3(t_i) + \frac{1}{24}l_i^3\phi_3''(t_i) + \frac{1}{4!}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,\phi_3^{(4)}(\xi_i)\,dt, \end{aligned}$$

where ξ_i is some intermediate point between t and t_i, and the change of variable (D.4) was used. Solving for $l_i\phi_3(t_i)$ and summing up all over the k intervals lead to

$$\sum_{i=1}^{k} l_i\phi_3(t_i) = \int_{y_0}^{y_k}\phi_3(t)\,dt - \frac{1}{24}\sum_{i=1}^{k} l_i^3\phi_3''(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,\phi_3^{(4)}(\xi_i)\,dt. \tag{D.32}$$

As to the third term on the right hand side of (D.32), it can be said that

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,\phi_3^{(4)}(\xi_i)\,dt\right| \leqslant \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\left\|\phi_3^{(4)}\right\|_\infty dt.$$

Integrating the right hand side of the last equation, and since each time $\phi_3(t)$ is differentiate an h in the denominator is obtained, it follows that

$$\begin{aligned}\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\left\|\phi_3^{(4)}\right\|_\infty dt &= \frac{1}{4!80}\left\|\phi_3^{(4)}\right\|_\infty\sum_{i=1}^{k}l_i^5\\ &= kO\left(\bar{l}^5\right)\left\|\phi_3^{(4)}\right\|_\infty\\ &= O\left(\frac{\bar{l}^4}{h^4}\right),\end{aligned}$$

and so,

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\phi_3^{(4)}(\xi_i)\,dt\right| = O\left(\frac{\bar{l}^4}{h^4}\right).$$

The second term on the right hand side of (D.32) can be written as

$$\sum_{i=1}^{k}l_i^3\phi_3''(t_i) = \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_i\phi_3''(t_i)+\overline{l^2}\sum_{i=1}^{k}l_i\phi_3''(t_i). \tag{D.33}$$

As in (D.32), the second term on the right hand side of (D.33) can be expressed as

$$\overline{l^2}\sum_{i=1}^{k}l_i\phi_3''(t_i) = \overline{l^2}\int_{y_0}^{y_k}\phi_3''(t)\,dt - \frac{1}{24}\overline{l^2}\sum_{i=1}^{k}l_i^3\phi_3^{(4)}(t_i) - \frac{1}{4!}\overline{l^2}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\phi_3^{(6)}\left(\xi_i'\right)dt.$$

As did in C.5, similar steps can be followed to prove that $\int_{y_0}^{y_k}\phi_3''(t)\,dt = 0$. As before,

$$\begin{aligned}\left|\overline{l^2}\sum_{i=1}^{k}l_i^3\phi_3^{(4)}(t_i)\right| &\leqslant kO\left(\bar{l}^2\right)O\left(\bar{l}^3\right)\left\|\phi_3^{(4)}\right\|_\infty\\ &\leqslant O\left(\frac{\bar{l}^4}{h^4}\right)\end{aligned}$$

and

$$\overline{l^2}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\phi_3^{(6)}\left(\xi_i'\right)dt \leqslant \overline{l^2}\left\|\phi_3^{(6)}\right\|_\infty\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,dt,$$

which in turn follows that

$$\begin{aligned}\overline{l^2}\left\|\phi_3^{(6)}\right\|_\infty \sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,dt &= \frac{1}{80}\overline{l^2}\left\|\phi_3^{(6)}\right\|_\infty \sum_{i=1}^{k} l_i^5 \\ &= \overline{l^2}kO\left(\frac{\bar{l}^5}{h^6}\right) \\ &= O\left(\frac{\bar{l}^6}{h^6}\right),\end{aligned}$$

i.e.,

$$\overline{l^2}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,\phi_3^{(6)}\left(\xi_i'\right)dt = O\left(\frac{\bar{l}^6}{h^6}\right),$$

so that

$$\overline{l^2}\sum_{i=1}^{k} l_i\phi_3''(t_i) = O\left(\frac{\bar{l}^4}{h^4}\right). \tag{D.34}$$

Regarding the first term on the right hand side of (D.33),

$$\begin{aligned}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_i\phi_3''(t_i) \;&=\; \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}\phi_3''(t)\,dt - \frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_i^3\phi_3^{(4)}(t_i) \\ &\quad -\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}(t-t_i)^4\,\phi_3^{(6)}\left(\xi_i'\right)dt.\end{aligned} \tag{D.35}$$

Due to (C.7), the second and third terms in (D.35) are $o\left(\frac{\bar{l}^4}{h^4}\right)$ and $o\left(\frac{\bar{l}^6}{h^6}\right)$, respectively. As to the first term,

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}\phi_3''(t)\,dt\right| &\leqslant \sum_{i=1}^{k}\left|\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}\phi_3''(t)\,dt\right| \\ &\leqslant \max_i\left|l_i^2-\overline{l^2}\right|\sum_{i=1}^{k}\left|\int_{y_{i-1}}^{y_i}\phi_3''(t)\,dt\right| \\ &\leqslant o\left(\bar{l}^2\right)\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}\left|\phi_3''(t)\right|dt \\ &= o\left(\bar{l}^2\right)\int_{y_0}^{y_k}\left|\phi_3''(t)\right|dt.\end{aligned}$$

As was done with ϕ_1, based on (D.8), $\phi_3''(t)$ may be written as

$$\phi_3''(t) = \phi_{30} + \phi_{31}\frac{1}{h} + \phi_{32}\frac{1}{h^2},$$

where ϕ_{30}, ϕ_{31}, ϕ_{32} are products of derivatives of $F'''(t)$ and $K^2\left(\frac{x-t}{h}\right)$. Note that if $t \notin [x-h, x+h]$, then $\left|\frac{x-t}{h}\right| > 1$ and $\phi_3''(t) = 0$. Otherwise,

$$\left|\phi_3''(t)\right| \leqslant \|\phi_{30}\|_\infty + \frac{1}{h}\|\phi_{31}\|_\infty + \frac{1}{h^2}\|\phi_{32}\|_\infty,$$

and since $\int_{y_0}^{y_k} |\phi_3''(t)|\, dt = \int_{x-h}^{x+h} |\phi_3''(t)|\, dt$, then

$$\begin{aligned}\int_{x-h}^{x+h} \left|\phi_3''(t)\right| dt &\leqslant \|\phi_{30}\|_\infty \int_{x-h}^{x+h} dt + \frac{1}{h}\|\phi_{31}\| \int_{x-h}^{x+h} dt + \frac{1}{h^2}\|\phi_{32}\| \int_{x-h}^{x+h} dt \\ &\leqslant 2h\|\phi_{30}\| + 2\|\phi_{31}\| + \frac{2}{h}\|\phi_{32}\| \\ &= O\left(\frac{1}{h}\right).\end{aligned}$$

Thus, the first term on the right hand side of (D.33) is

$$\left|\sum_{i=1}^{k}\left(l_i^2 - \overline{l}^2\right)\int_{y_{i-1}}^{y_i} \phi_3''(t)\, dt\right| \leqslant o\left(\frac{\overline{l}^2}{h}\right). \tag{D.36}$$

By (D.36) and (D.34), Eq. (D.33) becomes

$$\sum_{i=1}^{k} l_i^3 \phi_3''(t_i) = o\left(\frac{\overline{l}^2}{h}\right),$$

so that Eq. (D.32) is finally expressed as

$$\sum_{i=1}^{k} l_i \phi_3(t_i) = \int_{y_0}^{y_k} \phi_3(t)\, dt + o\left(\frac{\overline{l}^2}{h}\right),$$

and Eq. (D.31) as

$$D_1 = \int_{y_0}^{y_k} \phi_3(t)\, dt + O\left(\overline{l}\right).$$

Using the change of variable (D.17), and for a sufficiently large sample size n,

$$\begin{aligned}\int_{y_0}^{y_k} \phi_3(t)\, dt &= h\int K^2(r)\, f(x-hr)\, dr \\ &= h\int K^2(r)\left[f(x) - hrf'(x) + O\left(h^2\right)\right] dr \\ &= h\left[f(x)\, A(K) + O\left(h^2\right)\right];\end{aligned}$$

i.e.,

$$D_1 = hf\left(x\right)A\left(K\right) + o\left(h^2\right),$$

or, multiplying by $1/n$,

$$D = \frac{1}{n}hf\left(x\right)A\left(K\right) + o\left(\frac{h^2}{n}\right). \tag{D.37}$$

Let us now start working with the covariance term in (D.29). As a starting point, by (D.30) and the parity conditions (C.12),

$$p_ip_j = F'\left(t_i\right)F'\left(t_j\right)l_il_j + O\left(\bar{l}^4\right).$$

Define

$$E_1 = -\frac{n}{2}E = \sum_{i<j} K\left(\frac{x-t_i}{h}\right)K\left(\frac{x-t_j}{h}\right)\left[F'\left(t_i\right)F'\left(t_j\right)l_il_j + O\left(\bar{l}^4\right)\right],$$

and define $\phi_4\left(z_1, z_2\right) \equiv K\left(\frac{x-z_1}{h}\right)K\left(\frac{x-z_2}{h}\right)F'\left(z_1\right)F'\left(z_2\right)$. Then,

$$E_1 = \sum_{i<j} l_il_j\phi_4\left(t_i, t_j\right) + O\left(\bar{l}^2\right).$$

The partial derivatives of ϕ_4, $\partial\phi_4/\partial z_1$ and $\partial\phi_4/\partial z_2$, can be expressed as

$$\begin{aligned}\frac{\partial\phi_4}{\partial z_1} &= \phi_{40_{z_1}} + \frac{1}{h}\phi_{41_{z_1}}\\ \frac{\partial\phi_4}{\partial z_2} &= \phi_{40_{z_2}} + \frac{1}{h}\phi_{41_{z_2}},\end{aligned}$$

where $\phi_{40_{z_1}}$, $\phi_{41_{z_1}}$, $\phi_{40_{z_2}}$ and $\phi_{41_{z_2}}$ are functions of products of $K\left(\frac{x-z_1}{h}\right)$, $K\left(\frac{x-z_2}{h}\right)$, $F'\left(z_1\right)$, $F'\left(z_2\right)$ and derivatives. Thus, by multivariate Ostrowski's inequality, given in Eq. (C.20),

$$\begin{aligned}\left|l_il_j\phi_4\left(t_i,t_j\right) - \int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}\phi_4\left(z_1,z_2\right)dz_2dz_1\right| &\leqslant \frac{1}{4}\left[l_i^2l_j\left\|\frac{\partial\phi_4}{\partial z_1}\right\|_\infty + l_j^2l_i\left\|\frac{\partial\phi_4}{\partial z_2}\right\|_\infty\right]\\ &\leqslant \frac{1}{4}l_{max}^3\left[\left(\left\|\phi_{40_{z_1z_2}}\right\|_\infty + \frac{1}{h}\left\|\phi_{41_{z_1z_2}}\right\|_\infty\right)\right]\\ &= O\left(\frac{\bar{l}^3}{h}\right),\end{aligned}$$

where $\phi_{40z_1z_2}{=}\phi_{40_{z_1}} + \phi_{40_{z_2}}$ and $\phi_{41_{z_1z_2}}{=}\phi_{41_{z_1}} + \phi_{41_{z_2}}$.

Summing all over the $\frac{1}{2}\left(k^2-k\right)$ intervals,

$$\begin{aligned}\sum_{i<j}\left|l_i l_j \phi_4(t_i,t_j)-\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}\phi_4(z_1,z_2)\,dz_2dz_1\right| &\leqslant \frac{1}{2}\left(k^2-k\right)O\left(\frac{\bar{l}^3}{h}\right)\\ &= O\left(\frac{\bar{l}}{h}\right);\end{aligned}$$

i.e.,

$$\sum_{i<j} l_i l_j \phi_4(t_i,t_j)=\sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}\phi_4(z_1,z_2)\,dz_2dz_1+O\left(\frac{\bar{l}}{h}\right),$$

or, considering the definition of $\phi_4(z_{1,}z_2)$ and for a sufficiently large n,

$$\sum_{i<j} l_i l_j \phi_4(t_i,t_j)=\int\int_{z_1}^{\infty}K\left(\frac{x-z_1}{h}\right)K\left(\frac{x-z_2}{h}\right)F'(z_1)F'(z_2)\,dz_2dz_1.$$

For convenience, let us call

$$I=\int\int_{z_1}^{\infty}K\left(\frac{x-z_1}{h}\right)K\left(\frac{x-z_2}{h}\right)F'(z_1)F'(z_2)\,dz_2dz_1, \qquad \text{(D.38)}$$

and let us first take the most inner integral. Considering the change of variable (D.17), with z_2 instead of t, and a Taylor series expansion,

$$\begin{aligned}\int_{z_1}^{\infty}K\left(\frac{x-z_2}{h}\right)F'(z_2)\,dz_2 &= h\int_{-\infty}^{\frac{x-z_1}{h}}K(r)\,F'(x-hr)\,dr\\ &= h\int_{-\infty}^{\frac{x-z_1}{h}}K(r)\left[F'(x)-hrF''(x)+\right.\\ &\quad\left.\frac{1}{2}h^2r^2F'''(x)+o\left(h^2\right)\right]dr\\ &= hF'(x)\int_{-\infty}^{\frac{x-z_1}{h}}K(r)\,dr+O\left(h^2\right).\end{aligned}$$

Define $\mathbb{K}(u)=\int_{-\infty}^{u}K(r)\,dr$. Then, going back to (D.38),

$$I=\int K\left(\frac{x-z_1}{h}\right)F'(z_1)\left[hF'(x)\,\mathbb{K}\left(\frac{x-z_1}{h}\right)+O\left(h^2\right)\right]dz_1.$$

Considering the change of variable (D.17), with w instead of r and z_1 instead of t, and again, a Taylor expansion,

$$\begin{aligned}
I &= h\int K\left(w\right)\left[hF'\left(x\right)+O\left(h^{2}\right)\right]\left[hF'\left(x\right)\mathbb{K}\left(w\right)+O\left(h^{2}\right)\right]dw\\
&=\int\left[h^{2}F'^{2}\left(x\right)K\left(w\right)\mathbb{K}\left(w\right)+O\left(h^{3}\right)\right]dw\\
&=h^{2}F'^{2}\left(x\right)\int K\left(w\right)\mathbb{K}\left(w\right)dw+O\left(h^{3}\right).
\end{aligned}$$

Since $\mathbb{K}'\left(\mathfrak{x}\right)=K\left(\mathfrak{x}\right)$,

$$\begin{aligned}
I &= h^{2}F'\left(x\right)^{2}\left[\frac{1}{2}\left\{\mathbb{K}\left(w\right)\right\}^{2}\right]_{-\infty}^{\infty}+O\left(h^{3}\right)\\
&=h^{2}F'\left(x\right)^{2}\left[\frac{1}{2}\left\{\mathbb{K}^{2}\left(\infty\right)-\mathbb{K}^{2}\left(-\infty\right)\right\}\right]_{-\infty}^{\infty}+O\left(h^{3}\right)\\
&=\frac{1}{2}h^{2}F'\left(x\right)^{2}+O\left(h^{3}\right),
\end{aligned}$$

so that E_1 is

$$\begin{aligned}
E_{1} &= \frac{1}{2}h^{2}F'^{2}\left(x\right)+O\left(h^{3}\right)+O\left(\frac{\bar{l}}{h}\right)+O\left(\bar{l}^{2}\right)\\
&=\frac{1}{2}h^{2}F'^{2}\left(x\right)+o\left(h^{2}\right)+O\left(\frac{\bar{l}}{h}\right),
\end{aligned}$$

or, multiplying by $-2/n$ and considering that $F'\left(x\right)=f\left(x\right)$,

$$E=-\frac{1}{n}h^{2}f^{2}\left(x\right)+o\left(\frac{h^{2}}{n}\right)+O\left(\frac{\bar{l}}{h}\right). \tag{D.39}$$

Substituting (D.39) and (D.37) in (D.29),

$$\begin{aligned}
\mathbb{V}\left[\hat{f}_{h}^{g}\left(x\right)\right] &= \frac{1}{h^{2}}\left\{\frac{1}{n}hf\left(x\right)A\left(K\right)-\frac{1}{n}h^{2}f^{2}\left(x\right)+o\left(\frac{h^{2}}{n}\right)+O\left(\frac{\bar{l}}{h}\right)\right\}\\
&=\frac{1}{nh}f\left(x\right)A\left(K\right)+o\left(\frac{1}{nh}\right). \qquad \text{(D.40)}
\end{aligned}$$

Finally, squaring (D.28) and adding (D.40),

$$MSE_{g}=MSE\left[\hat{f}_{h}^{g}\left(x\right)\right]=\frac{1}{4}h^{4}\mu_{2}\left(K\right)^{2}f''^{2}\left(x\right)+\frac{1}{nh}f\left(x\right)A\left(K\right)+o\left(h^{4}\right)+o\left(\frac{1}{nh}\right).$$

The proof for the $MISE_g$ expression is parallel to the one for the MSE_g. The neglected terms have to be proved to be negligible again, when integrating all over the x-domain.

