UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Estatística e Investigación Operativa

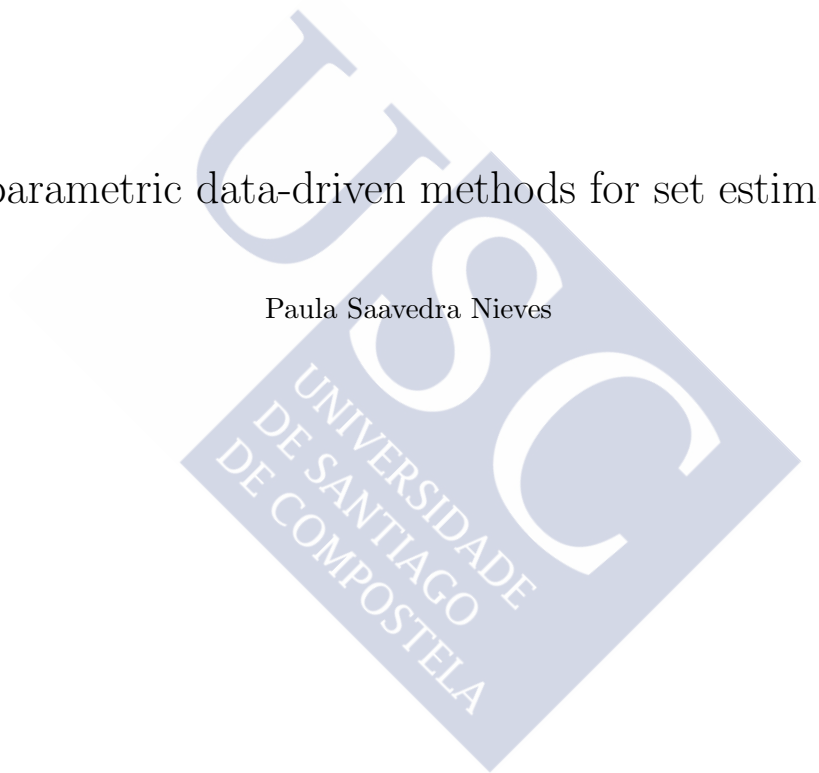
NONPARAMETRIC DATA-DRIVEN METHODS FOR SET ESTIMATION

Paula Saavedra Nieves

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Nonparametric data-driven methods for set estimation

Paula Saavedra Nieves



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Don Wenceslao González Manteiga, catedrático do Departamento de Estatística e Investigación Operativa da Universidade de Santiago de Compostela e Don Alberto Rodríguez Casal, titular de universidade no Departamento de Estatística e Investigación Operativa da Universidade de Santiago de Compostela, informan que a memoria titulada

NONPARAMETRIC DATA-DRIVEN METHODS FOR SET ESTIMATION

foi realizada baixo a súa dirección por Dona Paula Saavedra Nieves, estimando que a interesada se atopa en condicións de optar ao grao de Doutor, polo que solicitan que sexa admitida a trámite para a súa lectura e defensa pública.

Santiago de Compostela, 7 de Novembro de 2014

Os directores:

Prof. Dr. Wenceslao González Manteiga Prof. Dr. Alberto Rodríguez Casal

A doutoranda:

Paula Saavedra Nieves

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecementos

Gustaríame agradecerlles aos profesores Wenceslao González Manteiga e Alberto Rodríguez Casal o seu labor como directores desta tese doutoral. Moitas grazas por aceptarme como estudante de doutorado e pola atención e confianza recibidas durante esta etapa.

Grazas a todos os compañeiros do Departamento de Estatística e Investigación Operativa da Universidade de Santiago de Compostela polos momentos compartidos nestos anos. A Rosa e Pedro, por guiarme na miña primeira experiencia na docencia universitaria.

Un recordo moi cariñoso para Juan Carlos. Agradézoche o ben que te portaches comigo sempre.

Quero acordarme tamén da miña familia polo seu apoio incondicional. Moitas grazas a meus pais, Álex e Alfonso.

Finalmente, fago constar que este traballo foi financiado pola Universidade de Santiago de Compostela a través da convocatoria dos contratos predoutorais USC 2011, polo Proxecto MTM2008-03010 do Ministerio de Ciencia e Educación e pola rede IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences) of Belgian Science Policy.

Santiago de Compostela, Novembro de 2014
Paula Saavedra Nieves

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

Chapter 1

Introduction

1.1 Preliminaries

The problem of reconstructing a set in the Euclidean space from a random finite sample of points whose distribution is closely related to it can be considered from different points of view, including the estimation of supports, boundaries and level sets. This theory is known as *set estimation* and it has open a relatively new chapter of the statistics with important applications in cluster analysis (see Hartigan, 1975), quality control (see Devroye and Wise, 1980 or Baíllo et al., 2000) or image analysis to reconstruct, for example, the habitat of a plant or an animal species (see De Haan and Resnick, 1994). See Cuevas and Fraiman (2010) for a wide review on this topic.

The support estimation problem is established as the problem of estimating the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X with probability distribution $\mathbb{P}_X$ from independent and identically distributed observations, $\mathcal{X}_n = \{X_1, ..., X_n\}$, taken in it. From a practical point of view in Figure 1.1, is it possible to reconstruct the contour of Aral Sea in 2000 from a realization of $\mathcal{X}_{2000}$?

