TESIS DOCTORAL

Aportaciones teóricas y computacionales al análisis cluster, estimación de distribución y cálculo de esperanzas condicionales

Pablo Monfort Vinuesa

Departamento de Matemáticas

2014

TESIS DOCTORAL

Aportaciones teóricas y computacionales al análisis cluster, estimación de distribución y cálculo de esperanzas condicionales

Pablo Monfort Vinuesa
Departamento de Matemáticas

Conformidad de los supervisores:

Director:	Codirector:
Dr. D. José Enrique Chacón Durán	Dr. D. Agustín García Nogales

Badajoz, 2014

Theoretical and computational contributions to cluster analysis, distribution function estimation and conditional expectation calculation

Pablo Monfort Vinuesa
Department of Mathematics

Supervisors Approval:

Dr. José Enrique Chacón Durán Dr. Agustín García Nogales

Badajoz, 2014

Summary

In the Thesis we show how to use methods to analyze the asymptotic behavior of kernel distribution function estimators. Exact expressions for the mean integrated squared error in terms of the characteristic function of the distribution and the Fourier transform of the kernel are employed to obtain the limit value of the optimal bandwidth sequence in its greatest generality. The assumptions in our results are mild enough so that they are applicable when the kernel used in the estimator is a superkernel, or even the sinc kernel, and this allows to extract some interesting consequences, as the existence of a class of distributions for which the kernel estimator achieves a first-order improvement in efficiency over the empirical distribution function.

In the second part we develope a Monte Carlo method to approximate conditional expectations in a probabilistic framework motivated by a general result inspired by the Besicovitch covering theorem for differentiation of measures. The method is specially useful when densities are not available or are not easy to compute. The method is illustrated by means of various examples and can also be used in a statistical setting to approximate the conditional expectation given a sufficient statistic. In this paper it is used to compute the minimum risk equivariant estimator (MRE) of the location parameter of a general half-normal distribution since this estimator is described in terms of a conditional expectation for known values of the location and scale parameters. For the sake of completeness, an explicit expression for the minimum risk equivariant estimator of the scale parameter is given.

As far as we are aware, these estimators have not appeared before in the literature. Simulation studies to compare the behavior of the new estimators with those of maximum likelihood and unbiased estimators are presented.

Finally, we explore the performance of several automatic bandwidth selectors, originally designed for density gradient estimation, as data-based procedures for nonparametric, modal clustering. The key tool to obtain a clustering from density gradient estimators is the mean shift algorithm, which allows to obtain a partition not only of the data sample, but also of the whole space. The results of our simulation study suggest that most of the methods considered here, like cross validation and plug in bandwidth selectors, are useful for cluster analysis via the mean shift algorithm.

Resumen

En la presente Tesis exponemos cómo utilizar métodos para anlizar el comportamiento asintótico de estimadores núcleo de la función de distribución. Utilizamos expresiones exactas del error cuadrático integrado medio en términos de la función característica de la distribución y de la transformada de Fourier del núcleo para obtener el límite de la secuencia de anchos de banda óptimos en su modo más genérico. Las hipótesis requeridas en nuestros resultados son tan suaves que son aplicables en el caso de disponer de un supernúcleo como núcleo del estimador. Incluyendo incluso el caso de que el núcleo considerado del estimador sea el núcleo sinc. Esto último permite extraer algunas consecuencias interesantes como la existencia de un tipo de distribuciones para las que el estimador núcleo alcanza una mejora de primer grado en eficiencia con respecto a la función de distribución empírica.

En la segunda parte de la Tesis, desarrollamos un método de Monte Carlo para aproximar esperanzas condicionales en un marco probabilístico basándonos en un resultado fruto del teorema de Besicovitch para diferenciación de medidas. El método es especialmente útil cuando o no se dispone de las densidades o éstas no son calculables. El método es ilustrado con los cálculos de medias de varios ejemplos y puede ser también utilizado en un contexto estadístico para aproximar esperanzas condicionales dado un estadístico suficiente. Además, utilizamos dichos resultados para computar el estimador equivariante de mínimo riesgo (MRE) del parámetro de

localización de una distribución half-normal ya que este estimador es expresado en términos de una esperanza condicional para valores conocidos de los parámetros de localización y escala. Proporcionamos también una expresión explícita del estimador equivariante de mínimo riesgo del parámetro de escala. Por último, se exponen estudios de simulación para comparar el comportamiento de los nuevos estimadores con el insesgado y el de máxima verosimilitud.

Finalmente, analizamos el comportamiento de varios selectores automáticos de ancho de banda, originalmente diseñados para estimación de gradiente de densidades, como procedimientos para análisis no paramétrico de cluster modal basados en datos. La herramienta clave para obtener una clasificación cluster a partir de estimadores del gradiente de la densidad es el algoritmo mean shift, el cual permite obtener una partición, no sólo de los datos de la muestra, si no del espacio completo. Los resultados de este estudio de simulación apuntan a que la mayoría de los métodos aquí considerados, como los selectores de ancho de banda de validación cruzada o el plug-in, son útiles para realizar análisis cluster vía el algoritmo del mean shift.

Acknowledgments

En primer lugar me gustaría dar las gracias a mi director José Enrique sin cuyo apoyo, santa paciencia y explicaciones habría sido imposible que llegase a leer esta Tesis. Ha sido continua su comprensión a mis distintos ritmos de trabajo según mi situación laboral. También me gustaría felicitarte desde aquí por tener tantísimos conocimientos sobre Estadística y haber tenido siempre una idea tan clara de hacia dónde se dirigía la Tesis y en qué aspectos íbamos a investigar.

Gracias igualmente a Agustín que me permitió colaborar con él y Paloma para realizar todo el trabajo referente a la simulación y cálculos estadísticos del artículo que estaban desarrollando.

Também gostaria de agradecer ao Carlos. Provavelmente é uma das pessoas mais inteligentes e com grande conhecimento em Estatística que eu conheço. Mas não só isso, também é simples e um excelente hospedeiro. Muito obrigado por fazer as minhas estadias tão fáceis.

Por supuesto, gracias a mi familia y a las frecuentes preguntas de todos ellos de "¿Y cuándo terminas la Tesis entonces?" que han servido para que uno no se distraiga más de lo justamente necesario.

También me gustaría agradecer enormemente el apoyo recibido durante todos estos años a Cristina por haberme ayudado y guiado en todo el proceso

hacia la Tesis al tenerlo ella tan reciente. Seguiré visitando tu despacho para reírnos, tranquila.

También estoy agradecido a José, Javi y, muy especialmente, a Adrián por apoyarnos unos a otros en los momentos complicados tanto en la elaboración de la Tesis como con las complicaciones laborales de todos estos años.

Y no puedo olvidar, obviamente, a la persona gracias a la cual llegué a la Universidad y que hizo todo lo posible por lograr que progresase y me mantuviese en ella. Muchísimas gracias, Mariángeles.

Muchas gracias al resto de profesores que me han hecho más llevadera esta tarea en todos los campus por los que he pasado y, en especial a los de mi primer destino, Mérida. Muchas gracias Eva, José Luis, Emilio y Araceli.

Y por último gracias al recientemente fallecido Carlos Benítez por haber "fundado" un Departamento de Matemáticas así de acogedor donde uno ha podido sentirse tan a gusto durante tantos años.

Y ya puestos, gracias también a mis amigos más cercanos Javi, Jorge y Pilar por animarme estos años.

Gracias a todos.

Contents

0 Preface **1**

0.1 Historical Preface 1

0.2 Structure of the Thesis 3

1 Introduction **1**

1.1 Distribution function estimation 1

1.1.1 The empirical distribution function 1

1.1.2 The kernel distribution function estimator 3

1.1.3 MISE properties of kernel distribution function estimators 5

1.1.4 Optimal bandwidth selection 7

1.1.5 The contribution of this work 8

1.2 Approximation of conditional expectations 9

1.2.1 The problem of computing conditional expectations . 9

1.2.2 A natural Monte Carlo method 10

1.2.3 Applications for the General Half-Normal distribution 11

1.2.4 The contribution of this work 12

1.3 Nonparametric cluster analysis 12

1.3.1 The problem of cluster analysis 12

1.3.2 Clustering methodologies 13

1.3.3 The contribution of this work 14

I Distribution function estimation **16**

2 Fourier methods for smooth distribution function estimation **17**

2.1 Introduction 17
2.2 Main results 19
2.2.1 Limit behavior of the optimal bandwidth sequence . . 19
2.2.2 Sinc kernel distribution function estimator 22
2.3 Numerical examples 24
2.3.1 Example A.1 24
2.3.2 Example A.2 26
2.4 Proofs 27
2.5 References 31

II Conditional expectation approximation **35**

3 On equivariant estimation of the parameters of the general half-normal distribution making use of a Monte Carlo method to approximate conditional expectations **36**

3.1 Introduction 36
3.2 A Monte Carlo method to approximate conditional expectations 38
3.3 Equivariant estimation of the location parameter of the general half-normal distribution 43
3.4 Equivariant estimation of the scale parameter of the general half-normal distribution 50
3.5 References 56

III Cluster analysis **58**

4 A comparison of bandwidth selectors for mean shift clustering **59**

4.1 Introduction 59
4.2 Mean shift clustering 60
4.3 Bandwidth matrix selectors 62
4.4 Simulation study 65

4.5 Conclusion 71
4.6 Appendix 72
4.7 References 73

IV Conclusions and future research **76**

5 Conclusions **77**

6 Future Research **79**
6.1 R package 79
6.2 New methods for bandwidth selection 80
6.3 Integrated regression 80
6.4 Optimal bandwidth selection for other error measures 81
6.5 Fast Fourier Transform 81

A Simulation Programs **82**

Bibliography **109**

List of authors **118**

List of figures **121**

List of tables **123**

0 Preface

0.1 Historical Preface

We can structure this thesis in three different parts: distribution function estimation, conditional expectation calculation with application to equivariant estimation of the parameters of a general half-normal distribution and cluster analysis.

In relation to the first one, in 1956 and 1962 Rosenblatt and Parzen introduce the kernel density estimator and, on the basis of this one, Nadaraya (1964), Tiago de Oliveira (1963) and Watson and Leadbetter (1963) develope the kernel estimator for the distribution function as an alternative to the empirical estimator.

In 1973 Yamato proved in his paper some results about the uniform convergence of the kernel distribution function estimator. These aspects have been studied recently by Giné and Nickl (2009) and Chacón and Rodríguez-Casal (2010).

From that time to today many papers have studied these estimators proposing differents ways to choose the kernel K (Swanepoel (1988) and Jones (1990)) and, mainly, the optimal bandwidth h through cross-validation methods (Sarda (1993), Altman and Léger (1995) and Bowman et al. (1998)), plug-in methods (Altman and Léger (1995), Polansky and Baker (2000), Martins and Tenreiro (2003) and Tenreiro (2003)) and others.

The existence of a optimal bandwidth is proved by Tenreiro (2006) under very mild conditions satisfied for any finite-order kernel. In year 2007,

Chacón, Montanero and Nogales study the case for superkernels which is not included in Tenreiro's paper.

In the second part, the problem of estimating the parameters of the general half-normal distribution is considered. Recall that a half-normal distribution $HN(0,1)$ is the distribution of a random variable $X := |Z|$, where Z has a standard normal distribution. A general half-normal distribution $HN(\xi,\eta)$ is obtained from a half-normal distribution $HN(0,1)$ by a location-scale transformation: $HN(\xi,\eta)$ is the distribution of $Y = \xi + \eta X$. The classical paper Daniel (1959) introduces half-normal plots and the half-normal distribution, a special case of the folded and truncated normal distributions (see Johnson et al. (1994)). Bland et al. (1999) and Bland (2005) propose a so-called half-normal method to deal with relationships between measurement error and magnitude, with applications in medecine. Pewsey (2002) uses the maximum likelihood principle to estimate the parameters of the general half-normal distribution, and presents a brief survey on the general half-normal distribution, its relations with other well-known distributions and its usefulness in the analysis of highly skew data. Pewsey (2004) proposes bias-corrected versions of the maximum likelihood estimators. Nogales et al. (2011) deals with the problem of unbiased estimation for the general half-normal distribution.

Namely, the problem of determining the minimum risk equivariant (MRE) estimators of the location and scale parameters is explored in Chapter 3. Simulation studies are realized to compare the behavior of these estimators with maximum likelihood and unbiased estimators.

A natural probrabilistic Monte Carlo method to compute conditional expectations is used to approximate the MRE estimation of the location parameter because its expression involves two conditional expectations not easily computable. The method is justified by a theorem of Besicovitch (1945, 1946) on differentation of measures.

Cluster analysis, the third part of this thesis, has not been studied widely from a theoretical point of view. This field of the Statistics has been mainly developed by researchers from Computer Science and Statistics in a computational way.

One of the most important problems for this aspect is the lack of a global goal in cluster analysis. This lack of a global goal can be seen in the even today ambiguous and generic definitions of cluster analysis such as the definition given by Hand, Mannila and Smyth in 2001: "partitioning a data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups".

A first method in cluster analysis is the mean shift algorithm introduced by Fukunaga and Hostetler (1975) to estimate the gradient of a multivariate density. But this method is useful for clustering too as Silverman wrote in his known book Density estimation for Statistics and Data Analysis (1986).

This method was used again several years later in the field of Engineering with papers of Cheng (1995), Carreira-Perpiñán (2006) and Comaniciu and Meer (2002).

Due to the mean shift algorithm depends on a good choice of a bandwidth matrix, Chacón and Duong (2013) and Horová, Kolácek and Vopatová (2013) propose several automatic methods for bandwidth selectors.

Nowadays the cluster analysis is a current issue where a lot of scientists are researching due to the ability of existing computers. Since 1981, when W.H.E. Day published his paper "The complexity of computing metric distances between partitions" (Mathematical Social sciences, 1, 269-287), a wide range of clustering methods have been proposed in the literature. But the most of these methods have been focused on clustering for a group of points.

0.2 Structure of the Thesis

The structure of the present Thesis is as follows.

First of all we explain the motivation which have taken us to write this Thesis. Then, the three main articles are included.

In chapter I we present the article titled Fourier methods for smooth distribution function estimation (J.E. Chacón, P. Monfort and C. Tenreiro, Statistics and Probability Letters 84 (2014) 223-230). In this paper, Fourier transforms methods are used to analyzed the asymptotic behavior of kernel distribution function estimators. Morever, exact expressions for the mean

integrated squared error are given in terms of the Fourier transform of the kernel and the characteristic function of the distribution. These expressions are valid with superkernel and the sinc kernel and simulations in these cases are shown.

In chapter II we show the paper On equivariant estimation of the parameters of the general Half-Normal distribution making use of a Monte Carlo method to approximate conditional expectations where we study the problem of estimating the parameters of the general half-normal distribution. Namely, the problem of determining the minimum risk equivariant (MRE) estimators of the parameters is explored. Simulation studies are realized to compare the behavior of these estimators with maximum likelihood and unbiased estimators. A natural Monte Carlo method to compute conditional expectations is used to approximate the MRE estimation of the location parameter because its expression involves two conditional expectations not easily computables. The used Monte Carlo method is justified by a theorem of Besicovitch on differentiation of measures, and has been slightly modified to solve a sort of "curse of dimensionality" problem appearing in the estimation of this parameter.

In chapter II we show the paper A Monte Carlo method to approximate conditional expectations based on a theorem of Besicovitch: application to equivariamt estimation of the parameters of the general half-normal distribution (A.G. Nogales, P. Pérez and P. Monfort, arXiv:1306.1182 (2013)) where a Monte Carlo method is developed to approximate conditional expectation in a probabilistic framework. Examples to evaluate the minimum risk equivariant estimator of the location parameter of a general half-normal distribution are given.

In chapter III we display the paper titled A comparison of bandwidth selectors for mean shift clustering (J.E. Chacón and P. Monfort, arXiv:1310.7855 (2014)). In this paper we analyze the behavior of several automatic bandwidth selectors applied in modal clustering through the mean shift algorithm.

In next section, a brief discussion of the main results in this Thesis is done. Also are included some open questions and future research to develope. Finally, an appendix with the simulation programs used in the papers and a list of the references are shown to conclude the Thesis.

1 Introduction

This thesis is based on three clearly differentiated problems. A previous general framework for all of them will be set up in this chapter, and the contributions made on each of the topics will be described.

1.1 Distribution function estimation

1.1.1 The empirical distribution function

The problem of estimating an unknown distribution function F from a sample $X_1, ..., X_n$ of independent and identically distributed (iid) random variables with a common univariate probability distribution P and distribution function F, has been widely studied. The empirical (cumulative) distribution function is surely the most commonly used estimator. This empirical estimator is defined as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

where I_A is the indicator function for the set A.

It is easy to see that this estimator is consistent. In fact, for a fixed $x \in \mathbb{R}$ it is the minimum variance unbiased estimator of $F(x)$, with

$$\mathbb{E}[F_n(x)] = F(x)$$

and

$$Var[F_n(x)] = \frac{1}{n} F(x)(1 - F(x)).$$

Moreover, the Central Limit Theorem ensures that, for each fixed $x \in \mathbb{R}$, the empirical distribution is asymptotically normally distributed

$$\sqrt{n}\{F_n(x) - F(x)\} \to_d N\big(0, F(x)\{1 - F(x)\}\big).$$

Globally, surely the two most used measures of discrepancy are the uniform distance

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

and the (squared) L_2 distance, also know as Integrated Squared Error (ISE),

$$ISE(F_n) = \int_{\mathbb{R}} \{F_n(x) - F(x)\}^2 dx.$$

The paper by del Barrio, Cuesta-Albertos and Matrán (2000) includes a detailed review of the uses of these two distances in the context of testing for goodness of fit.

The importance of the empirical distribution function is well reflected in the following famous result:

Theorem 1.1 (Glivenko-Cantelli Theorem, 1933)**.** *Given $X_1, \ldots, X_n$ random variables i.i.d. with a common distribution function F. Then*

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \quad a.s.$$

as $n \to \infty$.

The Glivenko-Cantelli theorem tells us about consistency of F_n in $\mathbb{L}^\infty$. A further refinement was shown by Massart (1990), who found a tight value for the constant involved in an inequality discovered by Dvoretzky, Kiefer and Wolfowitz (1956).

Theorem 1.2 (Dvoretzky, Kiefer and Wolfowitz Inequality)**.** *Let be $X_1, \ldots, X_n$ random variables i.i.d. with distribution function F. Then, for each $n \in \mathbb{N}$ and each $\lambda \geq 0$,*

$$P\Big(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \lambda\Big) \leq 2 \exp\{-2\lambda^2\}.$$

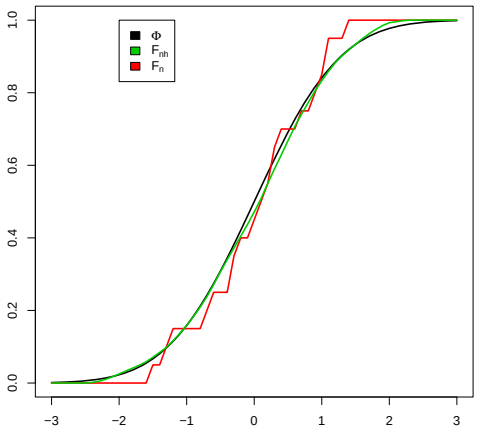

Figure 1.1: Comparison among Φ, F_n and F_{nh}

In relation to the Integrated Squared Error, most existing results deal with its expected value, which is commonly referred to as the Mean Integrated Squared Error (MISE), defined by

$$MISE\{F_n\} = \mathbb{E}\int_{\mathbb{R}} \{F_n(x) - F(x)\}^2 \, dx.$$

It is not hard to show that that

$$MISE\{F_n\} = \frac{1}{n}\int_{\mathbb{R}} F(x)\{1 - F(x)\} \, dx,$$

and hence that $MISE\{F_n\}$ is finite if the distribution has a finite first absolute moment, that is, if the condition $\int_{\mathbb{R}} |x| dF(x) < \infty$ holds.

1.1.2 The kernel distribution function estimator

However, the empirical distribution estimator presents some undesirable properties. One of the most important ones is that F_n is not a continuous function, and this could represent a disadvantage at the time of estimating a continuous distribution function. This problem is illustrated with a simple simulation based on a sample of size $n = 20$ from a standard normal $N(0, 1)$ in Figure 1.1.

The kernel estimator of a distribution function was introduced independently by Tiago de Oliveira (1963), Nadaraya (1964) and Watson and Leadbetter (1964) as a smooth alternative to the discontinuous empirical estimator. It is also represented in Figure 1.1.

To define this estimator we need two previous concepts.

Definition 1.1. *A* kernel k *is an integrable function with* $\int_{\mathbb{R}} k(x)\,dx = 1$.

We will mainly consider kernels which are also densities, namely, $k \geq 0$.

Definition 1.2. *Let* k *be a kernel and* $h > 0$ *a real number and assume that* $X_1, \ldots, X_n$ *are iid continuous random variables with common density* f*. The* kernel estimator *of* f*, with kernel* k *and bandwidth* h*, is defined as*

$$f_{nh}(x) \equiv f_{nh}(x; X_1, \ldots, X_n) = \frac{1}{nh} \sum_{i=1}^{n} k\Big(\frac{x - X_i}{h}\Big).$$

We observe that this estimator, according to the definition of kernel, is a density if $k \geq 0$.

By integrating this estimator we obtain en estimator of the distribution function, which is called kernel distribution function estimator.

Definition 1.3. *According to the previous notations and given the kernel density estimator* f_{nh} *for the density* f*, we define the* kernel distribution function estimator *as*

$$F_{nh}(x) = \int_{-\infty}^{x} f_{nh}(t)\,dt.$$

An equivalent definition for this estimator can be obtained by expanding the above definition.

$$\begin{aligned} F_{nh}(x) &= \int_{-\infty}^{x} f_{nh}(t)\,dt = \int_{-\infty}^{x} \frac{1}{nh} \sum_{i=1}^{n} k\Big(\frac{t - X_i}{h}\Big)\,dt \\ &= \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} k\Big(\frac{t - X_i}{h}\Big)\,dt = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\frac{x - X_i}{h}} k(y)\,dy \\ &= \frac{1}{n} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \end{aligned}$$

where we have used the change of variables $y = \frac{t-X_i}{h}$, and where the function $K(x) = \int_{-\infty}^{x} k(t)\,dt$ will be referred to as an integrated kernel.

So, we have a new equivalent definition for the distribution kernel estimator.

Definition 1.4. *Let us consider an integrated kernel K and a real number $h > 0$, and assume that $X_1, \ldots, X_n$ are iid continuous random variables with common distribution function F. We define the* kernel estimator *of F, with integrated kernel K and bandwidth h as*

$$F_{nh}(x) \equiv F_{nh}(x; X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big).$$

Obviously, when k is itself a density, then K is a proper distribution function so that the kernel estimator F_{nh} also inherits its properties.

The definition of F_{nh} can be extended for the case $h = 0$ following some results of Chacón and Rodríguez-Casal (2010), who showed that $F_{nh} = F_n$ when $h = 0$.

1.1.3 MISE properties of kernel distribution function estimators

The paper of Yamato (1973) contains a detailed study of the consitency properties of the kernel distribution function estimator with respect to the uniform distance.

However, surely due to its mathematical tractability, many more papers deal with kernel distribution function estimators from the point of view of MISE, as for example Azzalini (1981), Jones (1990) and, more recently, Tenreiro (2003), Chacón and Rodríguez-Casal (2010) or Mason and Swanepoel (2012).

Here we will also focus on the $MISE$ as a measure of the performance of the kernel distribution function estimator

$$MISE(h) \equiv MISE\{F_{nh}\} = \mathbb{E}\int_{\mathbb{R}} [F_{nh}(x) - F(x)]^2\,dx$$

Jones (1990) proved that asymptotically, with respect to the $MISE$ criterion, the optimal integrated kernel K is the distribution function corresponding to a uniform distribution:

$$K(x) = \begin{cases} 0 & \text{if} \quad x < -\sqrt{3} \\ \frac{x+\sqrt{3}}{2\sqrt{3}} & \text{if} \quad -\sqrt{3} \leq x < \sqrt{3} \\ 1 & \text{if} \quad \sqrt{3} \leq x \end{cases}$$

However, there are not significative differences between this integrated kernel and others, like the Gaussian one, given by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt, \qquad x \in \mathbb{R}$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}$, corresponding to the standard normal distribution.

Therefore, the problem is not an appropiate selection of the kernel K but choosing an optimal bandwidth h to minimize the $MISE$.

For a better understanding of the role that h plays in the performance of F_{nh}, it is helpful to decompose the $MISE$ as the sum of the integrated variance and the integrated squared bias: since the integrand is not negative, we can exchange the order of integration and expectation by applying the Fubini Theorem, obtaining

$$\begin{aligned} MISE(h) &= \mathbb{E}\Big[\int_{\mathbb{R}} \{F_{nh}(x) - F(x)\}^2\, dx \Big] = \int_{\mathbb{R}} \mathbb{E}[\{F_{nh}(x) - F(x)\}^2]\, dx \\ &= \int_{\mathbb{R}} Var\{F_{nh}(x)\}\, dx + \int_{\mathbb{R}} \{\mathbb{E}[\{F_{nh}(x)] - F(x)\}^2\, dx \\ &= IV(h) + ISB(h) \end{aligned}$$

where $IV(h)$ and $ISB(h)$ are called integrated variance and integrated squared bias.

Hence, given F, K and n, we obtain a real function

$$MISE : [0, \infty) \subset \mathbb{R} \to \mathbb{R}.$$

1.1.4 Optimal bandwidth selection

Let suppose that there is a bandwidth h_{0n} that minimizes this $MISE$ function; that is,

$$MISE(h_{0n}) \leq MISE(h), \quad \forall h > 0.$$

The value $h = h_{0n} \in (0, \infty)$ is the optimal value to use in the estimator F_{nh} in order to estimate the distribution. Though, it is obvious that this bandwidth depends on n and F, which is unknown from a statistical point of view. Then, the problem for choosing an optimal bandwidth is equivalent to the problem of the distribution estimation.

Regarding the empirical distribution function F_n, we showed before that $F_{nh} = F_n$ when $h = 0$. This correspondence also hold for the $MISE$, since

$$MISE(0) = \frac{1}{n} \int_{\mathbb{R}} F(x)[1 - F(x)]\, dx = MISE\{F_n\}.$$

In Figure 1.2 the function $MISE(h)$ is shown for the case where $n = 20$, $F = \Phi$ and K is the integrated kernel corresponding to a standard Gaussian distribution. The figure clearly shows that the error of the empirical estimator ($MISE(0)$) could be improved with a good choice of h. Moreover, the estimator F_{nh} has an error smaller than that of F_n not only for h_{0n}, but also for a wide range of values of h.

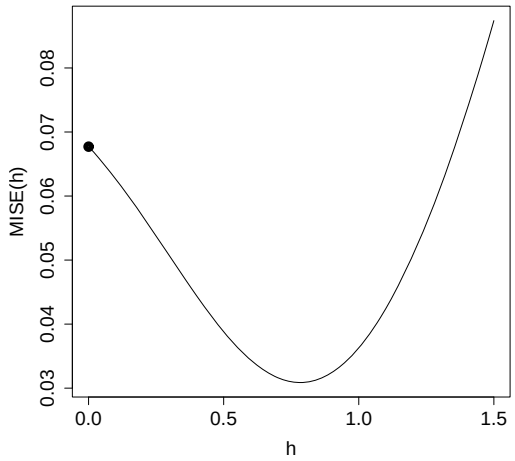

Figure 1.2: The $MISE$ as a function of the bandwidth

1.1.5 The contribution of this work

The paper by Tenreiro (2006) contains a precise result about the existence and asymptotic properties of the optimal bandwidth h_{0n}. His result, however, cannot be applied to the case where superkernels are used. Superkernels are defined as kernels whose Fourier transform is identically equal to one in a neighbourhood of the origin, and they are known to produce rate-adaptive kernel density estimators, as shown in Devroye (1992) and Chacón, Montanero and Nogales (2007).