□

Appendix E

Proofs and results of Chapter 4

E.1 Proof of Theorem 4.1

Proof. Let us consider the cases $i = j$ and $i \neq j$ separately. Then, Eq. (4.1) becomes

$$\hat{\psi}_u^g = \frac{1}{\eta^{u+1}} L^{(u)}(0) \sum_{i=1}^{k} w_i^2 + \frac{1}{\eta^{u+1}} \sum_{i \neq j} L^{(u)} \left(\frac{t_i - t_j}{\eta} \right) w_i w_j. \tag{E.1}$$

Recall that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{C}[X,Y]$. Since $(n_1, n_2, ..., n_k)$ is a multinomial random vector, applying the expectation operation to (E.1) gives

$$\mathbb{E}\left(\hat{\psi}_u^g\right) = \alpha_1 + \alpha_2 \tag{E.2}$$

where

$$\alpha_1 = \frac{1}{\eta^{u+1}} L^{(u)}(0) \left(\sum_{i=1}^{k} p_i^2 + \frac{1}{n} \sum_{i=1}^{k} p_i q_i \right) \tag{E.3}$$

and

$$\alpha_2 = \frac{1}{\eta^{u+1}} \sum_{i \neq j} L^{(u)} \left(\frac{t_i - t_j}{\eta} \right) p_i p_j \left(1 - \frac{1}{n} \right),$$

being $p_i = F(y_i) - F(y_{i-1})$, $q_i = 1 - p_i$.

On the one hand, by Eq. (C.10), and considering only the main term,

$$\begin{aligned} p_i^2 &= \left[F'(t_i) l_i + \frac{l_i^2}{8} \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1}) \right] \right]^2 \\ &= F'(t_i)^2 l_i^2 + \frac{1}{4} F'(t_i) l_i^3 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1}) \right] + \frac{1}{64} l_i^4 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1}) \right]^2. \end{aligned}$$

Then

$$\begin{aligned}\sum_{i=1}^{k} p_i^2 &= \sum_{i=1}^{k} F'(t_i)^2 l_i^2 + \frac{1}{4}\sum_{i=1}^{k} F'(t_i)\, l_i^3 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right] + \\ & \quad \frac{1}{64}\sum_{i=1}^{k} l_i^4 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]^2 .\end{aligned}$$

By Assumptions 4.2 and 4.4, and because $k = (y_k - y_0)/\bar{l}$,

$$\begin{aligned}\left|\sum_{i=1}^{k} l_i^4 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]^2\right| &\leqslant \sum_{i=1}^{k}\left| l_i^4 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]^2\right| \\ &\leqslant \sum_{i=1}^{k}\left|l_i^4\right| \left|\left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]^2\right| \\ &\leqslant kO\left(\bar{l}^4\right) \\ &\leqslant O\left(\bar{l}^3\right).\end{aligned}$$

Similarly,

$$\begin{aligned}\left|\sum_{i=1}^{k} F'(t_i)\, l_i^3 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]\right| &\leqslant \sum_{i=1}^{k}\left|F'(t_i)\, l_i^3 \left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]\right| \\ &\leqslant \sum_{i=1}^{k}\left|F'(t_i)\right|\left|l_i^3\right| \left|\left[F^{(2)}(\tau_i) - F^{(2)}(\tau_{i-1})\right]\right| \\ &\leqslant k\left\|F'\right\|_\infty O\left(\bar{l}^3\right) 2\left\|F''\right\|_\infty \\ &\leqslant O\left(\bar{l}^2\right),\end{aligned}$$

so

$$\sum_{i=1}^{k} p_i^2 = \sum_{i=1}^{k} F'(t_i)^2 l_i^2 + O\left(\bar{l}^2\right).$$

On the other hand, following similar steps,

$$\sum_{i=1}^{k} p_i q_i = \sum_{i=!}^{k} F'(t_i)\, l_i + O\left(\bar{l}\right).$$

Going back to (E.3),

$$\alpha_1 = \frac{1}{\eta^{u+1}} L^{(u)}(0) \left[\sum_{i=1}^{k} F'(t_i)^2 l_i^2 + O\left(\bar{l}^2\right) + \frac{1}{n} \sum_{i=1}^{k} F'(t_i) l_i + O\left(\frac{\bar{l}}{n}\right) \right]. \tag{E.4}$$

As to the first term in brackets in (E.4), it can be rewritten as

$$\sum_{i=1}^{k} F'(t_i)^2 l_i^2 = \sum_{i=1}^{k} \left(l_i - \bar{l}\right) l_i F'(t_i)^2 + \bar{l} \sum_{i=1}^{k} l_i F'\left(t_i^2\right). \tag{E.5}$$

Using Assumption 4.4, the first term on the right hand side in (E.5) can be bounded by

$$\begin{aligned} \left| \sum_{i=1}^{k} \left(l_i - \bar{l}\right) l_i F'(t_i)^2 \right| &\leqslant \sum_{i=1}^{k} \left| \left(l_i - \bar{l}\right) l_i F'(t_i)^2 \right| \\ &\leqslant \max_i \left| l_i - \bar{l} \right| k l_{max} \left\| F'^2 \right\|_\infty \\ &\leqslant o\left(\bar{l}\right). \end{aligned} \tag{E.6}$$

Regarding the second term on the right hand side of (E.5), by (C.19)

$$\left| l_i F'(t_i)^2 - \int_{y_{i-1}}^{y_i} F'(t)^2 \, dt \right| \leqslant \frac{1}{4} l_i^2 \mathfrak{L}_{F'^2}.$$

Summing up over all k intervals and by Eq. (C.4),

$$\begin{aligned} \sum_{i=1}^{k} \left| l_i F'(t_i)^2 - \int_{y_{i-1}}^{y_i} F'(t)^2 \, dt \right| &\leqslant \frac{1}{4} \mathfrak{L}_{F'^2} \frac{k}{k} \sum_{i=1}^{k} l_i^2 \\ &\leqslant \frac{1}{4} \mathfrak{L}_{F'^2} \frac{(y_k - y_0)}{\bar{l}} \overline{l^2} \\ &\leqslant O\left(\bar{l}\right). \end{aligned} \tag{E.7}$$

So, by (E.7) and (E.6), Eq. (E.5) is

$$\sum_{i=1}^{k} F'(t_i)^2 l_i^2 = \bar{l} \int_{y_0}^{y_k} F'(t)^2 \, dt + o\left(\bar{l}\right).$$

As to the third term in brackets in (E.4), again, by (C.19),

$$\left| F'(t_i) l_i - \int_{y_{i-1}}^{y_i} F'(t) \, dt \right| \leqslant \frac{1}{4} l_i^2 \mathfrak{L}_{F'},$$

and summing all over the k intervals,

$$\begin{aligned}\sum_{i=1}^{k}\left|F'\left(t_i\right)l_i-\int_{y_{i-1}}^{y_i}F'\left(t\right)dt\right| &\leqslant \frac{1}{4}\mathfrak{L}_{F'}\frac{k}{k}\sum_{i=1}^{k}l_i^2\\ &\leqslant \frac{1}{4}\mathfrak{L}_{F'}\frac{\left(y_k-y_0\right)}{\bar{l}}\overline{l^2}\\ &\leqslant O\left(\bar{l}\right),\end{aligned}$$

i.e.,

$$\begin{aligned}\sum_{i=1}^{k}F'\left(t_i\right)l_i &= \int_{y_0}^{y_k}F'\left(t\right)dt+O\left(\bar{l}\right)\\ &= \int_{y_0}^{y_k}f\left(t\right)dt+O\left(\bar{l}\right)\\ &= 1+O\left(\bar{l}\right).\end{aligned}$$

Going back to (E.4),

$$\begin{aligned}\alpha_1 &= \frac{1}{\eta^{u+1}}L^{(u)}\left(0\right)\left[\bar{l}\int_{y_0}^{y_k}F'\left(t\right)^2dt+o\left(\bar{l}\right)+\frac{1}{n}\left[1+O\left(\bar{l}\right)\right]+O\left(\frac{\bar{l}}{n}\right)\right]\\ &= \frac{1}{\eta^{u+1}}L^{(u)}\left(0\right)\left[\bar{l}A\left(f\right)+o\left(\bar{l}\right)\right]\\ &= O\left(\frac{\bar{l}}{\eta^{u+1}}\right).\end{aligned}\tag{E.8}$$

Concerning α_2, by Eq. (C.10) and just considering the main term, multiplying p_i and p_j gives

$$p_ip_j=F'\left(t_i\right)F'\left(t_j\right)l_il_j+O\left(\bar{l}^3\right),$$

so

$$\alpha_2=\left(1-\frac{1}{n}\right)\frac{1}{\eta^{u+1}}\left[\sum_{i\neq j}\Phi_1\left(t_i,t_j\right)l_il_j+O\left(\bar{l}^3\right)\sum_{i\neq j}L^{(u)}\left(\frac{t_i-t_j}{\eta}\right)\right],\tag{E.9}$$

where $\Phi_1\left(z_1,z_2\right)\equiv L^{(u)}\left(\frac{z_1-z_2}{\eta}\right)F'\left(z_1\right)F'\left(z_2\right)$.

As to the second term in brackets in Eq. (E.9),

$$\begin{aligned}\left|\sum_{i\neq j} L^{(u)}\left(\frac{t_i-t_j}{\eta}\right)\right| &\leqslant \sum_{i\neq j}\left|L^{(u)}\left(\frac{t_i-t_j}{\eta}\right)\right| \\ &\leqslant \left(k^2-k\right)\left\|L^{(u)}\right\|_{\infty} \\ &\leqslant \left(\frac{\left(y_k-y_0\right)^2}{\bar{l}^2}-\frac{\left(y_k-y_0\right)}{\bar{l}}\right)\left\|L^{(u)}\right\|_{\infty}.\end{aligned}$$

Then,

$$O\left(\bar{l}^3\right)\sum_{i\neq j} L^{(u)}\left(\frac{t_i-t_j}{\eta}\right)=O\left(\bar{l}\right).$$

Substituting in (E.9),

$$\alpha_2=\left(1-\frac{1}{n}\right)\frac{1}{\eta^{u+1}}\left[\sum_{i\neq j}\Phi_1\left(t_i,t_j\right)l_il_j+O\left(\bar{l}\right)\right].$$

Note that as the sample size increases, by Assumption 4.4, the number of intervals k increases and the average length $\bar{l}$ decreases. So, the first term in brackets can be expressed as

$$\sum_{i\neq j}\Phi_1\left(t_i,t_j\right)l_il_j=\int\int\Phi_1\left(z_1,z_2\right)dz_2dz_1+O\left(\bar{l}\right),$$

so that

$$\alpha_2=\left(1-\frac{1}{n}\right)\frac{1}{\eta^{u+1}}\left[\int\int\Phi\left(z_1,z_2\right)dz_2dz_1+O\left(\bar{l}\right)\right].$$

Now, putting the factor $\frac{1}{\eta^{u+1}}$ into the double integral in the last equation and defining $r=\frac{Z_1-Z_2}{\eta}$, integrating by parts and by convolution properties, α_2 becomes

$$\alpha_2=\left(1-\frac{1}{n}\right)\left[\int\int L\left(r\right)f\left(z_2+\eta r\right)f^{(u)}\left(z_2\right)dz_2dr+O\left(\frac{\bar{l}}{\eta^{u+1}}\right)\right]. \tag{E.10}$$

By Taylor series of $f\left(z_2+\eta r\right)$ around z_2,

$$\begin{aligned}\int\int L\left(r\right)f\left(z_2+\eta r\right)f^{(u)}\left(z_2\right)dz_2dr \quad = \quad & \int\int L\left(r\right)f^{(u)}\left(z_2\right)\left[f\left(z_2\right)+\eta rf'\left(z_2\right)+...\right. \\ & \left.+\frac{\eta^sr^s}{s!}f^{(s)}\left(z_2\right)+\frac{\eta^{s+1}r^{s+1}}{(s+1)!}f^{(s+1)}\left(\xi\right)\right]dz_2dr.\end{aligned}$$

Now, standard algebra and using that L is of order s (Assumption 4.1), lead to

$$\begin{aligned}\int\int L(r)\, f(z_2+\eta r)\, f^{(u)}(z_2)\, dz_2 dr &= \int f^{(u)}(z_2)\, f(z_2)\, dz_2 + \\ &\quad \frac{\eta^s}{s!}\mu_s(L)\int f^{(u)}(z_2)\, f^{(s)}(z_2)\, dz_2 + \\ &\quad O\left(\eta^{s+1}\right) \\ &= \psi_u + \frac{\eta^s}{s!}\mu_s(L)\,\psi_{u+s} + O\left(\eta^{s+1}\right). \qquad \text{(E.11)}\end{aligned}$$

Substituting (E.11) into (E.10) and multiplying by $\left(1-\frac{1}{n}\right)$,

$$\alpha_2 = \psi_u + \frac{\eta^s}{s!}\mu_s(L)\,\psi_{u+s} + O\left(\eta^{s+1}\right) + O\left(\frac{\bar{l}}{\eta^{u+1}}\right). \qquad \text{(E.12)}$$

So, by (E.12) and (E.8), the expectation (E.2) can be finally written as

$$\mathbb{E}\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}}{\eta^{u+1}}\right) + \psi_u + \frac{\eta^s}{s!}\sigma_L^s\psi_{u+s} + O\left(\eta^{s+1}\right),$$

from which, the bias is

$$\mathbb{B}\left(\psi_u^g\right) = \frac{\eta^s}{s!}\sigma_L^s\psi_{u+s} + O\left(\eta^{s+1}\right) + O\left(\frac{\bar{l}}{\eta^{u+1}}\right). \qquad \text{(E.13)}$$

Regarding the variance of (4.1), it is expressed as

$$\mathbb{V}\left(\hat{\psi}_u^g\right) = \frac{1}{\eta^{2u+2}}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{r=1}^{k}\sum_{v=1}^{k} L^{(u)}\left(\frac{t_i-t_j}{\eta}\right) L^{(u)}\left(\frac{t_r-t_v}{\eta}\right)\mathbb{C}\left(w_iw_j, w_rw_v\right). \qquad \text{(E.14)}$$

In Eq. (E.14), the covariance can be treated by considering all the different 7 cases in which indices are equal or unequal between each other. In each case, there could be equivalent situations, from which only one will be treated, as all of them are of the same order. In what follows, all those covariance cases will be expressed as moments of the multinomial distribution (Newcomer et al., 2008). The term corresponding to case a will be expressed as $\mathbb{V}_a\left(\hat{\psi}_u^g\right)$, such that $\mathbb{V}\left(\hat{\psi}_u^g\right) = 4\mathbb{V}_1\left(\hat{\psi}_u^g\right) + 4\mathbb{V}_2\left(\hat{\psi}_u^g\right) + \mathbb{V}_3\left(\hat{\psi}_u^g\right) + \mathbb{V}_4\left(\hat{\psi}_u^g\right) + 2\mathbb{V}_5\left(\hat{\psi}_u^g\right) + \mathbb{V}_6\left(\hat{\psi}_u^g\right) + 2\mathbb{V}_7\left(\hat{\psi}_u^g\right)$.

- Case 1: $i \neq j$ and $[i = r \neq s; j \neq v]$. Expressing covariance as expectations:

$$\begin{aligned}\mathbb{C}\left(w_iw_j, w_iw_v\right) &= \mathbb{E}\left(w_i^2w_jw_v\right) - \mathbb{E}\left(w_iw_j\right)\mathbb{E}\left(w_iw_v\right) \\ &= \frac{1}{n^4}\mathbb{E}\left(n_i^2n_jn_v\right) - \frac{1}{n^4}\mathbb{E}\left(n_in_j\right)\mathbb{E}\left(n_in_v\right).\end{aligned}$$

According to Newcomer et al. (2008), doing some algebra and substituting back in

(E.14), the main term is

$$\mathbb{V}_1\left(\hat{\psi}_u^g\right) \approx \frac{1}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \sum_{\substack{i,j,v=1 \\ i\neq j\neq v\neq i}}^{k} \gamma_1\left(t_i, t_j, t_v\right) l_i l_j l_v,$$

where

$$\gamma_1\left(z_1, z_2, z_3\right) = L^{(u)}\left(\frac{z_1 - z_2}{\eta}\right) L^{(u)}\left(\frac{z_1 - z_3}{\eta}\right) F'\left(z_1\right) F'\left(z_2\right) F'\left(z_3\right).$$

By Assumption 4.4, the sum in $\mathbb{V}_1\left(\hat{\psi}_u^g\right)$ can be approximated by a triple integral, so that

$$\mathbb{V}_1\left(\hat{\psi}_u^g\right) = \frac{1}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \int\int\int \gamma_1\left(z_1, z_2, z_3\right) dz_3 dz_2 dz_1 + o\left(\frac{1}{\eta^{2u+2} n}\right),$$

and finally,

$$\mathbb{V}_1\left(\hat{\psi}_u^g\right) = O\left(\frac{1}{\eta^{2u} n}\right). \qquad \text{(E.15)}$$

The other three situations like this occur when $i \neq j$, and $[i = v \neq r; j \neq r]$, $[j = r \neq v; i \neq v]$, or $[j = v \neq r; i \neq r]$.

- Case 2: $i \neq j$ and $r = v = i$.

$$\mathbb{C}\left(w_i w_j, w_i w_i\right) = \frac{1}{n^4}\mathbb{E}\left(n_i^3 n_j\right) - \frac{1}{n^4}\mathbb{E}\left(n_i n_j\right)\mathbb{E}\left(n_i^2\right).$$

Considering the moments of the multinomial distribution,

$$\mathbb{V}_2\left(\hat{\psi}_u^g\right) \approx \frac{1}{\eta^{2u+2}} O\left(\frac{1}{n}\right) L^{(u)}(0) \sum_{i\neq j} \gamma_2\left(t_i, t_j\right) l_i^2 l_j,$$

where $\gamma_2\left(z_1, z_2\right) = L^{(u)}\left(\frac{z_1 - z_2}{\eta}\right) F'^2\left(z_1\right) F'\left(z_2\right)$.