Geffroy (1964), Rényi and Sulanke (1963, 1964) are the first references on support estimation. Rényi and Sulanke (1963, 1964) studied this problem when S is convex in the two-dimensional case. They proposed a natural estimator, the convex hull of the sample points $\mathcal{X}_n$ and they studied its asymptotic behaviour. But, what happens if S is not convex? For instance, if the support S has more than one connected component then the convex hull of sample could not to be a good estimator. Therefore, to solve this limitation, two alternatives can be considered: no assumption is made on the shape of S a more flexible shape restriction than convexity is assumed on the shape of S. For the first approach, Chevalier (1976) and Devroye and Wise (1980) proposed a smoothed version of the sample $\mathcal{X}_n$ as an estimator for the support. The problem of support estimation was introduced by Devroye and Wise (1980) in connection with a

practical application, the detection of abnormal behavior of a system, plant or machine. Asymptotic results on the performance of the estimator were obtained, among others, by Chevalier (1976), Devroye and Wise (1980) and Korostelëv and Tsybakov (1993). In the second approach, Baíllo et al. (2000) and Baíllo and Cuevas (2001) assumed that S was connected and star-shaped, respectively, incorporating these prior informations on Devroye and Wise's estimator. Rodríguez-Casal (2007) studied first the estimation of an $r-$convex support with $r > 0$. The $r-$convexity assumption will be presented and studied in depth in the next sections.

Figure 1.1: A realization of $\mathcal{X}_{2000}$ on the Aral Sea (left). Aral Sea's image from the Moderate Resolution Imaging Spectroradiometer on NASA's Terra satellite in 2000 (center). Aral Sea's boundary (right).

When an important part of the support S is almost empty from the probabilistic point of view, estimating the support could not to be too interesting. In this case, if f denotes the density function of X then it could make sense to consider $t-$level sets of type

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geqslant t\} \tag{1.1}$$

where $t > 0$. However, in most of the applications, the practitioner needs to guarantee that the level set has a fixed probability content greater than or equal to $1 - \tau$ with $\tau \in (0, 1)$. So, the value of t is unknown and an alternative level set definition can be presented:

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geqslant f_\tau\} \tag{1.2}$$

where

$$f_\tau = \sup \left\{ y \in (0, \infty) : \int_{-\infty}^{\infty} f(t) \mathbb{I}_{\{f(t) \geqslant y\}} \geqslant 1 - \tau \right\}. \tag{1.3}$$

Studying regions with different probability concentrations can be useful, for instance, to analyze the possible spatial clustering of rare diseases. This kind of studies has grown in literature considerably, see Diggle (2013). A data set that consists of the residential coordinates for 322 cases diagnosed of chronic granulocytic leukemia in the North West of England between 1982 up to 1998 (inclusive) is showed in Figure 1.2. This real data set will be described in depth in next sections. For values of τ close to one, the level set $L(\tau)$ represents the domain concentrated around the greatest mode. However, if τ is close to zero then it represents the effective support of the density f, see Figure 1.2.

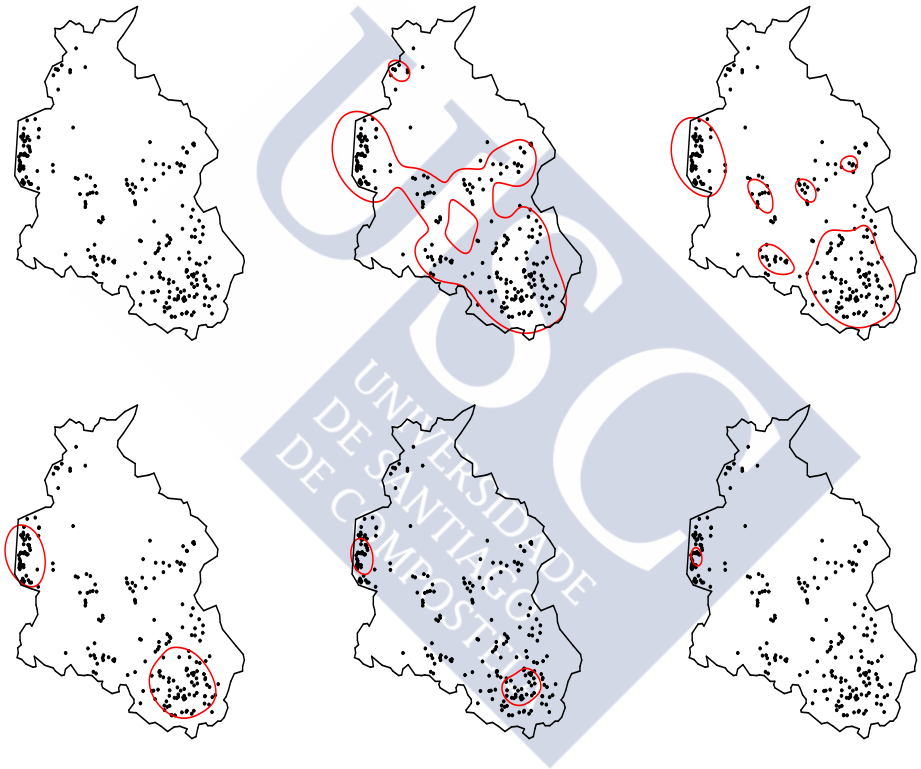


Figure 1.2: In the first row, distribution of 322 cases diagnosed of leukaemia on the North West of England (left), level set for $\tau = 0.05$ in red color (center) and $\tau = 0.25$ (right). In the second row, level set for $\tau = 0.5$ in red color (left), level set for $\tau = 0.75$ in red color (center) and $\tau = 0.95$ (right).