So we focused on the goal of extending Tenreiro's (2006) result and finding a generalization that could cover the case of a superkernel as well. This goal is achieved in Chapter 2.

The theoretical tools that we use to achieve that goal are based on Fourier transforms. This techniques are relatively straightforward to apply in the density case, by assuming that the density function is square integrable, but much more delicate arguments are needed to obtain valid expressions, in terms of characteristic functions, for the $MISE$ of the kernel distribution function estimator, since it makes no sense to assume that a distribution function is square integrable (no distribution function can satisfy such a requirement).

By expressing the $MISE$ in terms of characteristic functions we also obtained a result about the limir behavior of the optimal bandwidth sequence in its most general form, including the possibility of using superkernels and the sinc kernel in kernel distribution function estimation.

As a consequence, we also showed that there exists a class of distributions for which the kernel distribution estimator presents a first-order improvement over its empirical counterpart, opposite to the usual situation, where only second-order improvements are possible.

1.2 Approximation of conditional expectations

1.2.1 The problem of computing conditional expectations

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ be an n-dimensional random variable and $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ a random variable with finite mean.

In this context, we can define the conditional expectation $E(Y|X)$ as follows.

Definition 1.5. *The conditional expectation $E(Y|X)$ is defined as a random variable on $\mathbb{R}^n$ such that $\int_{X^{-1}(B)} Y\, dP = \int_B E(Y|X) dP^X$ for any Borel set B in $\mathbb{R}^n$, where P^X denotes the probability distribution of X.*

Although the existence of the conditional expectation is guaranteed via the Radon-Nikodym theorem, its computation is, generally, involved.

When the joint density f of Y and X is known, $E(Y|X = x)$ is the mean of the conditional distribution $P^{Y|X=x}$ of Y given $X = x$, whose density is $f(x, y)/f_X(x)$, where f_X denotes the marginal distribution of X. In this case the problem of computing a conditional expectation reduces to evaluating a mean, and many methods exist which can be used to do so.

The main problem appears when a joint density for X and Y is not available, or it is difficult to determine. In this case the problem of evaluating the conditional expectation can become an arduous problem. Nevertheless, this is still an interesting problem due to $y = E(Y|X = x)$ is the regression curve of Y given $X = x$.

1.2.2 A natural Monte Carlo method

We describe a natural Monte Carlo method, inspired by a Besicovitch theorem for the differentiation of measures, to evaluate such a conditional expectation in a probabilistic setting (see, for instance, Corollary 2.14 of Mattila (1995)). This theorem extends to Radon measures the classical Lebesgue Differentation Theorem. The method is really useful when densities are not available or they are not easy to compute.

Given two real random variables X and Y, the mentioned Monte Carlo method of approximation of the conditional expectation $E(Y|X = x)$ is based on the naive idea that one can approximate it from a sample $(x_i, y_i)_{1\leq i\leq n}$ by the mean of the y_i corresponding to points x_i lying in a narrow neighborhood of x. Despite the simplicity of the argument, and to the best of our knowledge, this result has not been described this way in the literature. When the joint density of X and Y is known, $E(Y|X = x)$ is the mean of the conditional distribution of Y given $X = x$, and the problem of compute a conditional expectation is reduced to the problem of computing a mean. At this point, it must be noticed that the mentioned Monte Carlo method does not rely on the existence of a joint density and could be specially useful to approximate conditional expectations when densities are not available (or are not easy to compute).

Although its nature is probabilistic, in a statistical framework we can provide additional guarantees on the method, since the obtained Monte Carlo approximation to the conditional expectation $E(Y|X = x)$ coincides with the value at the point x of the kernel estimator, introduced by Nadaraya and Watson, of the regression curve $y = E(Y|X = x)$ for the kernel $K(x) = I_{[-1,1]}(x)$ (see Nadaraya (1989), p. 115). From this point of view, ϵ plays the role of the bandwidth parameter. We refer to Härdle (1992, Ch. 5) for a detailed discussion on the important problem of the choice of the bandwidth. This way we establish a connection with the first chapter of this thesis.

1.2.3 Applications for the General Half-Normal distribution

The method is applied for us to evaluate the minimum risk equivariant estimator of the location parameter of a general half-normal distribution. This estimator is described in terms of two conditional expectations for known values of the location and scale parameters and its value is approximated using simulation.

We remind that a general half-normal distribution depends on two parameters as follows:

Definition 1.6. *Let Z denote a standard normal random variable. Then, $Y = \xi + \eta X$, where $X = |Z|$, is a general half-normal random variable with location and scale parameters ξ and η, respectively, and density*

$$f(y) = \frac{2}{\eta}\phi\left(\frac{y-\xi}{\eta}\right) = \frac{b}{\eta}exp\left\{\frac{-(y-\xi)^2}{2\eta^2}\right\}, y > \xi, -\infty < \xi < \infty, \eta > 0,$$

where $b = (2/\pi)^{1/2}$ and $\phi(\cdot)$ denotes the standard normal density.

As mentioned above, the main application of the Monte Carlo method for the approximation of conditional expectations is approximating the estimation of the location parameter of the general half-normal distribution, because it is defined in terms of a quotient of two not-easily-computable parameter-free conditional expectations given a $n-1$-dimensional statistic U. Some problems about "curse of dimensionality" appear when n is large because, in this case, it is not easy to find large samples of points lying in a small ball centered at a point $U(y)$. This is the reason to modify the Monte Carlo method for the approximation of conditional expectations taking advantage of the underlying distribution of Y (the general half-normal distribution) and the invariance properties of U. This could become an important scholium of the paper, as the ideas used here could be useful to deal with the "curse of the dimensionality" in similar situations.

Finally, we develop an explicit expression for the minimum risk equivariant estimator of the parameter η of a general half-normal distribution.

Although less interesting from the point of view of applications, for the sake of completeness, the MRE estimators of both parameters ξ and η are given when the other parameter is known. As far as we know, this had not been done before.

1.2.4 The contribution of this work

A Monte Carlo method to approximate conditional expectations in a probabilistic framework is motivated by a general result inspired by the Besicovitch covering theorem for differentiation of measures. The method is specially useful when densities are not available or are not easy to compute. The method is illustrated through various examples and can also be used in a statistical setting to approximate the conditional expectation given a sufficient statistic. In this thesis it is used to compute the minimum risk equivariant estimator (MRE) of the location parameter of a general half-normal distribution since this estimator is described in terms of a conditional expectation for known values of the location and scale parameters. For the sake of completeness, an explicit expression for the minimum risk equivariant estimator of the scale parameter is given. As far as we are aware, these estimators have not appeared before in the literature. Simulation studies to compare the behavior of the new estimators with those of maximum likelihood and unbiased estimators are presented in this thesis.

1.3 Nonparametric cluster analysis

1.3.1 The problem of cluster analysis

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939 and famously used by Cattell beginning in 1943 to treat theory classification in personality psychology.

According to the different notions for distance (Euclidean, Manhattan, Mahalanobis...) or cluster and other concepts and desired properties, a lot of different methods for cluster analysis can be developed.

1.3.2 Clustering methodologies

A rough classification of clustering methodologies can be made by grouping the different methods into three types: hierarchical clustering, centroid-based clustering and density-based clustering.

Strategies for hierarchical clustering generally fall into two types: agglomerative (each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy) and divisive (all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy). The results of hierarchical clustering are usually presented in a dendrogram.

When the number of final clusters is known, the centroid-based clustering methods are used. The most important representative of this group of methods is surely the k-means algorithm, although more robust variants have been developed; see, for instance, the excellent recent survey of García-Escudero *et al.* (2010).

Due to the fast development of computer power, a lot of methods have been created in recent years improving existing ones. They are necessary in a context of high-dimensional data clustering. Some of this recent methods focus on distribution-based clustering or density-based clustering, meaning that the whole distribution (as opposed to centroids) are involved in the clustering procedure.

The mean shift algorithm was introduced in 1975 by Fukunaga and Hostetler with the goal of estimating the gradient of a multivariate density. Later, Silverman (1986) highlighted this algorithm as an important application of kernel smoothing. But it was not studied again until years later through several authors such as Cheng (1995), Carreira-Perpiñán (2006) and Comaniciu and Meer (2002).

The algorithm can be used to iteratively shift each data point towards a local maximum of a kernel estimator of density, and two data points are then identified to belong to the same cluster if they have been moved to the same local maximum. This way, the mean shift algorithm induces a clustering of the data in a nonparametric way, based on a kernel estimator of the density.

1.3.3 The contribution of this work

Recently, several authomatic methods for bandwidth selection have been proposed for density gradient estimation; see Chacón and Duong (2013) and Horová, Kolácek and Vopatová (2013).

On the other hand, Chacón (2012) set up a population background for density-based clustering in which the ideal population goal is clearly identified, and also several distances between clusterings are proposed to measure the performance of a clustering method.

So our goal was to compare the newly introduced bandwidth selection methods for density gradient estimation, but in the context of mean shift clustering, making use of the recently proposed loss measures for density-based clustering to evaluate their performance. This goal is achieved in Chapter 4.

We use these automatic bandwidth selectors in order to obtain the clustering of the space that they induce via the mean shift algorithm. Then, we measure the distance between both clusterings of $\mathbb{R}^d$, the clustering obtained with each automatic bandwidth selector and the ideal clustering of the density.

We analyze the performance of several bandwidth selectors using the following distance between two clusterings introduced in Chacón (2012): given two clusterings $C = \{C_1, ..., C_r\}$ and $D = \{D_1, ..., D_s\}$ of a probability measure P, with $r \leq s$, the distance in measure between them is defined as

$$d_P(C, D) = \frac{1}{2} \min_{\sigma \in P_s} \sum_{i=1}^{s} P(C_i \triangle D_{\sigma(i)})),$$

where P_s denotes the set of permutations of $\{1, 2, ..., s\}$, the partition C has been enlarged by adding $s - r$ empty sets $C_{r+1} = ... = C_s = \emptyset$ if necessary, and $\triangle$ denotes the symmetric difference between two sets, namely $C \triangle D = (C \cap D^c) \cup (C^c \cap D)$.

As a result of our study, we conclude that none of the ten automatic bandwidth matrix selectors included in the study showed a consistent superior performance over the rest of the methods.

All the cross-validation, plug-in, smoothed cross-validation proposals, and the iterative method, are best for one of the models considered in our

study, but utterly fail to identify the cluster structure for one, two or even three of the remaining ones. This suggests that the problem of bandwidth selection for mean shift clustering, though related, is different from that of bandwidth selection for density gradient estimation, and presents its own peculiarities.

Part I

Distribution function estimation

2

Fourier methods for smooth distribution function estimation

2.1 Introduction

The kernel estimator of a distribution function was introduced independently by Tiago de Oliveira (1963), Nadaraya (1964) and Watson and Leadbetter (1964) as a smooth alternative to the empirical estimator. It is defined as the distribution function corresponding to the well-known kernel density estimator. Precisely, given independent real random variables $X_1, \ldots, X_n$ with common and unknown distribution function F, assumed to be absolutely continuous with density function f, the kernel estimator of $F(x)$ is

$$F_{nh}(x) = n^{-1} \sum_{j=1}^{n} K\big(h^{-1}(x - X_j)\big),$$

where $h > 0$ is the and the function K will be referred to as the integrated kernel, since it is assumed that $K(x) = \int_{-\infty}^{x} k(y)dy$ for some integrable function k, called kernel, having unit integral over the whole real line.

Classical references on kernel distribution function estimators include Yamato (1973), which provided mild necessary and sufficient conditions for its consistency in uniform norm, Azzalini (1981), Swanepoel (1988) and Jones (1990) on asymptotic squared error analysis of the estimator, or Sarda (1993), Altman and Léger (1995) and Bowman, Hall and Prvan (1998), and more recently Polansky and Baker (2000) and Tenreiro (2006), on data-driven bandwidth selection. There are also other recent papers

on different aspects of kernel distribution function estimation, like Tenreiro (2003), Swanepoel and Van Graan (2005), Janssen, Swanepoel and Veraverbeke (2007), Giné and Nickl (2009), Berg and Politis (2009), Chacón and Rodríguez-Casal (2010) or Mason and Swanepoel (2012). See Servien (2009) for a detailed survey on distribution function estimation, not limited to .

This paper is devoted to the study of the kernel distribution function estimator from the point of view of the mean integrated squared error,

$$\mathrm{MISE}(h) \equiv \mathrm{MISE}_n(h) = \mathbb{E} \int_{-\infty}^{\infty} \{F_{nh}(x) - F(x)\}^2 dx.$$

In this sense, the optimal bandwidth h_{0n} is the value of $h > 0$ minimizing $\mathrm{MISE}(h)$. The existence of such a bandwidth was proved in Theorem 1 of Tenreiro (2006) under very general assumptions, and Proposition 2 in the same paper showed that $h_{0n} \to 0$ whenever the Fourier transform of k is not identically equal to 1 on any neighbourhood of the origin. This condition can be considered mild as well, since it is satisfied for any finite-order kernel; however, it does not hold for a superkernel (see Chacón, Montanero and Nogales (2007)).

The purpose of this note is to show how to use Fourier transform techniques for the analysis of kernel distribution estimators. Particularly, expressing the MISE in terms of characteristic functions allows us to obtain a result on the limit behavior of the optimal bandwidth sequence in its most general form so that it also covers the case of a superkernel, and to explore its consequences showing the peculiar properties of the use of superkernels and the sinc kernel in kernel distribution function estimation. Precisely, it is shown in Section 2.2 that in some situations the sequence h_{0n} does not necessarily tend to zero. Moreover, we exhibit a class of distributions for which the kernel distribution estimator presents a first-order improvement over its empirical counterpart, opposite to the usual situation, where only second-order improvements are possible (see Remark 2.3). Our findings are illustrated in Section 2.3 through two representative examples.

2.2 Main results

Recall from Chacón and Rodríguez-Casal (2010) that the kernel distribution function estimator admits the representation

$$F_{nh}(x) = \int F_n(x - hz)dK(z), \tag{2.1}$$

where F_n denotes the empirical distribution function (here and below integrals without integration limits are meant over the whole real line). Using this, and standard properties of the empirical process, it is possible to obtain a decomposition of $\mathrm{MISE}(h) = \mathrm{IV}(h) + \mathrm{ISB}(h)$, where the integrated variance $\mathrm{IV}(h) = \int \mathrm{Var}\{F_{nh}(x)\}dx$ and the integrated squared bias $\mathrm{ISB}(h) = \int\{\mathbb{E}[F_{nh}(x)] - F(x)\}^2dx$ can be expressed in the following exact form:

$$\mathrm{IV}(h) = n^{-1}\iiint \big\{F\big(x - h(y \vee z)\big) - F(x - hy)F(x - hz)\big\}dK(y)dK(z)dx, \tag{2.2}$$

$$\mathrm{ISB}(h) = \iiint \{F(x - hy) - F(x)\}\{F(x - hz) - F(x)\}dK(y)dK(z)dx, \tag{2.3}$$

with $y \vee z$ standing for $\max\{y, z\}$.

Note that the representation (2.1) and the exact expressions (2.2) and (2.3) also make sense for $h = 0$, implying that the kernel distribution estimator reduces to the empirical distribution function for $h = 0$, for which the well-known MISE formula reads $\mathrm{MISE}(0) = \mathrm{IV}(0) = n^{-1}\int F(1 - F)$ whenever $\psi(F) = \int F(1 - F)$ is finite. Moreover, it is not hard to check that $\int |x|dF(x) < \infty$ and $\int |y\,k(y)|dy < \infty$ ensure that $\mathrm{MISE}(h)$ is finite for all $h > 0$, so those two minimal conditions will be assumed henceforth. Note that the required condition that F have a finite mean is slightly stronger than $\psi(F) < \infty$ since $\psi(F) \leq 2\int |x|dF(x)$.

2.2.1 Limit behavior of the optimal bandwidth sequence

Denote by φ_g the Fourier transform of a function g, defined as $\varphi_g(t) = \int e^{itx}g(x)dx$. As in Chacón, Montanero and Nogales (2007), the key to understand the limit behavior of the optimal bandwidth sequence is to use

Fourier transforms to express the MISE criterion. Abdous (1993) provided a careful account of the necessary conditions under which the MISE can be expressed in terms of Fourier transforms. The proof of his Proposition 2 implicitly derives formulas for ISB(h) and IV(h) in terms of φ_k and φ_f for $h > 0$. We reproduce this result here for completeness, and show that it can be extended to cover the case $h = 0$ as well.

Theorem 2.1. *If* $\int |x| dF(x) < \infty$ *and* $\int |y\, k(y)| dy < \infty$ *then, for all* $h \geq 0$, *the IV and ISB functions can be written as*

$$\mathrm{IV}(h) = (2\pi)^{-1} n^{-1} \int t^{-2} |\varphi_k(th)|^2 \{1 - |\varphi_f(t)|^2\} dt,$$
$$\mathrm{ISB}(h) = (2\pi)^{-1} \int t^{-2} |1 - \varphi_k(th)|^2 |\varphi_f(t)|^2 dt.$$

Particularly, note that for $h = 0$ the previous result yields a Parseval-like formula for distribution functions,

$$\psi(F) = \int F(1 - F) = (2\pi)^{-1} \int t^{-2} \{1 - |\varphi_f(t)|^2\} dt, \tag{2.4}$$

which can be useful to compute errors in an exact way in cases where F does not have a close expression but φ_f does, as it happens for instance for the normal distribution (see also Section 2.3 below). Moreover, we show in Lemma 2.1 that (2.4) remains valid for integrated kernels K. In the following it will be assumed that $\psi(K) > 0$, a property that immediately holds, using (2.4), whenever $\varphi_k(t) \in [0, 1]$ for all t. Note that, for density estimation, admissible kernels are precisely those whose Fourier transform satisfies that restriction (see (1988)).

The limit behavior of the optimal bandwidth sequence h_{0n} is determined in its greatest generality by the following constants, depending on the Fourier transforms of f and k: let C_f denote the smallest positive frequency from which φ_f is null along a proper interval and D_f the positive frequency from which φ_f is identically null (so that $C_f \leq D_f$, both possibly being infinite); also, denote S_k the greatest frequency such that φ_k is identically equal to one on $[0, S_k]$ and T_k the smallest frequency such that φ_k is not identically

equal to one on a subinterval of $[T_k, \infty)$, and note that $S_k \leq T_k$ with both possibly being zero. In mathematical terms,

$$\begin{aligned} C_f &= \sup\{r \geq 0 \colon \varphi_f(t) \neq 0 \text{ a.e. for } t \in [0, r]\} \\ D_f &= \sup\{t \geq 0 \colon \varphi_f(t) \neq 0\} \\ S_k &= \inf\{t \geq 0 \colon \varphi_k(t) \neq 1\} \\ T_k &= \inf\{r \geq 0 \colon \varphi_k(t) \neq 1 \text{ a.e. for } t \geq r\} \end{aligned}$$

Finally, define $h_* = \sup\{h \geq 0 \colon \mathrm{ISB}(h) = 0\}$. The following result shows the limit of the optimal bandwidth sequence h_{0n} in the common case where $C_f = D_f$ and $S_k = T_k$.

Theorem 2.2. *Assume that* $\int |x| dF(x) < \infty$, $\int |y\, k(y)| dy < \infty$ *and* $\psi(K) > 0$, *and suppose that* $C_f = D_f$ *and* $S_k = T_k$. *Then,* $h_{0n} \to S_k/D_f$ *as* $n \to \infty$ *and also* $h_* = S_k/D_f$.

A number of consequences can be extracted from Theorem 2.2:

Remark 2.1. *A kernel* k *with* $S_k > 0$ *is called a superkernel (see Chacón, Montanero and Nogales (2007)). If an integrated superkernel is used in the kernel distribution function estimator and the density* f *is such that* $D_f < \infty$ *(see Chacón, Montanero and Nogales (2007), and Section 2.3 below for examples of such distributions) then, contrary to the usual situation, the optimal bandwidth sequence* h_{0n} *does not tend to zero, but to the strictly positive constant* S_k/D_f. *Moreover, any positive constant can be the limit of an optimal bandwidth sequence, because modifying the scale of the density by taking* $f_a(x) = f(x/a)/a$, *for any* $a > 0$, *it follows that* $D_{f_a} = D_f/a$, *and hence the limit of the optimal bandwidth sequence equals* aS_k/D_f.

Remark 2.2. *Since* $h_* = S_k/D_f$, *the kernel estimator* F_{nh} *is unbiased for any fixed (i.e., not depending on* n*) choice of* $h \in [0, S_k/D_f]$. *If either* K *is not an integrated superkernel or the characteristic function* φ_f *does not have bounded support, then the only kernel distribution estimator with null ISB corresponds to* $h = 0$, *the empirical distribution function.*

Remark 2.3. *It is shown in the proof of Theorem 2.2 that for* $h \in [0, S_k/D_f]$ *the MISE of* F_{nh} *admits the exact expression* $\mathrm{MISE}(h) = n^{-1}\psi(F) - n^{-1}\psi(K)h$. *From this, it follows that for any fixed* $h \in (0, S_k/D_f]$ *the kernel estimator* F_{nh} *presents an asymptotic first-order reduction in MISE over the empirical estimator; that is, its MISE is of order* n^{-1} *as for the empirical estimator, yet with a strictly smaller constant (namely,* $\psi(F) - \psi(K)h < \psi(F)$*). As a result, over the class of distributions with* D_f *bounded by a constant (say,* $D_f \leq M$*) the kernel estimator with bandwidth* $h = S_k/M$ *is strictly more efficient than the empirical distribution function* F_n*. This is in contrast with the more common case (i.e.,* $S_k = 0$ *or* $D_f = \infty$*) where it is well-known that the asymptotic improvement of* F_{nh} *over* F_n *is only of second order, in the sense that* $\mathrm{MISE}(h_{0n})$ *admits the asymptotic representation* $n^{-1}\psi(F) - cn^{-p} + o(n^{-p})$ *for some* $p > 1$ *and* $c > 0$ *(see, e.g., Jones (1990), and Shao and Xiang (1997)).*

2.2.2 Sinc kernel distribution function estimator

In this section we consider the sinc kernel, defined by $\mathrm{sinc}(x) = \sin(x)/(\pi x)$ for $x \neq 0$ and $\mathrm{sinc}(0) = 1/\pi$. This function is not integrable, so the sinc kernel density estimator inherits this undesirable property, but such a defect can be corrected as described in Glad, Hjort and Ushakov (2003). Nevertheless, the sinc kernel is square integrable, and as such the sinc kernel density estimator achieves certain optimality properties with respect to the MISE (Davis, 1977), that make the sinc kernel useful for density estimation (see Glad, Hjort and Ushakov (2007), or Tsybakov (2009), Section 1.3).

Abdous (1993, Section 3) provided a careful study showing that it also makes sense to use the MISE criterion for kernel distribution function estimators based on the integrated sinc kernel. However, it is not so clear from his developments how the sinc kernel distribution function estimator is explicitly defined, nor the asymptotic properties of the optimal bandwidth sequence in this case, since Theorem 2.2 above can not be directly applied given that the sinc kernel is not integrable. This section contains a detailed treatment of these issues.

First, note that the definition of the integrated kernel $K(x) = \int_{-\infty}^{x} \mathrm{sinc}(z)dz$ has to be understood in the sense of Cauchy principal value, i.e. $K(x) =$

$\lim_{M\to\infty}\int_{-M}^{x}\operatorname{sinc}(z)dz$, because the integral is not Lebesgue-convergent. A simpler way to express such principal value is $K(x) = \frac{1}{2} + \operatorname{Si}(x)$, where $\operatorname{Si}(x) = \int_0^x \operatorname{sinc}(z)dz$ is the sine integral function (with the usual convention that $\int_a^b = -\int_b^a$ if $b < a$). This yields the following explicit form for the sinc kernel distribution function estimator:

$$F_{nh}^{\text{sinc}}(x) = \tfrac{1}{2} + n^{-1}\sum_{j=1}^{n}\operatorname{Si}\big(h^{-1}(x - X_j)\big). \tag{2.5}$$

An alternative, and perhaps more natural, derivation of (2.5) is found through the use of inversion formulas. The sinc kernel density estimator with bandwidth $h = 1/T$ is readily obtained from the inversion formula $f(x) = (2\pi)^{-1}\int e^{-itx}\varphi_f(t)dt$ by replacing φ_f with the empirical characteristic function $\varphi_n(t) = n^{-1}\sum_{j=1}^{n} e^{itX_j}$, conveniently truncated to get a finite integral $(2\pi)^{-1}\int_{-T}^{T} e^{-itx}\varphi_n(t)dt$ (see for instance Chiu, 1992, p. 774). An inversion formula relating F and φ_f is the so-called Gil-Pelaez formula $F(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty t^{-1}\Im\{e^{-itx}\varphi_f(t)\}dt$, with $\Im\{z\}$ standing for the imaginary part of a complex number z, which is valid for a continuous F in the principal value sense (Gurland, 1948). Reasoning as before, replacing φ_f with φ_n and restricting the domain of integration to $[0, 1/h]$, results in the same sinc kernel distribution function estimator shown in (2.5).

As a square integrable function, the Fourier transform of the sinc kernel is the indicator function of the interval $[-1, 1]$. In this sense, Abdous (1993) showed that the IV and ISB formulas of Theorem 2.1 above remain valid for the sinc kernel distribution estimator, as long as the square integrability of f is added to its assumptions, leading to the following simple exact MISE formula for $h > 0$:

$$\operatorname{MISE}(h) = (n\pi)^{-1}\int_0^{1/h} t^{-2}\{1 - |\varphi_f(t)|^2\}dt + \pi^{-1}\int_{1/h}^{\infty} t^{-2}|\varphi_f(t)|^2 dt. \tag{2.6}$$

Straightforward differentiation shows that the critical points of such a MISE function are located at any value $h_\diamond$ such that $|\varphi_f(1/h_\diamond)|^2 = (n+1)^{-1}$. This does not reveal, however, if such critical points are local minima or maxima. The following result shows the existence of a global minimizer h_{0n} of (2.6), and that Theorem 2.2 above remains valid for the sinc kernel estimator.

Note, again, that Theorem 2 of Tenreiro (2006) on the existence of h_{0n} can not be directly applied here because it relies on the assumption that the kernel function is integrable.

Theorem 2.3. *Assume that f is square integrable and $\int |x| dF(x) < \infty$. Then, there exists a bandwidth h_{0n} that minimizes the MISE of the sinc kernel distribution function estimator. Moreover, if $C_f = D_f$ then $h_{0n} \to 1/D_f$ as $n \to \infty$ and also $h_* = 1/D_f$.*

If it were integrable, the sinc kernel could be considered as a superkernel with $S_{\text{sinc}} = T_{\text{sinc}} = 1$, so from Theorem 2.3 it follows that all the remarks above about the limit behavior of h_{0n} and the optimal MISE for superkernel distribution function estimators can be equally applied to the sinc kernel distribution estimator.

2.3 Numerical examples

In this section we present some examples to further illustrate the usefulness and consequences of Theorems 2.1, 2.2 and 2.3 above.

2.3.1 Example A.1

In this example we consider the so-called Jackson-de la Vallé Poussin distribution F, with density function

$$f(x) = \frac{3}{4\pi}\left(\frac{\sin(x/2)}{x/2}\right)^4 = \frac{9 + 3\cos(2x) - 12\cos(x)}{2\pi x^4}$$

and whose characteristic function is shown in and (1971, p. 516) to be

$$\varphi_f(t) = \begin{cases} 1 - 3t^2/2 + 3|t|^3/4, & |t| \leq 1 \\ (2 - |t|)^3/4, & 1 \leq |t| \leq 2 \\ 0, & |t| \geq 2 \end{cases},$$

which implies that $C_f = D_f = 2$.