By Assumption 4.4,

$$\mathbb{V}_2\left(\hat{\psi}_u^g\right) = \frac{1}{\eta^{2u+2}} O\left(\frac{\bar{l}}{n}\right) \int\int \gamma_2\left(z_1, z_2\right) dz_2 dz_1 + \frac{1}{\eta^{2u+2}} o\left(\frac{\bar{l}}{n}\right),$$

so that, using the change of variable $r = \left(z_1 - z_2\right)/\eta$,

$$\mathbb{V}_2\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}}{\eta^{2u+1} n}\right). \qquad \text{(E.16)}$$

The other three situations like this are: $[i \neq j; r = v = j]$,$[i = j = r \neq v]$ and $[i = j = v \neq r]$.

- Case 3: $i = j = v = r$.

$$\mathbb{C}\left(w_i w_i, w_i w_i\right) = \mathbb{V}\left(w_i^2\right) = \frac{1}{n^4}\mathbb{E}\left(n_i^4\right) - \frac{1}{n^4}\mathbb{E}\left(n_i^2\right)^2.$$

Considering the moments of the multinomial distribution,

$$\mathbb{V}_3\left(\hat{\psi}_u^g\right) \approx \frac{1}{\eta^{2u+2}} L^{(u)}(0)^2 O\left(\frac{1}{n}\right) \sum_{i=1}^{k} \gamma_3\left(t_i\right) l_i^3,$$

where $\gamma_3(t) = F'^3(t)$.

Proceeding as in (E.5), and by Assumption 4.4,

$$\mathbb{V}_3\left(\hat{\psi}_u^g\right) = \frac{1}{\eta^{2u+2}} L^{(u)}(0)^2 O\left(\frac{\bar{l}^2}{n}\right) \int_{y_0}^{y_k} \gamma_3(t) dt + o\left(\frac{\bar{l}^2}{n\eta^{2u+2}}\right),$$

i.e.,

$$\mathbb{V}_3\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}^2}{n\eta^{2u+2}}\right). \tag{E.17}$$

There are no more situations like this.

- Case 4: $i = j \neq r = v$.

$$\mathbb{C}\left(w_i w_j, w_r w_v\right) = \mathbb{C}\left(w_i^2, w_r^2\right) = \frac{1}{n^4}\mathbb{E}\left(n_i^2 n_r^2\right) - \frac{1}{n^4}\mathbb{E}\left(n_i^2\right)\mathbb{E}\left(n_r^2\right).$$

Using the moments of the multinomial distribution,

$$\mathbb{V}_4\left(\hat{\psi}_u^g\right) \approx \frac{L^{(u)}\left(0\right)^2}{n\eta^{2u+2}} \sum_{i \neq r} \gamma_4\left(t_i, t_r\right) l_i^2 l_r^2,$$

where $\gamma_4\left(z_1, z_2\right) = F'(z_1)^2 F'\left(z_2\right)^2$. By Assumption 4.4,

$$\mathbb{V}_4\left(\hat{\psi}_u^g\right) = \frac{\bar{l}^2 L^{(u)}\left(0\right)^2}{n\eta^{2u+2}} \int\int \gamma_4\left(z_1, z_2\right) dz_1 dz_2 + o\left(\frac{\bar{l}^2}{n\eta^{2u+2}}\right),$$

which means that

$$\mathbb{V}_4\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}^2}{n\eta^{2u+2}}\right). \tag{E.18}$$

and there are no more situations like this.

- Case 5: $v \neq i = j \neq r \neq v$.

$$\mathbb{C}\left(w_i w_j, w_r w_v\right) = \mathbb{C}\left(w_i^2, w_r w_v\right) = \frac{1}{n^4}\mathbb{E}\left(n_i^2 n_r n_v\right) - \frac{1}{n^4}\mathbb{E}\left(n_i^2\right)\mathbb{E}\left(n_r n_v\right).$$

Moments of the multinomial distribution lead us to

$$\mathbb{V}_5\left(\hat{\psi}_u^g\right) \approx \frac{L^{(u)}(0)}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \sum_{v \neq i = j \neq r \neq v} \gamma_5\left(t_i, t_r, t_v\right) l_i^2 l_r l_v,$$

where $\gamma_5\left(z_1, z_2, z_3\right) = L^{(u)}\left(\frac{z_2 - z_3}{\eta}\right) F'\left(z_1\right)^2 F'\left(z_2\right) F'\left(z_3\right)$. Proceeding as in (E.5) and by Assumption 4.4,

$$\mathbb{V}_5\left(\hat{\psi}_u^g\right) = \frac{L^{(u)}(0)}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \bar{l} \int\int\int \gamma_5\left(z_1, z_2, z_3\right) dz_3 dz_2 dz_1 + o\left(\frac{\bar{l}}{n\eta^{2u+2}}\right),$$

so that,

$$\mathbb{V}_5\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}}{n\eta^{2u+1}}\right). \tag{E.19}$$

There is also one more situation like this, which is $i \neq j \neq r = v \neq i$.

- Case 6: $i \neq j \neq r \neq v$, and $r \neq i \neq v \neq j$ (i.e. $\#\{i, j, r, v\} = 4$)

$$\mathbb{C}\left(w_i w_j, w_r w_v\right) = \frac{1}{n^4}\mathbb{E}\left(n_i n_j n_r n_v\right) - \frac{1}{n^4}\mathbb{E}\left(n_i n_j\right)\mathbb{E}\left(n_r n_v\right)$$

and by the moments of the multinomial distribution,

$$\mathbb{V}_6\left(\hat{\psi}_u^g\right) \approx \frac{1}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \sum_{\#\{i,j,r,v\}=4} \gamma_6\left(t_i, t_j, t_r, t_v\right) l_i l_j l_r l_v,$$

where

$$\gamma_6\left(z_1, z_2, z_3, z_4\right) = L^{(u)}\left(\frac{z_1 - z_2}{\eta}\right) L^{(u)}\left(\frac{z_3 - z_4}{\eta}\right) F'\left(z_1\right) F'\left(z_2\right) F'\left(z_3\right) F'\left(z_4\right).$$

By Assumption 4.4,

$$\begin{aligned}\mathbb{V}_6\left(\hat{\psi}_u^g\right) &= \frac{1}{\eta^{2u+2}} O\left(\frac{1}{n}\right) \int\int\int\int \gamma_6\left(z_1, z_2, z_3, z_4\right) dz_4 dz_3 dz_2 dz_1 \\ &+ o\left(\frac{1}{n\eta^{2u+2}}\right),\end{aligned}$$

i.e.,

$$\mathbb{V}_6\left(\hat{\psi}_u^g\right) = O\left(\frac{1}{n\eta^{2u}}\right). \tag{E.20}$$

- Case 7: $i = r \neq j = v$.

$$\mathbb{C}\left(w_i w_j, w_r w_v\right) = \mathbb{C}\left(w_i w_j, w_i w_j\right) = \mathbb{V}\left(w_i w_j\right),$$

Using the moments of the multinomial distribution:

$$\mathbb{V}_7\left(\hat{\psi}_u^g\right) \approx \frac{1}{n\eta^{2u+2}} \sum_{i\neq j} \gamma_7\left(t_i, t_j\right) l_i^2 l_j,$$

where $\gamma_7\left(z_1, z_2\right) = L^{(u)}\left(\frac{z_1-z_2}{\eta}\right)^2 F'\left(z_1\right)^2 F'\left(z_2\right)$. By assumption 4.4,

$$\mathbb{V}_7\left(\hat{\psi}_u^g\right) = \frac{\bar{l}}{n\eta^{2u+2}} \int\int \gamma_7\left(z_1, z_2\right) dz_2 dz_1 + o\left(\frac{\bar{l}}{n\eta^{2u+2}}\right).$$

so that,

$$\mathbb{V}_7\left(\hat{\psi}_u^g\right) = O\left(\frac{\bar{l}}{n\eta^{2u+1}}\right). \tag{E.21}$$

There is one more case like this, which is $i = v \neq j = r$.

Considering equations (E.15) to (E.21) and Assumption 4.4, the order of (E.14) is

$$\mathbb{V}\left(\hat{\psi}_u^g\right) = O\left(\frac{1}{n\eta^{2u}}\right). \tag{E.22}$$

On the one hand, using Assumptions 4.3 and 4.4, although $\hat{\psi}_u^g$ is not an unbiased estimator, it is asymptotically unbiased, as can be seen in (E.13). On the other hand, by Assumption 4.3, Eq. (E.22) shows that variance asymptotically vanishes. This means that as n increases, $\hat{\psi}_u^g$ converges in quadratic mean to ψ_u, and the probability of the estimator being arbitrarily close to ψ_u converges to one. Thus, the weak consistency of (4.1) has been proved.

□

E.2 Proof of Theorem 4.2

Proof. Recall that, by definition,

$$\hat{f}_h^g(x) = \frac{1}{h}\sum_{i=1}^{k} w_i K\left(\frac{x-t_i}{h}\right) = \sum_{i=1}^{k} w_i K_h\left(x-t_i\right),$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ and $w_i = \frac{n_i}{n}$, where n_i is the number of data in the i-th interval and n is the sample size. Then, $MISE_g = MISE\left(\hat{f}_h^g\right)$ is

$$\begin{aligned} MISE_g &= \mathbb{E}\left[\int\left[\hat{f}_h^g(x) - f(x)\right]^2 dx\right] = \int \mathbb{E}\left\{\left[\hat{f}_h^g(x) - f(x)\right]^2\right\} dx \\ &= \int\left[\mathbb{E}\left[\hat{f}_h^g(x)\right] - f(x)\right]^2 dx + \int \mathbb{V}\left[\hat{f}_h^g(x)\right] dx. \end{aligned} \tag{E.23}$$

But

$$\mathbb{E}\left[\hat{f}_h^g(x)\right] = \sum_{i=1}^{k} \mathbb{E}(w_i) K_h(x - t_i) = \sum_{i=1}^{k} p_i K_h(x - t_i),$$

where $p_i = F(y_i) - F(y_{i-1})$, so

$$\int \left[\mathbb{E}\left[\hat{f}_h^g(x)\right] - f(x)\right]^2 dx = \int \left[\sum_{i=1}^{k} p_i K_h(x - t_i) - f(x)\right]^2 dx. \tag{E.24}$$

Besides,

$$\begin{aligned} \mathbb{V}\left[\hat{f}_h^g(x)\right] &= \sum_{i=1}^{k} \mathbb{V}(w_i) K_h(x - t_i)^2 + \\ &\quad 2\sum_{i<j}^{k} \mathbb{C}(w_i, w_j) K_h(x - t_i) K_h(x - t_j). \end{aligned} \tag{E.25}$$

Since $w_i = \frac{n_i}{n}$ and $(n_1, n_2, ..., n_k)$ is a random multinomial vector, then

$$\mathbb{V}(w_i) = \frac{1}{n} p_i (1 - p_i) \tag{E.26}$$

and, for $i < j$,

$$\mathbb{C}(w_i, w_j) = -\frac{1}{n} p_i p_j. \tag{E.27}$$

Thus, substituting (E.27) and (E.26) in (E.25),

$$\begin{aligned} \mathbb{V}\left[\hat{f}_h^g(x)\right] &= \frac{1}{n}\sum_{i=1}^{k} p_i (1 - p_i) K_h(x - t_i)^2 - \\ &\quad \frac{2}{n}\sum_{i<j} p_i p_j K_h(x - t_i) K_h(x - t_j) \\ &= \frac{1}{n}\sum_{i=1}^{k} p_i K_h(x - t_i)^2 - \frac{1}{n}\sum_{i,j=1}^{k} p_i p_j K_h(x - t_i) K_h(x - t_j), \end{aligned}$$

and so,

$$\int \mathbb{V}\left[\hat{f}_h^g(x)\right] dx = \frac{1}{n}\sum_{i=1}^{k} p_i \int K_h(x - t_i)^2 dx - \frac{1}{n}\sum_{i,j=1}^{k} p_i p_j \int K_h(x - t_i) K_h(x - t_j) dx. \tag{E.28}$$

Now, substituting (E.28) and (E.24) in (E.23),

$$\begin{aligned} MISE_g &= \int \left[\sum_{i=1}^{k} p_i K_h (x - t_i) - f(x) \right]^2 dx + \frac{1}{n} \sum_{i=1}^{k} p_i \int K_h (x - t_i)^2 \, dx - \\ & \frac{1}{n} \sum_{i,j=1}^{k} p_i p_j \int K_h (x - t_i) K_h (x - t_j) \, dx. \end{aligned} \tag{E.29}$$

The above expression can be simplified considering that

$$\begin{aligned} \int K_h (x - t_i)^2 \, dx &= \frac{1}{h^2} \int K \left(\frac{x - t_i}{h} \right)^2 dx \\ &= \frac{1}{h} \int K(u)^2 \, du \\ &= \frac{1}{h} A(K), \end{aligned} \tag{E.30}$$

where the change of variable $u = \frac{x - t_i}{h}$ was used. Also,

$$\begin{aligned} \int K_h (x - t_i) K_h (x - t_j) \, dx &= \frac{1}{h^2} \int K \left(\frac{x - t_i}{h} \right) K \left(\frac{x - t_j}{h} \right) dx \\ &= \frac{1}{h} \int K \left(\frac{t_j + hu - t_i}{h} \right) K(u) \, du \\ &= \frac{1}{h} \int K \left(u + \frac{t_j - t_i}{h} \right) K(u) \, du \\ &= \frac{1}{h} \int K \left(\frac{t_i - t_j}{h} - u \right) K(u) \, du \\ &= \frac{1}{h} K * K \left(\frac{t_i - t_j}{h} \right) = (K * K)_h (t_i - t_j), \end{aligned} \tag{E.31}$$

where the change of variable $u = \frac{x - t_j}{h}$ was used. Therefore, using (E.31) and (E.30) in (E.29),

$$MISE_g = \int \left[\sum_{i=1}^{k} p_i K_h (x - t_i) - f(x) \right]^2 dx + \frac{A(K)}{nh} - \frac{1}{n} \sum_{i,j=1}^{k} p_i p_j (K * K)_h (t_i - t_j), \tag{E.32}$$

which is expression (4.2).

□

E.2.1 Obtaining Equation (4.6)

When using the bootstrap method for selecting the bandwidth h, a pilot bandwidth ζ is used, which generally is asymptotically greater than h. Then, the estimator

$$\hat{f}_{\zeta}^{g}(x) = \sum_{i=1}^{k} w_i K_{\zeta}(x - t_i)$$

is considered as a reference density.

Next, a generic bootstrap sample $X_1^*, X_2^*, ..., X_n^*$ from $\hat{f}_{\zeta}^{g}$ is obtained and the bootstrap $MISE$ is defined as

$$MISE^* = MISE^*\left(\hat{f}_h^{g*}\right) = \mathbb{E}^*\left[\int \left(\hat{f}_h^{g*}(x) - \hat{f}_{\zeta}^{g}(x)\right)^2 dx\right],$$

where

$$\hat{f}_h^{g*}(x) = \sum_{i=1}^{k} w_i^* K_h(x - t_i),$$

and w_i^* are the observed proportion of data in the i-th interval when sampling from $\hat{f}_{\zeta}^{g}$. Following analog steps to those that yielded (E.32), it is obtained that

$$MISE^* = \int \left[\sum_{i=1}^{k} w_i^{\zeta} K_h(x - t_i) - \hat{f}_{\zeta}^{g}(x)\right]^2 dx + \frac{A(K)}{nh} - \frac{1}{n}\sum_{i,j=1}^{k} w_i^{\zeta} w_j^{\zeta} (K * K)_h (t_i - t_j), \tag{E.33}$$

where $w_i^{\zeta} = \mathbb{E}^*(w_i^*)$ is the proportion of data in the i-th interval of the reference density $\hat{f}_{\zeta}^{g}$; in other words, $w_i^{\zeta} = \hat{F}_{\zeta}(y_i) - \hat{F}_{\zeta}(y_{i-1})$, where

$$\begin{aligned}
\hat{F}_{\zeta}(y) &= \int_{-\infty}^{y} \hat{f}_{\zeta}^{g}(u)\, du \\
&= \sum_{i=1}^{k} w_i \int_{-\infty}^{y} K_{\zeta}(u - t_i) \\
&= \sum_{i=1}^{k} w_i \frac{1}{\zeta} \int_{-\infty}^{y} K\left(\frac{u - t_i}{\zeta}\right) du \\
&= \sum_{i=1}^{k} w_i \int_{-\infty}^{\frac{y - t_i}{\zeta}} K(v)\, dv \\
&= \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{y - t_i}{\zeta}\right),
\end{aligned}$$

and the change of variable $v = \frac{u-t_i}{\zeta}$ was considered. This is,

$$\begin{aligned} w_i^\zeta &= \sum_{j=1}^{k} w_j \mathbb{K}\left(\frac{y_i - t_j}{\zeta}\right) - \sum_{j=1}^{k} w_j \mathbb{K}\left(\frac{y_{i-1} - t_j}{\zeta}\right) \\ &= \sum_{j=1}^{k} w_j \left[\mathbb{K}\left(\frac{y_i - t_j}{\zeta}\right) - \mathbb{K}\left(\frac{y_{i-1} - t_j}{\zeta}\right)\right]. \end{aligned} \tag{E.34}$$