Two steps are necessary in order to reconstruct $L(\tau)$ in a fully data-driven way from the random sample $\mathcal{X}_n$ specified. First, the threshold f_τ must be estimated in order to ensure the probability content. Then, a method to reconstruct the level set must be selected. There are three methodologies for the estimation of level sets: Plug-in, excess

mass and hybrid. The choice of an algorithm depends on the geometric assumptions made on the shape of the level set just as support case. Next, these three methodologies are detailed briefly:

The *plug-in estimation* is the most natural choice to estimate $L(\tau)$ when no geometric information about the level set is available. It based on replacing f by a nonparametric estimator for the density function f_n in (1.2). Usually, f_n denotes the kernel estimator. Given $\mathcal{X}_n$, the kernel density estimator at point x is defined as

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(x - X_i\right), \tag{1.4}$$

where $K_H(z) = |H|^{-1/2}K(H^{-1/2}z)$, $|\ |$ represents the determinant, $K : \mathbb{R}^d \to \mathbb{R}$ denotes a kernel function (in what follows the Gaussian density) and H, a $(d \times d)-$dimensional symmetric positive definite matrix. The estimator defined in (1.4) is heavily dependent on the matrix H, see Wand and Jones (1995). Therefore, the practical problem of the plug-in methodology is the choice of this matrix. Unlike density estimation, the level set estimation has been considered in literature from many points of view but, in general, without deepening in methods for selecting H. In fact, this problem was first considered by Baíllo and Cuevas (2006) in the context of nonparametric statistical quality control. Singh et al. (2009) presented a plug-in procedure that is based on an empirical density estimator, the regular histogram. Later, Samworth and Wand (2010) derive an automatic bandwidth selection rule to estimate density level sets but only in the one-dimensional case.

The *excess mass estimation* assumes that the researcher has information a priori about the shape of the level set $G(t)$ defined in (1.1). Although they are not designed for estimating the level set $L(\tau)$ defined in (1.2), they can be adapted easily. This methodology was first proposed by Hartigan (1987) and Müller and Sawitzki (1987). Then, Polonik (1995) extended and investigated it in a very general framework. These algorithms are based on a quite simple idea: The set $G(t)$ maximizes the functional

$$H_t(B) = \mathbb{P}(B) - t\mu(B),$$

on the Borel sets B where $\mathbb{P}$ denotes the probability measure induced by f and μ, the Lebesgue measure. In addition, H_t can be estimated empirically. So, if $G(t)$ is assumed to belong to a family of sets then it could be reconstructed by maximizing the empirical version of the previous functional on the family considered. Consequently, unlike the plug-in approximation, excess mass methods do not need to smooth the sample $\mathcal{X}_n$ and, in addition, they impose geometric restrictions on the estimators.

The last and third methodology is a *hybrid* of the two previous ones. Just as the excess mass methods, the hybrid methodology assumes some shape restrictions on the

class of sets considered and, like the plug-in methods, it needs to smooth the data set. Walther (1997) proposed the granulometric smoothing method to reconstruct level sets $L(\tau)$ adapting the Devroye and Wise (1980)'s support estimator for level sets under $r-$convexity assumptions.

According to the previous ideas, most of the existing set estimators depend on smoothing parameters just as the nonparametric functional estimators. As we will see, some of them could be seen as shape indexes. Because of this, set estimation can be considered as the geometric counterpart of the classical theory of nonparametric functional estimation, see Simonoff (1996). However, one of the most important differences is related to the strong geometrical motivation behind set estimation. Since in this theory the target is reconstructing sets, rather than functions, it is natural that distances between sets, as well as the geometric properties concerning their shapes, play an significant role. Next, the specific mathematical tools in set estimation are presented in the Section 1.2. Concretely, we will define some useful distances in set estimation in Section 1.2.1 and the geometric shape conditions will be presented in Section 1.2.2. The rest of this chapter is organized as follows. In Section 1.3 the main results contained in this thesis are reviewed. Finally, the real and simulation data sets that will be used in this research work will be presented in Section 1.4.

1.2 Mathematical tools in set estimation

Some distances between sets and several interesting geometric shape conditions will be introduced in this section. Let $\mathbb{R}^d$ be Euclidean space and let $\|\ \|$ be the Euclidean norm. The complement of a set A, its closure, its interior and its boundary are denoted by A^c, $\overline{A}$, $\mathrm{Int}(A)$ and ∂A, respectively. The closed and open balls centered at x and with radius $r > 0$ are denoted by $B_r[x]$ and $B_r(x)$, respectively.

1.2.1 Distances between sets

In order to evaluate the quality of a set estimator, it is necessary to measure the distance between the estimator and the theoretical set. The most common distance between points in $\mathbb{R}^d$ is the Euclidean distance but the distance between sets is a different concept. The sets A and C in Figure 1.3 share a border and it could think that the distance between them is zero but A and C are quite different. Next, three distances will be presented on subsets of $\mathbb{R}^d$.

One of the most useful distances in set estimation is the distance in measure. It is defined on the bounded subsets of $\mathbb{R}^d$ which belong to the Borel $\sigma-$algebra.

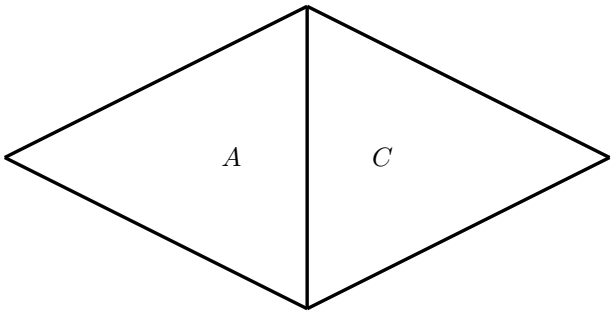


Figure 1.3: A and C have a common border.

Definition 1.2.1. *Let A and C be two bounded Borel sets. The distance in measure between A and C is defined by*

$$d_\mu(A, C) = \mu(A \triangle C),$$

where μ denotes the Lebesgue measure and $\triangle$, the symmetric difference (see Figure 1.4), that is,

$$A \triangle C = (A \backslash C) \cup (C \backslash A).$$

More generally, if f denotes a density function in $\mathbb{R}^d$ and A and C are two Borel sets (not necessarily bounded) then it is possible to define the distance

$$d_{\mu_f}(A, C) = \int_{A \triangle C} f(t)\ dt.$$

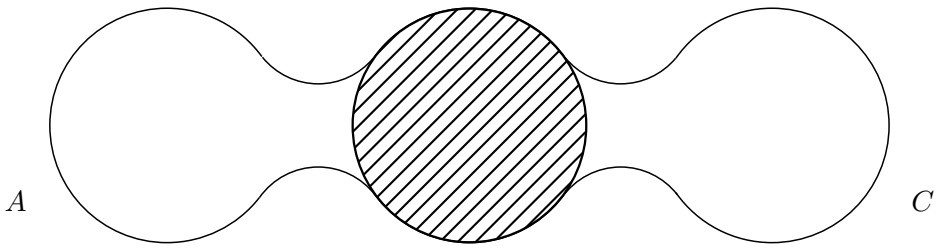


Figure 1.4: Symmetric difference between A and C in $\mathbb{R}^2$.