As shown in Theorems 2.2 and 2.3, since $C_f = D_f < \infty$ this distribution (or any of its rescalings $F_a(x) = F(x/a)$ with $a > 0$) represents a case where superkernel distribution function estimators are asymptotically more

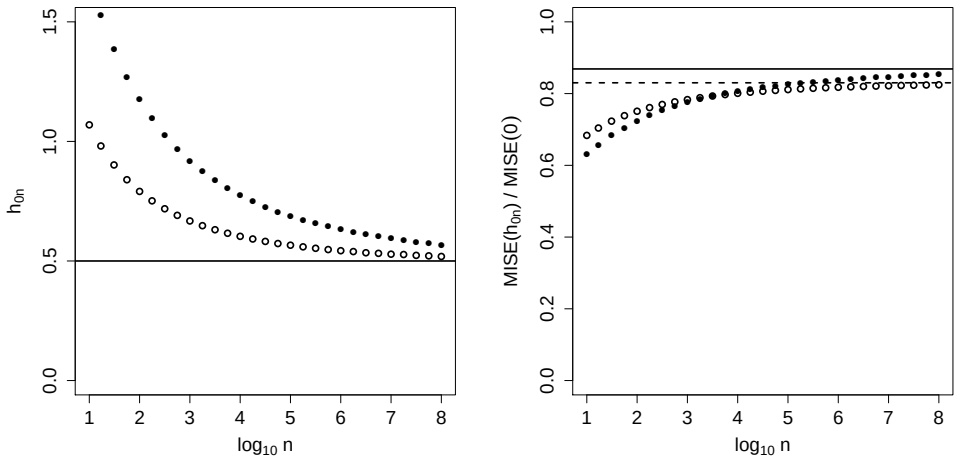

Figure 2.1: Optimal bandwidth sequence (left) and relative efficiency in MISE (right) for the estimation of the Jackson-de la Vallé Poussin distribution, as a function of $\log_{10} n$. The lines show the limit values. Solid circles and solid lines correspond to the trapezoidal superkernel and open circles and dashed lines correspond to the sinc kernel.

efficient than the empirical distribution function. To illustrate this fact, we include here a numerical comparison using two different superkernels: the sinc kernel and a proper superkernel, the trapezoidal superkernel given by $k(x) = (\pi x^2)^{-1}\{\cos x - \cos(2x)\}$, for which $S_k = T_k = 1$ (see Chacón, Montanero and Nogales, 2007).

It is not hard from Theorem 2.1 (for the trapezoidal kernel) and (2.6) (for the sinc kernel) to come up with an explicit formula for the exact MISE function in each case. These exact MISE calculations allow to numerically compute the optimal bandwidth sequences h_{0n} and the minimum MISE values. The optimal bandwidth sequences for both superkernel estimators are shown in Figure 3.2 (left) as a function of $\log_{10} n$, where it is already noticeable that they both have limit $1/2$, as predicted from theory.

The right graph in Figure 3.2 shows the relative efficiency of both superkernel estimators using optimal bandwidths with respect to the empirical estimator, namely $\mathrm{MISE}(h_{0n})/\mathrm{MISE}(0)$, together with their asymptotic values, given by $\mathrm{MISE}(S_k/D_f)/\mathrm{MISE}(0) = 1-\psi(K)S_k/\{\psi(F)D_f\}$. Using (2.4) it follows that $\psi(F) = (96\log 2 - 43)/(8\pi)$, and $\psi(K)$ equals $(4\log 2 - 2)/\pi$ and $1/\pi$ for the trapezoidal and the sinc kernel, respectively, resulting in asymptotic relative efficiencies of approximately 0.87 and 0.83 for the two

superkernel estimators, as reflected on Figure 3.2. For this distribution, the trapezoidal kernel is more efficient than the sinc kernel up to about sample size $n = 3000$, but asymptotically the sinc kernel is slightly more efficient. Both are markedly more efficient than the empirical distribution as was to be expected from asymptotic theory; besides, the gains are even more substantial for low and moderate sample sizes.

2.3.2 Example A.2

In this second example we make use of the MISE expressions in terms of characteristic functions to obtain exact MISE formulas for the case-study in which F corresponds to the $N(0, \sigma^2)$ distribution and the integrated kernel is either the standard normal distribution function Φ, or the integrated sinc kernel, and we compare both estimators.

For this specific example the exact MISE formula for the density estimation problem was provided in Fryer (1976) making use the convolution properties of the normal density function, which are also useful for deriving many other integral results for the normal density and its derivatives (see Aldershof et al., 1995).

However, convolution techniques seem to be of little use to find exact MISE expressions for kernel distribution function estimators in the normal case, where not even the estimation goal F has an explicit formula. For this problem, it is convenient to work with exact expressions in terms of characteristic functions. For instance, using (2.4) it immediately follows that the MISE for the empirical distribution function equals $n^{-1}\pi^{-1/2}\sigma$ and, similarly, it is not hard to show that for the kernel estimator with the normal kernel

$$\pi^{1/2}\mathrm{MISE}(h) = n^{-1}\{(h^2+\sigma^2)^{1/2} - h\} + \{(2h^2+4\sigma^2)^{1/2} - (h^2+\sigma^2)^{1/2} - \sigma\}$$

and with the sinc kernel

$$\pi\mathrm{MISE}(h) = (1+n^{-1})\big\{he^{-\sigma^2/h^2} + 2\sigma\sqrt{\pi}\Phi\big(\sigma\sqrt{2}/h\big)\big\} - n^{-1}h - (2+n^{-1})\sigma\sqrt{\pi}.$$

In Figure 3.3 we show the relative efficiency $\mathrm{MISE}(h_{0n})/\mathrm{MISE}(0)$ as a function of $\log_{10} n$ for $\sigma = 1$ for both kernel estimators with respect to the

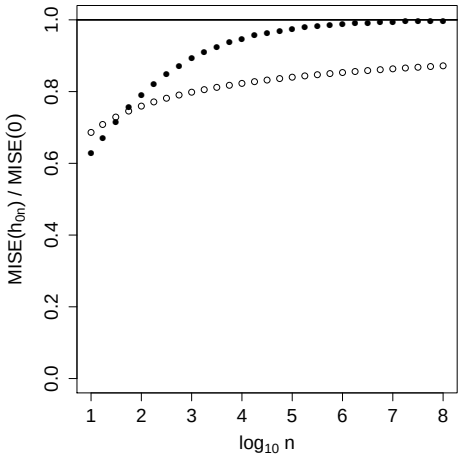

Figure 2.2: Relative efficiency in MISE for the estimation of standard normal distribution, as a function of $\log_{10} n$. The line shows the limit value. Solid circles correspond to the normal kernel and open circles correspond to the sinc kernel.

empirical distribution function. Here, all the three estimators are asymptotically equally efficient, in the sense that the relative efficiency converges to 1 as $n \to \infty$. However, it is clear that this convergence is much slower for the sinc kernel estimator, which is more efficient that the normal kernel estimator for sample sizes as low as $n = 50$.

2.4 Proofs

For $h > 0$, the statement of Theorem 2.1 is contained within the proof of Proposition 2 in Abdous (1993). Therefore, it only remains to show the case $h = 0$; i.e., Equation (2.4). This formula is valid in the more general situation where F is not necessarily a distribution function, but an integrated kernel with finite first order moment, as shown in the following lemma.

Lemma 2.1. *Suppose that* $K(x) = \int_{-\infty}^{x} k(y)dy$, *where* k *is an integrable function with* $\int k(y)dy = 1$ *and* $\int |yk(y)|dy < \infty$. *Then,*

$$\int K(x)\{1 - K(x)\}dx = (2\pi)^{-1} \int t^{-2}\{1 - |\varphi_k(t)|^2\}dt.$$

Proof. It is not hard to show that $K(x)\{1-K(x)\} = \int \{I_{[y,\infty)}(x)-K(x)\}^2 k(y)dy$, where I_A stands for the indicator function of a set A. Moreover, reasoning as in the proof of Proposition 2 in Abdous (1993), it follows that the condition $\int |yk(y)|dy < \infty$ guarantees that $\int |I_{[y,\infty)}(x)-K(x)|dx < \infty$ for all y, which implies that the function $G_y(x) = I_{[y,\infty)}(x)-K(x)$ is square integrable, since K is bounded (because $|K(x)| \leq \int |k(y)|dy$ for all x). Therefore, by Parseval's identity, $\int \{I_{[y,\infty)}(x) - K(x)\}^2 dx = (2\pi)^{-1} \int |\varphi_{G_y}(t)|^2 dt$. The Fourier transform of G_y is shown to be $(-it)^{-1}\{e^{ity} - \varphi_k(t)\}$, since splitting the integration region and using integration by parts,

$$\begin{aligned} -it\varphi_{G_y}(t) &= it \int_{-\infty}^{y} e^{itx} K(x)dx - it \int_{y}^{\infty} e^{itx}\{1 - K(x)\}dx \\ &= K(y)e^{ity} - \int_{-\infty}^{y} e^{itx} k(x)dx + \{1 - K(y)\}e^{ity} - \int_{y}^{\infty} e^{itx} k(x)dx \\ &= e^{ity} - \varphi_k(t). \end{aligned}$$

Thus, $\int \{I_{[y,\infty)}(x)-K(x)\}^2 dx = (2\pi)^{-1} \int t^{-2}\big[1+|\varphi_k(t)|^2-2\Re\big\{e^{-ity}\varphi_k(t)\big\}\big]dt$, where $\Re\{z\}$ denotes the real part of a complex number z. This finally leads to

$$\begin{aligned} \int K(x)\{1 - K(x)\}dx &= \iint \{I_{[y,\infty)}(x) - K(x)\}^2 k(y)dxdy \\ &= (2\pi)^{-1} \iint t^{-2}\big[1 + |\varphi_k(t)|^2 \\ &\quad - 2\Re\big\{e^{-ity}\varphi_k(t)\big\}\big]k(y)dydt \\ &= (2\pi)^{-1} \int t^{-2}\{1 - |\varphi_k(t)|^2\}dt, \end{aligned}$$

where the last line follows from the fact that $\int e^{-ity}\varphi_k(t)k(y)dy = \varphi_k(-t)\varphi_k(t) = |\varphi_k(t)|^2$. □

The proof of Theorem 2.2 is immediate from the following lemma.

Lemma 2.2. *Assume that F and K satisfy the assumptions of Theorem 2.2. Then,*

$$S_k/D_f \leq \inf_{n\in\mathbb{N}} h_{0n} \leq \limsup_{n\to\infty} h_{0n} \leq h_* \leq \min\{S_k/C_f, T_k/D_f\}.$$

Proof. First notice that $\mathrm{ISB}(h) = 0$ for all $h \in [0, S_k/D_f]$, since using Theorem 2.1

$$\begin{aligned} 0 \leq \pi\,\mathrm{ISB}(h) &= \int_0^\infty t^{-2}|1-\varphi_k(th)|^2|\varphi_f(t)|^2dt \\ &\leq \int_0^{S_k/h} t^{-2}|1-\varphi_k(th)|^2|\varphi_f(t)|^2dt + \int_{D_f}^\infty t^{-2}|1-\varphi_k(th)|^2|\varphi_f(t)|^2dt = 0, \end{aligned}$$

with the last equality due to the facts that $\varphi_k(th) = 1$ for $t \in [0, S_k/h]$ and $\varphi_f(t) = 0$ for $t \geq D_f$ by definition of S_k and D_f, respectively.

Therefore, for $h \in [0, S_k/D_f]$ the MISE reduces to the IV, and admits the exact expression $\mathrm{MISE}(h) = n^{-1}\{\psi(F) - \psi(K)h\}$ because, again using Theorem 2.1, noting the expression (2.4) for $\psi(F)$, taking into account the definition of S_k and D_f and making the change of variable $s = th$, we obtain

$$\begin{aligned} \mathrm{IV}(h) &= (n\pi)^{-1}\int_0^\infty t^{-2}|\varphi_k(th)|^2\{1-|\varphi_f(t)|^2\}dt \\ &= (n\pi)^{-1}\int_0^\infty t^{-2}\{1-|\varphi_f(t)|^2\}dt \\ &\quad - (n\pi)^{-1}\int_0^\infty t^{-2}\{1-|\varphi_k(th)|^2\}\{1-|\varphi_f(t)|^2\}dt \\ &= n^{-1}\psi(F) - (n\pi)^{-1}\int_{S_k/h}^\infty t^{-2}\{1-|\varphi_k(th)|^2\}dt \\ &= n^{-1}\psi(F) - n^{-1}\psi(K)h. \end{aligned}$$

Since the MISE function is linear in h with negative slope in $[0, S_k/D_f]$, its minimum has to be attached at some point greater than S_k/D_f, hence we obtain the first inequality.

On the other hand, reasoning as in Chacón et al. (2007) it is possible to show that $\mathrm{ISB}(h) > 0$ for $h > S_k/C_f$ and for $h > T_k/D_f$, thus yielding the last inequality. Finally, denote $h_L = \limsup_{n\to\infty} h_{0n}$ and assume that $h_L > h_*$. Then, the continuity of $\mathrm{ISB}(h)$ with respect to h (Tenreiro, 2006, Proposition 1) entails that there is a subsequence h_{0n_k} such that, as $k \to \infty$, $\mathrm{ISB}(h_{0n_k}) \to \mathrm{ISB}(h_L)$ with $\mathrm{ISB}(h_L) > 0$ since we are assuming $h_L > h_*$. But from (2.2) and (2.3) it immediately follows that, for every fixed h, $\mathrm{MISE}_{n_k}(h) \to \mathrm{ISB}(h)$ as $k \to \infty$, so we obtain that the following chain of inequalities

$$\begin{aligned}\mathrm{ISB}(h) &= \lim_{k\to\infty} \mathrm{MISE}_{n_k}(h) \geq \lim_{k\to\infty} \mathrm{MISE}_{n_k}(h_{0n_k}) \geq \lim_{k\to\infty} \mathrm{ISB}(h_{0n_k}) \\ &= \mathrm{ISB}(h_L) > 0\end{aligned}$$

is valid for every fixed h, implying that $\lim_{h\to 0} \mathrm{ISB}(h) \geq \mathrm{ISB}(h_L) > 0$, which contradicts Proposition 1 in Tenreiro (2006), where it is shown that $\mathrm{ISB}(h) \to 0$ as $h \to 0$. Hence, it should be $h_L \leq h_*$, as desired. □

Finally, we show the proof of Theorem 2.3. We focus only on the statement about the existence of the optimal bandwidth sequence, since the arguments showing the limit behavior can be adapted from the proof of Lemma 2.2 above.

Proof of Theorem 2.3. It is clear from (2.6) and (2.4) that $\lim_{h\to 0} \mathrm{MISE}(h) = n^{-1}\psi(F)$. Moreover, $\lim_{h\to\infty} \mathrm{MISE}(h) = \infty$, since $\varphi_f(0) = 1$ and by continuity it is possible to take $\delta > 0$ such that $|\varphi_f(t)|^2 > \frac{1}{2}$ for all $0 \leq t \leq \delta$, so this yields $\int_0^\infty t^{-2}|\varphi_f(t)|^2 dt \geq \frac{1}{2}\int_0^\delta t^{-2}dt = \infty$. These two limit conditions, together with the fact that $\mathrm{MISE}(h)$ is a continuous function, imply that the existence of a minimizer of the MISE is guaranteed if there is some $h_1 > 0$ such that $\mathrm{MISE}(h_1) < n^{-1}\psi(F)$. But from (2.6) we have

$$\mathrm{MISE}(h) - n^{-1}\psi(F) = -(n\pi)^{-1}h + (1 + n^{-1})\pi^{-1}\int_{1/h}^\infty t^{-2}|\varphi_f(t)|^2 dt$$

so that using the Riemann-Lebesgue lemma and the dominated convergence theorem, it follows that

$$\lim_{h\to 0} h^{-1}\{\mathrm{MISE}(h) - n^{-1}\psi(F)\} = -(n\pi)^{-1} < 0,$$

which entails that there is some $h_1 > 0$ fulfilling the aforementioned desired property. □

2.5 References

Abdous B. (1993). Note on the minimum mean integrated squared error of kernel estimates of a distribution function and its derivates. *Communications in Statistics Theory and Methods*, **22**, 603–609.

Aldershof, B., Marron, J.S., Park, B.U. and Wand, M.P. (1995) Facts about the Gaussian probability density function. *Applicable Analysis*, **59**, 289–306.

Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, **46**, 195–214.

Azzalini, A. (1981) A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, **68**, 326–328.

Berg, A. and Politis, D. (2009) CDF and survival function estimation with infinite-order kernels. *Electronic Journal of Statistics*, **3**, 1436–1454.

Bowman, A., Hall, P. and Prvan, T. (1998) Bandwidth selection for the smoothing of distribution functions. *Biometrika*, **85**, 799–808.

Butzer, P.L. and Nessel, R.J. (1971) *Fourier analysis and approximation.* Academic Press, New York.

Chacón, J.E., Montanero, J. and Nogales, A.G. (2007) A note on kernel density estimation at a parametric rate. *Journal of Nonparametric Statistics*, **19**, 13–21.

Chacón, J.E., Montanero, J., Nogales, A.G. and Pérez, P. (2007) On the existence an limit behavior of the optimal bandwidth for kernel density estimation. *Statistica Sinica*, **17**, 289–300.

Chacón, J.E. and Rodríguez-Casal, A. (2010) A note on the universal consistency of the kernel distribution function estimator. *Statistics and Probability Letters*, **80**, 1414–1419.

Chiu, S.-T. (1992) An automatic bandwidth selector for kernel density estimation. *Biometrika*, **79**, 771–782.

Cline, D.B.H. (1988) Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**, 1421–1427.

Davis, K.B. (1977) Mean integrated square error properties of density estimates. *Annals of Statistics*, **5**, 530–535.

Fryer, M.J. (1976) Some errors associated with the nonparametric estimation of density functions. *IMA Journal of Applied Mathematics*, **18**, 371–380.

Giné, E. and Nickl, R. (2009) An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, **143**, 569–596.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2003) Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, **30**, 415–427.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2007) Density estimation using the sinc kernel. Preprint Statistics No. 2/2007, Norwegian University of Science and Technology, Trondheim, Norway. Available at `http://www.math.ntnu.no/preprint/statistics/2007/`

Gurland, J. (1948) Inversion formulae for the distribution of ratios. *Annals of Mathematical Statistics*, **19**, 228–237.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2007) Modifying the kernel distribution function estimator towards reduced bias. *Statistics*, **41**, 93–103.

Jones, M.C. (1990) The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, **9**, 129–132.

Mason, D.M. and Swanepoel, J.W.H. (2012) Uniform in bandwidth limit laws for kernel distribution function estimators. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*, IMS Collections **9**, 241–253.

Nadaraya, E.A. (1964) Some new estimates for distribution functions. *Theory of Probability and Its Applications*, **15**, 497–500.

Polansky, A.M. and Baker, E.R. (2000) Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, **65**, 63–80.

Sarda, P. (1993) Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, **35**, 65–75.

Servien, R. (2009) Estimation de la fonction de répartition: revue bibliographique. *Journal de la Société Française de Statistique*, **150**, 84–104.

Shao, Y. and Xiang, X. (1997) Some extensions of the asymptotics of a kemel estimator of a distribution function. *Statist. Probab. Lett.*, **34**, 301–308.

Swanepoel, J.W.H. (1988) Mean integrated square error properties and optimal kernels when estimating a distribution function. *Communications in Statistics Theory and Methods*, **17**, 3785–3799.

Swanepoel, J.W.H. and Van Graan, F.C. (2005) A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics*, **32**, 551–562.

Tenreiro, C. (2003) On the asymptotic behavour of the ISE for automatic kernel distribution estimators. *Journal of Nonparametric Statistics*, **15**, 485–504.

Tenreiro, C. (2006) Asymptotic behaviour of multistage plug-in bandwidth selections for kernel distribution function estimators. *Journal of Nonparametric Statistics*, **18**, 101–116.

Tiago de Oliveira, J. (1963) Estatística de densidades: resultados assintóticos. *Revista da Faculdade de Ciências de Lisboa*, **9**, 111–206.

Tsybakov, A.B. (2009) *Introduction to Nonparametric Estimation.* Springer Science+Business Media, New York.

Watson, G.S. and Leadbetter, M.R. (1964) Hazard analysis II. *Sankhy ā Series A*, **26**, 101–116.

Yamato, H. (1973) Uniform convergence of an estimator of a distribution function. *Bulletin of Mathematical Statistics*, **15**, 69–78.

Part II

Conditional expectation approximation

3

On equivariant estimation of the parameters of the general half-normal distribution making use of a Monte Carlo method to approximate conditional expectations

3.1 Introduction

Let Z be a $N(0,1)$ random variable. The distribution of $X := |Z|$ is the so-called half-normal distribution. It will be denoted $HN(0,1)$ and its density function is

$$f_X(x) = \sqrt{\frac{2}{\pi}} \exp\left\{-\frac{1}{2}x^2\right\} I_{[0,+\infty[}(x).$$

A general half-normal distribution $HN(\xi,\eta)$ is obtained from $HN(0,1)$ by a location-scale transformation: $HN(\xi,\eta)$ is the distribution of $Y = \xi + \eta X$.

The classical paper Daniel (1959) introduces half-normal plots and the half-normal distribution, a special case of the folded and truncated normal distributions (see Johnson et al. (1994)). Bland et al. (1999) and Bland (2005) propose a so-called half-normal method to deal with relationships between measurement error and magnitude, with applications in medicine. Pewsey (2002) uses the maximum likelihood principle to estimate the parameters, and presents a brief survey on the general half-normal distribution, its relations with other well-known distributions and its usefulness in the analysis of highly skew data. Pewsey (2004) proposes bias-corrected versions of

the maximum likelihood estimators. Nogales et al. (2011) deals with the problem of unbiased estimation for the general half-normal distribution.

Here we consider the problem of equivariant estimation of the location and scale parameters, ξ and η, but first we provide a brief review of results for unbiased and maximum likelihood estimation appearing in the literature.

The density function of $HN(\xi, \eta)$ is

$$f_Y(y) = \frac{1}{\eta} f_X\left(\frac{y-\xi}{\eta}\right) = \frac{1}{\eta}\sqrt{\frac{2}{\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\xi}{\eta}\right)^2\right\} I_{[\xi,+\infty[}(y).$$

It is readily shown that

$$E(Y) = \xi + \eta\sqrt{\frac{2}{\pi}} \qquad \text{and} \qquad \mathrm{Var}(Y) = \frac{\pi-2}{\pi}\eta^2.$$

Let $Y_1, \ldots, Y_n$ be a sample of size n from a general half-normal distribution with unknown parameters, ξ and η. $Y_{1:n}$ denotes the minimum of $Y_1, \ldots, Y_n$. From the factorization criterion, we obtain that the expression $(\sum_{i=1}^n Y_i^2, \sum_{i=1}^n Y_i, Y_{1:n})$ is a sufficient statistic. Indeed, it is minimal sufficient, although not complete.

We write $Y_i = \xi + \eta X_i$, where $X_i = |Z_i|$, $1 \leq i \leq n$, $Z_1, \ldots, Z_n$ being a sample of the standard normal distribution $N(0,1)$. Throughout the paper, we also let

$$c_n := E(X_{1:n})$$

For $n \geq 2$, it is readily shown that $0 < c_n < \sqrt{\frac{2}{\pi}}$. In fact, the next lemma (Nogales et al. (2011)) yields an alternative expression and a refined bound for c_n. We write Φ for the standard normal cumulative distribution function.

Lemma 3.1. (i) $c_n = \int_0^\infty (2 - 2\Phi(t))^n \, dt$.

(ii) For $n \geq 1$, $c_n \leq \frac{1}{n}\sqrt{\frac{\pi}{2}} \leq \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2n}\right)$.

Notice also that $Y_{1:n} = \min_i Y_i = \xi + \eta X_{1:n}$ and $E(Y_{1:n}) = \xi + \eta c_n$.

The next proposition (Nogales et al. (2011)) yields unbiased estimators of the location and scale parameters, ξ and η. Both estimators are L-statistics and functions of the cited minimal sufficient statistic.

Proposition 3.1. (i) $\widetilde{\xi} := \frac{\sqrt{\frac{2}{\pi}}Y_{1:n} - c_n\bar{Y}}{\sqrt{\frac{2}{\pi}} - c_n}$ *is an unbiased estimator of the location parameter* ξ.

(ii) $\widetilde{\eta} := \frac{\bar{Y} - Y_{1:n}}{\sqrt{\frac{2}{\pi}} - c_n}$ *is an unbiased estimator of the scale parameter* η *whose distribution does not depend on* ξ.

Remark 3.1. We also have that the sample mean $\bar{Y}$ is an unbiased estimator of the mean $\xi + \eta\sqrt{\frac{2}{\pi}}$. Moreover, an unbiased estimator of η^2 is

$$\frac{\pi}{\pi - 2}\, S^2,$$

where $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the sample variance; notice that its distribution does not depend on ξ. $\bar{Y}$ and S^2 also are functions of the sufficient statistic given above. The reader is referred to Nogales et al. (2011) for these and other results about unbiased estimation of the parameters of the general half-normal distribution. □

Remark 3.2. Pewsey (2002) provides maximum likelihood estimates for each of the parameters ξ and η:

$$\widehat{\xi} := Y_{1:n}, \quad \widehat{\eta} := \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_{1:n})^2\right)^{1/2}$$

A large sample based bias-correction is used in Pewsey (2004) to improve the performance of the maximum likelihood estimators $\widehat{\xi}$ and $\widehat{\eta}$. □

3.2 A Monte Carlo method to approximate conditional expectations

In this section, we describe a natural Monte Carlo method to compute conditional expectations based on a theorem of Besicovitch on differentiation of measures. It will be used in the next section to approximate the MRE estimator of the location parameter ξ because its expression involves two conditional expectations not easy to compute.

We first recall briefly a theorem of Besicovitch (1945, 1946) for differentiation of measures (see, for instance, Corollary 2.14 of Mattila (1995)). This

theorem extend to Radon measures the classical Lebesgue Differentiation Theorem.

Theorem 3.1 (Besicovitch (1945, 1946))**.** *Let λ be a Radon measure on $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$ a locally λ-integrable function. Then*

$$\lim_{r \downarrow 0} \frac{1}{\lambda(B_r(x))} \int_{B_r(x)} f \, d\lambda = f(x)$$

for λ-almost all $x \in \mathbb{R}^n$, where $B_r(x)$ denotes the ball of center x and radius $r > 0$ for the norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$.

Now let $(\Omega, \mathcal{A}, P)$ be a probability space, $X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ be an n-dimensional random variable and $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be a real random variable with finite mean. The conditional expectation $E(Y|X)$ is defined as a random variable on $\mathbb{R}^n$ such that $\int_{X^{-1}(B)} Y \, dP = \int_B E(Y|X) dP^X$ for any Borel set B in $\mathbb{R}^n$, where P^X denotes the probability distribution of X.

Although the existence of the conditional expectation is guaranteed via the Radon-Nikodym theorem, its computation is, generally, involved. Nevertheless, according to the previous result, for P^X-almost every $x \in \mathbb{R}^n$,

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \frac{1}{P^X(B_\epsilon(x))} \int_{X^{-1}(B_\epsilon(x))} Y(\omega) \, dP(\omega) \\
&= \lim_{\epsilon \downarrow 0} \frac{1}{P^X(B_\epsilon(x))} \int_{B_\epsilon(x)} E(Y|X = x') \, dP^X(x') \\
&= E(Y|X = x)
\end{aligned}$$

By the Strong Law of Large Numbers, for almost every sequence (ω_i) in Ω, we have

$$P^X(B_\epsilon(x)) = \lim_k \frac{1}{k} \sum_{i=1}^{k} I_{B_\epsilon(x)}(X(\omega_i))$$

and

$$\int_{B_\epsilon(x)} E(Y|X = x') \, dP^X(x') = \lim_k \frac{1}{k} \sum_{i=1}^{k} I_{B_\epsilon(x)}(X(\omega_i)) Y(\omega_i)$$

where I_A denotes the indicator function of A. Observe that, for every $\epsilon > 0$, the rate of convergence is $1/\sqrt{n}$.