Then, the bootstrap integrated squared bias can be expressed as

$$\begin{aligned} \int \left[\sum_{i=1}^{k} w_i^\zeta K_h(x-t_i) - \sum_{i=1}^{k} w_i K_\zeta(x-t_i)\right]^2 dx &= \int \left\{\sum_{i=1}^{k} \left[w_i^\zeta K_h(x-t_i) - w_i K_\zeta(x-t_i)\right]\right\}^2 dx \\ &= \sum_{i,j=1}^{k} \int \left\{\left[w_i^\zeta K_h(x-t_i) - w_i K_\zeta(x-t_i)\right]\right. \\ &\quad \left.\left[w_j^\zeta K_h(x-t_j) - w_j K_\zeta(x-t_j)\right]\right\} dx \\ &= \sum_{i,j=1}^{k} \left\{w_i^\zeta w_j^\zeta \int K_h(x-t_i) K_h(x-t_j)\, dx - \right. \\ &\quad w_i^\zeta w_j \int K_\zeta(x-t_i) K_\zeta(x-t_j) - \\ &\quad w_i w_j^\zeta \int K_\zeta(x-t_i) K_h(x-t_j) + \\ &\quad \left. w_i w_j \int K_\zeta(x-t_i) K_\zeta(x-t_j)\right\}. \end{aligned} \tag{E.35}$$

Parallel to equation (E.31),

$$\int K_\zeta(x-t_i) K_\zeta(x-t_j)\, dx = (K * K)_\zeta (t_i - t_j). \tag{E.36}$$

Similarly,

$$\begin{aligned} \int K_h(x-t_i) K_\zeta(x-t_j)\, dx &= \int K_h(t_j + u - t_i) K_\zeta(u) \\ &= \int K_h(t_i - t_j - u) K_\zeta(u)\, du \\ &= (K_h * K_\zeta)(t_i - t_j), \end{aligned} \tag{E.37}$$

where the change of variable $u = x - t_j$ was used. Thus,

$$\int K_\zeta(x-t_i)\,K_h(x-t_j)\,dx = (K_\zeta * K_h)(t_i-t_j)\,. \tag{E.38}$$

Substituting (E.38), (E.37) and (E.36) in (E.35), and then in (E.33), a closed expression for the bootstrap $MISE$ is obtained,

$$\begin{aligned} MISE^* &= \sum_{i,j=1}^{k} w_i^\zeta w_j^\zeta (K*K)_h(t_i-t_j) - 2\sum_{i,j=1}^{k} w_i^\zeta w_j (K_h * K_\zeta)(t_i-t_j) \\ &\quad + \sum_{i,j=1}^{k} w_i w_j (K*K)_\zeta (t_i-t_j) + \frac{A(K)}{nh} \\ &\quad - \frac{1}{n}\sum_{i,j=1}^{k} w_i^\zeta w_j^\zeta (K*K)_\zeta (t_i-t_j)\,; \end{aligned}$$

i.e.,

$$\begin{aligned} MISE^* &= \frac{n-1}{n}\sum_{i,j=1}^{k} w_i^\zeta w_j^\zeta (K*K)_h(t_i-t_j) - 2\sum_{i,j=1}^{k} w_i^\zeta w_j (K_h * K_\zeta)(t_i-t_j) \\ &\quad + \sum_{i,j=1}^{k} w_i w_j (K*K)_\zeta (t_i-t_j) + \frac{A(K)}{nh}. \end{aligned} \tag{E.39}$$

Eq. (E.39) evaluates $MISE^*$ in h with no need of using Monte Carlo; i.e., without throwing different samples. Note that w_i^ζ, w_j^ζ, w_i, w_j and $(K*K)_\zeta(t_i-t_j)$ do not depend on h, and therefore they can be evaluated only once, while varying h to numerically approximate

$$h_{MISE^*} = \arg\min_h MISE^*.$$

If K were the Gaussian kernel, it is easy to see that $K*K$ is a $N(0,2)$, $(K*K)_h$ is a $N\left(0,2h^2\right)$, $K_h * K_\zeta$ is a $N\left(0,h^2+\zeta^2\right)$ and $K_\zeta * K_\zeta$ is a $N\left(0,2\zeta^2\right)$. As a consequence, in this particular case, Eq. (E.39) becomes

$$\begin{aligned} MISE^* &= \frac{n-1}{n}\sum_{i,j=1}^{k} w_i^\zeta w_j^\zeta K_{\sqrt{2}h}(t_i-t_j) - 2\sum_{i,j=1}^{k} w_i^\zeta w_j K_{\sqrt{h^2+\zeta^2}}(t_i-t_j) \\ &\quad + \sum_{i,j=1}^{k} w_i w_j K_{\sqrt{2}\zeta}(t_i-t_j) + \frac{A(K)}{nh}, \end{aligned}$$

which is Eq. (4.6).

E.3 Pilot bandwidth ζ_{opt}

For using the bootstrap bandwidth selector, a pilot bandwidth ζ is necessary. It has been well studied that this bandwidth ζ should be the one that gives the best approximation to the curvature of the density f, $A\left(f''\right)$. In other words, ζ should be the one that minimizes $\mathbb{E}\left\{\left[\hat{A}_\zeta\left(f''\right)-A\left(f''\right)\right]^2\right\}$. According to Cao (1990), $A\left(f''\right)$ may be approximated by

$$\hat{A}_\zeta\left(f''\right)\approx\frac{1}{n^2\zeta^6}\sum_{i\neq j}\int K''\left(\frac{x-x_i}{\zeta}\right)K''\left(\frac{x-x_j}{\zeta}\right)dx, \tag{E.40}$$

and also,

$$\begin{aligned}n^{-2}\zeta^{-6}\sum_{i\neq j}\int K''\left(\frac{x-x_i}{\zeta}\right)K''\left(\frac{x-x_j}{\zeta}\right)dx &= A\left(f''\right)-\zeta^2\mu_2\left(K\right)A\left(f'''\right) \quad \text{(E.41)}\\ &\quad +U_n+O\left(n^{-\frac{1}{2}}\right)+O\left(\zeta^3\right)\end{aligned}$$

where U_n is an U-statistic, such that

$$\mathbb{E}\left[U_n\right]=0 \tag{E.42}$$

and

$$\mathbb{V}\left[U_n\right]=2n^{-2}\zeta^{-9}A\left(f\right)A\left[K''*K''\right]+O\left(\zeta^3\right). \tag{E.43}$$

Thus,

$$A\left(f''\right)-\hat{A}_\zeta\left(f''\right)\approx\zeta^2\mu_2\left(K\right)A\left(f'''\right)-U_n+O\left(n^{-\frac{1}{2}}\right)+O\left(\zeta^3\right). \tag{E.44}$$

Squaring (E.44) and applying the expectation operator, the mean squared error is obtained,

$$\mathbb{E}\left\{\left[A\left(f''\right)-\hat{A}_\zeta\left(f''\right)\right]^2\right\}=\left\{\mathbb{E}\left[\hat{A}_\zeta\left(f''\right)\right]-A\left(f''\right)\right\}^2+\mathbb{V}\left[\hat{A}_\zeta\left(f''\right)\right],$$

and by (E.44) and (E.43),

$$\mathbb{E}\left\{\left[A\left(f''\right)-\hat{A}_\zeta\left(f''\right)\right]^2\right\}\approx\zeta^4\mu_2\left(K\right)^2A\left(f'''\right)^2+2n^{-2}\zeta^{-9}A\left(f\right)A\left[K''*K''\right]+O\left(n^{-1}\right)+O\left(\zeta^6\right).$$

Define

$$\chi\left(\zeta\right)=a\zeta^4+bn^{-2}\zeta^{-9}, \tag{E.45}$$

where $a = \mu_2(K)^2 A(f''')^2$ and $b = 2A(f)A[K'' * K'']$. Then, imposing that the first derivative of (E.45) equals zero,

$$\chi'(\zeta) = 4a\zeta^3 - 9bn^{-2}\zeta^{-10} = 0. \tag{E.46}$$

Now, solving for ζ, it is finally obtained

$$\zeta_{opt} = \left[\frac{9A(f)A[K'' * K'']n^{-2}}{4\mu_2(K)^2 A(f''')^2}\right]^{\frac{1}{13}}. \tag{E.47}$$

For having a practical expression for ζ_{opt}, the quantities $A[K'' * K'']$, $A(f)$ and $A(f''')^2$ need to be known or estimated. Assuming that f is a $N(\mu, \sigma^2)$ and K is a Gaussian kernel, it is easy to see that

$$\mu_2(K) = 1,$$

$$A(f) = \frac{1}{2\sigma\sqrt{\pi}},$$

$$A(f''') = \frac{15}{16\sigma^7\sqrt{\pi}},$$

and

$$A[K'' * K''] = \frac{11}{512\sqrt{2\pi}},$$

which substituting into (E.47) gives

$$\zeta_{opt} = \left(\frac{11}{200\sqrt{2}}\sigma^{13}n^{-2}\right)^{\frac{1}{13}} \approx 0.78\sigma n^{-\frac{2}{13}}. \tag{E.48}$$

In practice, given a sample of size n, σ can be estimated and plugged into (E.48). More detailed explanations about all expressions presented here can be found in Cao (1990, 1993).

E.4 On how to select the pilot bandwidth ζ for grouped data

For using the bootstrap bandwidth selector, it is necessary a pilot bandwidth ζ. It was shown in Appendix E.3 that, for continuous data, the optimal ζ can be given by Eq. (E.48). In this subsection, some guidelines for selecting ζ for grouped data will be obtained.

Note that Eq. (E.48) depends only on the sample size, n, and the standard deviation, σ, which for a given sample can be estimated. When estimating σ, there will be small differences whether proceeding with continuous or grouped data; nevertheless, although not exactly equal, results from (E.48) will be quite similar in any case.

Recall that ζ should be the one that better allows to estimate the curvature $A(f'')$. Then, the idea is to simulate how the optimal ζ $\left(\zeta_{opt_g}\right)$behaves for different grouping conditions (i.e., for different average lentghs) with respect to ζ obtained for continuous data (ζ_{opt}) via Eq. (E.48). Intuitively, it can be thought that as grouping becomes heavier, ζ needs to be larger in order to capture more information from the surroundings.

For performing the simulation, it was used the same normal mixture as in Section 3.4, namely, a normal mixture $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}$, where $\phi_{\mu,\sigma}$ is a $N\left(\mu,\sigma^2\right)$ density, $\alpha = (0.70, 0.22, 0.06, 0.02)$, $\mu = (207, 237, 277, 427)$ and $\sigma = (25, 20, 35, 50)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively. Also, the same three different sample sizes were used: 60, 240 and 960.

For each sample size, the simulation went as follows:

1. From f, simulate $B_0 = 200$ samples of size n and compute the average ζ_{opt}, $\bar{\zeta}_{opt}$.

2. Consider a grid of average lengths $\left(\bar{l}_1, \bar{l}_2, \bar{l}_3 ...\right)$. For $\bar{l}_i$, simulate a sample of size n and divide the data range into intervals such that the average length is $\bar{l}_i$. Then, consider the midpoints t_i as many times as n_i, the number of data at each interval.

3. For each of a grid of values $(\zeta_1, \zeta_2, \zeta_3, ...)$, estimate the curvature $A(f'')$ based on the grouped sample using the kernel density estimator (3.2), and keep the value of ζ that minimizes $\left[\hat{A}_\zeta(f'') - A(f'')\right]^2$. This is the value that was called ζ_{opt_g}.

4. Compute the ratio $\zeta_r = \frac{\zeta_{opt_g}}{\zeta_{opt}}$.

5. Follow Steps 2 to 4 for each $\bar{l}_i$, $i = 1, 2, 3, ...$

In Step 3, the curvature $A(f'')$ was estimated by means of the R function `dkde`, available in the package `kedd` (Guidoum, 2014). Given a sample, a kernel function and a bandwidth, `dkde` computes the r-th derivative of the kernel density estimator over a grid of values.

Considering $r = 2$, the grouped sample obtained in step 2 and the default Gaussian kernel, a grid of bandwidths $(\zeta_1, \zeta_2, \zeta_3, ...)$ was used, one at a time. Then, the estimates of the 2-nd derivative were squared, and the integral over the domain was approximated via Monte Carlo.

E.1 shows three different patterns of ζ_r versus ω, one for each sample size. Patterns for sample sizes 240 and 960 are quite similar. Roughly speaking, in these cases the message is that for $\omega \leqslant 0.075$, since ζ_r is on average 1, the bandwidth ζ_{opt_g} should be taken the same as ζ_{opt} for continuous data, while for $\omega > 0.075$, ζ_{opt_g} should be taken accordingly to the relationship given by the line with positive slope, which is practically the same in both cases. For sample size 60, the plot suggest that for $\omega \leqslant 0.10$, ζ_{opt_g} should be taken as around $0.8\zeta_{opt}$, while for $\omega > 0.10$, ζ_{opt_g}should be taken following the line with positive

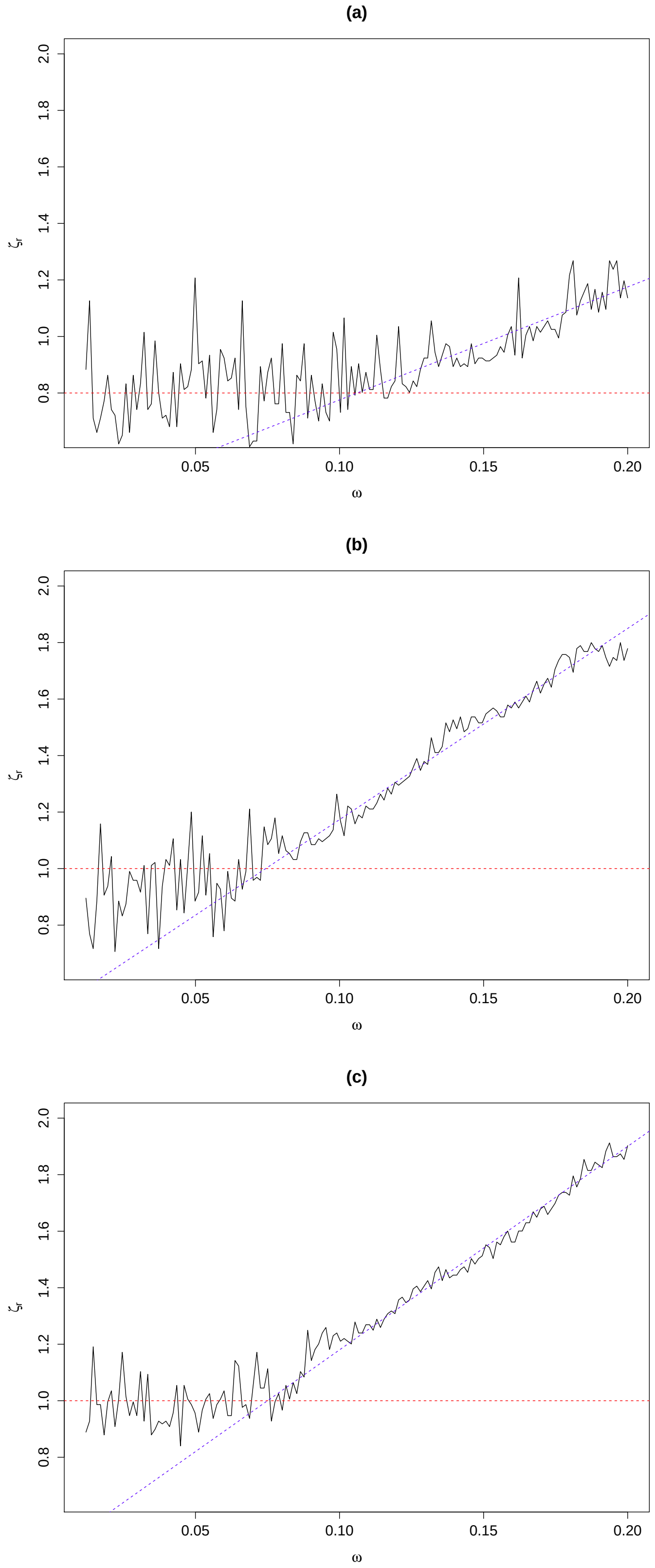

Figure E.1: Plots showing ζ_r versus ω for sample sizes (a) 60, (b) 240, (c) 960.

slope. These results confirm the preliminary idea: for large ω (i.e., heavy grouping), ζ_{opt_g} should be somewhat larger than ζ_{opt}.

The plots also suggest that the pattern between ζ_r and ω may change at some value between sample size 60 and 240. From this, practical guidelines for selecting ζ_{opt_g} can be stated as follows: for sample sizes 150 or below, choose $\zeta_{opt_g} \approx 0.8\zeta_{opt}$ whenever $\omega \leqslant 0.10$, and $\zeta_{opt_g} \approx \zeta_{opt}\,(4\omega + 0.4)$ otherwise. For sample sizes over 150, select $\zeta_{opt_g} \approx \zeta_{opt}$ for $\omega \leqslant 0.075$, and $\zeta_{opt_g} \approx \zeta_{opt}\,(7\omega + 0.5)$ for $\omega > 0.075$.