Remark 1.2.2. *d_μ is not a metric. For instance, if B_1 and B_2 are two bounded Borel sets such that B_1 differs from B_2 only by a finite set of points then $B_1 \neq B_2$ but $d_\mu(B_1, B_2) = 0$.*

From an intuitive point of view, $d_{\mu_f}(A, C)$ represents the probability that an observation from a random variable with density f belongs only to one of the two sets A and C. In general, d_{μ_f} gives more weight in regions where data tends to be denser.

Another alternative for measuring the distance between two sets is provided by the Hausdorff distance. It is defined over the space of the nonempty compact subsets in a given metric space. In particular, over the $d-$dimensional Euclidean space, $\mathbb{R}^d$.

Definition 1.2.3. *Let $A, C \subset \mathbb{R}^d$ be two sets. The Minkowski addition is defined by*

$$A \oplus C = \{a + c : a \in A,\ c \in C\}.$$

The Minkowski subtraction is defined by

$$A \ominus C = \{x : x + C \subset A\}$$

where $x + C$ denotes $\{x\} \oplus C$. For $\delta \in \mathbb{R}$,

$$\delta A = \{\delta a : a \in A\}.$$

Definition 1.2.4. *Let A and C be nonempty compact subsets of $\mathbb{R}^d$. The Hausdorff distance between A and C is defined by*

$$d_H(A, C) = \max\left\{\sup_{a \in A} d(a, C), \sup_{c \in C} d(c, A)\right\},$$

where

$$d(a, C) = \inf\{\|a - c\| : c \in C\}.$$

Equivalently,

$$d_H(A, C) = \inf\{\varepsilon > 0 : A \subset C \oplus B_\varepsilon(0),\ C \subset A \oplus B_\varepsilon(0)\}.$$

Figure 1.5 illustrates the difference between the usual Euclidean distance and the Hausdorff distance.

Remark 1.2.5. *The Hausdorff distance was defined over the collection of nonempty compact subsets of $\mathbb{R}^d$. It can be proved that d_H is a metric, see Section 2.4 in Edgar (1990) or Section 1.4 in Matheron (1975) for more details.*

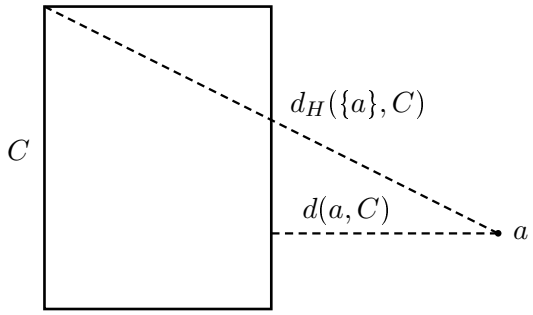


Figure 1.5: Hausdorff distance between $\{a\}$ and C.

Hausdorff distance quantifies the physical proximity between two sets whereas the distance in measure is useful to quantify their similarity in content. According to the previous comments, these distances between sets will be used to measure the error of the estimation when a set is reconstructed from a random sample of points $\mathcal{X}_n$.

1.2.2 Shape conditions

In the most general cases, no geometric conditions are imposed on the theoretical set that is going to be estimated. However, more sophisticated estimators could be considered if some additional information is given on it. Next, the family of convex and $r-$convex sets and the family of sets that satisfies rolling free conditions will be presented.

Definition 1.2.6. *A set $A \subset \mathbb{R}^d$ is said to be convex if for every pair of points $x,\ y \in A$ and for all $\gamma \in [0,1]$, it is verified that $\gamma x + (1-\gamma)y \in A$.*

According to the Definition 1.2.6, in the one-dimensional case, the only convex and compact sets are the intervals $[a, b]$ with $a \leqslant b$. Figure 1.6 shows two sets in $\mathbb{R}^2$. The first one is convex but the second one is not convex.

Definition 1.2.7. *Let $A \subset \mathbb{R}^d$ be a set. The convex hull of A is defined as the intersection of all convex sets in $\mathbb{R}^d$ containing A. It is denoted by conv(A).*

Then, the convex hull of a set A is the smallest convex set containing A. Of course, if A is convex then $A = \text{conv}(A)$. The convexity assumption may be too restrictive in practice. For example, it is not satisfied for sets with more than one connected components. A more general geometric condition is the $r-$convexity with $r > 0$ that generalizes the convexity property, see Definition 1.2.8.

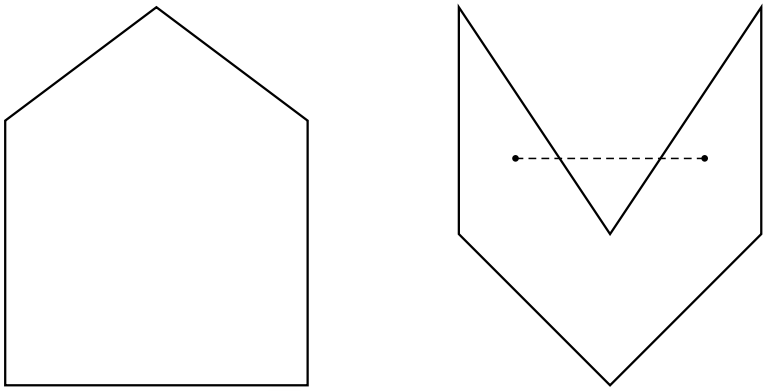


Figure 1.6: Convex set (left). Nonconvex set (right).