Hence, we have proved the following result:

Theorem 3.2. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ be an n-dimensional random variable and $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be a real random variable with finite mean. Then, for P^X-almost every $x \in \mathbb{R}^n$ and almost every sequence (ω_i) in Ω, we have*

$$E(Y|X = x) = \lim_{\epsilon \downarrow 0} \lim_{k} \frac{\sum_{i=1}^{k} I_{B_\epsilon(x)}(X(\omega_i))Y(\omega_i)}{\sum_{i=1}^{k} I_{B_\epsilon(x)}(X(\omega_i))}.\text{„}$$

This theorem yields a means of approximating the conditional expectation of Y given X. The following simple example illustrates the method.

Example 3.1. Let (X, Y) be a bivariate normal random variable with null mean such that $\mathrm{Var}\,(X) = \mathrm{Var}\,(Y) = 1$ and $\mathrm{Cov}\,(X, Y) = 0.5$. In this case, there is no need for an approximation to the conditional expectation of Y given $X = x$ because it is $x/2$. The conditional distribution of Y given $X = x$ is $N(\frac{1}{2}x, \frac{1}{2}\sqrt{3})$. Applying the proposed method to evaluate $E(Y|X = 1)$, given a small $\epsilon > 0$, we may choose a sample $(x_i, y_i)_{1 \le i \le k}$ from the joint distribution of X and Y and approximate $E(Y|X = 1)$ by

$$\frac{\sum_{i=1}^{k} I_{[1-\epsilon,1+\epsilon]}(x_i)y_i}{\sum_{i=1}^{k} I_{[1-\epsilon,1+\epsilon]}(x_i)}. \tag{1}$$

Taking $\epsilon = 0.1, 0.01$ and samples from the joint distribution of X and Y with sample sizes k large enough to obtain $m = m(k) = \sum_{i=1}^{k} I_{[1-\epsilon,1+\epsilon]}(x_i) = 100, 1000, 5000$, we obtained the approximations for $E(Y|X = 1)$ summarized in Table 1 and Figure 1; 100 replications of each simulation have been conducted to obtain the table and the figure. Namely, taking $m = 1000$, for instance, the value 0.493947 appearing in the table as an approximation of $E(Y|X = 1)$ when $\epsilon = 0.1$ is the mean of the 100 values of the quotient (1) obtained after 100 replications of the experiment of choosing a k-sized sample $(x_i, y_i)_{1 \le i \le k}$ of the joint distribution of (X, Y), k being large enough to get $m = m(k) = 1000$. Table 1 also includes the "mean squared error" (MSE) calculated from these 100 values: the format used for a typical entry

in the table is $E(Y|X=1) \pm \text{MSE}$. The box-plot of the figure describes the distribution of these 100 values (a dotted red line represents the mean).

m	100	1000	5000
$\epsilon = 0.1$	0.505885 ± 0.006395	0.493947 ± 0.000815	0.497892 ± 0.000128
$\epsilon = 0.01$	0.503655 ± 0.007165	0.499826 ± 0.000716	0.499471 ± 0.000150

Table 3.1: Approximation of $E(Y|X=1) \pm \text{MSE}$ as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$

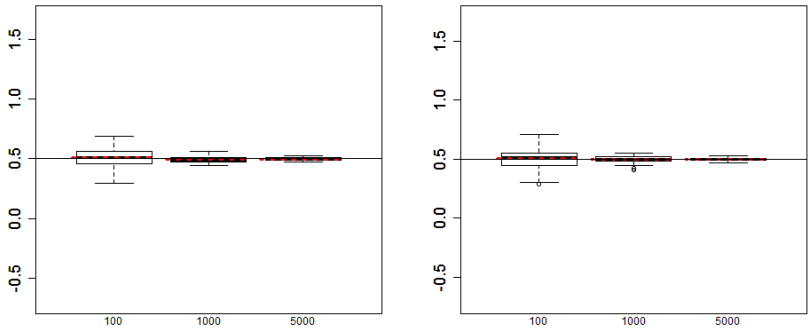

Figure 3.1: Box plots of the approximations of $E(Y|X=1)$ as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$.

Remark 3.3. The described method of Monte Carlo approximation to the conditional expectation $E(Y|X=x)$ is based on the naïve idea that one can approximate it from a sample $(x_i, y_i)_{1 \leq i \leq n}$ by the mean of the y_i corresponding to points x_i lying in a narrow neighborhood of x. From a probabilistic point of view, the method has been justified by the mentioned theorem of Besicovitch on differentiation of measures. When the joint density of X and Y is known, $E(Y|X=x)$ is the mean of the conditional distribution of Y given $X=x$, and the problem of compute a conditional expectation is reduced to the problem of computing a mean. Notice that the existence of a joint density is not required by the method and it could be specially useful when densities are not available or are not easy to compute (see the next example). □

Example 3.2. (Example 1, continuation) A similar simulation study has been performed to approximate the conditional expectation $E(V|U = 0.5)$, where $V = \sin(X \cdot Y)$ and $U = \cos(X^2 + Y^2)$; the obtained results are:

m	100	1000	5000
$\epsilon = 0.1$	0.127650 ± 0.001890	0.127280 ± 0.000202	0.124169 ± 0.000025
$\epsilon = 0.01$	0.123063 ± 0.001620	0.125869 ± 0.000153	0.1252856 ± 0.000031

Table 3.2: Approximation of $E(V|U = 0.5) \pm S^2$ (S^2 is the sample variance) as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$

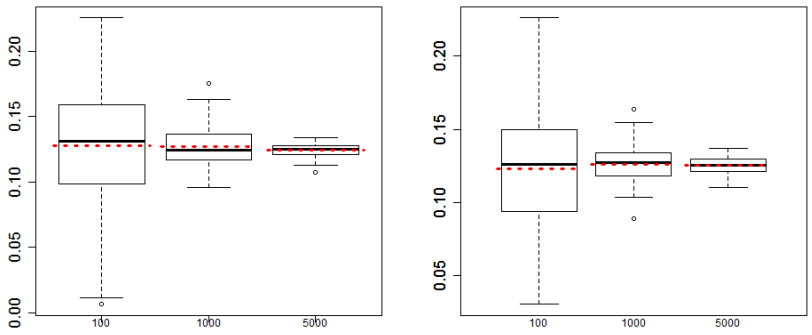

Figure 3.2: Box plots of the approximations of $E(V|U = 0.5)$ as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$

Remark 3.4. In a statistical framework, we can provide additional guarantees on the method, since the obtained Monte Carlo approximation to the conditional expectation $E(Y|X = x)$ coïncides with the value at the point x of the kernel estimator (the Nadaraja-Watson estimator) of the regression curve $y = E(Y|X = x)$ for the kernel $K(x) = I_{[-1,1]}(x)$ (see Nadaraya (1989), p. 115). From this point of view, ϵ plays the role of the bandwidth parameter. We refer to Härdle (1992, Ch. 5) for a detailed discussion on the important problem of the choice of the bandwidth. In this paper, the main application of the Monte Carlo method for the approximation of conditional expectations is given in the next section to approximate the estimation of

the location parameter of the general half-normal distribution, beacuse it is defined in terms of a quotient of two not-easily-computable parameter-free conditional expectations given a $(n-1)$-dimensional statistic U. Some "curse of dimensionality problem" appears when n is large because, in this case, it is not easy to find large samples of points lying in a small ball centered at a point $U(y)$. This is why we had to modify the Monte Carlo method for the approximation of conditional expectations taking advantage of the underlying distribution of Y (the general half-normal distribution) and the invariance properties of U. This could become an important scholium of the paper, as the ideas used here could be useful to deal with the "curse of dimensionality problem" in similar situations. □

3.3 Equivariant estimation of the location parameter of the general half-normal distribution

In this section we consider the problem of determining the minimum risk equivariant estimator of the location parameter ξ of the general half-normal distribution $HN(\xi, \eta)$ when the scale parameter η is unknown. We cannot provide an explicit expression for this estimator, since it is described in terms of two conditional expectations that had to be approximated by simulation.

To achieve this goal, an R program was developed based on the method of computing conditional expectations described in the previous section. In fact, the method has been slightly modified to solve a sort of "curse of dimensionality" problem.

We consider the scale-location family of densities

$$f_{(\xi,\eta)}(y_1, ..., y_n) = \frac{1}{\eta^n} f\left(\frac{y_1 - \xi}{\eta}, ..., \frac{y_n - \xi}{\eta}\right),$$

where

$$f(y_1, ..., y_n) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right\} I_{[0,+\infty[}(y_{1:n}).$$

This family remains invariant under transformations of the form $g_{a,b}(y_1, ..., y_n) = (a + by_1, ..., a + by_n)$, $a \in \mathbb{R}$, $b > 0$.

To estimate the location parameter ξ when the scale parameter η is unknown, we have the next result, a direct consequence of classical equivariant

estimation theory (see Lehmann (1983)). First, recall that an estimator T of the location parameter is equivariant if $T(a + bx_1, \ldots, a + bx_n) = a + bT(x_1, \ldots, x_n)$, for all $a \in \mathbb{R}$ and all $b > 0$.

Proposition 3.2. *When the loss function $W_2(x; \xi, \eta) = \eta^{-2}(x - \xi)^2$ is considered, the MRE estimator $\mathring{\xi}$ of ξ is*

$$\mathring{\xi} = T_0^* - (\rho \circ U)T_1^*$$

where

$$T_0^* = \bar{Y}, \quad T_1^* = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \bar{Y}|$$

$$U = \left(\frac{Y_1 - Y_n}{Y_{n-1} - Y_n}, \ldots, \frac{Y_{n-2} - Y_n}{Y_{n-1} - Y_n}, \frac{Y_{n-1} - Y_n}{|Y_{n-1} - Y_n|} \right),$$

$$\rho = \frac{E_{\xi=0,\eta=1}(T_0^* T_1^* | U)}{E_{\xi=0,\eta=1}(T_1^{*2} | U)}$$

Remark 3.5. T_0^* can be replaced by any other equivariant estimator of ξ (i.e., satisfying $T_0^*(a + by_1, \ldots, a + by_1) = a + bT_0^*(y_1, \ldots, y_1)$ for every $a \in \mathbb{R}$, $b > 0$), and T_1^* can be replaced by any positive estimator of η satisfying $T_1^*(a + by_1, \ldots, a + by_1) = bT_1^*(y_1, \ldots, y_1)$ for every $a \in \mathbb{R}$, $b > 0$. □

A simulation study has been performed to investigate the behavior of the minimum risk equivariant estimator $\mathring{\xi}$. In it, we used 100 simulations with sample sizes $n = 100, 1000, 5000$ from the $HN(10, 4)$ distribution, obtaining the results summarized in Table 3 and Figure 3 (see below how we have made use of the method of approximation of conditional expectations to obtain the values of the Tables 2 and 3).

m	100	1000	5000
$\epsilon = 0.1$	9.811914 ± 2.307873	9.835290 ± 1.881063	9.862105 ± 0.158635
$\epsilon = 0.01$	9.840835 ± 0.687243	9.827150 ± 3.501158	9.907112 ± 0.027826

Table 3.3: Approximations of $\mathring{\xi} \pm$ MSE as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$

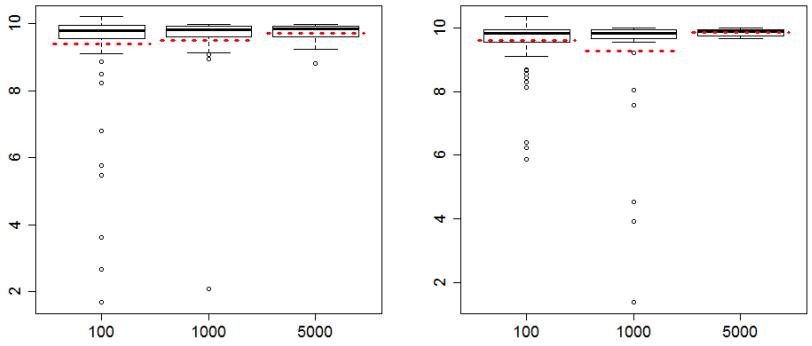

Figure 3.3: Box plots of the approximations of $\mathring{\xi} \pm \text{MSE}$ as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$

To compare the behavior of the unbiased estimator $\tilde{\xi}$, the maximum likelihood estimator $\hat{\xi}$ and the minimum risk equivariant estimator $\mathring{\xi}$, we used 100 simulations with sample sizes $n = 100, 1000, 5000$ from the $HN(10, 4)$ distribution, obtaining the results summarized in Table 4 and Figure 4:

	m	100	1000	5000
$\tilde{\xi}$	$\epsilon = 0.1$	9.997350 ± 0.003289	9.999356 ± 0.000020	10.000823 ± 0.000001
	$\epsilon = 0.01$	9.999662 ± 0.003443	9.999989 ± 0.000025	10.001050 ± 0.000002
$\hat{\xi}$	$\epsilon = 0.1$	10.047256 ± 0.005455	10.004383 ± 0.000039	10.000823 ± 0.000001
	$\epsilon = 0.01$	10.049128 ± 0.005752	10.005005 ± 0.000050	10.001050 ± 0.000002
$\mathring{\xi}$	$\epsilon = 0.1$	9.412753 ± 2.307873	9.517691 ± 1.881063	9.732626 ± 0.158635
	$\epsilon = 0.01$	9.603969 ± 0.687243	9.274164 ± 3.501158	9.867600 ± 0.027826

Table 3.4: Approximations of $\tilde{\xi} \pm \text{MSE}$, $\hat{\xi} \pm \text{MSE}$ and $\mathring{\xi} \pm \text{MSE}$ as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$

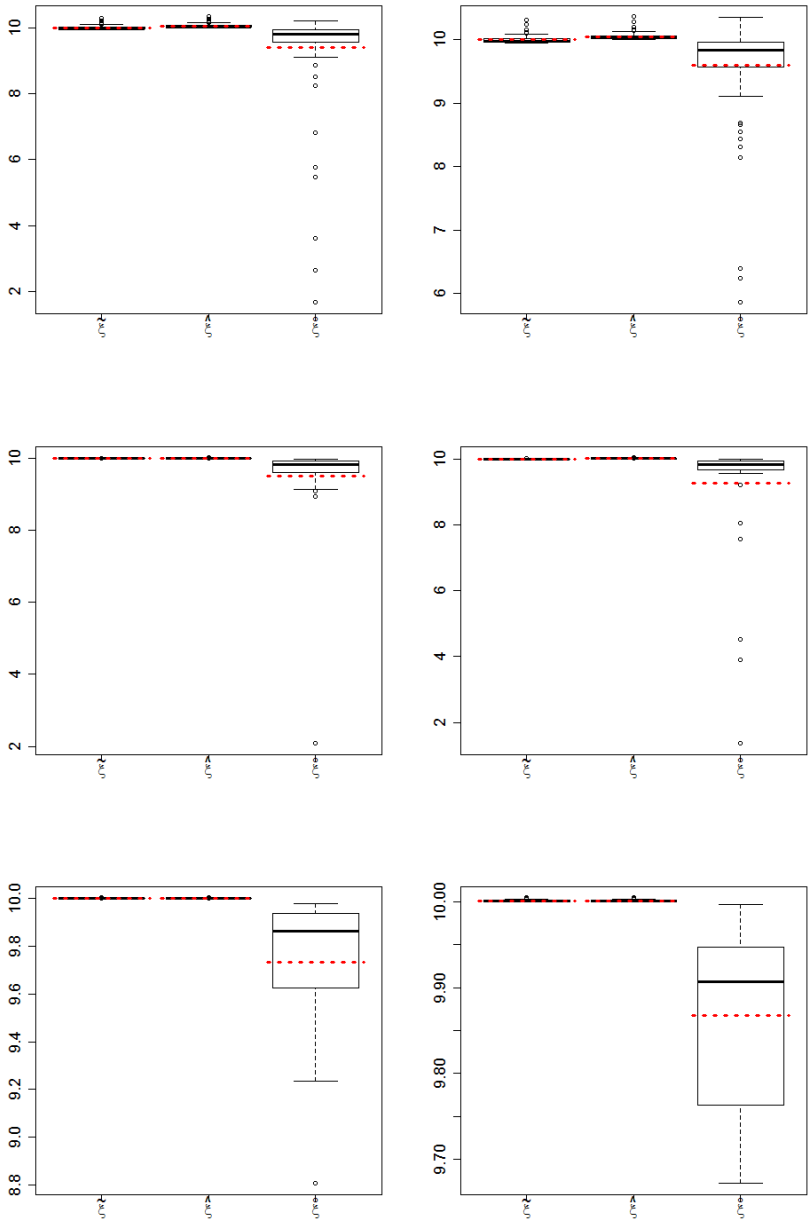

Figure 3.4: Box plots of the approximations of $\tilde{\xi}$, $\hat{\xi}$ and $\mathring{\xi}$ as a number of simulations, $m = 100, 1000, 5000$, for $\epsilon = 0.1$ and $\epsilon = 0.01$ function of the

Table 4 and Figure 4 illustrate the biased character of the maximum likelihood estimator $\hat{\xi}$ and the minimum risk equivariant estimator $\mathring{\xi}$. As

expected, the behavior of this approximation to the MRE estimator is worse than those of the unbiased estimator $\tilde{\xi}$ or the maximum likelihood estimator $\hat{\xi}$. However, this method provides a way to proceed when other estimation methods are not available.

Let us describe in more details the ideas used in these simulations. For a sample $y = (y_1, \ldots, y_n)$, $n = 100, 1000, 5000$, of the distribution $HN(10, 4)$, we have

$$\rho(U(y)) = \lim_{\epsilon \to 0} \frac{N_\epsilon}{D_\epsilon}$$

where

$$N_\epsilon = \int_{A_\epsilon(y)} f(y')dy', \quad D_\epsilon = \int_{A_\epsilon(y)} g(y')dy',$$
$$f(y') = T_0^*(y')T_1^*(y') \exp\left\{-\frac{1}{2}\|y'\|_2^2\right\}, \quad g(y') = T_1^*(y')^2 \exp\left\{-\frac{1}{2}\|y'\|_2^2\right\},$$
$$A_\epsilon(y) = \{y' \in [0,10]^n : \max_{1 \le i \le n-1} |U_i(y') - U_i(y)| \le \epsilon\}.$$

Now, take a sample S of $A_\epsilon(y)$ and approximate N_ϵ and D_ϵ by

$$\frac{1}{\text{card}\,(S)} \sum_{y' \in S} f(y') \quad \text{and} \quad \frac{1}{\text{card}\,(S)} \sum_{y' \in S} g(y'),$$

respectively. So, $\rho(U(y))$ can be approximated by

$$C(y) := \frac{\sum_{y' \in S} f(y')}{\sum_{y' \in S} g(y')}$$

and $\mathring{\xi}(y)$ is approximated by $D(y) := T_0^*(y) - C(y)T_1^*(y)$.

To approximate $C(y)$, a first idea would be to divide the interval $[0, 10]$ in multiple subintervals of small length $\epsilon > 0$ and consider the grid in the interval $[0, 10]^n$ formed by the n-power set of the ends of these subintervals (we have restricted ourselves to the interval [0,10] because the functions $f(y)$ and $g(y)$ are almost null when one of the coordinates of the vector y is greater than 10). The sample S would then be formed by the grid nodes that are in A_ϵ. The main problem with this approach is that the size m of the sample S is very small: it becomes smaller as n increases, because of the so-called "curse of dimensionality" problem. In order to avoid this problem and obtain a sample size m large enough for S (given n, we take $m = 100n$),

we have used the following algorithm, a modification of the described Monte Carlo method to approximate conditional expectations that hinges on the use of the invariance of U under scale and location transformations. Namely:

Step A. Let $n \in \mathbb{N}$ and be $y = (y_1, \ldots, y_n)$ a n-sized sample of the distribution $HN(10, 4)$. For $1 \leq i \leq n-2$, let $a_i := \frac{y_1 - y_n}{y_{n-1} - y_n}$ and take $0 < \epsilon < \min\{0.1, \min_{1 \leq i \leq n-2} |a_i|\}$.

Step A.1. At this stage we choose at random $100 \cdot n$ vectors $v^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$, $1 \leq j \leq 100n$, in $\mathbb{R}^n$ such that $\max_{1 \leq i \leq n-1} |U_i(v^{(j)}) - U_i(y)| \leq \epsilon$ as follows:

A.1.1. Make $j = 1$.

A.1.2. Take $v_{n-1}^{(j)}, v_n^{(j)}$ at random in $[0, 10]$ such that $v_{n-1}^{(j)} - v_n^{(j)}$ has the same sign as $y_{n-1} - y_n$. (So, the last coordinates of $U(v^{(j)})$ and $U(y)$ are the same).

A.1.3. For $1 \leq i \leq n-2$ take $v_i^{(j)}$ at random on the interval determined by $v_n^{(j)} + (v_{n-1}^{(j)} - v_n^{(j)})(a_i - \epsilon)$ and $v_n^{(j)} + (v_{n-1}^{(j)} - v_n^{(j)})(a_i + \epsilon)$. (So $|U_i(v^{(j)}) - U_i(y)| \leq \epsilon$).

A.1.4. Make $j = j + 1$ a go back to Step 1 until $100n$ vectors $v^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$, $1 \leq j \leq 100n$ are obtained.

Step A.2. Since the vectors $v^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$, $1 \leq j \leq 100n$, do not lie necessarily in $[0, 10]^n$ (so neither in $A_\epsilon(y)$), we can done some random location-scale trasnformations to put them into $[0, 10]^n$. These transformations do not modify the required fact that $\max_{1 \leq i \leq n-1} |U_i(v^{(j)}) - U_i(y)| \leq \epsilon$.

A.2.1. If $v_{i_0}^{(j_0)} < 0$ for some i_0, j_0, we define $u_i^{(j)} = v + v_i^{(j)}$, $1 \leq i \leq n$, $1 \leq j \leq 100n$, where v is choosen at random between $-\min_{1 \leq i \leq n, 1 \leq j \leq 100n} v_i^{(j)}$ and $1 - \min_{1 \leq i \leq n, 1 \leq j \leq 100n} v_i^{(j)}$. Otherwise, $u_i^{(j)} = v_i^{(j)}$, $1 \leq i \leq n$, $1 \leq j \leq 100n$.

A.2.2. Each vector $u^{(j)}$ is divided by $\max_{1 \leq i \leq n} u_i^{(j)}$ and multiplied by a random number choosen in $[0, 10]$ to obtain the vector $w^{(j)}$.

A.2.3. Take $S = \{w^{(j)} : 1 \leq j \leq 100n\}$ and approximate $C(y)$ by

$$\frac{\sum_{j=1}^{100n} f(w^{(j)})}{\sum_{j=1}^{100n} g(w^{(j)})}$$

and $D(y)$ by $T_0^*(y) - C(y)T_1^*(y)$.

Step B. Finally, following the process designed in Step A, we choose $k := 100$ random samples $y^{(i)}$ of size n from the $HN(10, 4)$ distribution and approximate the mean and the mean squared error of $\mathring{\xi}$ by

$$\frac{1}{k}\sum_{i=1}^{k} D(y^{(i)}) \quad \text{and} \quad \frac{1}{k}\sum_{i=1}^{k}(D(y^{(i)}) - 10)^2, \quad \text{respectively.}$$

Remark 3.6. Notice that both $\hat{\xi}$ and $\tilde{\xi}$ are equivariant estimators of the location parameter ξ. So they have greater risk for the loss function W_2 than $\mathring{\xi}$. Hence, in the previous simulation study, the MSE of $\mathring{\xi}$ should have been smaller than the MSE of $\hat{\xi}$ and $\tilde{\xi}$. That has not been the case because, for the MRE estimator, we have not real estimates of ξ, but approximations of these estimates obtained by a modification of the Monte Carlo method of computing conditional expectations appearing as the numerator and denominator of a quotient. But this is a possible issue to approximate minimum risk estimations of a location parameter, and a possible way avoid the "curse dimensionality problem". □

Remark 3.7. Although less interesting from the perspective of real applications, for completeness we now consider the problem of estimating the scale parameter ξ when the location parameter η is known, say $\eta = \eta_0$. In this case, the joint density of $Y_1, \ldots, Y_n$ is

$$f_\xi(y_1, \ldots, y_n) = \frac{1}{\eta_0^n}\sqrt{\frac{2}{\pi}}^n \exp\left\{-\frac{1}{2\eta_0^2}\sum_{i=1}^{n}(y_i - \xi)^2\right\} I_{[\xi,+\infty[}(y_{1:n}),$$

where $y_{1:n} := \min\{y_1, \ldots, y_n\}$. This family remains invariant under translations of the form $g_a(y_1, \ldots, y_n) = (y_1 - a, \ldots, y_n - a)$.

The equivariant estimator of minimum mean squared error of the location parameter ξ is

$$T_1 = \bar{Y} - \frac{\eta_0}{\sqrt{2\pi n}} \frac{\exp\left\{-\frac{n}{2\eta_0^2}\left(Y_{1:n} - \bar{Y}\right)^2\right\}}{\Phi\left[\frac{\sqrt{n}}{\eta_0}\left(Y_{1:n} - \bar{Y}\right)\right]}.$$

In fact, for the loss function $W_2'(\xi, x) = (x - \xi)^2$, the MRE estimator of the location parameter ξ is the Pitman estimator

$$T_1(y_1, \ldots, y_n) = \frac{\int_{-\infty}^{+\infty} u f_0(y_1 - u, ..., y_n - u)du}{\int_{-\infty}^{+\infty} f_0(y_1 - u, ..., y_n - u)du}.$$

For $y \in \mathbb{R}^n$, we write $\bar{y}$ for the mean of $y_1, \ldots, y_n$. After some algebraic manipulations, we obtain:

$$\int_{-\infty}^{+\infty} u f_0(y_1 - u, ..., y_n - u)du =$$
$$\left(\frac{\sqrt{2}}{\eta_0\sqrt{\pi}}\right)^n \exp\left\{-\frac{1}{2\eta_0^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} \frac{\eta_0}{\sqrt{n}}$$
$$\times \left[-\frac{\eta_0}{\sqrt{n}} \exp\left\{-\frac{n}{2\eta_0^2}(y_{1:n} - \bar{y})^2\right\} + \bar{y}\sqrt{2\pi}\,\Phi\left(\frac{\sqrt{n}}{\eta_0}(y_{1:n} - \bar{y})\right)\right]$$

and

$$\int_{-\infty}^{+\infty} f_0(y_1 - u, ..., y_n - u)du =$$
$$\left(\frac{\sqrt{2}}{\eta_0\sqrt{\pi}}\right)^n \exp\left\{-\frac{1}{2\eta_0^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} \frac{\eta_0}{\sqrt{n}}\sqrt{2\pi}\,\Phi\left[\frac{\sqrt{n}}{\eta_0}(y_{1:n} - \bar{y})\right]$$

and the statement follows easily from these expressions. □

3.4 Equivariant estimation of the scale parameter of the general half-normal distribution

Unlike what happens with the location parameter ξ, for the scale parameter η an explicit expression for the MRE estimator is obtained.

Recall that an estimator T of the scale parameter η is equivariant if $T(a+bx_1,\ldots,a+bx_n) = bT(x_1,\ldots,x_n)$, for all $a \in \mathbb{R}$ and all $b > 0$.

Proposition 3.3. *When using the loss function $W_1(x;\xi,\eta) = \eta^{-2}(x-\eta)^2$, the MRE estimator $\mathring{\eta}$ of η is*

$$\mathring{\eta}(y) = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}\frac{t(n+1)\left(\left[\sqrt{\frac{n(n+1)}{n-1}}\frac{\bar{y}-y_{1:n}}{S(y)},\infty\right[\right)}{t(n+2)\left(\left[\sqrt{\frac{n(n+2)}{n-1}}\frac{\bar{y}-y_{1:n}}{S(y)},\infty\right[\right)}S(y).$$

where $t(n)$ denotes the Student's t-distribution with n degrees of freedom and S^2 is the sample variance.