Appendix F

Proof of Theorem 5.1

Proof. Applying the expectation operator to (5.1),

$$\begin{aligned}\mathbb{E}\left[\hat{F}_h^g\left(x\right)\right] &= \mathbb{E}\left[\sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x-t_i}{h}\right)\right] \\ &= \sum_{i=1}^{k} \mathbb{E}\left[w_i \mathbb{K}\left(\frac{x-t_i}{h}\right)\right] \\ &= \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) \mathbb{E}\left[w_i\right] \\ &= \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) p_i, \qquad \text{(F.1)}\end{aligned}$$

where $p_i = F\left(y_i\right) - F\left(y_{i-1}\right)$.

Using a Taylor expansion of p_i around t_i, as in (C.10), by the parity conditions (C.12), substituting into (F.1) gives

$$\mathbb{E}\left[\hat{F}_h^g\left(x\right)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) \left\{F'\left(t_i\right) l_i + F'''\left(t_i\right) \frac{l_i^3}{24} + \frac{1}{4!}\mathfrak{o}_3\right\},$$

where $\mathfrak{o}_3 = F^{(4)}\left(\xi_i\right)\left(\frac{l_i}{2}\right)^4 - F^{(4)}\left(\xi_{i-1}\right)\left(-\frac{l_i}{2}\right)^4$. Then,

$$\begin{aligned}\mathbb{E}\left[\hat{F}_h^g\left(x\right)\right] \quad = \quad & \sum_{i=1}^{k} l_i F'\left(t_i\right) \mathbb{K}\left(\frac{x-t_i}{h}\right) + \frac{1}{24}\sum_{i=1}^{k} l_i^3 F'''\left(t_i\right) \mathbb{K}\left(\frac{x-t_i}{h}\right) + \\ & \frac{1}{4!}\sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) \mathfrak{o}_3. \qquad \text{(F.2)}\end{aligned}$$

Due to eq. (C.13), it follows that

$$\begin{aligned}\left|\sum_{i=1}^{k}\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathfrak{o}_3\right| &\leqslant \sum_{i=1}^{k}\left|\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathfrak{o}_3\right| \\ &\leqslant \|\mathbb{K}\|_\infty\sum_{i=1}^{k}|\mathfrak{o}_3| \\ &\leqslant \|\mathbb{K}\|_\infty kO\left(\bar{l}^5\right) \\ &= O\left(\bar{l}^4\right),\end{aligned}$$

so, defining $H_1(t)=F'(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$ and $H_2(t)=F'''(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$,

$$\mathbb{E}\left[\hat{F}_h^g(x)\right]=\sum_{i=1}^{k}l_iH_1(t_i)+\frac{1}{24}\sum_{i=1}^{k}l_i^3H_2(t_i)+O\left(\bar{l}^4\right). \tag{F.3}$$

Considering the first term on the right hand side of (F.3), taking the integral over the i-th interval, and using a Taylor expansion, gives

$$\begin{aligned}\int_{y_{i-1}}^{y_i}H_1(t)\,dt &= \int_{y_{i-1}}^{y_i}\left[H_1(t_i)+H_1'(t_i)(t-t_i)+\frac{1}{2}H_1''(t_i)(t-t_i)^2+\right. \\ &\left.\frac{1}{3!}H_1'''(t_i)(t-t_i)^3+\frac{1}{4!}H_1^{(4)}(\xi_i)(t-t_i)^4\right]dt.\end{aligned}$$

Using the change of variable $s=t-t_i$ and the parity conditions (C.12),

$$\int_{y_{i-1}}^{y_i}H_1(t)\,dt=l_iH_1(t_i)+\frac{1}{24}l_i^3H_1''(t_i)+\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}H_1^{(4)}(\xi_i)(t-t_i)^4\,dt,$$

and summing all over the k intervals and reordering,

$$\sum_{i=1}^{k}l_iH_1(t_i)=\int H_1(t)\,dt-\frac{1}{24}\sum_{i=1}^{k}l_i^3H_1''(t_i)-\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}H_1^{(4)}(\xi_i)(t-t_i)^4\,dt. \tag{F.4}$$

Bounding the third term on the right hand side of (F.4),

$$\begin{aligned}\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(4)}\left(\xi_i\right)\left(t-t_i\right)^4 dt\right| &\leqslant \frac{1}{4!}\left\|H_1^{(4)}\right\|_\infty \sum_{i=1}^{k}\left[\frac{s^5}{5}\right]_{-\frac{l_i}{2}}^{\frac{l_i}{2}} \\ &\leqslant \frac{1}{4!80}\left\|H_1^{(4)}\right\|_\infty \sum_{i=1}^{k} l_i^5 \\ &\leqslant k l_{max}^5 \frac{1}{4!80}\left\|H_1^{(4)}\right\|_\infty \qquad \text{(F.5)} \\ &= O\left(\frac{\bar{l}^4}{h^4}\right), \qquad \text{(F.6)}\end{aligned}$$

where it was used the result (C.2) and the fact that each derivative of H_1 takes a $1/h$ out of the expression.

Now, working on the second term on the right hand side of (F.4), it can be open out as

$$\sum_{i=1}^{k} l_i^3 H_1''\left(t_i\right) = \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i H_1''\left(t_i\right) + \overline{l^2}\sum_{i=1}^{k} l_i H_1''\left(t_i\right). \qquad \text{(F.7)}$$

The second term on the right hand side of (F.7) can be expressed as

$$\overline{l^2}\sum_{i=1}^{k} l_i H_1''\left(t_i\right) = \overline{l^2}\int H_1''(t)\,dt - \frac{\overline{l^2}}{24}\sum_{i=1}^{k} l_i^3 H_1^{(4)}\left(t_i\right) - \frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}\left(\xi_i\right)\left(t-t_i\right)^4 dt.$$

Bounding $\frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}\left(\xi_i\right)\left(t-t_i\right)^4 dt$,

$$\begin{aligned}\left|\frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}\left(\xi_i\right)\left(t-t_i\right)^4 dt\right| &\leqslant \frac{\overline{l^2}}{4!}\left\|H_1^{(6)}\right\|_\infty \sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}\left(t-t_i\right)^4 dt \\ &\leqslant \frac{\overline{l^2}}{4!80}\left\|H_1^{(6)}\right\|_\infty \sum_{i=1}^{k} l_i^5 \\ &\leqslant \frac{\overline{l^2}}{4!80}\left\|H_1^{(6)}\right\|_\infty k l_{max}^5 \\ &= O\left(\frac{\bar{l}^6}{h^6}\right),\end{aligned}$$

where results (C.4) and (C.2) were used. Using similar arguments,

$$\left|\sum_{i=1}^{k} l_i^3 H_1^{(4)}\left(t_i\right)\right| \leqslant O\left(\frac{\bar{l}^4}{h^4}\right).$$

In turn, bounding $\overline{l^2}\int H_1''(t)\,dt$ results in

$$\begin{aligned}\left|\overline{l^2}\int H_1''(t)\,dt\right| &\leqslant O\left(\bar{l}^2\right)\left|\int H_1''(t)\,dt\right| \\ &\leqslant O\left(\bar{l}^2\right)\int\left|H_1''(t)\right|dt \qquad \text{(F.8)}\end{aligned}$$

Let us consider an explicit expression for H_1'', which is

$$H_1''(t)=\mathbb{K}\left(\frac{x-t}{h}\right)F'''(t)-\frac{1}{h}2F''(t)\,K\left(\frac{x-t}{h}\right)+\frac{1}{h^2}F'(t)\,K'\left(\frac{x-t}{h}\right).$$

By Assumption 5.1, when $\frac{x-t}{h}<-1$ then

$$\mathbb{K}\left(\frac{x-t}{h}\right)=K'\left(\frac{x-t}{h}\right)=K\left(\frac{x-t}{h}\right)=0,$$

so in this case,

$$H_1''(t)=0. \qquad \text{(F.9)}$$

When $\frac{x-t}{h}>1$, then

$$K'\left(\frac{x-t}{h}\right)=K\left(\frac{x-t}{h}\right)=0$$

and

$$\mathbb{K}\left(\frac{x-t}{h}\right)=1,$$

so in this case,

$$H_1''(t)=F'''(t). \qquad \text{(F.10)}$$

As a consequence,

$$\begin{aligned}\int_{-\infty}^{\infty}\left|H_1''(t)\right|dt &= \int_{-\infty}^{x-h}\left|H_1''(t)\right|dt+\int_{x-h}^{x+h}\left|H_1''(t)\right|dt \\ &= \int_{-\infty}^{x-h}\left|F''''(t)\right|dt+ \\ &\quad \int_{x-h}^{x+h}\left|\mathbb{K}\left(\frac{x-t}{h}\right)F'''(t)-\frac{1}{h}2F''(t)\,K\left(\frac{x-t}{h}\right)+\right. \\ &\quad \left.\frac{1}{h^2}F'(t)\,K'\left(\frac{x-t}{h}\right)\right|dt,\end{aligned}$$

hence,

$$\begin{aligned}\int_{-\infty}^{\infty}\left|H_1''(t)\right|dt &\leqslant \int_{-\infty}^{x+h}\left|F''''(t)\right|dt + 2\left\|F''\right\|_\infty \int_{-1}^{1} K(u)\,du + \frac{2}{h}\left\|F'\right\|_\infty \left\|K'\right\|_\infty \\ &\leqslant \int_{-\infty}^{\infty}\left|F'''(t)\right|dt + 2\left\|F''\right\|_\infty + \frac{2}{h}\left\|F'\right\|_\infty \left\|K'\right\|_\infty \\ &= O\left(\frac{1}{h}\right).\end{aligned}$$

Thus, from eq. (F.8),

$$\begin{aligned}\left|\overline{l^2}\int H_1''(t)\,dt\right| &\leqslant O\left(\bar{l}^2\right)O\left(\frac{1}{h}\right) \\ &= O\left(\frac{\bar{l}^2}{h}\right).\end{aligned}$$

Updating (F.7), gives

$$\sum_{i=1}^{k} l_i^3 H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) + O\left(\frac{\bar{l}^2}{h}\right). \tag{F.11}$$

For bounding the first term on the right hand side of (F.11), realize that by previous elaborations,

$$\begin{aligned}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) \quad = \quad & \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1''(t)\,dt - \frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i) - \\ & \frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)(t - t_i)^4\,dt.\end{aligned} \tag{F.12}$$

Using result (C.7), the last two terms can be bounded as

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i)\right| &\leqslant \max_i\left|l_i^2 - \overline{l^2}\right| k l_{max}^3 \left\|H_1^{(4)}\right\|_\infty \\ &= o\left(\frac{\bar{l}^4}{h^4}\right)\end{aligned} \tag{F.13}$$

and

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1^{(6)}\left(\xi_i\right)\left(t-t_i\right)^4dt\right|\leqslant\frac{1}{4!80}\max_i\left|l_i^2-\overline{l^2}\right|kl_{max}^5\left\|H_1^{(6)}\right\|_\infty$$
$$=o\left(\frac{\bar{l}^6}{h^6}\right). \qquad (F.14)$$

For bounding $\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''\left(t\right)dt$, note that

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''\left(t\right)dt\right|&\leqslant\max_i\left|l_i^2-\overline{l^2}\right|\sum_{i=1}^{k}\left|\int_{y_{i-1}}^{y_i}H_1''\left(t\right)dt\right|\\&\leqslant o\left(\bar{l}^2\right)\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}\left|H_1''\left(t\right)\right|dt\\&\leqslant o\left(\bar{l}^2\right)\int\left|H_1''\left(t\right)\right|dt,\end{aligned} \qquad (F.15)$$

which is a similar expression to (F.8). Using the same arguments as before, Eq. (F.15) becomes

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''\left(t\right)dt\right|=o\left(\frac{\bar{l}^2}{h}\right). \qquad (F.16)$$

Considering equations (F.12), (F.13), (F.14), (F.15) and (F.16),

$$\sum_{i=1}^{k}l_i^3H_1''\left(t\right)=O\left(\frac{\bar{l}^2}{h}\right). \qquad (F.17)$$

By eqs. (F.17) and (F.5), Eq. (F.4) is

$$\sum_{i=1}^{k}l_iH_1\left(t_i\right)=\int H_1\left(t\right)dt+O\left(\frac{\bar{l}^2}{h}\right). \qquad (F.18)$$

Integrating by parts and a change of variable lead to

$$\int H_1\left(t\right)dt=\int F\left(x-hu\right)K\left(u\right)du.$$

Using a Taylor expansion on F and by kernel properties,

$$\int H_1\left(t\right)dt=F\left(x\right)+\frac{h^2}{2}F''\left(x\right)\mu_2\left(K\right)+O\left(h^4\right);$$

i.e., from (F.18),

$$\sum_{i=1}^{k} l_i H_1(t_i) = F(x) + \frac{h^2}{2} F''(x) \mu_2(K) + O\left(h^4\right) + O\left(\frac{\bar{l}^2}{h}\right). \tag{F.19}$$

Regarding the second term on the right hand side of (F.3),

$$\sum_{i=1}^{k} l_i^3 H_2(t_i) = \sum_{i=1}^{k} \left(l_i^2 - \overline{l^2}\right) l_i H_2(t_i) + \overline{l^2} \sum_{i=1}^{k} l_i H_2(t_i). \tag{F.20}$$

The first term on the right hand side of (F.20) can be easily bounded usign (C.2) and (C.7) as

$$\begin{aligned} \left|\sum_{i=1}^{k} \left(l_i^2 - \overline{l^2}\right) l_i H_2(t_i)\right| &\leqslant \sum_{i=1}^{k} \left|\left(l_i^2 - \overline{l^2}\right) l_i H_2(t_i)\right| \\ &\leqslant \max_i \left|l_i^2 - \overline{l^2}\right| l_{max} k \left\|H_2\right\|_\infty \\ &= o\left(\bar{l}^2\right), \end{aligned} \tag{F.21}$$

since $\left\|H_2\right\|_\infty = O(1)$. As to the second term, note that by Ostrowski's inequality (C.19),

$$\left| l_i H_2(t_i) - \int_{y_{i-1}}^{y_i} H_2(t)\, dt \right| \leqslant \frac{1}{4} \mathfrak{L}_{H_2} l_i^2,$$

which summing up all over the k intervals and considering (C.4) and the fact that $\mathfrak{L}_{H_2} = \left\|H_2'\right\|_\infty$, lead to

$$\begin{aligned} \sum_{i=1}^{k} \left| l_i H_2(t_i) - \int_{y_{i-1}}^{y_i} H_2(t)\, dt \right| &\leqslant \frac{1}{4} \mathfrak{L}_{H_2} \frac{k}{k} \sum_{i=1}^{k} l_i^2 \\ &= \frac{1}{4} \mathfrak{L}_{H_2} k \overline{l^2} \\ &= O\left(\frac{\bar{l}}{h}\right), \end{aligned}$$

which in turn, by (C.4), implies that

$$\begin{aligned} \overline{l^2} \sum_{i=1}^{k} l_i H_2(t_i) &= \left[\bar{l}^2 + o\left(\bar{l}^2\right)\right] \left[\int H_2(t)\, dt + O\left(\frac{\bar{l}}{h}\right)\right] \\ &= \bar{l}^2 \int H_2(t)\, dt + o\left(\bar{l}^2\right). \end{aligned}$$

Integrating by parts and a change of variable lead to

$$\int H_2(t)\, dt = \int F''(x - hu) K(u)\, du.$$

Now, by a Taylor expansion on F'' and simplifying due to the kernel K properties,

$$\int H_2(t)\, dt = F''(x) + \frac{h^2}{2} F^{(4)}(x) \mu_2(K) + O\left(h^3\right),$$

so that

$$\overline{l^2} \sum_{i=1}^{k} l_i H_2(t_i) = \bar{l}^2 \left[F''(x) + \frac{h^2}{2} F^{(4)}(x) \mu_2(K) + O\left(h^3\right) \right] + o\left(\bar{l}^2\right). \tag{F.22}$$

Using (F.22) and (F.21), Eq. (F.20) is

$$\sum_{i=1}^{k} l_i^3 H_2(t_i) = \bar{l}^2 \left[F''(x) + \frac{h^2}{2} F^{(4)}(x) \mu_2(K) + O\left(h^3\right) \right] + o\left(\bar{l}^2\right). \tag{F.23}$$

So, joining (F.23) and (F.19) into (F.3),

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = F(x) + \frac{h^2}{2} F''(x) \mu_2(K) + O\left(h^4\right) + O\left(\frac{\bar{l}^2}{h}\right) + \left[\frac{\bar{l}^2}{24} F''(x) + o\left(\bar{l}^2\right)\right],$$

or simply, using Assumption 5.4,

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = F(x) + \frac{h^2}{2} F''(x) \mu_2(K) + o\left(h^2\right),$$

from which, the bias is

$$\mathbb{B}\left[\hat{F}_h^g(x)\right] = \frac{1}{2} h^2 F''(x) \mu_2(K) + o\left(h^2\right). \tag{F.24}$$

Regarding the variance, applying this operator to (5.1) gives

$$\begin{aligned}
\mathbb{V}\left[\hat{F}_h^g(x)\right] &= \mathbb{V}\left[\sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x - t_i}{h}\right)\right] \\
&= \sum_{i=1}^{k} \mathbb{V}\left[w_i \mathbb{K}\left(\frac{x - t_i}{h}\right)\right] + 2 \sum_{i<j} \mathbb{C}\left[w_i \mathbb{K}\left(\frac{x - t_i}{h}\right), w_j \mathbb{K}\left(\frac{x - t_j}{h}\right)\right] \\
&= \sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x - t_i}{h}\right) \mathbb{V}(w_i) + 2 \sum_{i<j} \mathbb{K}\left(\frac{x - t_i}{h}\right) \mathbb{K}\left(\frac{x - t_j}{h}\right) \mathbb{C}(w_i, w_j).
\end{aligned}$$