Definition 1.2.8. *A closed set $A \subset \mathbb{R}^d$ is said to be $r-$convex for some $r > 0$ if*

$$A = C_r(A),$$

where

$$C_r(A) = \bigcap_{\{B_r(x): B_r(x) \cap A = \emptyset\}} (B_r(x))^c$$

denotes the $r-$convex hull of A.

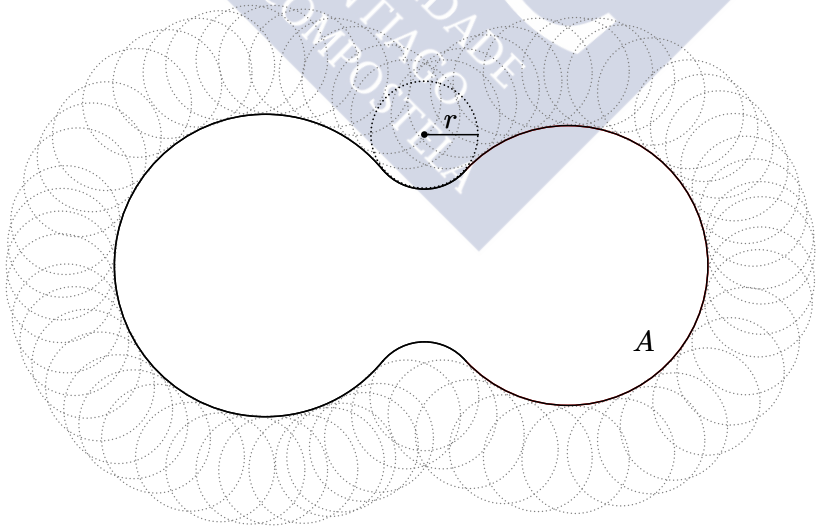


Figure 1.7: The set A is equal to $C_r(A)$. Therefore, A is $r-$convex.

According to Figure 1.7, the value of the parameter r is related to the shape of the set A. Furthermore, the $r-$convex hull of a set A generalizes in a natural way the concept of convex hull. The first one is calculated as the intersection of the complements of open balls with radius r which do not intersect A. The closure of the second one coincides with the intersection of all closed half spaces containing A.

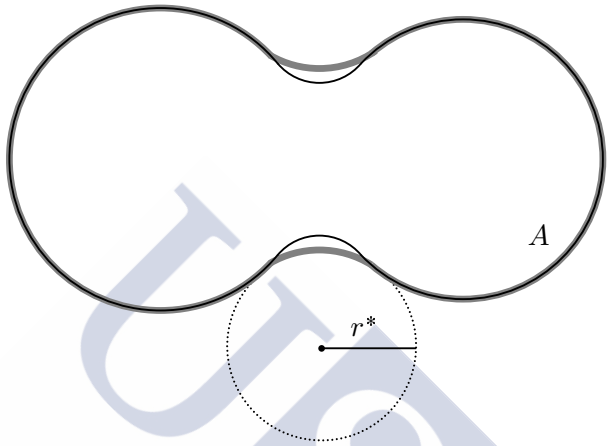


Figure 1.8: The set A in black is not equal to $C_{r*}(A)$ in gray. Therefore, A is not r^*-convex.

From a geometric point of view, the $r-$convex hull is related to the erosion and dilation operators. It is verified that

$$C_r(A) = (A \oplus B_r(0)) \ominus B_r(0),$$

where $\oplus$ and $\ominus$ denotes the Minkowski operators, see Definition 1.2.3. Figure 1.9 shows the process of dilation and erosion for a set $A \subset \mathbb{R}^2$.

Many interesting properties are satisfied by the $r-$convex hull. For instance, $C_r(A) \subset C_{r*}(A)$ for all $r \leqslant r^*$, compare Figures 1.7 and 1.8. In addition, it is easy to prove that if a set A is closed and convex then it will be $r-$convex for all $r > 0$. The reciprocal property will be also satisfied if the interior of the convex hull of A is not empty, see Walther (1999). Figure 1.10 shows these relationships for three interesting examples.

Walther (1997) studied the relationship between the $r-$convexity, the Serra's regular model and the free rolling condition. In Definition 1.2.9 the Serra's model is defined. For more details about Serra's regular model, see Serra (1984).

Definition 1.2.9. *Serra's regular model is the class of compact sets A that are morphologically open and closed with respect to the compact ball $B_r[0]$ of radius r for some $r > 0$, that is,*

$$A = (A \ominus B_r[0]) \oplus B_r[0] = (A \oplus B_r[0]) \ominus B_r[0].$$

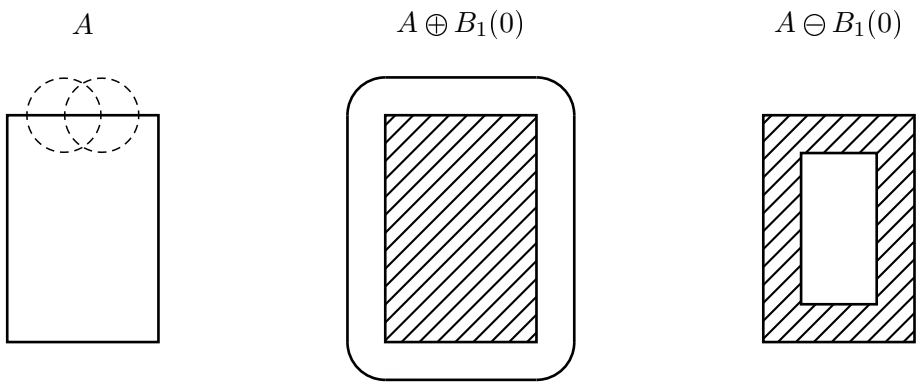


Figure 1.9: Dilation and erosion for a set $A \subset \mathbb{R}^2$.