***Proof* 1.** The MRE estimator of the scale parameter η, when using the loss function W_1, is

$$\mathring{\eta}(y) = \frac{\int_0^{+\infty} v^n f'(vy'_1,...,vy'_{n-1})dv}{\int_0^{+\infty} v^{n+1} f'(vy'_1,...,vy'_{n-1})dv},$$

where f' is the joint density when $\eta = 1$ of $Y'_i := Y_i - Y_n$, $1 \le i \le n-1$, and $y'_i := y_i - y_n$, $1 \le i \le n-1$.

Notice that

$$\begin{aligned} f'(y'_1,...,y'_{n-1}) &= \int_{-\infty}^{+\infty} f(y_1+t,...,y_n+t)dt \\ &= \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} y_i^2 + \frac{n}{2}\bar{y}^2\right\} \int_{-y_{1:n}}^{\infty} \exp\left\{-\frac{n}{2}(t+\bar{y})^2\right\} dt \\ &= \frac{1}{\sqrt{n}}\left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}(n-1)S^2(y)\right\} \int_{\sqrt{n}(\bar{y}-y_{1:n})}^{\infty} \exp\left\{-\frac{1}{2}u^2\right\} du. \end{aligned}$$

Hence, for $k \in \mathbb{N}$, applying Fubini's Theorem after a suitable change of

variables in the inner integral,

$$I_k(y) := \int_0^\infty v^k f'(vy'_1, ..., vy'_{n-1})dv$$

$$= \frac{1}{\sqrt{n}}\left(\frac{2}{\pi}\right)^{\frac{n}{2}} \int_0^\infty v^k \exp\left\{-\frac{1}{2}(n-1)v^2S^2(y)\right\} \int_{\sqrt{n}(\bar{y}-y_{1:n})}^\infty \exp\left\{-\frac{1}{2}u^2\right\} dudv$$

$$= \frac{1}{\sqrt{n}}\left(\frac{2}{\pi}\right)^{\frac{n}{2}} \int_{\sqrt{n}(\bar{y}-y_{1:n})}^\infty J_k(t,y)dt.$$

where

$$J_k(t,y) := \int_0^\infty v^{k+1} \exp\left\{-\frac{1}{2}v^2(t^2+(n-1)S^2(y))\right\} dv = \frac{2^{k/2}\Gamma\left(\frac{k+2}{2}\right)}{(t^2+(n-1)S^2(y))^{\frac{k+2}{2}}}.$$

where, for $t \geq \sqrt{n}(\bar{y} - y_{1:n})$, we have made the change of variables $w = \frac{1}{2}v^2(t^2 + (n-1)S^2(y))$.

So,

$$I_k(y) = \frac{1}{\sqrt{n}}\left(\frac{2}{\pi}\right)^{\frac{n}{2}} 2^{k/2}\Gamma\left(\frac{k+2}{2}\right) \int_{\sqrt{n}(\bar{y}-y_{1:n})}^\infty \frac{dt}{(t^2+(n-1)S^2(y))^{\frac{k+2}{2}}}$$

$$= \frac{2^{\frac{n+k}{2}}\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{n}\pi^{\frac{n-1}{2}}(n-1)^{\frac{k+1}{2}}S(y)^{k+1}} t(k+1)\left(\left[\sqrt{\frac{n(k+1)}{n-1}}\frac{\bar{y}-y_{1:n}}{S(y)}, \infty\right[\right).$$

Finally

$$\mathring{\eta}(y) = \frac{I_n(y)}{I_{n+1}(y)} = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} \frac{t(n+1)\left(\left[\sqrt{\frac{n(n+1)}{n-1}}\frac{\bar{y}-y_{1:n}}{S(y)}, \infty\right[\right)}{t(n+2)\left(\left[\sqrt{\frac{n(n+2)}{n-1}}\frac{\bar{y}-y_{1:n}}{S(y)}, \infty\right[\right)} S(y).$$

□

Remark 3.8. A simulation study has been performed to compare the behavior of the unbiased estimator $\tilde{\eta}$, the maximum likelihood estimator $\hat{\eta}$ and the MRE estimator $\mathring{\eta}$ using 1000 simulated random samples of size $n = 10, 20, 30$ from the $HN(10, 4)$ distribution. The results obtained for the means and the mean squared errors of the three estimators are presented in Table 5 and Figure 5 (as before, a dotted red line represents the mean).

n	$\tilde{\eta}$	$\hat{\eta}$	$\mathring{\eta}$
10	3.996009 ± 1.052443	3.520680 ± 0.952987	3.568520 ± 0.929288
20	3.996575 ± 0.526328	3.760888 ± 0.458780	3.795590 ± 0.450882
30	4.015727 ± 0.324161	3.845478 ± 0.294937	3.871677 ± 0.291209

Table 3.5: Sample mean and MSE of the estimators calculated using 1000 random samples of size n from the $HN(10, 4)$ distribution

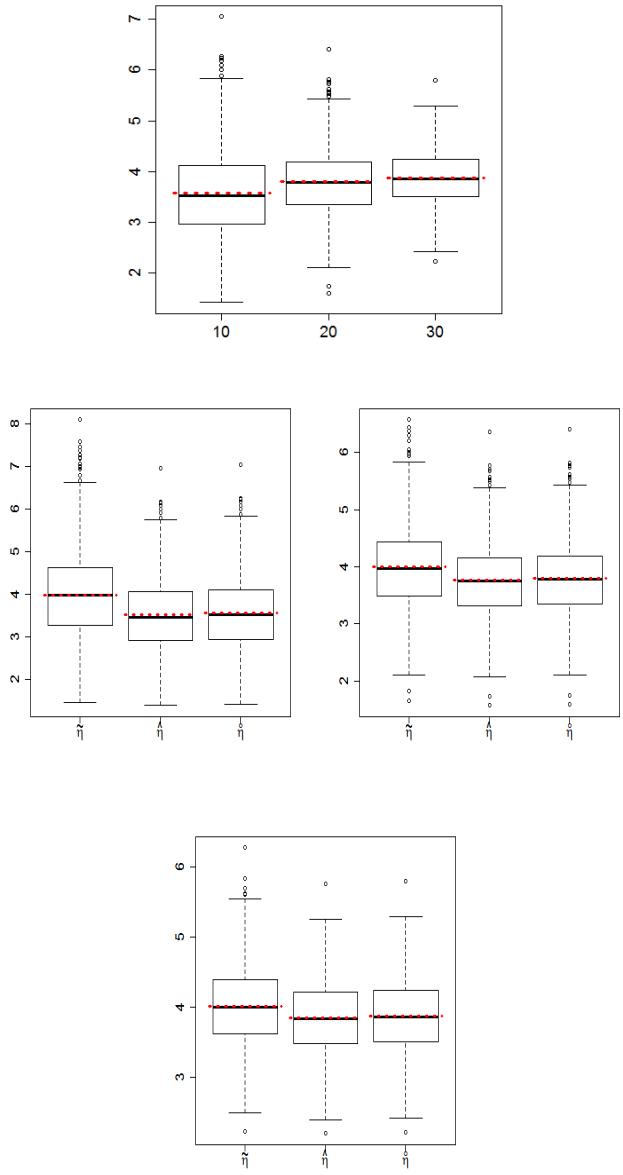

Figure 3.5: Box plots for the estimator $\mathring{\eta}$ for sample sizes $n = 10, 20, 30$ and for the estimators $\tilde{\eta}$, $\hat{\eta}$ y $\mathring{\eta}$ for sample sizes $n = 10, 20, 30$, respectively

Notice that both $\hat{\eta}$ and $\tilde{\eta}$ are equivariant estimators of the scale parameter η. So they have greater risk for the loss function W_1 than $\mathring{\eta}$. Hence (see Table 5 and Figure 5), in the previous simulation study, the MSE of $\mathring{\eta}$

is smaller than the MSE of $\hat{\eta}$ and $\tilde{\eta}$. □

Remark 3.9. Although less interesting from the perspective of real applications, for completeness we now consider the problem of estimating the scale parameter η when the location parameter ξ is known, say $\xi = \xi_0$. After the shift $(y_1, \ldots, y_n) \mapsto (y_1 - \xi_0, \ldots, y_n - \xi_0)$, the statistical model remains invariant under the transformations (dilations) of the form $(y_1, \ldots, y_n) \mapsto (ay_1, \ldots, ay_n)$, for $a > 0$. For the loss function $W_1'(\eta, x) = (x - \eta)^2/\eta^2$, the MRE estimator of the scale parameter η is

$$T_2 = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{2}\Gamma(\frac{n+2}{2})}\sqrt{\sum_{i=1}^{n}(Y_i - \xi_0)^2} = \frac{B(\frac{n+1}{2}, \frac{1}{2})}{\sqrt{2\pi}}\sqrt{\sum_{i=1}^{n}(Y_i - \xi_0)^2},$$

where Γ and B denote Euler's gamma and beta functions. In fact, for the loss function W_1', the MRE estimator of η is

$$T_2(y_1, \ldots, y_n) = \frac{\displaystyle\int_0^\infty v^n h_1(v(y_1 - \xi_0), ..., v(y_n - \xi_0))dv}{\displaystyle\int_0^\infty v^{n+1} h_1(v(y_1 - \xi_0), ..., v(y_n - \xi_0))dv},$$

where

$$h_1(y_1, \ldots, y_n) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right\} I_{[0,+\infty[}(y_{1:n}).$$

To simplify the notation, we assume without loss of generality that $\xi_0 = 0$. The change of variable $t = \frac{1}{2}\sum_{i=1}^{n} y_i^2 v^2$ leads to, for $k = n, n+1$,

$$\int_0^\infty v^k h_1(vy_1, ..., vy_n)dv = 2^{\frac{n+k-1}{2}}\pi^{-\frac{n}{2}}\left(\sum_{i=1}^{n} y_i^2\right)^{-\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right) I_{[0,+\infty[}(y_{1:n}),$$

and the assertion then follows easily.

Note also that, when $\xi = \xi_0$,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \xi_0)^2$$

is the minimum variance unbiased estimator of η^2. This is a consequence of the Lehmann-Scheffé Theorem and the facts that $\sum_{i=1}^{n}(Y_i - \xi_0)^2$ is a suffi-

cient and complete statistic and $\eta^{-2}\sum_{i=1}^{n}(Y_i-\xi_0)^2$ has a $\chi^2(n)$ distribution. A little more work shows that

$$\frac{\Gamma(\frac{n}{2})}{\sqrt{2}\Gamma(\frac{n+1}{2})}\sqrt{\sum_{i=1}^{n}(Y_i-\xi_0)^2}=\frac{B(\frac{n}{2},\frac{1}{2})}{\sqrt{2\pi}}\sqrt{\sum_{i=1}^{n}(Y_i-\xi_0)^2}$$

is the minimum variance unbiased estimator of η. □

3.5 References

Besicovitch, A.S. , A general form of the covering principle and relative differentiation of additive functions, I, *Proceedings of the Cambridge Philosophical Society* **41**, (1945), 103-110.

Besicovitch, A.S., A general form of the covering principle and relative differentiation of additive functions, II, *Proceedings of the Cambridge Philosophical Society* **42**, (1946), 205-235.

Bland, J.M., *The half-normal distribution method for measurement error: two case studies*, Unpublished talk available on http://www-users.york.ac.uk/ mb55/talks/halfnor.pdf, 2005).

Bland J.M., Altman D.G., Measuring agreement in method comparison studies, *Stat Methods Med Res.* **8**, (1999), 135-160.

Daniel, C., Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics* **1**, (1959), 311–341.

Johnson, N.L., Kotz, S., Balakrishnan, N., *Continuous Univariate Distributions*, Vol. 1, 2nd Ed.; Wiley: New York, 1994.

Härdle, W., *Applied Nonparametric Regression*, Econometric Society Monographs, 19, Cambridge University Press, Cambridge, 1992.

Lehamnn, E.L., *Theory of Point Estimation*, Wiley, 1983.

Mattila, P., *Geometry of sets and measures in euclidean spaces*, Cambridge University Press, New York, 1995.

Nadaraya, E.A., *Nonparametric Estimation of Probability Densities and Regressión Curves*, Kluwer Academic Publisher, 1989.

Nogales, A.G., Pérez, P., Unbiased Estimation for the General Half-Normal Distribution, *Comm. Statist. Theory Methods* (2011), to appear.

Pewsey, A., Large-sample inference for the general half-normal distribution, *Comm. Statist. Theory Methods* **31**, (2002), 1045–1054.

Pewsey, A., Improved likelihood based inference for the general half-normal distribution, *Comm. Statist. Theory Methods* **33**, (2004), 197–204.

Wiper, M. P., Girón, F. J., Pewsey, A., Objective Bayesian inference for the half-normal and half-t distributions *Comm. Statist. Theory Methods* **37**, (2008), 3165–3185.

Part III

Cluster analysis

4

A comparison of bandwidth selectors for mean shift clustering

4.1 Introduction

The mean shift algorithm was introduced by Fukunaga and Hostetler in a seminal paper in 1975, with the goal of estimating the gradient of a multivariate density. They also showed that their algorithm can be helpful for many applications in several pattern recognition problems, and particularly pointed out its usefulness for clustering and data filtering.

Even if this algorithm was highlighted in the popular book by Silverman (1986) as an important application of kernel smoothing, it remained relatively neglected in the Statistics literature, until it was "re-discovered" by Cheng (1995), Carreira-Perpiñán (2006) and Comaniciu and Meer (2002) for its applications in Engineering. Some recent contributions that make use of the mean shift algorithm, either explicitly or implicitly, are Li, Ray and Lindsay (2007), Ozertem and Erdogmus (2011), Genovese, Perone-Pacifico, Verdinelli and Wasserman (2012) or Chacón and Duong (2013).

Being closely related to the problem of density gradient estimation, the mean shift algorithm inherits its dependence on the choice of a suitable bandwidth matrix. It was only recently (see Chacón and Duong (2013) and Horová, Kolácek and Vopatová (2013)) that automatic methods for bandwidth selection for density gradient estimation were proposed. The goal of this paper is to provide a comparative study of the performance of these automatic bandwidth selectors, not with respect to the problem of

density gradient estimation, but regarding the clustering of the space that they induce via the mean shift algorithm.

The rest of the paper is organized as follows. In Section 2 below the clustering procedure derived from the mean shift algorithm is introduced. A brief review of the existing bandwidth matrix selectors for density gradient estimation is contained in Section 3. The details of the simulation study comparing these methodologies are given in Section 4 and some conclusions are discussed in Section 5. Finally, we show in an Appendix the ascending property of the mean shift algorithm with an unconstrained bandwidth matrix.

4.2 Mean shift clustering

Let us consider a probability density $f\colon \mathbb{R}^d \to \mathbb{R}$, and denote by $\mathsf{D}f$ its gradient vector, so that with the usual column notation for vectors $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ we have

$$\mathsf{D}f = \frac{\partial f}{\partial \boldsymbol{x}} = \Big(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\Big)^\top,$$

with $^\top$ standing for the transpose operator.

The mean shift algorithm is a variant of the well-known gradient ascent algorithm which is usually employed to find the local maxima of a given function. Explicitly, given any starting point $\boldsymbol{y}_0 \in \mathbb{R}^d$, the mean shift algorithm iteratively constructs a sequence $(\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots)$ according to the following updating mechanism

$$\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \mathbf{A}\mathsf{D}f(\boldsymbol{y}_j)/f(\boldsymbol{y}_j), \tag{4.1}$$

where $\mathbf{A}$ is a $d \times d$ positive definite matrix conveniently chosen to guarantee the convergence of the sequence. The only difference with the usual gradient ascent algorithm is that (4.1) uses the normalized gradient $\mathsf{D}f/f$ to accelerate the convergence when the starting point belongs to a low-density zone.

Since the shift at every step is done approximately along the gradient direction it follows that the limit point of the mean shift sequence should

be a local maximum of f (i.e., a mode of the density). This induces a clustering scheme in which any two points are said to belong to the same cluster whenever the sequences constructed from them as starting points converge to the same mode of f. In that case, it is also common to say that the two points belong to the same domain of attraction of such local maximum, and this type of clustering is called *modal clustering*.

Moreover, since the mean shift algorithm is applicable with any point in $\mathbb{R}^d$ as starting point, eventually this clustering scheme induces a partition of the whole space $\mathbb{R}^d$ into disjoint clusters. This partition, built up from the knowledge of the density f, will be referred to as the ideal population clustering. A precise definition of this ideal population clustering can be found in Chacón (2012).

When the density f is unknown, but a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from f is observed instead, the mean shift algorithm (4.1), with the density and the density gradient estimated from the sample, yields a data-based clustering of the whole space $\mathbb{R}^d$.

The goal of most clustering methodologies is not to partition $\mathbb{R}^d$, but only the data sample. Nevertheless, it is clear that by partitioning the whole space the mean shift algorithm induces, in particular, a clustering of the data by assigning two data points to the same cluster if they belong to the same component of the aforementioned partition of $\mathbb{R}^d$.

The density and density gradient estimators considered here are of kernel type. The kernel density estimator has the form

$$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_i),$$

where the kernel K is a spherically symmetric d-variate density function, the bandwidth matrix $\mathbf{H}$ is symmetric and positive definite, and we have used the re-scaling notation $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$ (see Wand and Jones (1995), Chapter 4). Then, following Chacón, Duong and Wand (2011), the density gradient estimator is just the gradient of the kernel density estimator, given by

$$\mathsf{D}\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \mathsf{D}K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}) = n^{-1} |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \sum_{i=1}^{n} \mathsf{D}K\big(\mathbf{H}^{-1/2}(\boldsymbol{x} - \mathbf{X})\big)$$

Being symmetric, the kernel K can be expressed as $K(\boldsymbol{x}) = \frac{1}{2}k(\boldsymbol{x}^\top \boldsymbol{x})$, where the function $k\colon \mathbb{R}_+ \to \mathbb{R}$ is known as the profile of K (see Comaniciu and Meer (2002)). Furthermore, the kernel K is usually assumed to be smooth and unimodal, so that $g(x) = -k'(x) \geq 0$. Thus, following the ideas of Fukunaga and Hostetler (1975), Chacón and Duong (2013) showed that a sensible estimator of the normalized gradient $\mathsf{D}f(\boldsymbol{x})/f(\boldsymbol{x})$ is $\mathbf{H}^{-1}\mathbf{m}_{\mathbf{H}}(\boldsymbol{x})$, where the term $\mathbf{m}_{\mathbf{H}}(\boldsymbol{x}) = \sum_{i=1}^n \omega_{i,\mathbf{H}}(\boldsymbol{x})\mathbf{X}_i - \boldsymbol{x}$ is known as the *mean shift*. It is the difference between a weighted mean of the data and $\boldsymbol{x}$, with the weights $\omega_{i,\mathbf{H}}(\boldsymbol{x})$ defined as

$$\omega_{i,\mathbf{H}}(\boldsymbol{x}) = \frac{g\big(M_{\mathbf{H}}(\boldsymbol{x}, \mathbf{X}_i)\big)}{\sum_{\ell=1}^n g\big(M_{\mathbf{H}}(\boldsymbol{x}, \mathbf{X}_\ell)\big)},$$

where $M_{\mathbf{H}}$ denotes the Mahalanobis distance $M_{\mathbf{H}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^\top \mathbf{H}^{-1}(\boldsymbol{x}-\boldsymbol{y})$. Hence, by plugging this estimate in (4.1) and taking $\mathbf{A} = \mathbf{H}$, the updating mechanism of the data-based mean shift algorithm simply reads

$$\boldsymbol{y}_{j+1} = \sum_{i=1}^n \omega_{i,\mathbf{H}}(\boldsymbol{y}_j)\mathbf{X}_i. \tag{4.2}$$

Originally, Fukunaga and Hostetler (1975) developed the mean shift algorithm using a constrained bandwidth matrix consisting of a scalar h^2 times the identity matrix, $h > 0$, and Comaniciu and Meer (2002) showed that for this constrained form the choice of $\mathbf{A} = \mathbf{H}$ guarantees that the mean shift sequence is convergent, as long as the kernel K has a convex and monotonically decreasing profile k. In the Appendix below we show that (4.2), the unconstrained version of the mean shift algorithm, is also convergent.

4.3 Bandwidth matrix selectors

As it is common for all kernel smoothing methods, the performance of mean shift clustering is highly influenced by the choice of the bandwidth matrix. Since the element having the biggest impact on the performance of the mean shift algorithm appears to be the density gradient, it seems reasonable that a bandwidth matrix chosen to obtain a good kernel density gradient estimate could lead to an appealing clustering via the mean shift algorithm.

Surprisingly, the literature tackling the problem of automatic, data-based bandwidth matrix selection for kernel density gradient estimation is quite scant and recent. We are aware only of two contributions dealing with this problem: Chacón and Duong (2013) and Horová, Kolácek and Vopatová (2013). In both papers the measure to evaluate the performance of the kernel density gradient estimator $\mathsf{D}\hat{f}_{\mathbf{H}}$ is the mean integrated squared error, defined as

$$\mathrm{MISE}(\mathbf{H}) = \int_{\mathbb{R}^d} \|\mathsf{D}\hat{f}_{\mathbf{H}}(\boldsymbol{x}) - \mathsf{D}f(\boldsymbol{x})\|^2 d\boldsymbol{x},$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^d$. With this goal in mind, the optimal bandwidth for kernel density gradient estimation is taken to be $\mathbf{H}_{\mathrm{MISE}}$, the minimizer of the MISE function over the class of all symmetric positive definite matrices.

In Chacón and Duong (2013), three bandwidth matrix selectors were proposed for kernel estimation of the r-th derivative of a multivariate density f, for arbitrary r. They are defined as the minimizers of certain criteria which aim to estimate the MISE. These criteria can be shown to generalize the well-known cross validation (CV), plug-in (PI) and smooth cross validation (SCV) methodologies proposed earlier for the base case of univariate density estimation (i.e., $d = 1$ and $r = 0$). In the case of the density gradient (arbitrary d and $r = 1$), these criteria can be written as

$$\begin{aligned}
\mathrm{CV}(\mathbf{H}) &= -n^{-2} \sum_{i,j=1}^{n} \nabla^2 K_{2\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j) + 2[n(n-1)]^{-1} \sum_{i \neq j} \nabla^2 K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j) \\
\mathrm{PI}(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\big\{\mathbf{H}^{-1}\mathbf{R}(\mathsf{D}K)\big\} \\
&\quad - \tfrac{1}{4}\{(\mathrm{vec}^\top \mathbf{I}_d) \otimes (\mathrm{vec}^\top \mathbf{H}) \otimes (\mathrm{vec}^\top \mathbf{H})\} n^{-2} \sum_{i,j=1}^{n} \mathsf{D}^{\otimes 6} K_{\mathbf{G}}(\mathbf{X}_i - \mathbf{X}_j) \\
\mathrm{SCV}(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\big\{\mathbf{H}^{-1}\mathbf{R}(\mathsf{D}K)\big\} \\
&\quad - n^{-2} \sum_{i,j=1}^{n} \nabla^2 \big\{K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}}\big\}(\mathbf{X}_i - \mathbf{X}_j),
\end{aligned}$$

respectively. Here, $\nabla^2 = \sum_{i=1}^{d} (\partial^2/\partial x_i^2)$ is the Laplace operator, tr denotes the trace operator, vec is the vectorization operator that transforms a matrix into a vector by stacking the columns of the matrix one underneath the other, $\otimes$ denotes the Kronecker product, $\mathbf{R}(\mathsf{D}K) = \int_{\mathbb{R}^d} \mathsf{D}K(\boldsymbol{x})\mathsf{D}K(\boldsymbol{x})^\top d\boldsymbol{x}$ is a $d \times d$

matrix, the vector $\mathsf{D}^{\otimes 6} K_{\mathbf{G}} \in \mathbb{R}^{d^6}$ includes all 6-th order partial derivatives of $K_{\mathbf{G}}$, arranged in a particular order (see Chacón and Duong (2010)), and $\mathbf{G}$ is a pilot bandwidth matrix. Computation of these criteria is not simple, but efficient implementations were proposed in Chacón and Duong (2014).

In Horová, Kolácek and Vopatová (2013) an iterative method (IT) was proposed to treat the cases $r = 0$ and $r = 1$ for arbitrary d. These authors noted that the asymptotic approximation of the optimal bandwidth $\mathbf{H}_{\mathrm{MISE}}$, so-called $\mathbf{H}_{\mathrm{AMISE}}$, can be characterized as the solution of a particular equation involving the unknown density f. So a sensible choice for the bandwidth is introduced as the solution of a data-based estimate of this equation, which for $r = 1$ can be written as

$$(d+2)n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\big\{\mathbf{H}^{-1}\mathbf{R}(\mathsf{D}K)\big\} \\ 4n^{-2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} \nabla^2\big\{K_{4\mathbf{H}} - 2K_{3\mathbf{H}} + K_{2\mathbf{H}}\big\}(\mathbf{X}_i - \mathbf{X}_j) = 0.$$

Again, the computational details to obtain the solution of this equation are not simple, and an iterative method to solve it (hence the name of this bandwidth selector) is proposed in Horová, Kolácek and Vopatová (2013).

All these methodologies focus on the most general form for the bandwidth matrix $\mathbf{H}$, which is only required to be symmetric and positive definite. Other popular choices for the bandwidth matrix include constrained forms such as $\mathbf{H}$ being diagonal, $\mathbf{H} = \mathrm{diag}(h_1^2, \ldots, h_d^2)$, or the parametrization using a single bandwidth $h > 0$ so that $\mathbf{H} = h^2\mathbf{I}_d$, with $\mathbf{I}_d$ denoting the $d \times d$ identity matrix.

The thorough study of Wand and Jones (1993) reported that for density estimation, in general, the diagonal parametrization results in a small loss of efficiency, but the single-bandwidth estimator should not be blindly used for unscaled multivariate data (see also Chacón (2009)). For density derivative estimation, Chacón, Duong and Wand (2011) showed that the loss of efficiency due to the use of simpler bandwidth matrix parametrizations can be even more severe. However, the goal of cluster analysis is quite different from that of density estimation, so that not very precise density estimates may equally lead to nearly optimal clusterings (see Chacón (2012), Figure 6, for

an illustration of this phenomenon), so in principle the simpler parametrizations should not be completely discarded. In fact, the very simple diagonal bandwidth proposal of Azzalini and Torelli (2007) was shown to produce good results in Chacón and Duong (2013). Therefore, unconstrained but also diagonal bandwidth matrices will be considered in the simulation study below.

4.4 Simulation study

The main goal of this paper is to provide an empirical comparison of the performance of several bandwidth selection methods for mean shift clustering.

Five true models are analyzed in the study, which cover a wide variety of cluster shapes. Two of these densities are normal mixture densities; hence a parametric cluster analysis of these two models, by fitting an estimated density through maximum likelihood, would probably yield quite good results (see Chacón and Duong (2013), and references therein). But to exploit the nonparametric nature of the mean shift approach we also include three densities with more intricate features which are not likely to be accurately recovered in a parametric setup. Figure 4.1 shows the true densities and the ideal population clusterings associated to each of these models, along with the names which we will use to refer to them. A precise definition of these models can be found in Chacón and Duong (2013).

The automatic bandwidth selectors compared in this study are the CV, PI and SCV bandwidths proposed in Chacón and Duong (2013) for density gradient estimation, the IT method introduced in Horová, Kolácek and Vopatová (2013), the normal-scale bandwidth (NS) for density gradient estimation introduced in Chacón, Duong and Wand (2011), and the simple proposal AT of Azzalini and Torelli (2007), which shrinks the diagonal normal-scale bandwidth for density estimation by a factor 3/4 (hence, it could be considered as a diagonal variant of the previous one). For the CV, PI, SCV and IT methods we also considered their respective diagonal versions, which are obtained by minimizing (or solving, in the case of IT) the objective criteria over the class of all positive definite diagonal matrices.