Considering that $(n_1, n_2, \ldots, n_k)$ is multinomial random vector, and since $w_i = n_i/n$, the last equation can be rewritten as

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) - \frac{2}{n}\sum_{i<j}\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)p_ip_j. \quad \text{(F.25)}$$

Since $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i, as in (C.10) and by parity conditions (C.12), the first term on the right hand side of (F.25) (except a factor $1/n$) can be written as

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) = \sum_{i=1}^{k}l_iH_3(t_i) + O\left(\bar{l}\right), \quad \text{(F.26)}$$

where $H_3(t) = \mathbb{K}^2\left(\frac{x-t}{h}\right)F'(t)$. Integrating H_3 over the i-th interval and using a Taylor expansion, gives

$$\begin{aligned}\int_{y_{i-1}}^{y_i} H_3(t)\,dt &= \int_{y_{i-1}}^{y_i}\left[H_3(t_i) + (t-t_i)H_3'(t_i) + \frac{1}{2}H_3''(t_i)(t-t_i)^2 + \right. \\ &\quad \left. \frac{1}{3!}H_3'''(t_i)(t-t_i)^3 + \frac{1}{4!}H^{(4)}(\xi_i)(t-t_i)^4\right]dt.\end{aligned}$$

Taking the variable $s = t - t_i$ and by parity condtions (C.12),

$$\int_{y_{i-1}}^{y_i} H_3(t)\,dt = l_iH_3(t_i) + \frac{1}{24}l_i^3H_3''(t_i) + \frac{1}{4!}\int_{y_{i-1}}^{y_i}H_3^{(4)}(\xi_i)(t-t_i)^4\,dt,$$

and summing all over the k intervals and reordering,

$$\sum_{i=1}^{k}l_iH_3(t_i) = \int H_3(t)\,dt - \frac{1}{24}\sum_{i=1}^{k}l_i^3H_3''(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}H_3^{(4)}(\xi_i)(t-t_i)^4\,dt. \quad \text{(F.27)}$$

Following parallel steps after (F.4), it is easy to see that

$$\begin{aligned}\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}H_3^{(4)}(\xi_i)(t-t_i)^4\,dt\right| &\leqslant \frac{1}{4!}\left\|H_3^{(4)}\right\|_\infty\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}(t-t_i)^4\,dt \\ &\leqslant \frac{1}{4!80}\left\|H_3^{(4)}\right\|_\infty\sum_{i=1}^{k}l_i^5 \\ &= O\left(\frac{\bar{l}^4}{h^4}\right) \quad \text{(F.28)}\end{aligned}$$

and

$$\sum_{i=1}^{k} l_i^3 H_3''(t_i) = O\left(\frac{\bar{l}^2}{h}\right). \tag{F.29}$$

Considering (F.29), (F.28) and (F.27), Eq. (F.26) transforms into

$$\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) p_i\left(1-p_i\right) = \int H_3(t)\,dt + O\left(\bar{l}\right). \tag{F.30}$$

As before, using integration by parts and the change of variable $u = (x-t)/h$,

$$\begin{aligned}\int H_3(t)\,dt &= 2\int F(x-hu)\,K(u)\,\mathbb{K}(u)\,du\\ &= 2\int \mathbb{K}(u)\,K(u)\left[F(x) - huF'(x) + \frac{1}{2}h^2u^2F''(\xi)\right]du\\ &= 2\int \mathbb{K}(u)\,K(u)\,F(x)\,du - 2h\int \mathbb{K}(u)\,K(u)\,F'(x)\,udu + O\left(h^2\right),\end{aligned}$$

where ξ is a value between x and $x-hu$.

Note that $\frac{d}{du}\mathbb{K}^2(u) = 2\mathbb{K}(u)\,\mathbb{K}'(u)$ and that $\mathbb{K}'(u) = K(u)$, so $\frac{d}{du}\mathbb{K}^2(u) = 2\mathbb{K}(u)\,K(u)$. Then,

$$\int H_3(t)\,dt = F(x) - hF'(x)\,C_0 + O\left(h^2\right),$$

where $C_0 = 2\int \mathbb{K}(u)\,K(u)\,udu$. Substituting the last expression into (F.30) gives

$$\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) p_i\left(1-p_i\right) = F(x) - hF'(x)\,C_0 + O\left(h^2\right), \tag{F.31}$$

since by Assumption 5.4, $\bar{l} = o\left(h^2\right)$.

Let us turn back to eq. (F.25). Because $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i, as in (C.10) and by parity conditions (C.12), the second term on the right hand side of (F.25) (except a factor $-2/n$) can be written as

$$\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right) p_i p_j = \sum_{i<j} H_4(t_i,t_j)\,l_i l_j + O\left(\bar{l}^2\right), \tag{F.32}$$

where $H_4(z_1,z_2) = \mathbb{K}\left(\frac{x-z_1}{h}\right)\mathbb{K}\left(\frac{x-z_2}{h}\right)F'(z_1)\,F'(z_2)$.

Considering the second order Taylor expansion around (t_i,t_j) and by parity conditions (C.12),

$$\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 \;=\; H_4(t_i,t_j)\,l_i l_j + \frac{\mathfrak{T}_0}{2}, \tag{F.33}$$

where

$$\begin{aligned} \mathfrak{T}_0 &= \int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}\left[\frac{\partial^2 H_4}{\partial z_1^2}(\xi_1,\xi_2)(z_1-t_i)^2 + 2\frac{\partial^2 H_4}{\partial z_1\partial z_2}(\xi_1,\xi_2)(z_1-t_i)(z_2-t_j)\right. \\ &\left.+\frac{\partial^2 H_4}{\partial z_2^2}(\xi_1,\xi_2)(z_2-t_j)^2\right]dz_2dz_1 \end{aligned}$$

Summing all over the $k(k-1)/2$ enclosures in (F.33) and reordering,

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) = \sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 - \frac{1}{2}\sum_{i<j}\mathfrak{T}_0. \qquad \text{(F.34)}$$

The second term on the right hand side of (F.34), can be easily bounded:

$$\begin{aligned} \left|\frac{1}{2}\sum_{i<j}\mathfrak{T}_0\right| &\leqslant \frac{1}{2}\sum_{i<j}\left(\frac{l_i^3 l_j}{12}\left\|\frac{\partial^2 H_4}{\partial z_1^2}\right\|_\infty + 2\frac{l_i^2 l_j^2}{16}\left\|\frac{\partial^2 H_4}{\partial z_1\partial z_2}\right\|_\infty + \frac{l_i l_j^3}{12}\left\|\frac{\partial^2 H_4}{\partial z_2^2}\right\|_\infty\right) \\ &= O\left(\frac{k^2 l_{max}^4}{h^2}\right) \\ &= O\left(\frac{\bar{l}^2}{h^2}\right). \end{aligned}$$

As a consequence,

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) = \sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 + O\left(\frac{\bar{l}^2}{h^2}\right). \qquad \text{(F.35)}$$

On the other hand,

$$\begin{aligned} \sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 &= \frac{1}{2}\sum_{i\neq j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 \\ &= \frac{1}{2}\sum_{i,j=1}^{k}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 \\ &\quad -\frac{1}{2}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i}\int_{y_{i-1}}^{y_i} H_4(z_1,z_2)\,dz_2dz_1 \\ &= \frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2dz_1 + O(\bar{l}). \qquad \text{(F.36)} \end{aligned}$$

Now, using (F.36) and (F.35),

$$\sum_{i<j} l_i l_j H_4\left(t_i, t_j\right) = \frac{1}{2}\int\int H_4\left(z_1, z_2\right) dz_2 dz_1 + O\left(\bar{l}\right) + O\left(\frac{\bar{l}^2}{h^2}\right). \qquad \text{(F.37)}$$

Integration by parts and two changes of variable $[u_1 = \left(x - z_1\right)/h$, $u_2 = \left(x - z_2\right)/h]$ lead to

$$\frac{1}{2}\int\int H_4\left(z_1, z_2\right) dz_2 dz_1 = \frac{1}{2}\left[\int F\left(x - hu\right) K\left(u\right) du\right]^2.$$

A Taylor expansion around x gives,

$$\begin{aligned}\frac{1}{2}\int\int H_4\left(z_1, z_2\right) dz_2 dz_1 &= \frac{1}{2}\left[F\left(x\right) + \frac{1}{2}h^2 F''\left(x\right)\mu_2\left(K\right) + O\left(h^4\right)\right]^2 \\ &= \frac{1}{2}\left[F^2\left(x\right) + O\left(h^2\right)\right], \qquad \text{(F.38)}\end{aligned}$$

so that, considering (F.38), (F.37) and Assumption 5.4, Eq. (F.32) becomes

$$\sum_{i<j} \mathbb{K}\left(\frac{x - t_i}{h}\right)\mathbb{K}\left(\frac{x - t_j}{h}\right) p_i p_j = \frac{1}{2}F^2\left(x\right) + O\left(h^2\right). \qquad \text{(F.39)}$$

Now, putting back (F.39) and (F.31) in (F.25) and simplifying,

$$\mathbb{V}\left[\hat{F}_h^g\left(x\right)\right] = \frac{1}{n}F\left(x\right)\left[1 - F\left(x\right)\right] - \frac{h}{n}F'\left(x\right)C_0 + O\left(\frac{h^2}{n}\right). \qquad \text{(F.40)}$$

Joining (F.40) and (F.24), it is obtained

$$\begin{aligned}MSE\left[\hat{F}_h^g\left(x\right)\right] \quad = \quad & \frac{1}{4}h^4 F''\left(x\right)^2 \mu_2\left(K\right)^2 + \frac{1}{n}F\left(x\right)\left[1 - F\left(x\right)\right] - \frac{h}{n}F'\left(x\right)C_0 \\ & + O\left(\frac{h^2}{n}\right) + o\left(h^4\right). \qquad \text{(F.41)}\end{aligned}$$

Finally, dealing with the integrated versions of the terms coming up in the proof of (F.41), it can be obtained the following asymptotic expression for $AMISE$,

$$AMISE\left[\hat{F}_h^g\right] = \frac{1}{4}h^4 \mu_2\left(K\right)^2 A\left(f'\right) + \frac{1}{n}\int F\left(x\right)\left[1 - F\left(x\right)\right] dx - \frac{h}{n}C_0,$$

which corresponds with just integrating the leading terms in (F.41).

□

Appendix G

Resumen extenso

Esta tesis surge a partir de un problema planteado por investigadores del Instituto de Agricultura Sostenible (CSIC), en Córdoba, España, al equipo de Modelización, Optimización e Inferencia Estadística (MODES) de la Universidade da Coruña (UDC). El problema en cuestión era sobre cómo predecir con cierta precisión la emergencia de malas hierbas, con el problema añadido de que debido a las condiciones experimentales, los datos disponibles se obtenían de manera agrupada.

Las malas hierbas se caracterizan por ser persistentes, muy competitivas y por disminuir el rendimiento de los cultivos, lo que tiene consecuencias negativas tanto en lo económico como en lo social. Resulta evidente, por una parte, la importancia de combatir de manera eficiente a las malas hierbas mediante programas de erradicación apropiados. Por otra parte, la eficiencia de estos programas depende en gran medida de una buena capacidad de predicción de la emergencia de este tipo de plantas.

El problema de modelizar y predecir la emergencia de malas hierbas no es nuevo, y en el ámbito de la malherbología se ha abordado mediante el ajuste de modelos paramétricos no lineales de regresión, tales como los modelos de regresión Logística, de Gompertz o de Weibull. El ajuste se hace considerando como variable dependiente el número acumulado de plantas emergidas, y como variable independiente, el tiempo hidrotermal acumulado. Sin embargo, este enfoque tiene algunos inconvenientes. Uno de ellos es que los métodos paramétricos no son lo suficientemente flexibles para capturar ciertas características complejas que pueden aparecer en los valores observados de emergencia acumulada, tales como saltos abruptos o la presencia de colas pesadas. Por otra parte, los valores de la emergencia acumulada obtenida monitorizando tiempos hidrotermales acumulados consecutivos, no son independientes estadísticamente, con lo cual se ve afectada la obtención de intervalos de confianza y tests de hipótesis que requieren que los residuos no estén correlacionados. Además, aunque la experiencia puede ayudar a elegir el tipo de modelo adecuado, si el modelo elegido no describiera adecuadamente la emergencia acumulada, existe el peligro de que el análisis proporcione conclusiones erróneas. No obstante, en la literatura científica de la malherbología no se suele considerar lo anterior, siendo el ajuste del modelo de regresión el principal objetivo.

La propuesta alternativa a los modelos paramétricos es la aproximación no paramétrica, que no exige modelo alguno a las variables aleatorias en consideración, dejando que sean los datos los que hablen "por sí mismos". Así, los métodos no paramétricos proveerían mayor flexibilidad en la descripción y la estimación de la emergencia de las malas hierbas.

Además de la aproximación no paramétrica, el problema se puede abordar desde el punto de vista de la búsqueda de estructura en los datos, ya sea mediante la estimación de la función de densidad o de distribución. En efecto, si no hubiera limitaciones debidas a la monitorización y se pudiera observar el tiempo hidrotermal acumulado al momento de la emergencia de cada planta, lo lógico sería plantear el problema en términos de sólo una variable aleatoria (tiempo hidrotermal acumulado a la emergencia), con la cual se podría estimar la función de densidad, o bien, dado que la emergencia acumulada es una función creciente cuyos valores se encuentran entre 0 y 1, también es lógico pensar en ella como un problema de estimación de la distribución.

Una herramienta no paramétrica clásica para estimar la estructura de un conjunto de datos es la suavización tipo núcleo, ya sea para estimar la densidad o la distribución. Sin embargo, tal como se mencionó al principio, los datos obtenidos por los malherbólogos complica su utilización, pues sólo se conoce el número de emergencias entre dos observaciones consecutivas del tiempo hidrotermal acumulado. Así, para usar los estimadores núcleo de la densidad o de la distribución es necesario modificarlos de alguna manera para obtener estimaciones de dichas funciones a partir de datos agrupados.

Esta tesis comienza con una revisión de los principales conceptos usados en malherbología, así como en la estimación no paramétrica de la densidad y la distribución. Se presentan el histograma y la función de distribución empírica como estimadores que aunque tienen ciertas buenas propiedades, sus estimaciones tienen la desventaja de ser escalonadas.

Ante la relativa aspereza de las estimaciones del histograma y de la función de distribución empírica, la estimación tipo núcleo resulta idónea para obtener esa suavidad extra que siempre es deseable en las estimaciones de las funciones de densidad y distribución. Desde el punto de vista matemático, la estimación tipo núcleo es relativamente sencilla, lo que permite estudiar sus propiedades estadísticas con cierta facilidad. Por tal razón, esta tesis incluye un capítulo con los principales resultados sobre la estimación tipo núcleo de la densidad y de la distribución, resaltando sus propiedades asintóticas y, dada su importancia en el desempeño de los estimadores, se incluye también una revisión de los principales selectores de ventana usados tradicionalmente.

La primera contribución de esta investigación comienza con una propuesta para modificar el bien conocido estimador tipo núcleo de la densidad, $\hat{f}_h$, (ecuación 2.1) para poder usarlo con datos agrupados, dando como resultado el estimador $\hat{f}^g_h$, (capítulo 3, ecuación 3.2). Puesto que los datos están contenidos (agrupados) en intervalos, se propone considerar a los puntos medios de los intervalos y repetirlos tantas veces como datos haya al interior de éstos. En esencia, esta propuesta es una forma de desagregar los datos, y tiene la ventaja de ser sencilla, fácil de implementar y hasta intuitiva. En efecto, ante la falta de información adicional, parece razonable elegir el punto medio del intervalo para representar la localización de un dato al interior de éste.

Un elemento clave para el buen desempeño del estimador núcleo de la densidad es elegir adecuadamente el parámetro de suavización o ventana. Cuando los datos disponibles son completos (no agrupados), los desarrollos clásicos permiten obtener una expresión para la ventana óptima desde el punto de vista del error cuadrático medio integrado asintótico ($AMISE$), es decir, una ventana $AMISE$ óptima, h_{AMISE} (ecuación (2.12)). Ya que la estructura del estimador modificado para datos agrupados es básicamente la misma, un primer impulso sería considerar esa misma expresión de la ventana. Sin embargo, al considerar los puntos medios de los intervalos en la expresión del estimador modificado, se introduce un error en las mediciones, un error que puede llegar a ser notablemente más grande que el error experimental que, en principio, cabría esperar en cualquier medición no agrupada. Por lo anterior, es pertinente preguntarse en qué medida sería correcto considerar la ventana h_{AMISE} en el caso del estimador núcleo modificado.

Una reflexión sobre el tema permite concluir que lo anterior dependerá del grado de agrupación de los datos. Es decir, que cuanto más pequeños sean los intervalos, más cerca se está del caso de datos completos y, por lo tanto, el error que se pueda cometer al utilizar la ventana h_{AMISE} para datos completos será cada vez más despreciable. Dicho desde el punto de vista contrario, existe un límite máximo para el grado de agrupación de los datos, a partir del cual el error introducido al considerar los puntos medios se vuelve importante, y justo a partir de ese límite sería incorrecto utilizar esta ventana $AMISE$ óptima. Esto plantea dos objetivos: el primero, demostrar rigurosamente bajo qué supuestos es posible usar la ventana h_{AMISE} en el caso de datos agrupados. Como es habitual, esos supuestos deberán estar expresados en términos de la función que se quiera estimar, de la función núcleo utilizada y de la relación asintótica entre el parámetro de suavizado y el tamaño muestral, pero también se deberá considerar alguna cantidad representativa del grado de agrupación de los datos. El segundo objetivo consiste en encontrar, en la práctica, cuál es ese límite en la agrupación de los datos.