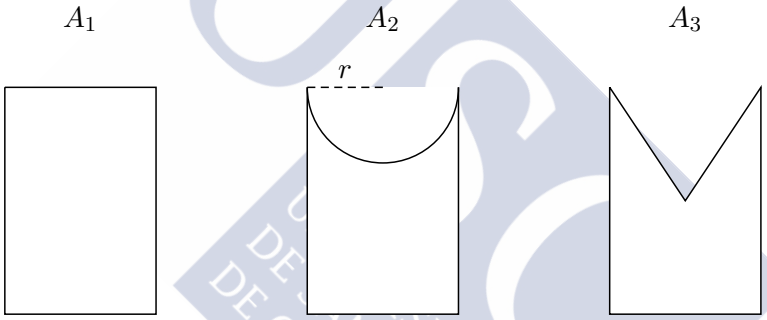


Figure 1.10: A_1 is convex and, so, $r-$convex for all $r > 0$. A_2 is not convex but it is $r-$convex. A_3 is not convex and it is not $r-$convex, for all $r > 0$.

Next, the free rolling condition and some interesting comments about it are introduced. It can be seen as a sort of geometric smoothness statement. For a detailed discussion of these issues, see Walther (1997, 1999).

Definition 1.2.10. *Let $A \subset \mathbb{R}^d$ be a closed set and $r > 0$. A ball of radius r is said to roll freely in A if for each boundary point $b \in \partial A$ there exists $x \in \mathbb{R}^d$ such that $b \in B_r[x] \subset A$.*

Figure 1.11 shows two sets verifying the free rolling property.

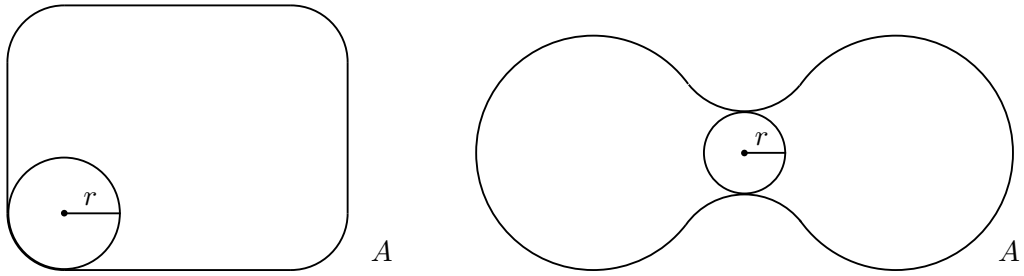


Figure 1.11: A ball of radius r rolls freely in $A \subset \mathbb{R}^2$.

The relationship between these three geometric families of sets can be found in Theorem 1.2.11. Indeed, Walther (1997) provided a generalization of the Blaschke's Rolling Theorem that gives an exact geometric characterization of Serra's regular model in terms of $r-$convexity, free rolling conditions or smoothing restrictions for boundaries.

Theorem 1.2.11. *(Walther, 1997) Let $A \neq \varnothing$ be a compact subset of $\mathbb{R}^d$ and $\alpha_0 > 0$. Then the following are equivalent:*

1. *$(A \oplus lB_1[0]) \ominus A = A$ with $l \in [0, \alpha_0)$ and $(A \ominus lB_1[0]) \oplus A = A$ with $l \in [0, \alpha_0)$.*
2. *A and $\overline{A^c}$ are α_0-convex and $int(A_i) \neq \varnothing$ for each path-connected component $A_i \subset A$.*
3. *A ball of radius l rolls freely inside each path-connected component of A and $\overline{A^c}$ for all $0 \leqslant l \leqslant \alpha_0$.*
4. *∂A is $(d-1)-$dimensional submanifold in $\mathbb{R}^d$ with the outward pointing unit normal vector $\eta(s)$ at $s \in \partial A$ satisfying the Lipschitz condition*

$$\|\eta(s) - \eta(t)\| \leqslant \frac{1}{\alpha_0}\|s - t\| \text{ for all } s,\ t \in \partial A.$$

   Moreover, for some $\alpha_0 > 0$ the preceding is equivalent to:

5. *A belongs to Serra's regular model.*

Figures 1.12 and 1.13 show a $r-$convex set which does not belong to Serra's regular model. Similarly, Figure 1.14 shows a set which belongs to Serra's regular model.

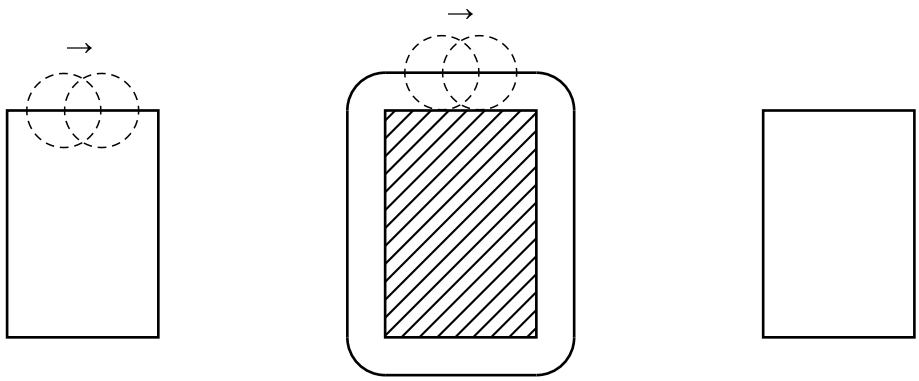


Figure 1.12: $A \subset \mathbb{R}^2$ (left). $A \oplus B_r[0]$ (center). $(A \oplus B_r[0]) \ominus B_r[0]$ (right).