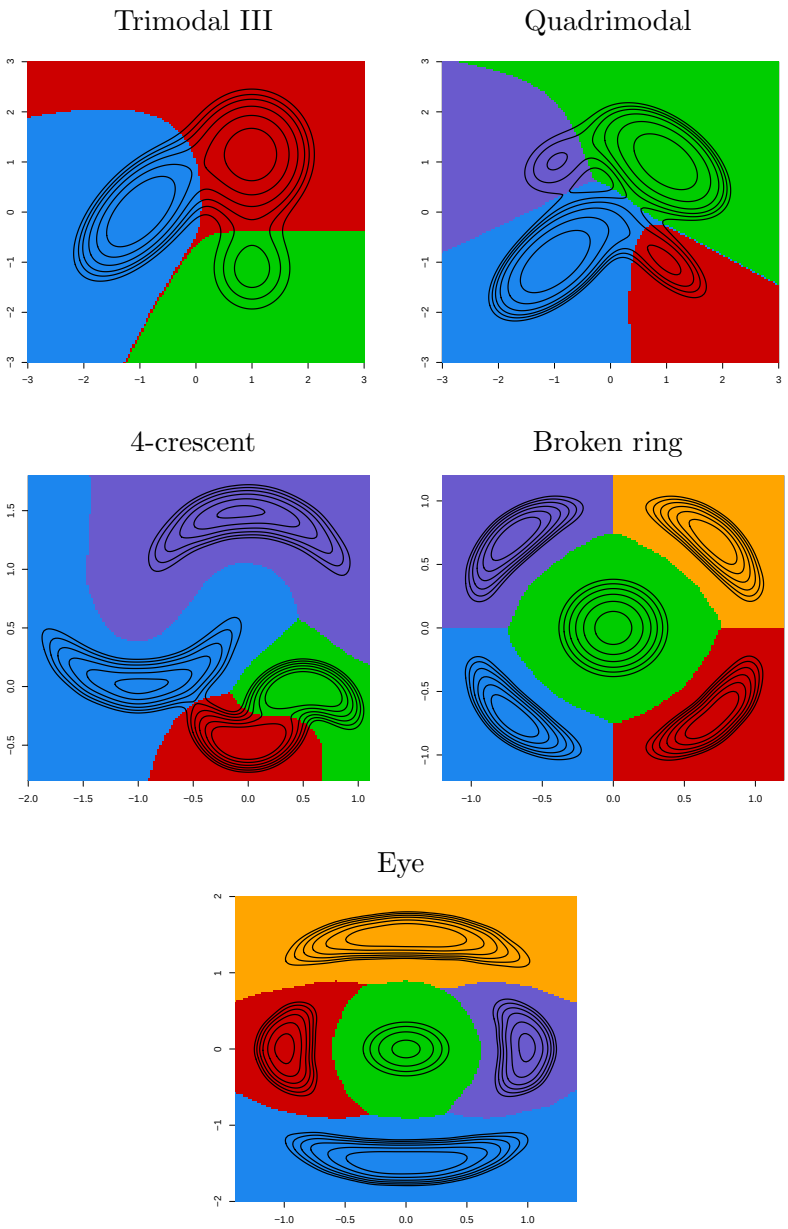

Figure 4.1: The five true density models included in the simulation study, with the ideal population clustering shown in different colors.

These are denoted by adding 'D' to their initials (i.e., CVD, PID, SCVD and ITD), while their unconstrained counterparts are represented by CVU, PIU, SCVU and ITU, respectively.

The measure of the performance of each of these methods is completely different than that employed in Chacón and Duong (2013). There, different clusterings of the data were compared by means of the adjusted Rand index criterion, introduced in Hubert and Arabie (1985). Here, the interest is not to compare different clusterings of the data, but clusterings of the whole space $\mathbb{R}^d$. Therefore, it is necessary to use a distance between clusterings of $\mathbb{R}^d$, and we will use the *distance in measure* proposed in Chacón (2012).

This distance is defined as follows: given two clusterings $\mathcal{C} = \{C_1, \ldots, C_r\}$ and $\mathcal{D} = \{D_1, \ldots, D_s\}$ of a probability measure P, with $r \leq s$, the distance in measure between them is defined as

$$d_P(\mathcal{C}, \mathcal{D}) = \frac{1}{2} \min_{\sigma \in \mathcal{P}_s} \sum_{i=1}^{s} P(C_i \triangle D_{\sigma(i)}),$$

where $\mathcal{P}_s$ denotes the set of permutations of $\{1, 2, \ldots, s\}$, the partition $\mathcal{C}$ has been enlarged by adding $s - r$ empty sets $C_{r+1} = \cdots = C_s = \varnothing$ if necessary, and $\triangle$ denotes the symmetric difference between two sets, namely $C \triangle D = (C \cap D^c) \cup (C^c \cap D)$.

Even if the true densities are known in a simulation study, the exact value of $d_P(\mathcal{C}, \mathcal{D})$ is difficult to compute in practice. We used a fine enough grid defined over a large rectangle chosen to contain at least 0.999 probability mass. This regular grid is ruled in rectangles by considering a tiny rectangle centered at each grid point with its sides of length half the distance to the next grid point in each coordinate direction. Each grid point is assigned to one cluster via the mean shift algorithm, and hence every cluster can be approximated by the union of tiny rectangles surrounding the grid points that are labeled to belong to it. By computing the probability mass of each tiny rectangle and adding up the contributions corresponding to the rectangles that approximate each symmetric difference we obtain an approximation of $P(C_i \triangle D_j)$ for each $i, j = 1, \ldots, s$. Finally, finding the minimum over all the permutations $\mathcal{P}_s$ is known as a linear sum assignment problem, and efficient algorithms to solve it are shown, e.g., in Papadimitriou and Steiglitz (1982).

One hundred samples of size $n = 500$ from each density in the study were drawn. For each of these samples all the ten bandwidth selectors NS, AT, CVU, CVD, PIU, PID, SCVU, SCVD, ITU and ITD were computed and a partition in clusters of the whole space was obtained through the mean shift algorithm. Finally, we recorded the distance in measure between such data-based partitions and the ideal population clustering. The sample medians and interquartile ranges (IQR) of these distances over the 100 samples are summarized in Table 4.1. The median and the IQR were preferable over the more usual mean and standard deviation because the distribution of these random distances in measure was generally skewed and contained some outliers.

In view of Table 4.1 it is clear that no bandwidth selector is uniformly preferable over the others. However, it seems clear that NR and AT nearly always induced a poor clustering (an exception is NR for the broken ring model). The reason for this bad performance could be partially explained by Table 4.2. There it is shown the distribution of the number of clusters obtained by each method along the 100 simulation runs for each density model (some unusually large number of clusters have been omitted for clarity). In Table 4.2 it is possible to appreciate that both NR and AT normally induce a number of clusters smaller than those appearing in the true model, which can be interpreted as a well-known oversmoothing effect, due to the fact that these two bandwidth selectors are based on a normal reference rule. As noted before, an exception is the broken ring model, where NR correctly identifies 5 clusters in all the cases. We acknowledge, however, that the density clustering approach of Azzalini and Torelli (2007) is not based on the mean shift algorithm; the goal here was to test if their appealingly simple bandwidth proposal were also suitable for mean shift clustering.

The performance of the IT methods is somehow erratic. ITU is the best method for the trimodal mixture density, and has moderately good results for the quadrimodal mixture density as well, but both ITU and ITD are unable to deal with more complicated features like those appearing in the last three density models. Again, a partial explanation for this is provided in Table 4.2, where it is shown that ITU, and especially ITD, tend to partition the space in only one cluster, thus presenting a highly oversmoothed

Model	Bandwidth selector				
	NR	AT	CVU	CVD	PIU
Trimodal III	6.37e-02	6.37e-02	1.04e-03	2.05e-04	9.96e-05
	(6.36e-02)	(5.15e-06)	(5.29e-02)	(3.42e-02)	(4.60e-03)
Quadrimodal	9.70e-02	9.68e-02	**1.63e-02**	4.75e-02	3.16e-02
	(2.34e-04)	(1.34e-04)	(4.35e-02)	(4.59e-02)	(4.65e-02)
4-crescent	1.21e-01	2.44e-01	2.03e-21	2.24e-21	2.44e-03
	(0.00e+00)	(0.00e+00)	(5.16e-22)	(5.34e-04)	(1.49e-02)
Broken ring	3.62e-14	2.81e-01	3.72e-14	3.36e-14	**2.58e-14**
	(6.12e-15)	(9.57e-02)	(9.90e-15)	(4.81e-15)	(5.33e-15)
Eye	2.67e-02	2.67e-02	1.61e-15	2.44e-02	**3.24e-16**
	(4.51e-17)	(1.04e-17)	(2.67e-02)	(2.67e-02)	(7.03e-17)

Model	Bandwidth selector				
	PID	SCVU	SCVD	ITU	ITD
Trimodal III	9.26e-05	3.19e-02	6.37e-02	**7.02e-05**	1.04e-04
	(1.27e-03)	(6.36e-02)	(6.36e-02)	(1.98e-02)	(5.19e-02)
Quadrimodal	4.74e-02	9.70e-02	9.68e-02	5.24e-02	9.79e-02
	(4.77e-02)	(3.09e-04)	(2.13e-04)	(4.82e-02)	(2.05e-01)
4-crescent	3.44e-03	**1.90e-21**	**1.84e-21**	1.21e-01	3.70e-01
	(3.91e-02)	(2.42e-05)	(5.51e-22)	(1.23e-01)	(0.00e+00)
Broken ring	**2.53e-14**	3.20e-14	3.12e-14	3.77e-01	3.77e-01
	(6.12e-15)	(4.59e-15)	(4.19e-15)	(0.00e+00)	(0.00e+00)
Eye	**3.37e-16**	4.40e-16	4.50e-16	3.24e-01	3.24e-01
	(3.20e-04)	(2.54e-17)	(3.34e-17)	(0.00e+00)	(0.00e+00)

Table 4.1: Sample median and (interquartile range) for the distance in measure between the data-based clusterings induces by each bandwidth selection method and the ideal population clustering along 100 simulation runs of sample size $n = 500$ of each distribution. The significantly best methods for each model are marked in bold font.

Trimodal

	No. of clusters						
H	1	2	**3**	4	5	6	7
NR	1	54	43	2	0	0	0
AT	1	90	8	0	0	0	0
CVU	1	11	34	28	12	9	1
CVD	0	13	41	24	12	2	3
PIU	1	14	51	23	10	1	0
PID	0	10	48	31	9	0	0
SCVU	4	46	45	5	0	0	0
SCVD	1	50	45	3	0	0	0
ITU	2	20	57	18	3	0	0
ITD	1	19	62	14	2	0	0

Quadrimodal

	No. of clusters						
H	1	2	3	**4**	5	6	7
NR	0	87	13	0	0	0	0
AT	0	97	3	0	0	0	0
CVU	0	14	28	31	16	7	2
CVD	0	19	44	27	8	1	0
PIU	0	20	30	30	13	5	0
PID	0	18	36	31	12	1	0
SCVU	0	83	16	1	0	0	0
SCVD	0	86	14	0	0	0	0
ITU	0	43	39	13	3	1	0
ITD	38	50	10	2	0	0	0

Eye

	No. of clusters						
H	1	2	3	4	**5**	6	7
NR	0	0	0	100	0	0	0
AT	0	0	1	98	1	0	0
CVU	0	0	0	44	52	3	1
CVD	0	0	0	30	42	12	3
PIU	0	0	0	0	78	18	3
PID	0	0	0	0	67	28	4
SCVU	0	0	0	0	99	1	0
SCVD	0	0	0	0	99	1	0
ITU	81	8	8	3	0	0	0
ITD	100	0	0	0	0	0	0

Broken ring

	No. of clusters						
H	1	2	3	4	**5**	6	7
NR	0	0	0	0	100	0	0
AT	48	33	12	6	1	0	0
CVU	0	0	9	2	86	2	1
CVD	0	0	0	0	98	1	0
PIU	0	0	0	0	95	5	0
PID	0	0	0	0	92	7	1
SCVU	0	0	0	0	98	2	0
SCVD	0	0	0	0	100	0	0
ITU	100	0	0	0	0	0	0
ITD	100	0	0	0	0	0	0

4-crescent

	No. of clusters								
H	1	2	3	**4**	5	6	7	8	9
NR	0	8	82	7	2	0	1	0	0
AT	0	99	1	0	0	0	0	0	0
CVU	0	0	1	81	18	0	0	0	0
CVD	15	1	0	63	17	0	3	0	0
PIU	0	0	0	13	28	37	18	2	2
PID	0	0	0	12	23	43	14	5	3
SCVU	0	0	0	74	24	2	0	0	0
SCVD	0	0	0	76	21	2	1	0	0
ITU	22	16	45	9	4	2	0	0	0
ITD	100	0	0	0	0	0	0	0	0

Table 4.2: Distribution of the number of clusters for each clustering method along the five density models. The true number of clusters is marked in bold.

estimate. The fact that both methods found only one cluster in all the cases for the broken ring model is the reason why the IQR of the distribution of their distances in measure is exactly zero (the distance in measure is a constant variable in this case).

The PI bandwidth selectors induce the clusterings with lowest distance in measure for the broken ring and eye models, and are close to the best performance in the two normal mixture density models, ranking second to best in terms of median error. They fail, however, to capture the features of the 4-crescent density, with a tendency to find more clusters than present (as seen in Table 4.2). The CV bandwidths perform disappointingly in the case of the trimodal mixture density, but CVU ranks first for the quadrimodal density model and both CVU and CVD obtain moderately good results for the densities with complicated features, frequently finding the right number of clusters. Both SCV proposals are probably the best ones concerning the right number of clusters for the densities with complicated features, and indeed they have the best marks for the 4-crescent model, but their behaviour is far from optimal for the normal mixture densities, with performances close to NR and AT.

With respect to the unconstrained-diagonal bandwidth dilemma, our study seems to suggest that diagonal bandwidths perform worse than their unconstrained counterparts in most of the cases. However, perhaps the use of diagonal matrices should not be blindly discarded, since indeed in some cases their performance is comparable or even slightly better than that of the unconstrained ones, but with a clearly smaller computational cost.

4.5 Conclusion

We explored here the influence of the bandwidth matrix in the mean shift algorithm from the point of view of modal clustering. Due to the crucial influence of the density gradient estimate in the mean shift algorithm we analyzed the practical performance of ten bandwidth selectors originally designed for density gradient estimation.

None of the ten automatic bandwidth matrix selectors showed a consistent superior performance over the rest of the methods in our simulation

study, but surely neither NR nor AT can be recommended for general use. All the CV, PI, SCV and IT proposals are best for one of the models, but utterly fail to identify the cluster structure for one, two or even three of the remaining ones. This suggests that the problem of bandwidth selection for mean shift clustering, though related, is different from that of bandwidth selection for density gradient estimation, and presents its own peculiarities, which undoubtedly deserve to be studied in further detail.

Since CVU and PIU are the only methods that failed solely for one of the density models, any of these two bandwidth matrix selectors would represent a cautious recommendation in practice, out of the ten methods studied here.

4.6 Appendix

Here it is shown that when the profile of the kernel is a bounded, convex, non-increasing, differentiable function, then the mean shift is an ascending algorithm; that is, the points of the sequence $(\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots)$ obtained through the mean shift algorithm attain sequentially increasing values of the estimated density $\hat{f}_{\mathbf{H}}$, so that the sequence $\big(\hat{f}_{\mathbf{H}}(\boldsymbol{y}_0), \hat{f}_{\mathbf{H}}(\boldsymbol{y}_1), \hat{f}_{\mathbf{H}}(\boldsymbol{y}_2), \ldots\big)$ is convergent.

Proof. Notice that since $K(\boldsymbol{x}) = \frac{1}{2}k(\boldsymbol{x}^\top \boldsymbol{x})$ it follows that $2|\mathbf{H}|^{1/2}K_{\mathbf{H}}(\boldsymbol{x} - \mathbf{X}_i) = k\big(M_{\mathbf{H}}(\boldsymbol{x}, \mathbf{X}_i)\big)$. Therefore,

$$2n|\mathbf{H}|^{1/2}\big\{\hat{f}_{\mathbf{H}}(\boldsymbol{y}_{j+1}) - \hat{f}_{\mathbf{H}}(\boldsymbol{y}_j)\big\} = \sum_{i=1}^{n}\big\{k\big(M_{\mathbf{H}}(\boldsymbol{y}_{j+1}, \mathbf{X}_i)\big) - k\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big)\big\}.$$

Then, following Comaniciu and Meer (2002), the convexity of the profile k implies that

$$\begin{aligned} k\big(M_{\mathbf{H}}(\boldsymbol{y}_{j+1}, \mathbf{X}_i)\big) - k\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big) \\ \geq k'\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big)\big\{M_{\mathbf{H}}(\boldsymbol{y}_{j+1}, \mathbf{X}_i) - M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big\}. \end{aligned}$$

Hence, expanding the difference between the two Mahalanobis distances we obtain

$$2n|\mathbf{H}|^{1/2}\big\{\hat{f}_{\mathbf{H}}(\boldsymbol{y}_{j+1}) - \hat{f}_{\mathbf{H}}(\boldsymbol{y}_j)\big\} \geq -\sum_{i=1}^{n} g\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big) \times \big\{\boldsymbol{y}_{j+1}^\top \mathbf{H}^{-1}\boldsymbol{y}_{j+1} - \boldsymbol{y}_j^\top \mathbf{H}^{-1}\boldsymbol{y}_j - 2(\boldsymbol{y}_{j+1} - \boldsymbol{y}_j)^\top \mathbf{H}^{-1}\mathbf{X}_i\big\}.$$

But definition (4.2) of the updating step entails that $\sum_{i=1}^n g\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big)\mathbf{X}_i = \sum_{i=1}^n g\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big)\boldsymbol{y}_{j+1}$ so that it is possible to replace $\mathbf{X}_i$ for $\boldsymbol{y}_{j+1}$ in the last term of the previous display, and simplify to get

$$2n|\mathbf{H}|^{1/2}\big\{\hat{f}_{\mathbf{H}}(\boldsymbol{y}_{j+1}) - \hat{f}_{\mathbf{H}}(\boldsymbol{y}_j)\big\} \geq \sum_{i=1}^{n} g\big(M_{\mathbf{H}}(\boldsymbol{y}_j, \mathbf{X}_i)\big) M_{\mathbf{H}}(\boldsymbol{y}_j, \boldsymbol{y}_{j+1}) \geq 0,$$

so the sequence $\big(\hat{f}_{\mathbf{H}}(\boldsymbol{y}_0), \hat{f}_{\mathbf{H}}(\boldsymbol{y}_1), \hat{f}_{\mathbf{H}}(\boldsymbol{y}_2), \dots\big)$ is non-decreasing and bounded, hence convergent. □

4.7 References

Azzalini, A. and Torelli, N. (2007) Clustering via nonparametric density estimation. *Stat. Comput.*, **17**, 71–80.

Carreira-Perpiñán, M.Á. (2006) Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal.*, **22**, 1318–1323.

Chacón, J.E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.* **37**, 249–265.

Chacón, J.E. (2012) Clusters and water flows: a novel approach to modal clustering through Morse theory. *arXiv preprint arXiv:1212.1384.*

Chacón, J.E. and Duong, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398.

Chacón, J.E. and Duong, T. (2013) Bandwidth selection for multivariate density derivative estimation, with applications to clustering and bump hunting. *Electron. J. Statist.*, **7**, 499–532.

Chacón, J.E. and Duong, T. (2013) Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate Gaussian density. *arXiv preprint arXiv:1310.2559.*

Chacón, J.E., Duong, T. and Wand, M.P. (2011) Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**, 807–840.

Cheng, Y. (1995) Mean shift, mode seeking, and clustering. *IEEE T. Pattern Anal.*, **17**, 790–799.

Comaniciu, D. and Meer, P. (2002) Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal.*, **24**, 603–619.

Fukunaga, K. and Hostetler, L.D. (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory*, **21**, 32–40.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012) Nonparametric ridge estimation. *arXiv preprint arXiv:1212.5156.*

Horová, I., Koláček, J. and Vopatová, K. (2013) Full bandwidth matrix selectors for gradient kernel density estimate. *Comput. Statist. Data Anal.*, **57**, 364–376.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

Li, J., Ray, S. and Lindsay, B.G. (2007) A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, **8**, 1687–1723.

Ozertem, U. and Erdogmus, D. (2011) Locally defined principal curves and surfaces. *J. Mach. Learn. Res.*, **12**, 241–274.

Papadimitriou, C. and Steiglitz, K. (1982) *Combinatorial Optimization: Algorithms and Complexity.* Prentice Hall, Englewood Cliffs.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC, London.

Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.*, **88**, 520–528.

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

Part IV

Conclusions and future research

5 Conclusions

Some of the main results obtained in this Thesis are the following:

- We use Fourier transform techniques to obtain exact expressions for the mean integreated squared error of kernel distribution function estimators. These expressions are then employed to analyze the asymptotic behavior of the kernel estimators optimal bandwidth sequence in its greatest generality, including the cases where superkernels and the sinc kernel are used. The results thus obtained can guide our steps in order to develope automatic bandwidth selectors in the future.

- We show the existence of classes of distributions for which the kernel distribution estimator presents a first-order improvement over its empirical counterpart, opposite to the usual situation, where only second-order improvements are possible.

- In relation to the half-normal distribution, we have developed an explicit expression for the minimum risk equivariant (MRE) estimator of the scale parameter and we have determined the MRE estimators of scale and location parameters when the other one is unknown.

- We have modified a natural Monte Carlo method of approximation of conditional expectations in order to approximate the MRE estimation of the location parameter of a half-normal distribution. Several simulations have been performed to analyze the behavior of this two methods obtaining a wide variety of graphics about this.

- The mean shift algorithm is used as the basis for nonparametric, density-based clustering. We combine recent advances in multivariate bandwidth selection for density gradient estimation with a novel population background proposal for cluster analysis to compare the performance of several nonparametric clustering methods in practice.
- Our conclusions suggest that the problem of bandwidth selection for mean shift clustering presents its own peculiarities and deserves further study. None of of the existing bandwidth selection methods dominated all the others, but the cross-validation and plug-in selectors for density gradient estimation show promise as automatic standards for nonparametric mean shift clustering.

6 Future Research

We enumerate the most important questions and points that have appeared during the development of this dissertation which will be studied in the near future.

6.1 R package

Nowadays, there are several important packages in R in relation to kernel estimation. Some of them are KernSmooth (Wand, 2006), sm (Bowman and Azzalini, 2007) and ks (Duong, 2013).

The KernSmooth package is the first important package for kernel estimation. In this package, Wand implements functions for kernel smoothing and density estimation corresponding to the book *Kernel Smoothing* (Wand and Jones, 1995).

The sm package provides most of the smoothing methods for nonparametric regression and density estimation published in the book *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* (1997), by A. Azzalini and A. Bowman.

The ks package is focuses on kernel density estimation for multivariate data. In this package several tools are implemented such as diagonal and unconstrained data-driven bandwidth matrices for kernel density estimation and selectors for 1- to 6-dimensional data.

We are planning to create a new R package with several necessary and unavailable functions in kernel estimation. We are interested in implementing modal clustering methods such as the mean shift algorithm. Besides, it will be useful to include in the development of this package several variants of the mean shift algorithm like median shift, medioid shift and others.

To improve its applicability it would be convenient to use Improved Fast Gauss Transform (Raykar, Duraiswami and Zhao, 2010) to compute each step in the mean shift algorithm. To our knowledge, this fast transform has not been implemented in R to date.

6.2 New methods for bandwidth selection

In Chapter 2 of this thesis we have developed several new formulas to quantify the MISE of the kernel distribution estimator in terms of characteristic functions. Using these expressions we want to propose new bandwidth selectors based on these formulas.

For instance, replacing the characteristic function of the distribution with its empirical counterpart we would obtain an unbiased estimator of the MISE (perhaps shifted by a quantity independent of the bandwidth). This would lead us to a new kind of cross-validation bandwidth selector for kernel distribution function estimation. It would be interesting then, to compare this new cross-validation criterion with that of Bowman, Hall and Prvan (1998).

Looking further afield, if the proposed selector behaves like other cross-validation bandwidths, which tend to be quite variable, it would be interesting to investigate if one could modify this selector as in Chiu (1992), to obtain a more stabilized version.

6.3 Integrated regression

In a seminal paper, Stute (1997) proposed to use the integrated regression function for testing the goodness of fit of a parametric regresion model. The proposed procedures are useful to test if the regression function belongs to a specific family or not.

However, little is known about the estimation of this integrated regression function. Stute (1997) showed that it can be empirically estimated, in a way that makes it look like the natural analogue of the distribution function in the regression setting.

It would be interesting to study the problem of the estimation of the integrated regression function in greater detail, and particularly, to propose alternative smooth estimators, since the empirical one resembles the empirical distribution function in the sense that it is not continuous. In this context, not only kernel but also local polynomial estimators could be proposed and analyzed.

6.4 Optimal bandwidth selection for other error measures

The methodology used in most of this thesis for the problem of distribution function estimation is based on the MISE. Little is known, however, about the problem of selecting the bandwidth in kernel distribution function estimation with respect to the uniform error. Most papers dealing with this error measure are concerned only with conistency results.

We are interested in developing methods to obtain optimal bandwidth selectors in the case where the performance of the kernel distribution function estimator is measured by the uniform error.

6.5 Fast Fourier Transform

Fast implementations of kernel density estimators based on the use of the fast Fourier transform are commonly used in statistical sofware. A faster implementation for the kernel distribution estimator would be a really desirable tool.

Along the writing of the Chapter 2 of this thesis we realized that there are also explicit inversion formulas that are useful to express a distribution function from its characteristic function (namely, the so-called Gil-Pelaez formula). Using these formulas it would be possible to develop a Fast Fourier Transform implementation of the kernel distribution function estimator.

A

Simulation Programs

The programs for the simulations given in the papers of this Thesis have been done through the statistical software and programming environment R.