El teorema 3.1, cuya demostración se incluye en el apéndice D, enuncia de manera rigurosa cuáles son los supuestos que se deben satisfacer para que la ventana h_{AMISE} sea la misma en el caso de datos completos que en el caso de datos agrupados. Como ya se adelantaba, los supuestos están expresados en términos de la función que se desea estimar (en este caso, la densidad f), la función núcleo, el parámetro de suavizado, el tamaño muestral, y como medida del grado de agrupación de los datos, se ha considerado la longitud promedio de los intervalos, $\bar{l}$. Además, se ha incluido una cota para la máxima diferencia absoluta entre la longitud de los intervalos y la longitud media de los mismos, que permite controlar el grado de variabilidad en la longitud de los intervalos (véanse los supuestos 3.1 a 3.4 en la sección 3.3). Este resultado es importante, pues demuestra que cuando se cumplen esos supuestos, aunque los datos estén agrupados, la expresión del $AMISE$ es la misma que para el caso de datos no agrupados (compárense las ecuaciones (3.5) y (2.11)) . Lo anterior significa que los términos no dominantes en la expresión del error cuadrático medio integrado ($MISE$) (que en el caso de datos agrupados dependen de su grado de agrupación, mediante $\bar{l}$) se han desvanecido lo suficientemente rápido al teóricamente aumentar el tamaño muestral, razón por la cual el término dominante (esto es, el $AMISE$) coincide con el $AMISE$ en el caso de datos completos. No obstante, si esos supuestos no se satisfacen, los términos no dominantes del $MISE$ siguen siendo importantes en el caso de datos agrupados, con lo que el $AMISE$ es una mala aproximación del $MISE$.

El razonamiento anterior lleva a distinguir dos tipos de agrupaciones: agrupación ligera y agrupación pesada. En la agrupación ligera, la expresión de la ventana h_{AMISE} coincide en el caso de datos agrupados y no agrupados. En el caso de agrupación pesada, dado que el $AMISE$ no es una buena aproximación del $MISE$, no sería recomendable usar dicha ventana. Surge de manera natural preguntarse cuál es, en la práctica, ese límite de agrupamiento de los datos que define la frontera entre agrupación ligera y pesada, y que permite el uso (o no) de la ventana h_{AMISE} en el caso de datos agrupados.

Otro aspecto importante concerniente a la ventana h_{AMISE} es el de las cantidades de las que depende. Además de depender del tamaño muestral y de cantidades que dependen de la función núcleo elegida, también depende de la curvatura de la función de densidad, $A(f'')$. En la práctica, lo que se hace es considerar una estimación de ésta y sustituirla en la expresión h_{AMISE}, obteniendo así un selector plug-in. Así, aunque en teoría, y siempre que se cumplan las condiciones necesarias, la ventana h_{AMISE} es la misma para datos completos que para datos agrupados, en la práctica existe una sutil diferencia: la estimación de la curvatura se hace, en un caso, con datos completos; en el otro, con datos agrupados (véanse las ecuaciones (3.6), (3.7) y (3.8)).

Para evaluar el desempeño del estimador núcleo de la densidad para datos agrupados usando el selector plug-in, se consideraron tres tamaños muestrales (60, 240 y 960) y se realizaron dos estudios de simulación (sección 3.4). Simulando muestras de esos tamaños muestrales a partir de una mixtura de normales, en el primer estudio se consideraron dos escenarios: a) En el primer escenario (S1) se simuló que la longitud de los intervalos se reducía de manera rápida al aumentar el tamaño muestral. b) En el segundo (S2), la longitud de los intervalos se reducía de manera muy lenta conforme el tamaño muestral aumentaba. Es decir, el primer escenario simulaba una transición rápida a intervalos pequeños (agrupación ligera) al aumentar el tamaño muestral, mientras que el segundo escenario simulaba una transición muy lenta (agrupación pesada).

Los resultados confirmaron lo que la teoría predecía. Conforme el tamaño muestral aumenta, si las agrupaciones son ligeras, el selector plug-in es una buena opción para aproximarse a la ventana que minimiza el $MISE$ en el caso de datos agrupados, h_{MISE_g}, obteniéndose cada vez mejores estimaciones de la densidad. En cambio, si al aumentar el tamaño muestral las agrupaciones se mantienen gruesas (pesadas), el selector plug-in es incapaz de dar resultados cercanos a la ventana h_{MISE_g}, por lo que, de usarse, se obtendrían malas estimaciones de la densidad.

En el segundo estudio de simulación se consideró un sólo tamaño muestral (240) y se consideraron diferentes conjuntos de intervalos con longitudes promedio que iban desde valores muy pequeños (bastantes intervalos) hasta valores relativamente grandes (pocos intervalos), y se obtuvo el $MISE_g$ del estimador para diferentes valores de ventana h. Los resultados mostraron que existe un rango específico de valores de $\bar{l}$ para los cuales, usando la ventana que minimiza el $MISE_g$, el estimador $\hat{f}_h^g$ tiene su mejor desempeño. Además, en la simulación se encontró que el selector plug-in proveía buenas aproximaciones de la ventana h_{MISE_g} en esa región específica de valores de $\bar{l}$, mostrando de manera preliminar que, en la práctica, el valor de $\bar{l}$ que definía la frontera entre agrupación ligera y pesada era de alrededor de 0.06 veces el rango de los datos, r. Lo anterior se puso a prueba considerando datos reales adecuadamente agrupados, obteniéndose que el selector plug-in proveía buenos resultados hasta valores de $\bar{l}$ de alrededor de $0.075r$, valor muy cercano al obtenido en las simulaciones.

Como ya se mencionó, el selector plug-in requiere estimar la curvatura de la densidad, $A\left(f''\right)$, por lo que era importante demostrar la consistencia del estimador de la curvatura para datos agrupados, $\hat{A}_g$ (capítulo 4). Bajo las condiciones adecuadas, la consistencia se enuncia en el teorema 4.1, mismo que se demuestra en el apéndice E.1. Sin embargo, aunque de este resultado se infiere la consistencia del selector plug-in, hay que recordar que éste sólo provee buenos resultados cuando la agrupación es ligera. Así, quedaba pendiente la propuesta de un selector de ventana para los casos de agrupación pesada.

La solución que se propone es un selector bootstrap. A partir de una expresión exacta del $MISE$ para datos agrupados (véase el teorema 4.2, cuya demostración se encuentra en el apéndice E.2), mediante procedimientos paralelos se deriva una expresión exacta para el $MISE$ versión bootstrap, $MISE^*$(ecuación 4.4). A partir de esta última, mediante algunos desarrollos es posible obtener una expresión operativa para el $MISE^*$, dada por la ecuación (4.5). Así, el selector bootstrap propuesto se basa en evaluar numéricamente esta última ecuación, y seleccionar la ventana h que la minimice. En el caso de que se use un núcleo Gaussiano, la ecuación (4.5) se convierte en la ecuación (4.6) (véase el apéndice E.2.1).

El selector bootstrap depende inicialmente de una estimación piloto de la densidad, para lo cual es necesario contar con una ventana piloto ζ. La teoría demuestra que la ventana piloto ζ debe ser aquella que minimice el promedio de la diferencia cuadrática entre la curvatura $A\left(f''\right)$ y su estimación núcleo usando ζ, $\hat{A}_\zeta\left(f''\right)$. Asumiendo que la densidad es normal y que el núcleo es Gaussiano, es posible demostrar que la ventana piloto se puede obtener mediante la ecuación (4.7). Para obtener reglas de selección de ζ para datos agrupados, se realizaron algunos estudios de simulación cuyos resultados pueden verse en los los apéndices E.3 y E.4.

Para evaluar el desempeño del estimador $\hat{f}_h^g$ usando el selector bootstrap, se realizaron algunos estudios de simulación (véanse los detalles en la sección 4.2). Se consideraron los mismos escenarios que en el caso del selector plug-in, S1 y S2. Los resultados fueron bastante buenos, ya que el selector bootstrap demostró proveer buenas aproximaciones de la ventana h_{MISE_g} en ambos escenarios, a diferencia del selector plug-in, que solo provee buenos resultados en S1. Ciertamente, puede notarse una ligera ventaja del selector plug-in sobre el selector bootstrap en S1, pero aún así, los resultados del selector bootstrap resultaron bastante competitivos. En resumen, la recomendación es usar el selector plug-in cuando la agrupación sea ligera y el tamaño muestral sea mediano o grande. En cualquier otro caso (es decir, agrupación pesada y cualquier tamaño muestral), se recomienda el selector bootstrap.

Las guías previas sobre la frontera entre agrupación ligera y pesada se confirman. Se dice que hay agrupación ligera cuando $\bar{l} \approx 0.075r$ o menos, y la agrupación pesada comienza a partir de $\bar{l} \approx 0.075r$. Si no hay certeza de qué tipo de agrupación está presente en los datos, la recomendación es usar el selector bootstrap.

El otro enfoque no paramétrico planteado es el de la estimación tipo núcleo de la distribución (capítulo 5). Este estimador se obtiene directamente al integrar el estimador tipo núcleo de la densidad para datos agrupados, $\hat{f}_h^g$, (ecuación 3.2), dando como resultado el estimador de la distribución, $\hat{F}_h^g$, (ecuación (5.1)), que por construcción puede usarse con datos agrupados.

Al igual que en el caso del estimador núcleo de la distribución para datos completos, $\hat{F}_h$, (ecuación (2.41)), el estimador $\hat{F}_h^g$ requiere seleccionar la ventana de manera adecuada para tener un buen desempeño. En este sentido, en el caso de datos completos, una posibilidad es considerar la ventana h_{AMISE_F}, dada por la ecuación (2.50). Dado que ambos estimadores están definidos de manera similar, cabe preguntarse en qué circunstancias sería correcto utilizar la ventana h_{AMISE_F} en el caso de datos agrupados.

Haciendo un ejercicio similar al realizado con el estimador $\hat{f}_h^g$, el teorema 5.1 (demostrado en el apéndice F) establece de manera rigurosa bajo qué supuestos el $AMISE$ de ambos estimadores, $\hat{F}_h$ y $\hat{F}_h^g$, coinciden y, por lo tanto, la expresión de la ventana $AMISE$ óptima, h_{AMISE_F}, es la misma. Salvo el orden de diferenciabilidad de la función núcleo, nótese que los supuestos son esencialmente los mismos. Este teorema premite obtener la ecuación (5.3), que en efecto, es la misma que (2.50).

Al igual que en el caso del estimador $\hat{f}_h^g$, se debe estimar el funcional $A(f')$ con datos agrupados para que la ecuación (5.3) sea utilizable en la práctica, ya que el resto de cantidades son conocidas o calculables. Procediendo según Polansky y Baker (véase la sección

2.2.3, ecuación (2.56)), se obtiene una estimación de $A(f')$, $\hat{A}_{PB_g}(f)$, que al sustituir en (5.3) se convierte en el selector plug-in $\hat{h}_{PB_g}$, (ecuación 5.4).

Mediante estudios de simulación se analizó el desempeño del estimador núcleo de la distribución para datos agrupados usando el selector $\hat{h}_{PB_g}$, considerando los mismos tamaños muestrales, escenarios de agrupación ligera y pesada, S1 y S2, y demás características (véanse los detalles en la sección 5.3). A diferencia del estimador $\hat{f}_h^g$, las curvas del $MISE_g$ del estimador $\hat{F}_h^g$ mostraron diferencias mínimas para valores ligeramente diferentes de la ventana h que minimiza el $MISE_g$. En otras palabras, que cuando se quiere estimar la distribución, el margen para equivocarnos en el valor de la ventana h óptima es algo más grande que en el caso de la densidad, con lo cual, las estimaciones de la distribución se ven menos afectadas que las estimaciones de la densidad. Este resultado es interesante, pues sugiere que a pesar de que los datos son agrupados, la estimación de la distribución tipo núcleo es un procedimiento más robusto, en el sentido de que es menos sensible a ligeras desviaciones del valor de la ventana h seleccionada con respecto a la ventana óptima.

Los resultados de la simulación considerando diferentes grados de agrupación y un sólo tamaño muestral refuerzan lo anterior. Al igual que en el caso del estimador de la densidad $\hat{f}_h^g$, es posible identificar una región de agrupación para la cual el selector $\hat{h}_{PB_g}$ da buenas aproximaciones de la ventana óptima que minimiza el $MISE_g$, y además, esta región de agrupación es más amplia que en el caso de la estimación de la densidad. Es decir, en el caso de la densidad, ahí donde el selector plug-in falla, en el caso de la distribiución aún da buenos resultados. Esta propiedad le da al estimador $\hat{F}_h^g$ cierta ventaja sobre el estimador $\hat{f}_h^g$, ya que su selector plug-in resiste más en casos de agrupación pesada. Esto permite que el estimador $\hat{F}_h^g$ tenga un buen desempeño en niveles de agrupamiento en los que el estimador $\hat{f}_h^g$ necesita el uso de selectores de ventana más elaborados. Por lo anterior, en el caso de la estimación de la distribución tipo núcleo con datos agrupados no es tan necesaria la propuesta de un selector alternativo como en el caso de la estimación de la densidad. Sin embargo, es evidente que sería de gran ventaja proponer un selector de ventana más preciso. Este podría ser un tópico interesante en investigaciones futuras.

La última parte de esta tesis (capítulo 6) muestra la aplicación de los estimadores y los selectores propuestos a conjuntos de datos de emergencia reales. Lo primero que se hizo fue usar el estimador $\hat{f}_h^g$ para estimar la estructura de los tres conjuntos de datos. A pesar de que los datos mostraban agrupación pesada, eligiendo la ventana adecuadamente (de acuerdo con los criterios para distinguir entre agrupación ligera y pesada, obtenidos en los capítulos anteriores), la estimación tipo núcleo mostró ser una herramienta efectiva para encontrar la estructura de los datos.

A partir de esas estimaciones de la densidad se propusieron modelos apropiados de mixturas de normales para describir a los datos de emergencia, lo que permitió realizar estudios de simulación en donde los selectores plug-in y bootstrap fueron puestos a prueba bajo diferentes condiciones de agrupación. Los resultados confirmaron lo que ya se había encontrado previamente en este trabajo, a saber: 1) Que cuando la agrupación es ligera ($\bar{l} < 0.075r$), pueden usarse ambos selectores, plug-in y bootstrap, aunque es ligeramente preferibe usar el plug-in. 2) En caso de que la agrupación sea pesada ($\bar{l} > 0.075r$), no se recomienda el uso del selector plug-in, recomendándose absolutamente el uso del selector bootstrap. Prcoediendo de esta manera, el error de estimación de la densidad permanece bajo control en ambos casos, pero lo más importante es que permanece razonablemente acotado en casos de agrupación pesada, permitiendo así al estimador $\hat{f}_h^g$ detectar la estructura de los datos (aunque sea de manera parcial o aproximada) incluso en casos de agrupación muy pesada.

El primer estudio de simulación mostró también que algunas densidades son más difíciles de estimar que otras. Las que que resultan más difíciles de estimar son aquellas que tienen múltiples modas o áreas alternadas de alta y baja densidad, lo cual es esperable: si la propia agrupación de los datos oculta información valiosa de su estructura, la pérdida de información es más pronunciada en casos de mayor curvatura. Lo anterior, y el hecho de que sólo se está considerando una ventana a lo largo de todo el dominio, hace más difícil la estimación de este tipo de densidades. No obstante, este problema se resuelve hasta cierto punto usando el selector de ventana adecuado según el grado de agrupación de los datos.

Por otra parte, se realizó una comparación de la bondad de ajuste entre algunos métodos paramétricos de regresión no lineal (Logística, Gompertz y Weibull) y el estimador tipo núcleo de la distribución para datos agrupados, $\hat{F}_h^g$. Los resultados mostraron que a menos de que la distribución sea relativamente suave y sigmoidal, los métodos de regresión mencionados pueden tener serios problemas para describir algunos detalles de la distribución de los datos. En cambio, el método no paramétrico demostró ser una buena opción en general, con resultados bastante competitivos en casos tanto de funciones claramente sigmoidales como en casos de funciones más curvas.

Más aún, la comparación de la bondad de ajuste entre las regresiones paramétricas mencionadas y el estimador tipo núcleo de la distribución también se realizó mediante estudios de simulación. Simulando las mismas condiciones de agrupación presentes en los tres conjuntos de datos, este estudio corroboró que, en promedio, el estimador no paramétrico de la distribución tiene un desempeño bastante competitivo o en algunos casos incluso mejor que los métodos no paramétricos habitualmente usados en malherbología. Así, la estimación tipo núcleo de la distribución para datos agrupados es una opción válida para describir la relación entre al emergencia de las malas hierbas y el tiempo hidrotermal acumulado. Además, su flexibilidad resultó ser muy útil para describir estructuras que no son tan sencillas, como aquellas que tienen más de una moda o tienen características o variaciones sutiles en la densidad. Los modelos paramétricos, debido a su rigidez, se muestran bastante limitados para describir estos últimos casos.