Figure 1.13: $A \subset \mathbb{R}^2$ (left). $A \ominus B_r[0]$ (center). $(A \ominus B_r[0]) \oplus B_r[0]$ (right).

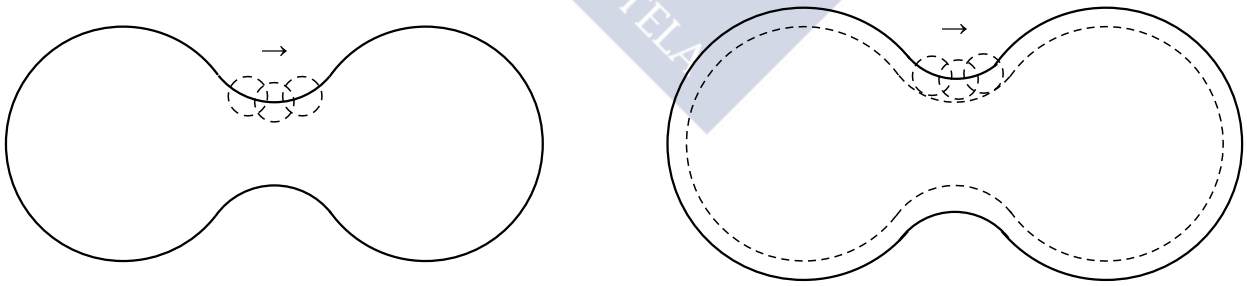


Figure 1.14: $A \subset \mathbb{R}^2$ (left). $A \oplus B_r[0]$ (right).

According to Section 1.1, many methods for reconstructing density level sets or supports assume geometric restrictions on the shape of the set to be estimated. So, the geometric properties defined in this section will be really used in this work.

1.3 A brief overview of the main results

The aim of this section is to summarize the main results achieved in this research. Our interest has been focused on the support and density level set data-driven estimation. This problem is not an easy task because the performance of the estimator depends on the geometric assumptions made on the shape of the set to be estimated. Two new algorithms will be proposed to estimate $r-$convex supports and density level sets.

In Chapter 2, the existing data-driven algorithms in literature for estimating the support and density level sets will be reviewed. It will be showed that in both estimation theories, the estimators typically depend on a sequence of smoothing parameters. Therefore, the theoretical results will make special emphasis on asymptotic properties, especially consistency and convergence rates. Chapter 2 is organized as follows. The literature about support estimation will be presented in Section 2.1. First, the general case, when no geometric assumptions are made on the shape of the support is considered. Then, the reconstruction of convex and $r-$convex supports is studied. They could be estimated as the convex and $r-$convex hulls of sample points, respectively. In that point, it is motivated the need to estimate first the smoothing parameter r if reconstructing $r-$convex supports in a data-driven way is the goal. The main theoretical results about level set estimation will be presented in Section 2.2 by describing in depth the three methodologies for reconstructing density level sets. In particular, the hybrid method proposed by Walther (1997) adapts a specific support estimator to the context of level sets. Two new proposals will be presented in this work for reconstructing convex and $r-$convex density level sets. These two new hybrid algorithms are adaptations of two support estimators too. Just as it has been discussed for the support, it will be show that the estimation of r will be necessary to reconstruct r-convex density level sets in an automatic way. One interesting open question concerns the practical performance of the methods of these three groups of methodologies. Their practical behaviour is compared through an extensive simulation study. We focus on the one-dimensional setting so as to include some methods that do not have multidimensional counterparts. The results are presented in Section 2.3. First, the most competitive algorithms of each methodology are identified. Plug-in methods will be compared in Section 2.3.1. Excess mass algorithms will be studied in Section 2.3.2 and hybrid methodology in Section 2.3.3. A final comparison of the most competitive methods in each group is showed in Section 2.3.4. Chapter 2 closes with some general and useful conclusions for practitioners elaborated from the obtained simulation results in Section 2.3.5.

Chapter 3 focuses on the data-driven reconstruction of $r-$convex supports. According to the previous comments, the estimation of the parameter r is necessary. This problem was first considered by Mandal and Murthy (1997) in the literature. A new

automatic algorithm will be proposed for selecting it from the data under the hypothesis that the sample is uniformly generated. Chapter 3 is organized as follows. The need of estimating of the smoothing parameter r is motivated again in Section 3.1. Some graphical tools for selecting r are presented. They can be useful in a first approximation. As it has been told before, it is necessary to determine the optimal value of r to be estimated. This definition is given in Section 3.2. The estimator for this parameter will be presented in Section 3.3 and its consistency will be proved too. The resulting automatic support estimator obtained from this smoothing parameter is presented in Section 3.4. It is shown that the estimator proposed is able to achieve the same convergence rates as the convex hull for estimating convex sets but under a much more flexible smoothness condition. The numerical and computational aspects to estimate the smoothing parameter are detailed in Section 3.5. This chapter closes with a brief simulation study and a real data example. In Section 3.6, the behaviors of our new proposal and the Mandal and Murthy's method are compared through a simulation study. Finally, a real example is considered. It is tested if the Aral Sea has lost water in the last years in Section 3.7.

Once the good behavior of our hybrid method for estimating $r-$convex density level sets, with fixed r, was checked in Section 2.3, the need of estimating the parameter r appears again. In Chapter 4, a new data-driven algorithm to estimate it will be proposed. As far as we know, this problem is considered by first time in this research work. Chapter 4 is organized as follows. The problem is introduced for level sets $G(t)$ defined in (1.1) in Section 4.1. In Section 4.2, the optimal value of r to be estimated is established. An estimator for it is defined in Section 4.3. Its consistency is proved in Section 4.4. As consequence, an automatic and consistent level set estimator can be calculated from the smoothing parameter estimator. This estimator is presented in Section 4.5 and its convergence rates are obtained too. The numerical and computational questions to estimate the optimal parameter are exposed in Section 4.6. Finally, a real data set will be analyzed in Section 4.7. The distribution of 233 cases of diagnosed chronic granulocytic leukemia and 988 controls on the North West of England is considered. It is interesting to know if the clusters of these two data sets are similar in order to detect the areas where the incidence of the illness is higher.