Simulations in Chapter 3

Conditional Expectation Calculation

The function simulacionBinormal generates an approximation of $E(Y|X = 1)$ and the boxplots of the Figure 1 for (X, Y) such as in Example 1. There are 3 inputs: epsilon, numerosimulaciones (the number of replications) and tamanomuestra (the sample size).

```
library(mvtnorm)

simulacionBinormal<-function(epsilon=0.1,numerosimulaciones=50,tamanomuestra
=c(10,20,30)){

a<-c()
esperanza<-c()

for(i in tamanomuestra){
   muestrafinal<-matrix(,1,2)
   while(dim(muestrafinal)[1]<(i*numerosimulaciones+1)){
      muestra<-rmvnorm(1,mean=c(0,0),sigma=matrix(c(1,0.5,0.5,1),2,2))
      if((muestra[1]<(1+epsilon))&&(muestra[1]>(1-epsilon))){
         muestrafinal<-rbind(muestrafinal,muestra)
```

```
      }
   }

   for(z in 0:(numerosimulaciones-1)){
      esperanza<-c(esperanza,sum(muestrafinal[1+z*i+1:i,2])/i)
   }
}

png(filename=paste("Graph_Binormal_Eps",epsilon,"and",numerosimulaciones,
"rep.png"))

plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7),ylim=c(min(esperanza)-1,
max(esperanza)+1))

for(s in 0:(length(tamanomuestra)-1)){
   estim<-esperanza[ s*numerosimulaciones+1:numerosimulaciones ]
   boxplot(estim,add=T,at=s+1,width=1,cex.axis=1.5)
   segments(s+1-0.45,mean(estim),s+1+0.45,mean(estim),col=2,lwd=3,lty=3)
   mtext(tamanomuestra[s+1],1,at=s+1)
   a<-c(a,mean(estim))
}

abline(h=0.5)

dev.off()

save(a,esperanza,file=paste("Sim_Binormal_Eps",epsilon,"_",numerosimulaciones,
"rep.RData",sep=""))

}

simulacionBinormal(0.01,100,c(100,1000,5000))
simulacionBinormal(0.1,100,c(100,1000,5000))
```

The function simulacionBinormalSenCos generates an approximation of $E(V|U = 0.5)$ and the boxplots of the Figure 2 for (U,V) such as in the example 2. There are 3 inputs: epsilon, numerosimulaciones (the number of replications) and tamanomuestra (the sample size).

```
library(mvtnorm)
```

```
simulacionBinormalSenCos<-function(epsilon=0.1,numerosimulaciones=50,tamanomuestra
=c(10,20,30)){

esperanza<-c()
for(i in tamanomuestra){
   muestrafinal<-matrix(,1,2)
   while(dim(muestrafinal)[1]<(i*numerosimulaciones+1)){
      muestra<-rmvnorm(1,mean=c(0,0),sigma=matrix(c(1,0.5,0.5,1),2,2))
      if(((cos(sum(muestra^2)))<(0.5+epsilon))&&((cos(sum(muestra^2)))
      >(0.5-epsilon))){
         muestrafinal<-rbind(muestrafinal,muestra)
      }
   }
   for(z in 0:(numerosimulaciones-1)){
      esperanza<-c(esperanza,sum(sin(muestrafinal[1+z*i+1:i,1]*muestrafinal
      [1+z*i+1:i,2]))/i)
   }
}

png(filename=paste("Graph_BinormalSenCos_Eps",epsilon,"and",numerosimulaciones,
"rep.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7),ylim=c(min(esperanza)-1,
max(esperanza)+1))
a<-c()

for(s in 0:(length(tamanomuestra)-1)){
   estim<-esperanza[ s*numerosimulaciones+1:numerosimulaciones ]
   boxplot(estim,add=T,at=s+1,width=1,cex.axis=1.5)
   segments(s+1-0.45,mean(estim),s+1+0.45,mean(estim),col=2,lwd=3,lty=3)
   mtext(tamanomuestra[s+1],1,at=s+1)
   a<-c(a,mean(estim))
}

dev.off()

save(a,esperanza,file=paste("Sim_BinormalSenCos_Eps",epsilon,"_",
numerosimulaciones,"rep.RData",sep=""))

}

simulacionBinormalSenCos(0.1,100,c(100,1000,5000))
simulacionBinormalSenCos(0.01,100,c(100,1000,5000))
```

Parameters Estimation

The function simulacionHN generates an Excel archive where several parameters are saved in each column. Some of these parameters are $\tilde{\xi}$, $\overset{\circ}{\xi}$, $\hat{\xi}$, $\tilde{\eta}$, $\overset{\circ}{\eta}$ and $\hat{\eta}$. There are 6 inputs: the parameters ξ and η of the half-normal, tamanovector (sample size), numeromuestras (number of replications), cuantosyprima (number of vectors used to calculating the approximations) and ep (epsilon).

```
library(Matrix)
library(nor1mix)
library(XLConnect)

simulacionHN<-function(xi=0,eta=1,tamanovector,numeromuestras=1,cuantosyprima
=100,ep){

nombre<-paste("resultados_xi",xi,"_eta",eta,"_",numeromuestras,"muestras_de
_tamano",tamanovector,"con",cuantosyprima,"valores_y_prima_y_",ep,"_como_
epsilon.xlsx",sep="")

wb<-loadWorkbook(nombre,create=T)
createSheet(wb,name="resultados")
nombresvariables<-c("cn","ey","vary","xitilde","etatilde","xigorro","etagorro",
"t0estrella","t1estrella","pho","xiyprima","etacirculo")

writeWorksheet(wb,t(nombresvariables),sheet=1,startRow=1,startCol=1,header=F,
rownames=F)

saveWorkbook(wb)

norma<-function(v,p=2){
   return(  sum(abs(v)^p)^(1/p)   )
}

for(contador in 1:numeromuestras){
   z<-rnorm(tamanovector)
   x<-abs(z)
   y<-xi+eta*x

   cnfuncion<-function(x,n){
      (2-2*pnorm(x))^n
   }
```

```
cn<-integrate(cnfuncion,0,3,n=tamanovector)$value
ey<-xi+eta*sqrt(2/pi)
vary<-((pi-2)/pi)*eta^2
xitilde<-(  (sqrt(2/pi)*min(y)-cn*mean(y) )   /  (sqrt(2/pi)-cn)     )
etatilde<- (mean(y)-min(y))/(sqrt(2/pi)-cn)
xigorro<-min(y)
etagorro<-sqrt(sum((y-min(y))^2)/tamanovector)
etacirculo<-(gamma((tamanovector+1)/2)/gamma((tamanovector+2)/2))*sqrt
((tamanovector-1)/2)*sqrt(var(y))*((1-pt(sqrt((tamanovector*(tamanovector+1))
/(tamanovector-1))*((mean(y)-min(y))/sqrt(var(y))),tamanovector+1)))/((1-pt(
sqrt((tamanovector*(tamanovector+2))/(tamanovector-1))*((mean(y)-min(y))
/sqrt(var(y))),tamanovector+2)))

t0estrella<-mean(y)
t1estrella<-(sum(abs(y-mean(y)))/tamanovector)
u<-(y[1:tamanovector-1]-y[tamanovector])/c(rep(y[tamanovector-1]
-y[tamanovector],tamanovector-2),abs(y[tamanovector-1]-y[tamanovector]))


f<-function(y){
   mean(y)*(sum(abs(y-mean(y))))/length(y)*exp(-0.5*sum(y^2))
}

g<-function(y){
   ((sum(abs(y-mean(y))))/length(y))^2*exp(-0.5*sum(y^2))
}

transformacion<-function(x){
   return(c((x[1:(length(x)-2)]-x[length(x)])/(x[length(x)-1]-x[length(x)]),
   (x[length(x)-1]-x[length(x)])/(abs(x[length(x)-1]-x[length(x)]))))
}

muestrainicial<-matrix(0,cuantosyprima,tamanovector)

transformadadey<-transformacion(y)
epsilon<-min(ep,min(abs(transformadadey[1:(tamanovector-2)])))

for(i in 1:cuantosyprima){

   z2<-runif(2,0,10)
   if(sign(z2[1]-z2[2])==sign(y[tamanovector-1]-y[tamanovector])){
      muestrainicial[i,c(tamanovector-1,tamanovector)]<-z2
   } else{
      muestrainicial[i,c(tamanovector-1,tamanovector)]<-c(z2[2],z2[1])
   }
```

```
      for(j in 1:(tamanovector-2)){
         aux1<-muestrainicial[i,tamanovector]+(muestrainicial[i,tamanovector-1]
         -muestrainicial[i,tamanovector])*(transformadadey[j]-epsilon)
         aux2<-muestrainicial[i,tamanovector]+(muestrainicial[i,tamanovector-1]
         -muestrainicial[i,tamanovector])*(transformadadey[j]+epsilon)
         muestrainicial[i,j]<-runif(1,min(aux1,aux2),max(aux1,aux2))
      }
   }

   d<-min(muestrainicial)

   if(d<0){
      asumar<-runif(1,-d,1-d)
      muestrainicial<-muestrainicial+asumar
   }

   maximo2<-apply(muestrainicial,1,max)
   muestrainicial<-muestrainicial/maximo2*runif(1,0,10)

   nepsilon<-apply(muestrainicial,1,f)
   depsilon<-apply(muestrainicial,1,g)
   pho<-sum(nepsilon)/sum(depsilon)
   xicomplicado<-t0estrella-pho*t1estrella

   haches<-c(cn,ey,vary,xitilde,etatilde,xigorro,etagorro,t0estrella,
   t1estrella,pho,xicomplicado,etacirculo)

   wb<-loadWorkbook(nombre,create=F)
   writeWorksheet(wb,t(haches),sheet=1,startRow=contador+1,startCol=1,
   header=F,rownames=F)
   saveWorkbook(wb)

   cat("Simulacion ",contador, " de ", numeromuestras,"\n")

}
}
```

When the Excel archives have been created with the function simulacionHN, all the boxplots and graphs included in this thesis in the figures 3, 4 and 5 are drawn with next program.

```
library(XLConnect)
```

```
wb1<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano100con10000valores
_y_prima_y_0.1_como_epsilon.xlsx",create=F)
wb2<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano100con10000valores
_y_prima_y_0.01_como_epsilon.xlsx",create=F)
wb3<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano1000con10000valores
_y_prima_y_0.1_como_epsilon.xlsx",create=F)
wb4<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano1000con10000valores
_y_prima_y_0.01_como_epsilon.xlsx",create=F)
wb5<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano5000con10000valores
_y_prima_y_0.1_como_epsilon.xlsx",create=F)
wb6<-loadWorkbook("resultados_xi10_eta4_100muestras_de_tamano5000con10000valores
_y_prima_y_0.01_como_epsilon.xlsx",create=F)


b1<-readWorksheet(wb1,sheet="Resultados")
b2<-readWorksheet(wb2,sheet="Resultados")
b3<-readWorksheet(wb3,sheet="Resultados")
b4<-readWorksheet(wb4,sheet="Resultados")
b5<-readWorksheet(wb5,sheet="Resultados")
b6<-readWorksheet(wb6,sheet="Resultados")


mediasapintar<-apply(b1,2,mean,na.rm=T)[11]
mediasapintar1<-apply(b1,2,mean,na.rm=T)[c(4,6,11)]
mediasapintareta1<-apply(b1,2,mean,na.rm=T)[c(5,7,12)]
apply(b1,2,median,na.rm=T)
apply((b1[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


mediasapintar[2]<-apply(b2,2,mean,na.rm=T)[11]
mediasapintar2<-apply(b2,2,mean,na.rm=T)[c(4,6,11)]
mediasapintareta2<-apply(b2,2,mean,na.rm=T)[c(5,7,12)]
apply(b2,2,median,na.rm=T)
apply((b2[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


mediasapintar[3]<-apply(b3,2,mean,na.rm=T)[11]
mediasapintar3<-apply(b3,2,mean,na.rm=T)[c(4,6,11)]
mediasapintareta3<-apply(b3,2,mean,na.rm=T)[c(5,7,12)]
apply(b3,2,median,na.rm=T)
apply((b3[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


mediasapintar[4]<-apply(b4,2,mean,na.rm=T)[11]
mediasapintar4<-apply(b4,2,mean,na.rm=T)[c(4,6,11)]
```

```
mediasapintareta4<-apply(b4,2,mean,na.rm=T)[c(5,7,12)]
apply(b4,2,median,na.rm=T)
apply((b4[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


mediasapintar[5]<-apply(b5,2,mean,na.rm=T)[11]
mediasapintar5<-apply(b5,2,mean,na.rm=T)[c(4,6,11)]
mediasapintareta5<-apply(b5,2,mean,na.rm=T)[c(5,7,12)]
apply(b5,2,median,na.rm=T)
apply((b5[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


mediasapintar[6]<-apply(b6,2,mean,na.rm=T)[11]
mediasapintar6<-apply(b6,2,mean,na.rm=T)[c(4,6,11)]
mediasapintareta6<-apply(b6,2,mean,na.rm=T)[c(5,7,12)]
apply(b6,2,median,na.rm=T)
apply((b6[,c(4,6,11)]-10)^2,2,mean,na.rm=T)


tamanomuestra<-c(100,1000,5000)


### Boxplot A1. (several xi's)
png(filename=paste("HN_tamano ",100,"_epsilon ",0.1,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b1[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar1[s+1],s+1+0.45,mediasapintar1[s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Boxplot A2. (several xi's)
png(filename=paste("HN_tamano ",100,"_epsilon ",0.01,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b2[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar2[s+1],s+1+0.45,mediasapintar2[s+1],col=2,
   lwd=3,lty=3)
}
```

```
dev.off()


### Boxplot A3. (several xi's)
png(filename=paste("HN_tamano ",1000,"_epsilon ",0.1,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b3[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar3[s+1],s+1+0.45,mediasapintar3[s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Boxplot A4. (several xi's)
png(filename=paste("HN_tamano ",1000,"_epsilon ",0.01,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b4[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar4[s+1],s+1+0.45,mediasapintar4[s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Boxplot A5. (several xi's)
png(filename=paste("HN_tamano ",5000,"_epsilon ",0.1,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b5[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar5[s+1],s+1+0.45,mediasapintar5[s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Boxplot A6. (several xi's)
png(filename=paste("HN_tamano ",5000,"_epsilon ",0.01,".png"))
plot.new()
```

```
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b6[,c(4,6,11)],names=c(expression(tilde(xi)),expression(hat(xi)),
expression(ring(xi))),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar6[s+1],s+1+0.45,mediasapintar6[s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Boxplot B1. (xi ring. epsilon 0.1)
png(filename=paste("HN_nuevo xi_epsilon ",0.1," and ",100,"replications.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(cbind(b1[,11],b3[,11],b5[,11]),names=c(100,1000,5000),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar[2*s+1],s+1+0.45,mediasapintar[2*s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Grafico B2. (xi ring. epsilon 0.1)
png(filename=paste("HN_nuevo xi_epsilon ",0.01," and ",100,"replications.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(cbind(b2[,11],b4[,11],b6[,11]),names=c(100,1000,5000),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar[2*s+2],s+1+0.45,mediasapintar[2*s+2],col=2,
   lwd=3,lty=3)
}
dev.off()


### Grafico C1. (several eta's)
png(filename=paste("HN_Etas con tamano ",100,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b1[,c(5,7,12)],names=c(expression(tilde(eta)),expression(hat(eta)),
expression(ring(eta))),cex.axis=1.5)
# Pintar las medias:
for(s in 0:2){
segments(s+1-0.45,mediasapintareta1[s+1],s+1+0.45,mediasapintareta1[s+1],col=2,
lwd=3,lty=3)
}
```

```
dev.off()



### Grafico C2. (several eta's)
png(filename=paste("HN_Etas con tamano ",1000,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b3[,c(5,7,12)],names=c(expression(tilde(eta)),expression(hat(eta)),
expression(ring(eta))),cex.axis=1.5)
# Pintar las medias:
for(s in 0:2){
segments(s+1-0.45,mediasapintareta3[s+1],s+1+0.45,mediasapintareta3[s+1],col=2,
lwd=3,lty=3)
}
dev.off()


### Grafico C3. (several eta's)
png(filename=paste("HN_Etas con tamano ",5000,".png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(b5[,c(5,7,12)],names=c(expression(tilde(eta)),expression(hat(eta)),
expression(ring(eta))),cex.axis=1.5)
# Pintar las medias:
for(s in 0:2){
segments(s+1-0.45,mediasapintareta5[s+1],s+1+0.45,mediasapintareta5[s+1],col=2,
lwd=3,lty=3)
}
dev.off()


### Grafico D1. (eta tilde)
png(filename=paste("HN_eta tilde and ",100,"replications.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(cbind(b1[,5],b3[,5],b5[,5]),names=c(100,1000,5000),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar[2*s+1],s+1+0.45,mediasapintar[2*s+1],col=2,
   lwd=3,lty=3)
}
dev.off()
```

```
### Grafico D2. (eta hat)
png(filename=paste("HN_eta gorro and ",100,"replications.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(cbind(b1[,7],b3[,7],b5[,7]),names=c(100,1000,5000),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar[2*s+1],s+1+0.45,mediasapintar[2*s+1],col=2,
   lwd=3,lty=3)
}
dev.off()


### Grafico D3. (eta ring)
png(filename=paste("HN_eta circulo and ",100,"replications.png"))
plot.new()
plot.window(xlim=c(0.3,length(tamanomuestra)+0.7))
boxplot(cbind(b1[,12],b3[,12],b5[,12]),names=c(100,1000,5000),cex.axis=1.5)
for(s in 0:2){
   segments(s+1-0.45,mediasapintar[2*s+1],s+1+0.45,mediasapintar[2*s+1],col=2,
   lwd=3,lty=3)
}
dev.off()
```

Simulations in Chapter 4

Population clusters for normal mixture densities

The pop.clust function draws the population clusters for a normal mixture density. It has the following inputs: mus, Sigmas and props (means, variance matrix and proportions for the normal mixture) and several graphical parameters (gsize, fact, M, pc and ce).

```
pop.clust<-function(mus,Sigmas,props,gsize=50,fact=3,M=NULL,pc=15,ce=1.5){
    K<-length(props)
    if(K==1){mus<-matrix(mus,nrow=1)}

    if(is.null(M)){
    limits<-numeric()
    for(i in 1:K){
        mu<-mus[i,]
```

```
    sigma<-Sigmas[(2*i-1):(2*i),]
    ev <- eigen(sigma, symmetric = TRUE)
    s <- fact*sqrt(rev(sort(ev$values)))
    V <- t(ev$vectors[, rev(order(ev$values))])
    x <- s[1]
    y <- s[2]
    xy <- cbind(c(x, -x, 0, 0), c(0, 0, y, -y))
    xy <- xy %*% V
    xy <- sweep(xy, MARGIN = 2, STATS = mu, FUN = "+")
    limits<-rbind(limits,xy)
    }
xl1<-min(limits[,1]);xl2<-max(limits[,1])
yl1<-min(limits[,2]);yl2<-max(limits[,2])
}
else{
xl1<-yl1<--M;xl2<-yl2<-M
}


xs<-seq(xl1,xl2,length=gsize)
ys<-seq(yl1,yl2,length=gsize)
xys<-expand.grid(xs,ys)
fxys<-dmvnorm.mixt(xys,mus=mus,Sigmas=Sigmas,props=props)
zs<-invvec(fxys)
clus.labs<-MS.muestral(x=xys,mus=mus,Sigmas=Sigmas,props=props)$labels
contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
.8,.85,.9,.95)),
xaxs="i",yaxs="i",xlab="",ylab="")
classPlotColors <- c("dodgerblue2", "red3", "green3", "slateblue",
"orange", "skyblue1", "forestgreen", "steelblue4", "gray", "brown",
"black")
points(xys,col=classPlotColors[clus.labs],pch=pc,cex=ce)
contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
.8,.85,.9,.95)),
add=TRUE)
}
```

Population clusters for densities 3, 4 and 5

These three functions (pop.clust3, pop.clust4 and pop.clust5) calculate the population clusters for 4 crescent, broken ring and eye densities. They have 3 inputs: n, gsize and pc. They are similar to the previous functions.

```
library(ks)
library(mclust)



h3<-function(x,Nruns=10){
    h0<-0
    h1<-1
    n0<-Inf
    n1<-1
    for(i in 1:Nruns){
        hmed<-(h0+h1)/2
        nmed<-MS.kde(x=x,H=hmed^2*diag(2))$nclus
        if(nmed>4){h0<-hmed;n0<-nmed}
        if(nmed<=4){h1<-hmed;n1<-nmed}
    }
    return(h1)
}


pop.clust3<-function(n=100,gsize=50,pc=15){


   png(filename="Densidad3.png",width = 4*480, height = 4*480)


    x<-r4cresc(n)$x
    h<-h3(x)
    xs<-seq(-2,1,length=gsize)
    ys<-seq(-0.8,1.6,length=gsize)
    xys<-as.matrix(expand.grid(xs,ys))
    fxys<-kde(x=x,H=h^2*diag(2),eval.points=xys)$estimate
    zs<-invvec(fxys)
    msx<-MS.kde(x=x,H=h^2*diag(2),eval.points=xys)
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
    .8,.85,.9,.95)),
    xaxs="i",yaxs="i",xlab="",ylab="")
    classPlotColors <- c("dodgerblue2", "red3", "green3", "slateblue",
    "orange", "skyblue1", "forestgreen", "steelblue4", "gray", "brown",
    "black")
    widthx<-xs[2]-xs[1]
    widthy<-ys[2]-ys[1]
    for(i in 1:nrow(xys)){
        rect(xys[i,1]-widthx/2,xys[i,2]-widthy/2,xys[i,1]+widthx/2,xys[i,2]+
        widthy/2,
```

```
        border=NA,col=classPlotColors[(msx$labels)[i]])
    }
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
    .8,.85,.9,.95)),
    add=TRUE)

    dev.off()

    return(list(x=x,msx=msx,h=h))
}



h4<-function(x,Nruns=10){
    h0<-0
    h1<-1
    n0<-Inf
    n1<-1
    for(i in 1:Nruns){
        hmed<-(h0+h1)/2
        nmed<-MS.kde(x=x,H=hmed^2*diag(2))$nclus
        if(nmed>5){h0<-hmed;n0<-nmed}
        if(nmed<=5){h1<-hmed;n1<-nmed}
    }
    return(h1)
}


pop.clust4<-function(n=100,gsize=50,pc=15){



    x<-rbrokenring(n)$x
    h<-h4(x)
    xs<-seq(-1.2,1.2,length=gsize)
    ys<-seq(-1.2,1.2,length=gsize)
    xys<-as.matrix(expand.grid(xs,ys))
    fxys<-kde(x=x,H=h^2*diag(2),eval.points=xys)$estimate
    zs<-invvec(fxys)
    msx<-MS.kde(x=x,H=h^2*diag(2),eval.points=xys)
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
    .8, .85,.9,.95)),xaxs="i",yaxs="i",xlab="",ylab="")
    classPlotColors <- c("dodgerblue2", "red3", "green3", "slateblue",
    "orange", "skyblue1", "forestgreen", "steelblue4", "gray", "brown",
    "black")
```

```
    widthx<-xs[2]-xs[1]
    widthy<-ys[2]-ys[1]
    for(i in 1:nrow(xys)){
        rect(xys[i,1]-widthx/2,xys[i,2]-widthy/2,xys[i,1]+widthx/2,xys[i,2]+
        widthy/2,border=NA,col=classPlotColors[(msx$labels)[i]])
    }
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,.8,
    .85, .9,.95)),add=TRUE)


    return(list(x=x,msx=msx,h=h))
}



h5<-function(x,Nruns=10){
    h0<-0
    h1<-1
    n0<-Inf
    n1<-1
    for(i in 1:Nruns){
        hmed<-(h0+h1)/2
        nmed<-MS.kde(x=x,H=hmed^2*diag(2))$nclus
        if(nmed>5){h0<-hmed;n0<-nmed}
        if(nmed<=5){h1<-hmed;n1<-nmed}
    }
    return(h1)
}


pop.clust5<-function(n=100,gsize=50,pc=15){


    x<-r4crescring(n)$x
    h<-h5(x)
    xs<-seq(-1.4,1.4,length=gsize)
    ys<-seq(-2,2,length=gsize)
    xys<-as.matrix(expand.grid(xs,ys))
    fxys<-kde(x=x,H=h^2*diag(2),eval.points=xys)$estimate
    zs<-invvec(fxys)
    msx<-MS.kde(x=x,H=h^2*diag(2),eval.points=xys)
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
    .8,.85,.9,.95)),
    xaxs="i",yaxs="i",xlab="",ylab="")
    classPlotColors <- c("dodgerblue2", "red3", "green3", "slateblue",
```

```
    "orange", "skyblue1","forestgreen", "steelblue4", "gray", "brown",
    "black")
    widthx<-xs[2]-xs[1]
    widthy<-ys[2]-ys[1]
    for(i in 1:nrow(xys)){
        rect(xys[i,1]-widthx/2,xys[i,2]-widthy/2,xys[i,1]+widthx/2,xys[i,2]+
        widthy/2,
        border=NA,col=classPlotColors[(msx$labels)[i]])
    }
    contour(xs,ys,zs,drawlabels=FALSE,lwd=2,levels=quantile(fxys,c(.7,.75,
    .8,.85,.9,.95)),
    add=TRUE)


    return(list(x=x,msx=msx,h=h))
}
```

Mean Shift partitions for the 5 densities

The function called particiones_MS_5densidades creates a tridimensional matrix with dimensions num_rep, 5, 11. Num_rep is the number of repetitions, 5 are the total densities used (trimodal III, quatrimodal, 4 crescent, broken ring and eye) and 11 are the number of bandwidth selectors used (Hlscv, Hscv, Hpi, ..., Hpi.diag). In this matrix, the clusters labels of the samples through the MS.kde function are saved.