Por último, puede decirse que los resultados son alentadores. Por una parte, se ha cumplido el objetivo de estimar suavemente tanto la función de densidad como la función de distribución cuando los datos se presentan agrupados, lo que para los malherbólogos significa la posibilidad de calcular con mayor precisión las probabilidades de emergencia de las malas hierbas. Por otra parte, se ha realizado un trabajo teórico y formal bastante completo en relación con las propiedades asintóticas de los estimadores propuestos, así como estudios de simulación para verificar el desempeño tanto de los estimadores como de los selectores de ventana propuestos. En particular, los selectores de ventana han probado ser efectivos en diferentes escenarios de agrupación, manteniendo el error de estimación realtivamente bajo control. Finalmente, ha sido posible establecer algunas guías de uso para distinguir en la práctica si un conjunto de datos presenta agrupación ligera o pesada, y a partir de eso actuar en consecuencia eligiendo el selector de ventana apropiado.

Quedan varias líneas de investigación por explorar. Por ejemplo: 1) Considerar criterios más complejos de desagregación de datos, lo que implicaría considerar tratamientos matemáticos más complejos. 2) Buscar un selector tipo plug-in tanto no sólo para el caso de agrupación ligera, sino pesada, tanto en el caso de la estimación de la densidad como de la distribución. 3) Considerar formas más elaboradas y efectivas de estimar la curvatura, dada su importancia para los selectores plug-in y bootstrap. 4) Además del selector plug-in, proponer selectores de ventana alternativos en el caso de la estimación tipo núcleo de la distribución para datos agrupados, lo que presumiblemente mejoraría de manera notable el desempeño de este estimador en casos de agrupación pesada o muy pesada. 5) Estudiar la aplicación de los métodos no paramétricos propuestos en esta investigación considerando datos procedentes de otras áreas del conocimiento, así como considerar datos con estructuras más complejas. 6) Considerar otras aproximaciones al problema planteado, como la regresión isotónica no paramétrica, o modelos más complejos como aquellos basados en procesos de Poisson no homogéneos.

Bibliography

Abramson, I.S. (1982), 'On bandwidth variation in kernel estimates -a square root law', *The Annals of Statistics* **9**, 168–176.

Akaike, H. (1954), 'An approximation to the density function', *Annals of the Institute of Statistical Mathematics* **6**, 127–132.

Aldershof, B. (1991), Estimation of integrated squared density derivatives, PhD thesis, University of North Carolina, Chapel Hill.

Alemseged, Y., R.E. Jones and R.W. Medd (2001), 'A farmer survey of weed management and herbicide resistance problems of winter crops in Australia', *Plant Protection Quarterly* **16**, 21–25.

Altman, N. and C. Léger (1995), 'Bandwidth selection for kernel distribution function estimation', *Journal of Statistical Planning and Inference* **46**, 195–214.

Anastassiou, G.A. (1995), 'Ostrowski type inequalities', *Proceedings of the American Mathematical Society* **123**, 3775–3781.

Anastassiou, G.A. (1997), 'Multivariate Ostrowski type inequalities', *Acta Mathematica Hungarica* **76**, 267–278.

Bates, D. and D. Watts (1988), *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

Bowman, A., P. Hall and T. Prvan (1998), 'Bandwidth selection for the smoothing of distribution functions', *Biometrika* **85**, 799–808.

Bowman, A.W. (1984), 'An alternative method of cross-validation for the smoothing of density estimates', *Biometrika* **71**, 353–360.

Bradford, K.J. (2002), 'Applications of hydrothermal time to quantifying and modeling seed germination and dormancy', *Weed Science* **50**, 248–260.

Cao, R. (1990), Aplicaciones y nuevos resultados del método bootstrap en la estimación no paramétrica de curvas, PhD thesis, Universidade de Santiago de Compostela.

Cao, R. (1993), 'Bootstrapping the mean integrated squared error', *Journal of Multivariate Analysis* **45**, 137–160.

Cao, R., M. Francisco-Fernández, A. Anand, F. Bastida and J.L. Gonzalez-Andujar (2011), 'Computing statistical indices for hydrothermal time using weed emergence data', *Journal of Agricultural Science* **149**, 701–712.

Chen, S. (2000), 'Probability density function estimation using Gamma kernels', *Annals of the Institute of Statistical Mathematics* **52**, 471–480.

Chiu, S.T. (1991*a*), 'Bandwidth selection for kernel density estimation', *The Annals of Statistics* **19**, 1883–1905.

Chiu, S.T. (1991*b*), 'Some stabilized bandwidth selectors for nonparametric regression', *The Annals of Statistics* **19**, 1528–1546.

Chiu, S.T. (1992), 'An automatic bandwidth selector for kernel density estimation', *Biometrika* **79**, 771–782.

Coit, D.W. and K.A. Dey (1999), 'Analysis of grouped data from field-failure reporting systems', *Reliability Engineering & System Safety* **65**, 95–101.

Colbach, N., C. Durr, J. Roger-Estrade and J. Caneill (2005), 'How to model the effects of farming practices on weed emergence', *Weed Research* **45**, 2–17.

Delow, J.J. and B.R. Milne (1986), 'Control of phalaris paradoxa in wheat', *Australian Weeds* **3**, 22–23.

Devroye, L. and L. Györfi (1985), *Nonparametric Density Estimation: The L1 View*, John Wiley.

Faraway, J.J. and M. Jhun (1990), 'Bootstrap choice of bandwidth for density estimation', *Journal of the American Statistical Association* **85**, 1119–1122.

Fernández-Quintanilla, C., L. Navarrete, J.L. González-Andújar, A. Fernández and M.J. Sánchez (1986), 'Seedling recruitment and age-specific survivorship and reproduction in populations of Avena Sterilis L. ssp. Ludoviciana', *Journal of Applied Ecology* **23**, 945–955.

Fix, E. and J.L. Hodges (1951), Discriminatory analysis-nonparametric discrimination: consistency properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas. Project no. 21-29-004.

Forcella, F., R.L. Benech Arnold, R. Sánchez and C.M. Ghersa (2000), 'Modelling seedling emergence', *Field Crops Research* **67**, 123–139.

González-Andújar, J.L. and M. Saavedra (2003), 'Spatial distribution of annual grass weed populations in winter cereal', *Crop Protection* **22**, 629–633.

González-Andujar, J.L., M. Francisco-Fernández, R. Cao, M. Reyes, J.M. Urbano, F. Forcella and F. Bastida (2015), 'A comparative study between non-linear regression and non-parametric approaches for modelling phalaris paradoxa seedling emergence', *Weed Research* . Manuscript submitted for publication.

Grundy, A.C. (2003), 'Predicting weed emergence: a review of approaches and future challenges', *Weed Research* **43**, 1–11.

Guidoum, A.C. (2014), *kedd: Kernel estimator and bandwidth selection for density and its derivatives.* R package version 1.0.1.

Guo, S. (2005), 'Analyzing grouped data with hierarchical linear modeling', *Children and Youth Services Review* **27**, 637–652.

Haj Seyed Hadi, M.R. and J.L. Gonzalez-Andujar (2009), 'Comparison of fitting weed seedling emergence models with nonlinear regression and genetic algorithm', *Computers and Electronics in Agriculture* **65**, 19 – 25.

Hall, P. (1982), 'The influence of rounding errors on some nonparametric estimators of a density and its derivatives', *SIAM Journal on Applied Mathematics* **42**, 390–399.

Hall, P. (1987), 'On Kullback-Leibler loss and density estimation', *The Annals of Statistics* **15**, 1491–1519.

Hall, P. and J.S. Marron (1987*a*), 'Estimation of integrated squared density derivatives', *Statistics & Probability Letters* **6**, 109–115.

Hall, P. and J.S. Marron (1987*b*), 'Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation', *Probability Theory and Related Fields* **74**, 567–581.

Hall, P. and J.S. Marron (1991), 'Local minima in cross-validation functions', *Journal of the Royal Statistical Society. Series B* **53**, 245–252.

Hall, P. and M.P. Wand (1996), 'On the accuracy of binned kernel density estimators', *Journal of Multivariate Analysis* **56**, 165–184.

Heidenreich, N.B., A. Schindler and S. Sperlich (2013), 'Bandwidth selection for kernel density estimation: a review of fully automatic selectors', *Advances in Statistical Analysis* **97**, 403–433.

Hodges, J.L. and E.L. Lehman (1956), 'The efficiency of some nonparametric competitors to the *t*-test', *The Annals of Mathematical Statistics* **13**, 324–335.

Härdle, W. (1991), *Smoothing Techniques, With Implementations in S*, Springer, New York.

Härdle, W. and D.W. Scott (1992), 'Smoothing in by weighted averaging using rounded points', *Computational Statistics* **7**, 97–128.

Hunter, E.A., C.A. Glasbeya and R.E.L. Naylor (1984), 'The analysis of data from germination tests', *The Journal of Agricultural Science* **102**, 207–213.

Izquierdo, J., J.L. González-Andújar, F. Bastida, J.A. Lezaún and M.J. Sánchez del Arco (2009), 'A thermal time model to predict corn poppy (Papaver Rhoeas) emergence in cereal fields', *Weed Science* **57**, 660–664.

Janssen, P., J.S. Marron, N. Veraverbeke and W. Sarle (1995), 'Scale measures for bandwidth selection', *Journal of Nonparametric Statistics* **5**, 359–380.

Jiménez-Hidalgo, M.J., M. Saavedra and L. García-Torres (1997), Phalaris brachystachys en cultivos de cereales, *in* F.X. Sans and C. Fernández-Quintanilla, eds, 'Biología de las malas hierbas de España', Phytoma-España.

Jones, M.C. (1990), 'The performance of kernel density functions in kernel distribution function estimation', *Statistics and Probability Letters* **9**, 129–132.

Jones, M.C. (1993), 'Simple boundary correction for kernel density estimation', *Statistics and Computing* **3**, 135–146.

Jones, M.C., J.S. Marron and S.J. Sheather (1996*a*), 'A brief survey of bandwidth selection for density estimation', *Journal of the American Statistical* **91**, 401–407.

Jones, M.C., J.S. Marron and S.J. Sheather (1996*b*), 'Progress in data-based bandwidth selection for kernel density estimation', *Computational Statistics* **11**, 337–381.

Jones, M.C. and P.J. Foster (1996), 'A simple nonnegative boundary correction method for kernel density estimation', *Statistica Sinica* **6**, 1005–1013.

Jones, M.C. and S.J. Sheather (1991), 'Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives', *Statistics & Probability Letters* **11**, 511–514.

Kullback, S. and R.A. Leibler (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**, 79–86.

Leblanc, M.L., D.C. Cloutier, K.A. Stewart and C. Hameld (2003), 'The use of thermal time to model common lambsquarters (Chenopodium album) seedling emergence in corn', *Weed Science* **51**, 718–724.

Leguizamón, E.S., C. Fernández-Quintanilla, J. Barroso and J.L. González-Andújar (2005), 'Using thermal and hydrothermal time to model seedling emergence of Avena Sterilis ssp. Ludoviciana in Spain', *Weed Research* **45**, 149–156.

Loftsgaarden, D.O. and C.P. Quesenberry (1965), 'A nonparametric estimate of a multivariate density function', *The Annals of Mathematical Statistics* **36**, 1049–1051.

Marron, J.S. (1992), Bootstrap bandwidth selection, *in* R.LePage and L.Billard, eds, 'Exploring the limits of Bootstrap', pp. 249–262.

Marron, J.S. (1993), 'Discussion of: 'Practical performance of several data driven bandwidth selectors' by Park and Turlach', *Computational Statistics* **8**, 17–19.

Marron, J.S. and A.B. Tsybakov (1995), 'Visual error criteria for qualitative smoothing', *Journal of the American Statistical Association* **90**, 499–507.

Marron, J.S. and D. Ruppert (1994), 'Transformations to reduce boundary bias in kernel density estimation', *Journal of the Royal Statistical Society. Series B 56* **4**, 653–671.

McGiffen, M., K. Spokas, F. Forcella, D. Archer, S. Poppe and R. Figueroa (2008), 'Emergence prediction of common groundsel (Senecio Vulgaris)', *Weed Science* **56**, 58–65.

Mächler, M. (2013), *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods).* R package version 1.1-4.

Minoiu, C. and S.G. Reddy (2009), 'Estimating poverty and inequality from grouped data: How well do parametric methods perform?', *Journal of Income Distribution* **18**, 160–178.

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

Naylor, R.E.L. (1981), 'An evaluation of various germination indices for predicting differences in seed vigour in Italian Ryegrass', *Seed Science and Technology* **9**, 593–600.

Newcomer, J.T., N.K. Neerchal and J.G. Morel (2008), Computation of higher order moments from two multinomial overdispersion likelihood models, Technical report, Department of Statistics, University of Maryland.

Ostrowski, A. (1938), 'Über die Absolutabweichung einer differentiebaren Funktion von ihrem Integralmittelwert', *Commentarii Mathematici Helvetici* **10**, 226–227.

Park, B.U. and J.S. Marron (1992), 'On the use of pilot estimators in bandwidth selection', *Journal of Nonparametric Statistics* **1**, 231–240.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**, 1065–1076.

Pinheiro, J.C. and D.M. Bates (2000), *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

Pipper, C.B. and C. Ritz (2007), 'Checking the grouped data version of the Cox model for interval-grouped survivla data', *Scandinavian Journal of Statistics* **34**, 405–418.

Polansky, A.M. and E.R. Baker (2000), 'Multistage plug-in bandwidth selection for kernel distribution function estimates', *Journal of Statistical Computation and Simulation* **65**, 63–80.

Quintela-del-Río, A. and G. Estévez-Pérez (2012), 'Nonparametric kernel distribution function estimation with kerdiest: An R package for bandwidth choice and applications', *Journal of Statistical Software* **50**, 1–21.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Reyes, M., M. Francisco-Fernández and R. Cao (2015*a*), 'Bandwidth selection in kernel density estimation for interval-grouped data'. Manuscript in preparation.

Reyes, M., M. Francisco-Fernández and R. Cao (2015*b*), 'Nonparametric kernel density estimation for general grouped data', *Journal of Nonparametric Statistics* . Manuscript submitted for publication.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics* **27**, 832–837.

Royo-Esnal, A., J. Torra, J.A. Conesa, F. Forcella and J. Recasens (2010), 'Modeling the emergence of three arable bedstraw (Galium) species', *Weed Science* **58**, 10–15.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel estimators', *Scandinavian Journal of Statistics* **9**, 65–78.

Ruppert, D. and D.B.H. Cline (1994), 'Bias reduction in kernel density estimation by smoothed empirical transformations', *The Annals of Statistics* **22**, 185–210.

Sarda, P. (1993), 'Smoothing parameter selection for smooth distribution function', *Journal of Statistical Planning and Inference* **35**, 65–75.

Scaillet, O. (2004), 'Density estimation using inverse and reciprocal inverse gaussian kernels', *Journal of Nonparametric Statistics* **16**, 217–226.

Schutte, B.J., E.E. Regnier, S.K. Harrison, J.T. Schmoll, K. Spokas and F. Forcella (2008), 'A hydrothermal seedling emergence model for Giant Ragweed (Ambrosia trifida)', *Weed Science* **56**, 555–560.

Scott, D.W. (1979), 'On optimal and data-based histograms', *Biometrika* **66**, 605–610.

Scott, D.W. (1985), 'Average shifted histograms: effective nonparametric density estimators in several dimensions', *The Annals of Statistics* **13**, 1024–1040.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley.

Scott, D.W. and G.R. Terrell (1987), 'Biased and unbiased cross-validation in density estimation', *Journal of the American Statistical Association* **82**, 1131–1146.

Scott, D.W. and S.J. Sheather (1985), 'Kernel density estimation with binned data', *Communications in Statistics - Theory and Methods* **14**, 1353–1359.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Sheather, S.J. and M.C. Jones (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society. Series B* **53**, 683–690.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Champman and Hall.

Spokas, K. and F. Forcella (2009), 'Software tools for weed seed germination modelling', *Weed Science* **57**, 216–227.

Swanepoel, J.W.H. (1988), 'Mean integrated squared error properties and optimal kernels when estimating a distribution function', *Communications in Statistics - Theory and Methods* **17**, 3785–3799.

Taylor, C.C. (1989), 'Bootstrap choice of the smoothing parameter in kernel density estimation', *Biometrika* **76**, 705–712.

Terrel, G.R. (1990), 'The maximal smoothing principle in density estimation', *Journal of the American Statistical Association* **85**, 470–477.

Titterington, D. (1983), 'Kernel-based density estimation using censored, truncated or grouped data', *Communications in Statistics - Theory and Methods* **12**, 2151–2167.

Turnbull, B.W. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society. Series B (Methodological)* **38**, 290–295.

Van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press.

Wand, M. (2013), *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. R package version 2.23-10.

Wand, M.P., J.S. Marron and D. Ruppert (1991), 'Transformations in density estimation (with comments)', *Journal of the American Statistical Association* **86**, 343–361.

Wand, M.P. and M.C. Jones (1995), *Kernel Smoothing*, Chapman and Hall.

Wand, M.P. and W.R. Schucany (1990), 'Gaussian-based kernels', *The Canadian Journal of Statistics* **18**, 197–204.

Woodroofe, M. (1970), 'On choosing a delta-sequence', *The Annals of Mathematical Statistics* **41**, 1665–1671.