There are several inputs such as densidad (the five densities labeled with numbers from 1 to 5), numero_repeticiones (number of repetions), tamano_muestra (each sample size)

There are several previous functions to set the parameters of some of the densities.

```
    mwweights = function(d)
    {
     switch(d,
     "1"=c(3,3,1)/7,
     "2"=c(1,3,1,3)/8)
    }

    mwmeans = function(d)
```

```
    {
     switch(d,
     "1"=rbind(c(-1,0), c(1,2/sqrt(3)), c(1, -2/sqrt(3))),
     "2"=rbind(c(-1,1), c(-1,-1),c(1,-1), c(1, 1)))
    }

    mwsdeviations = function(d)
    {
     switch(d,
     "1"=1/25*rbind(invvech(c(9,63/10,49/4)),invvech(c(9,0,49/4)),
     invvech(c(9,0,49/4))),

     "2"=rbind(invvech(c(4/9,8/45,4/9)),invvech(c(4/9,12/45,4/9)),
     invvech(c(4/9, -28/90,4/9)),invvech(c(4/9,-4/18,4/9))))

    }

limite_malla_x = function(d)
    {
     switch(d,
     "1"=c(-3,3),
     "2"=c(-3,3),
     "3"=c(-2,1.1),
     "4"=c(-1.2,1.2),
     "5"=c(-1.4,1.4))
    }

limite_malla_y = function(d)
    {
     switch(d,
     "1"=c(-3,3),
     "2"=c(-3,3),
     "3"=c(-0.8,1.8),
     "4"=c(-1.2,1.2),
     "5"=c(-2,2))
    }



particiones_MS_5densidades<-function(densidad=1:5,numero_repeticiones=100,
tamano_muestra=500,gsize=150){

   contador<-0

   for(d in densidad){
```

```
nombresinextension<-paste("Resultados_densidad_",d,sep="")
set.seed(346536)
res<-array(,dim=c(numero_repeticiones,gsize^2,11))
dimnames(res)[[3]]<-c("lscv","scv","pi","nr", "it","it.diag",
"it.single", "at","lscv_diag","scv_diag","pi_diag")


puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],
length=gsize)
puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],
length=gsize)

malla<-expand.grid(puntos_eje_x,puntos_eje_y)

contador<-contador+1

cat("Densidad ",d,"\n")
for(num_rep in 1:numero_repeticiones){

   setTxtProgressBar(barra_progreso,num_rep/numero_repeticiones)

   if(d==1 || d==2){
      muestra<-rmvnorm.mixt(tamano_muestra,mus=mwmeans(d),
      Sigmas=mwsdeviations(d),props=mwweights(d))
   }
   if(d==3){
      muestra<-r4cresc(tamano_muestra)$x
   }
   if(d==4){
      muestra<-rbrokenring(tamano_muestra)$x
   }
   if(d==5){
      muestra<-r4crescring(tamano_muestra)$x
   }

   hlscv_optimo<-Hlscv(muestra,deriv.order=1)
   res[num_rep,,1]<-MS.kde(muestra,hlscv_optimo,as.matrix(malla))
   $labels

   hscv_optimo<-Hscv(muestra,deriv.order=1,pilot="dunconstr")
   res[num_rep,,2]<-MS.kde(muestra,hscv_optimo,as.matrix(malla))
   $labels

   hpi_optimo<-Hpi(muestra,deriv.order=1,pilot="dunconstr")
```

```
        res[num_rep,,3]<-MS.kde(muestra,hpi_optimo,as.matrix(malla))
        $labels

        hnr_optimo<-(4/6)^(2/8)*var(muestra)*tamano_muestra^(-2/8)
        res[num_rep,,4]<-MS.kde(muestra,hnr_optimo,as.matrix(malla))
        $labels

        hiterativo_optimo<-Hit.r1(muestra,obj=FALSE)
        res[num_rep,,5]<-MS.kde(muestra,hiterativo_optimo,as.matrix(malla))
        $labels

        hiterativo_diag_optimo<-Hit.diag.r1(muestra,obj=FALSE)
        res[num_rep,,6]<-MS.kde(muestra,hiterativo_diag_optimo,
        as.matrix(malla))$labels

        hiterativo_single_optimo<-Hit.single.r1(muestra,obj=FALSE)
        res[num_rep,,7]<-MS.kde(muestra,hiterativo_single_optimo,
        as.matrix(malla))$labels

        h_azzalini_torelli_optimo<-3/4*(1/tamano_muestra)^(1/6)*
        diag(apply(muestra,2,sd))
        res[num_rep,,8]<-MS.kde(muestra,h_azzalini_torelli_optimo,
        as.matrix(malla))$labels

        hlscv_diag_optimo<-Hlscv.diag(muestra,deriv.order=1)
        res[num_rep,,9]<-MS.kde(muestra,hlscv_diag_optimo,as.matrix(malla))
        $labels

        hscv_diag_optimo<-Hscv.diag(muestra,deriv.order=1,pilot="dscalar")
        res[num_rep,,10]<-MS.kde(muestra,hscv_diag_optimo,as.matrix(malla))
        $labels

        hpi_diag_optimo<-Hpi.diag(muestra,deriv.order=1,pilot="dscalar")
        res[num_rep,,11]<-MS.kde(muestra,hpi_diag_optimo,as.matrix(malla))
        $labels

        save(res,file=paste(nombresinextension,".RData",sep=""))

    }

    cat("\n")


}
```

```
}
```

Distance in measure for normal mixture densities

```
library(ks)
library(mvtnorm)
library(clue)

probsquares.mixt<-function(lowers,uppers,gsizes,mus,Sigmas,props){

    xs<-seq(lowers[1],uppers[1],length=gsizes[1])
    ys<-seq(lowers[2],uppers[2],length=gsizes[2])
    xys<-as.matrix(expand.grid(xs,ys))
    widthx<-xs[2]-xs[1]
    widthy<-ys[2]-ys[1]

    xs2<-c(xs[1]-widthx/2,xs+widthx/2)
    ys2<-c(ys[1]-widthy/2,ys+widthy/2)
    xys2<-as.matrix(expand.grid(xs2,ys2))
    prob.squares<-numeric(nrow(xys2))
    for(l in 1:length(prob.squares)){
        prob.squares[l]<-pmvnorm.mixt(upper=xys2[l,],mus=mus,Sigmas=Sigmas,
        props=props,lower=c(-Inf,-Inf))
        }
    prob.squares<-matrix(prob.squares,nrow=gsizes[1]+1,ncol=gsizes[2]+1,
    byrow=TRUE)
    prob.squares<-prob.squares[(gsizes[1]+1):1,]
    prob.squares<-prob.squares[-(gsizes[1]+1),-1]+prob.squares[-1,-(gsizes
    [2]+1)]-prob.squares[-(gsizes[1]+1),-(gsizes[2]+1)]-prob.squares[-1,-1]
    psvector<-numeric()
    for (i in 1:gsizes[1]){
        psvector <- c(psvector, prob.squares[gsizes[1]+1-i,])


    }
    return(psvector)
}



probsquares.kde<-function(lowers,uppers,gsizes,x,h){
    n<-nrow(x)
```

```
    xs<-seq(lowers[1],uppers[1],length=gsizes[1])
    ys<-seq(lowers[2],uppers[2],length=gsizes[2])
    xys<-as.matrix(expand.grid(xs,ys))
    widthx<-xs[2]-xs[1]
    widthy<-ys[2]-ys[1]


    xs2<-c(xs[1]-widthx/2,xs+widthx/2)
    ys2<-c(ys[1]-widthy/2,ys+widthy/2)
    xys2<-as.matrix(expand.grid(xs2,ys2))
    prob.squares<-rep(0,nrow(xys2))
    for(i in 1:n){
        prob.squares<-prob.squares+pnorm(q=xys2[,1],mean=x[i,1],sd=h)*
        pnorm(q=xys2[,2],
        mean=x[i,2],sd=h)/n
        }
    prob.squares<-matrix(prob.squares,nrow=gsizes[1]+1,ncol=gsizes[2]+1,
    byrow=TRUE)
    prob.squares<-prob.squares[(gsizes[1]+1):1,]
    prob.squares<-prob.squares[-(gsizes[1]+1),-1]+prob.squares[-1,-(gsizes
    [2]+1)]-prob.squares[-(gsizes[1]+1),-(gsizes[2]+1)]-prob.squares[-1,-1]
    psvector<-numeric()
    for (i in 1:gsizes[1]){
        psvector <- c(psvector, prob.squares[gsizes[1]+1-i,])


    }
    return(psvector)
}




dmeas.probs<-function(labels1,labels2,probs){
    if(length(labels1)!=length(labels2)||length(labels2)!=length(probs)){
      stop("Error: the vectors of labels and probabilities all have to be of the
      same length")
    }

    nclus1<-length(unique(labels1))
    nclus2<-length(unique(labels2))

    Psymdif<-matrix(0,nrow=nclus1,ncol=nclus2)
    for(i in 1:nclus1){for(j in 1:nclus2){
```

```
        symdif.ind<-(((labels1==i)+(labels2==j))==1)*1


        if(sum(symdif.ind>0)){
            Psymdif[i,j]<-sum(probs[symdif.ind])
        }
    }}

    if(nclus1<nclus2){
        prob.clus2<-numeric(nclus2)
        for(k in 1:nclus2){
            prob.clus2[k]<-sum(probs[labels2==k])
        }
        Psymdif<-rbind(Psymdif,matrix(rep(prob.clus2,nclus2-nclus1),
        nrow=nclus2-nclus1,ncol=nclus2,byrow=TRUE))
    }
    if(nclus1>nclus2){
        prob.clus1<-numeric(nclus1)
        for(k in 1:nclus1){
            prob.clus1[k]<-sum(probs[labels1==k])
        }
        Psymdif<-cbind(Psymdif,matrix(rep(prob.clus1,nclus1-nclus2),
        ncol=nclus1-nclus2,nrow=nclus1,byrow=FALSE))
    }


    permutacion_solucion_minima<-solve_LSAP(Psymdif)
    resultado<-sum(Psymdif[cbind(seq_along(permutacion_solucion_minima),
    permutacion_solucion_minima)])
    return(resultado/2)
}
```

```
pmvnorm.mixt<-function (lower=-Inf,upper=Inf, mus, Sigmas, props = 1)
{
    if (!(identical(all.equal(sum(props), 1), TRUE)))
        stop("Proportions don't sum to one\n")

    d<-length(lower)
```

```
    if (missing(mus))
        mus <- rep(0, d)
    if (missing(Sigmas))
        Sigmas <- diag(d)
    if (identical(all.equal(props[1], 1), TRUE)) {
        if (is.matrix(mus))
            mus <- mus[1, ]
        dens <- pmvnorm(lower=lower,upper=upper, mean = mus, sigma = Sigmas
        [1:d,], algorithm=GenzBretz(maxpts = 25000, abseps = 10^-10, releps
        = 10^-10))
    }
    else {
        k <- length(props)
        dens <- 0
        for (i in 1:k) dens <- dens + props[i] * pmvnorm(lower=lower, upper=
        upper, mean = mus[i,], sigma = Sigmas[((i - 1) * d + 1):(i * d), ],
        algorithm=GenzBretz(maxpts = 25000, abseps = 10^-10, releps = 10^-10))
    }
    return(dens)
}
```

Calculation of distances between clusters

The function called calculo_de_distancias calculates the distances between clusters saved in the RData (Resultados_densidad_ 1 to 5) with the function particiones_MS_5densidades and clusters saved in the RData (p1 to p5).

These distances are saved in a tridimensional matrix (matriz_distancias) in the archive Matriz_de_distancias.RData.

```
calculo_de_distancias<-function(){

gsize<-150

p1<-load("p1.RData")
etiquetas1verdad<-p1$msx$labels

p2<-load("p2.RData")
etiquetas2verdad<-p2$msx$labels

p3<-load("p3.RData")
etiquetas3verdad<-p3$msx$labels
```

```
p4<-load("p4.RData")
etiquetas4verdad<-p4$msx$labels

p5<-load("p5.RData")
etiquetas5verdad<-p5$msx$labels



load("Resultados_densidad_1.RData")
etiquetas1estimadas<-res

load("Resultados_densidad_2.RData")
etiquetas2estimadas<-res

load("Resultados_densidad_3.RData")
etiquetas3estimadas<-res

load("Resultados_densidad_4.RData")
etiquetas4estimadas<-res

load("Resultados_densidad_5.RData")
etiquetas5estimadas<-res

numero_muestras<-dim(etiquetas1estimadas)[1]

nombres_columnas<-c("lscv","scv","pi","nr", "it","it.diag","it.single","at",
"lscv_diag","scv_diag","pi_diag")
matriz_distancias<-array(,dim=c(numero_muestras,length(nombres_columnas),5))
dimnames(matriz_distancias)[[2]]<-nombres_columnas



d<-1

   puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],length=gsize)
   puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],length=gsize)
   malla<-expand.grid(puntos_eje_x,puntos_eje_y)
   prob_normal_mix<-p1$psf


   for(i in 1:numero_muestras){
      for(k in 1:11){
         matriz_distancias[i,k,1]<-dmeas.probs(etiquetas1verdad,
         etiquetas1estimadas[i,,k],prob_normal_mix)
      }
```

```
   }
   save(matriz_distancias,file="Matriz_de_distancias.RData")


d<-2

   puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],length=gsize)
   puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],length=gsize)
   malla<-expand.grid(puntos_eje_x,puntos_eje_y)
   prob_normal_mix<-p2$psf

   for(i in 1:numero_muestras){
      for(k in 1:11){
         matriz_distancias[i,k,2]<-dmeas.probs(etiquetas2verdad,
         etiquetas2estimadas[i,,k],prob_normal_mix)
      }
   }
   save(matriz_distancias,file="Matriz_de_distancias.RData")


d<-3

   puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],length=gsize)
   puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],length=gsize)
   malla<-expand.grid(puntos_eje_x,puntos_eje_y)
   prob_squares_kde<-p3$psf

   for(i in 1:numero_muestras){
      for(k in 1:11){
         matriz_distancias[i,k,3]<-dmeas.probs(etiquetas3verdad,
         etiquetas3estimadas[i,,k],prob_squares_kde)
      }
   }
   save(matriz_distancias,file="Matriz_de_distancias.RData")


d<-4

   puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],length=gsize)
   puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],length=gsize)
   malla<-expand.grid(puntos_eje_x,puntos_eje_y)
   prob_squares_kde<-p4$psf

   for(i in 1:numero_muestras){
      for(k in 1:11){
```

```
        matriz_distancias[i,k,4]<-dmeas.probs(etiquetas4verdad,
        etiquetas4estimadas[i,,k],prob_squares_kde)
      }
   }
   save(matriz_distancias,file="Matriz_de_distancias.RData")


d<-5

   puntos_eje_x<-seq(limite_malla_x(d)[1],limite_malla_x(d)[2],length=gsize)
   puntos_eje_y<-seq(limite_malla_y(d)[1],limite_malla_y(d)[2],length=gsize)
   malla<-expand.grid(puntos_eje_x,puntos_eje_y)
   prob_squares_kde<-p5$psf

   for(i in 1:numero_muestras){
      for(k in 1:11){
        matriz_distancias[i,k,5]<-dmeas.probs(etiquetas5verdad,
        etiquetas5estimadas[i,,k],prob_squares_kde)
      }
   }
   save(matriz_distancias,file="Matriz_de_distancias.RData")



}
```

Bibliography

Abdous B. (1993). Note on the minimum mean integrated squared error of kernel estimates of a distribution function and its derivates. *Communications in Statistics Theory and Methods*, **22**, 603–609.

Aldershof, B., Marron, J.S., Park, B.U. and Wand, M.P. (1995) Facts about the Gaussian probability density function. *Applicable Analysis*, **59**, 289–306.

Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, **46**, 195–214.

Azzalini, A. (1981) A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, **68**, 326–328.

Azzalini, A. and Bowman, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, Oxford University Press, Oxford.

Azzalini, A. and Torelli, N. (2007) Clustering via nonparametric density estimation. *Stat. Comput.*, **17**, 71–80.

Barrio, E. del, Cuesta-Albertos, J.A. and Matrán, C. (2000) Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, **9**, 1-96.

Berg, A. and Politis, D. (2009) CDF and survival function estimation with infinite-order kernels. *Electronic Journal of Statistics*, **3**, 1436–1454.

Besicovitch, A.S. (1945) A general form of the covering principle and relative differentiation of additive functions, I, *Proceedings of the Cambridge Philosophical Society*, **41**, 103-110.

Besicovitch, A.S. (1946) A general form of the covering principle and relative differentiation of additive functions, II, *Proceedings of the Cambridge Philosophical Society*, 42, 205-235.

Bland, J.M. (2005) The half-normal distribution method for measurement error: two case studies, unpublished talk available on *http://www-users.york.ac.uk/ mb55/talks/halfnor.pdf*.

Bland, J.M. and Altman, D.G. (1999) Measuring agreement in method comparison studies, *Stat. Methods Med. Res.*, **8**, 135-160.

Bouchard, B., Ekeland, I. and Touzi, N. (2004) On the Malliavin approach to Monte Carlo approximation of conditional expectations, *Finance Stochast.*, **8**, 45-71.

Bowman, A., Hall, P. and Prvan, T. (1998) Bandwidth selection for the smoothing of distribution functions. *Biometrika*, **85**, 799–808.

Butzer, P.L. and Nessel, R.J. (1971) *Fourier analysis and approximation*. Academic Press, New York.

Cantelli, F.P. (1933) Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, **4**, 421-424.

Carreira-Perpiñán, M.Á. (2006) Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal.*, **22**, 1318–1323.

Cattell, R.B. (1943) The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, **38**, 476-506.

Chacón, J.E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.* **37**, 249–265.

Chacón, J.E. (2012) Clusters and water flows: a novel approach to modal clustering through Morse theory. *arXiv preprint arXiv:1212.1384*.

Chacón, J.E. and Duong, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398.

Chacón, J.E. and Duong, T. (2013) Bandwidth selection for multivariate density derivative estimation, with applications to clustering and bump hunting. *Electron. J. Statist.*, **7**, 499–532.

Chacón, J.E. and Duong, T. (2014) Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate Gaussian density. To appear in *Statistics and Computing.*

Chacón, J.E., Duong, T. and Wand, M.P. (2011) Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**, 807–840.

Chacón, J.E., Montanero, J. and Nogales, A.G. (2007) A note on kernel density estimation at a parametric rate. *Journal of Nonparametric Statistics*, **19**, 13–21.

Chacón, J.E., Montanero, J., Nogales, A.G. and Pérez, P. (2007) On the existence an limit behavior of the optimal bandwidth for kernel density estimation. *Statistica Sinica*, **17**, 289–300.

Chacón, J.E. and Rodríguez-Casal, A. (2010) A note on the universal consistency of the kernel distribution function estimator. *Statistics and Probability Letters*, **80**, 1414–1419.

Cheng, Y. (1995) Mean shift, mode seeking, and clustering. *IEEE T. Pattern Anal.*, **17**, 790–799.

Chiu, S.-T. (1992) An automatic bandwidth selector for kernel density estimation. *Biometrika*, **79**, 771–782.

Cline, D.B.H. (1988) Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**, 1421–1427.

Comaniciu, D. and Meer, P. (2002) Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal.*, **24**, 603–619.

Daniel, C. (1959) Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics*, **1**, 311-341.

Davis, K.B. (1977) Mean integrated square error properties of density estimates. *Annals of Statistics*, **5**, 530–535.

Day, W.H.E. (1986) The complexity of computing metric distances between partitions. *Mathematical Social sciences*, **1**, 269-287.

Devroye, L. (1992) A note on the usefulness of superkernels in density estimation. *Annals of Statistics*, **20**, 2037–2056.

Driver, H.E. and Kroeber, A.L. (1932) Quantitative expression of cultural relationships. *University of California Publications in American Archeology and Ethnology*, **31**, 211-256.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, **27**, 642-669.

Fryer, M.J. (1976) Some errors associated with the nonparametric estimation of density functions. *IMA Journal of Applied Mathematics*, **18**, 371–380.

Fukunaga, K. and Hostetler, L.D. (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory*, **21**, 32–40.

García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo, A. (2010) A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**, 89–109.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012) Nonparametric ridge estimation. *arXiv preprint arXiv:1212.5156*.

Giné, E. and Nickl, R. (2009) An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, **143**, 569–596.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2003) Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, **30**, 415–427.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2007) Density estimation using the sinc kernel. Preprint Statistics No. 2/2007, Norwegian University of Science and Technology, Trondheim, Norway. Available at `http://www.math.ntnu.no/preprint/statistics/2007/`

Glivenko, V. (1933) Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, **4**, 92-99.

Gurland, J. (1948) Inversion formulae for the distribution of ratios. *Annals of Mathematical Statistics*, **19**, 228–237.

Hand, D., Mannila, H. and Smith, P. (2001) *Principles of data mining*. The MIT Press, Massachusetts.

Horová, I., Koláček, J. and Vopatová, K. (2013) Full bandwidth matrix selectors for gradient kernel density estimate. *Comput. Statist. Data Anal.*, **57**, 364–376.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2007) Modifying the kernel distribution function estimator towards reduced bias. *Statistics*, **41**, 93–103.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994) *Continuous univariate distributions*, Vol. 1. Wiley, New York.

Jones, M.C. (1990) The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, **9**, 129–132.

Lehmann, E.L. (1983) *Theory of point estimation*. Wiley, New York.

Lehmann, E.L. (1986) *Testing Statistical Hypotheses*. Wiley, New York.

Li, J., Ray, S. and Lindsay, B.G. (2007) A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, **8**, 1687–1723.

Lindqvist, B. H. and Taraldsen, G. (2005) Monte Carlo conditioning on a sufficient statistic, *Biometrika*, **92**, 451-464.

Martins, A.P. and Tenreiro, C. (2003) Multistage plug-in bandwidth selection for kernel distribution function estimators, *X Annual Meeting of Statistical Portuguese Society 2002*, 415-426.

Mason, D.M. and Swanepoel, J.W.H. (2012) Uniform in bandwidth limit laws for kernel distribution function estimators. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*, IMS Collections **9**, 241–253.

Masry, E. (1989) Nonparametric estimation of conditional probability densities and expectations of stationary process: strong consistency and rates, *Stochastic Process and their Applications*, **32**, 109-127.

Massart, P. (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality, *Annals of Probability*, **18**, 1269-1293.

Mattila, P. (1995) *Geometry of sets and measures in euclidean spaces*. Cambridge University Press, New York.

Metropolis, N, and Ulam, S. (1949) The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335-341.

Nadaraya, E.A. (1964) Some new estimates for distribution functions. *Theory of Probability and Its Applications*, **15**, 497–500.

Nogales, A.G., Pérez, P., Unbiased estimation for the general half-normal distribution, *Comm. Statist. Theory Methods*, to appear.

Ozertem, U. and Erdogmus, D. (2011) Locally defined principal curves and surfaces. *J. Mach. Learn. Res.*, **12**, 241–274.

Papadimitriou, C. and Steiglitz, K. (1982) *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, New Jersey.

Parzen, E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

Pewsey, A. (2002) Large-sample inference for the general half-normal distribution, *Comm. Statist. Theory Methods*, **31**, 1045-1054.

Pewsey, A. (2004) Improved likelihood based inference for the general half-normal distribution, *Comm. Statist. Theory Methods*, **33**, 197-204.

Polansky, A.M. and Baker, E.R. (2000) Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, **65**, 63–80.

Raykar, V.C., Duraiswami, R. and Zhao, L.H. (2010) Fast computation of kernel estimators. *J. Comput. Graph. Statist.*, **19**, 205–200.

Rosenblatt, M (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

Sarda, P. (1993) Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, **35**, 65–75.

Servien, R. (2009) Estimation de la fonction de répartition: revue bibliographique. *Journal de la Société Française de Statistique*, **150**, 84–104.

Shao, Y. and Xiang, X. (1997) Some extensions of the asymptotics of a kemel estimator of a distribution function. *Statist. Probab. Lett.*, **34**, 301–308.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, London.

Stute, W. (1997) Nonparametric model checks for regression. *The Annals of Statistics*, **25**, 613-641.

Swanepoel, J.W.H. (1988) Mean integrated square error properties and optimal kernels when estimating a distribution function. *Communications in Statistics Theory and Methods*, **17**, 3785–3799.

Swanepoel, J.W.H. and Van Graan, F.C. (2005) A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics*, **32**, 551–562.

Tenreiro, C. (2003) On the asymptotic behavour of the ISE for automatic kernel distribution estimators. *Journal of Nonparametric Statistics*, **15**, 485–504.

Tenreiro, C. (2006) Asymptotic behaviour of multistage plug-in bandwidth selections for kernel distribution function estimators. *Journal of Nonparametric Statistics*, **18**, 101–116.

Tiago de Oliveira, J. (1963) Estatística de densidades: resultados assintóticos. *Revista da Faculdade de Ciências de Lisboa*, **9**, 111–206.

Tryon, R.C. (1939) *Cluster Analysis*. Ann. Arbor: Edward Bros, Michigan.

Tsybakov, A.B. (2009) *Introduction to Nonparametric Estimation*. Springer Science+Business Media, New York.

Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.*, **88**, 520–528.

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

Watson, G.S. and Leadbetter, M.R. (1964) Hazard analysis II. *Sankhy ā Series A*, **26**, 101–116.

Yamato, H. (1973) Uniform convergence of an estimator of a distribution function. *Bulletin of Mathematical Statistics*, **15**, 69–78.

Zubin, J.A. (1938) A technique for measuring likemindedness. *Journal of Abnormal and Social Psychology*, **33**, 508-516.

List of authors

Abdous, 20, 22, 23, 27, 28
Aldershof, 26
Altman, 1, 17
Arabie, 67
Azzalini, 5, 17, 65, 68, 79

Baker, 1, 17
Berg, 18
Besicovitch, 2, 10, 12, 38, 41
Bland, 2, 36
Bowman, 1, 17, 79, 80
Butzer, 24

Cantelli, 2
Carreira-Perpiñán, 3, 13, 59
Cattell, 12
Chacón, 1–3, 5, 8, 14, 18, 19, 21, 25, 29, 59, 61–65, 67
Cheng, 3, 13, 59
Chiu, 23, 80
Cline, 20
Comaniciu, 3, 13, 59, 62, 72
Cuesta-Albertos, 2

Daniel, 2, 36
Davis, 22
Day, 3
Devroye, 8
Driver, 12
Duong, 3, 14, 59, 61–65, 67, 79
Duraiswami, 80
Dvoretzky, 2

Erdogmus, 59

Fryer, 26
Fukunaga, 3, 13, 59, 62

García-Escudero, 13
Genovese, 59
Gil-Pelaez, 23
Giné, 1, 18
Glad, 22

Glivenko, 2
Gurland, 23

Härdle, 10, 42
Hall, 17, 80
Hand, 3
Hjort, 22
Horová, 3, 14, 59, 63–65
Hostetler, 3, 13, 59, 62
Hubert, 67

Janssen, 18
Johnson, 2, 36
Jones, 1, 5, 6, 17, 22, 61, 64, 79

Kiefer, 2
Kolácek, 3, 14, 59, 63–65
Kroeber, 12

Léger, 1, 17
Leadbetter, 1, 4, 17
Lehmann, 44
Li, 59
Lindsay, 59

Mannila, 3
Martins, 1
Mason, 5, 18
Massart, 2
Matrán, 2
Mattila, 10, 38
Meer, 3, 13, 59, 62, 72
Montanero, 2, 8, 18, 19, 21, 25

Nadaraya, 1, 4, 10, 17, 42
Nessel, 24
Nickl, 1, 18
Nogales, 2, 8, 18, 19, 21, 25, 37, 38

Ozertem, 59

Papadimitriou, 67
Parzen, 1
Perone-Pacifico, 59
Pewsey, 2, 36, 38
Polansky, 1, 17
Politis, 18
Prvan, 17, 80

Ray, 59
Raykar, 80
Rodríguez-Casal, 1, 5, 18, 19
Rosenblatt, 1

Sarda, 1, 17
Servien, 18
Shao, 22
Silverman, 3, 13, 59
Smyth, 3
Steiglitz, 67
Stute, 80, 81
Swanepoel, 1, 5, 17, 18

Tenreiro, 1, 5, 8, 17, 18, 24, 29, 30
Tiago de Oliveira, 1, 4, 17
Torelli, 65, 68
Tryon, 12
Tsybakov, 22

Ushakov, 22

Van Graan, 18
Veraverbeke, 18
Verdinelli, 59

Vopatová, 3, 14, 59, 63–65

Wand, 61, 64, 65, 79
Wasserman, 59
Watson, 1, 4, 10, 17, 42
Wolfowitz, 2

Xiang, 22

Yamato, 1, 5, 17

Zhao, 80
Zubin, 12

List of Figures

1.1 Comparison among Φ, F_n and F_{nh} 3
1.2 The $MISE$ as a function of the bandwidth 8

2.1 Optimal bandwidth sequence (left) and relative efficiency in MISE (right) for the estimation of the Jackson-de la Vallé Poussin distribution, as a function of $\log_{10} n$. The lines show the limit values. Solid circles and solid lines correspond to the trapezoidal superkernel and open circles and dashed lines correspond to the sinc kernel. 25
2.2 Relative efficiency in MISE for the estimation of standard normal distribution, as a function of $\log_{10} n$. The line shows the limit value. Solid circles correspond to the normal kernel and open circles correspond to the sinc kernel. 27

3.1 Box plots of the approximations of $E(Y|X = 1)$ as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$. . . 41
3.2 Box plots of the approximations of $E(V|U = 0.5)$ as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$ 42
3.3 Box plots of the approximations of $\mathring{\xi} \pm$ MSE as a function of the number of simulations, m, for $\epsilon = 0.1$ and $\epsilon = 0.01$ 45

3.4 Box plots of the approximations of $\tilde{\xi}$, $\hat{\xi}$ and $\mathring{\xi}$ as a number of simulations, $m = 100, 1000, 5000$, for $\epsilon = 0.1$ and $\epsilon = 0.01$ function of the 46

3.5 Box plots for the estimator $\mathring{\eta}$ for sample sizes $n = 10, 20, 30$ and for the estimators $\tilde{\eta}$, $\hat{\eta}$ y $\mathring{\eta}$ for sample sizes $n = 10, 20, 30$, respectively 54

4.1 The five true density models included in the simulation study, with the ideal population clustering shown in different colors. 66

List of Tables

3.1 Approximation of $E(Y|X = 1) \pm$ MSE as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$ 41

3.2 Approximation of $E(V|U = 0.5) \pm S^2$ (S^2 is the sample variance) as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$ 42

3.3 Approximations of $\mathring{\xi} \pm$ MSE as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$ 44

3.4 Approximations of $\tilde{\xi} \pm$ MSE, $\hat{\xi} \pm$ MSE and $\mathring{\xi} \pm$ MSE as a function of the number of simulations, m, for $\epsilon = 0.1, 0.01$ . . 45

3.5 Sample mean and MSE of the estimators calculated using 1000 random samples of size n from the $HN(10, 4)$ distribution 53

4.1 Sample median and (interquartile range) for the distance in measure between the data-based clusterings induces by each bandwidth selection method and the ideal population clustering along 100 simulation runs of sample size $n = 500$ of each distribution. The significantly best methods for each model are marked in bold font. 69

4.2 Distribution of the number of clusters for each clustering method along the five density models. The true number of clusters is marked in bold. 70