1.4 Data sets and models

In order to introduce the problems of support and density level set estimation have been considered two real data sets in Section 1.1. In addition and according to Section 1.3, the results of two simulations studies are showed in Chapters 2 and 3. Next, these real data sets and the models considered for simulations are presented.

1.4.1 Real data sets

The Aral Sea and the locations of cases of diagnosed leukaemia and controls in the North West of England will be used in the rest of this research work. Next, these two real data sets will be presented. For the first one, the procedure of generating uniform samples on some water regions of the Aral Sea in 2000 and 2011 will be explained. The second data set will be described in depth. The geographical locations of these two data sets are showed in Figure 1.15.

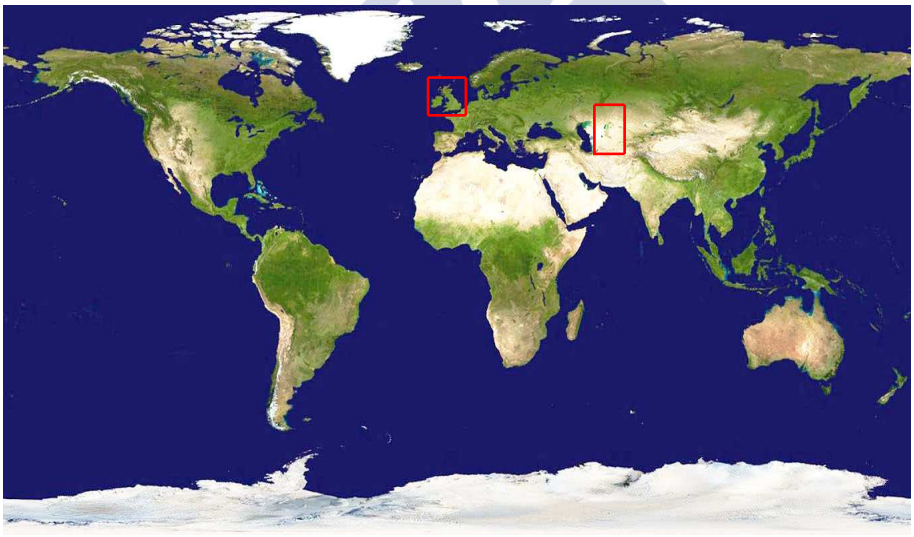


Figure 1.15: Geographical locations of Aral sea and the North West of England, see red rectangles in Asia and Europe, respectively.

Support estimation and Aral Sea. Aral Sea was once the fourth largest lakes in the world. However, water withdrawals for irrigation have devastated the Aral Sea revealing a geology of irregular cliffs. Jiménez and Yukich (2011) analyzed a complex waterfront of the Aral Sea for estimating surface integrals related to mean and standard

deviation of their height. A series of images from the Moderate Resolution Imaging Spectroradiometer on NASA's Terra satellite documents the changes in the water level of the Aral Sea. At the start of this series in 2000, the lake was already a fraction of its 1960 extent. The Northern Aral Sea had separated from the Southern Aral Sea and this part had split into eastern and western lobes that remained tenuously connected at both ends. In the rest of this work, we will consider the last one. Figure 1.16 (left) shows the Southern Aral Sea in 2011. The region of our interest is the area corresponding to the western lobe. It is delimited with red color.

Figure 1.16: Aral sea in 2011 (left). Areal sea's region of interest in 2000 and in 2011, respectively (center). Fisher linear discriminant (right).

Specifically, the study is focused on the photographs in Figure 1.16 (center). They are available on the website of the Earth Observatory, NASA. These two photographs show the Aral Sea in 2011 and 2000 and they will used to generate uniform random samples on these two water areas. Some steps are necessaries to do it:

1. *Image digitizing.* The two jpg files of the original color images in Figure 1.16 (center) have been digitized in an array of 1116×659 pixels. The information stored in every pixel consists of a vector (x_1, x_2, x_3) indicating the level of primary colors at that point. To build photos in Figure 1.16 (center), the pixels from 540 to 725 and from 59 to 385 in the original images to determinate the red rectangle have been selected.

2. *Image identification and cleaning.* The images of interest in Figure 1.16 (center)

must be treated in order to clearly decide the precise shape of the water area. The problem is to decide whether or not a pixel in the picture corresponds to water area. Fisher linear discriminant function was used by basing only on the color coordinates of every point. To put this in more precise terms, two large samples of pixels have been taken in the water and in the land area. Then the classical linear discrimination method was applied to classify the remaining points. The error rate was not appreciable. The result of this automatic discrimination-based treatment is shown in Figure 1.16 (right) where the water area has been colored in black.

From the previous discrimination method, it is possible to generate many uniform samples of size n, in particular, on the two water regions in Figure 1.16 (center) or, in general, on any one.

Density level set estimation for leukaemia data in the North West of England. As has been introduced in Section 1.1, spatial clustering of rare diseases has grown in recent years, in part prompted by increasing concerns over possible links between disease and sources of environmental pollution. See for instance, Besag and Newell (1991) or Diggle (2013). In this work, the particular case of the North West of England will be analyzed. It is one of the most important industrial and commercial regions of the United Kingdom with a diverse range of heavy and light engineering. The health of the people in this region of England is poor in comparison with other regions in both the United Kingdom and parts of Europe. The North West is currently tackling significant health challenges such as cancer, teenage pregnancies, heart disease, obesity, social inequity within the region and the affects of excessive drinking.

The North West of England has five distinctive sub-regions, Cheshire, Merseyside, Cumbria, Lancashire and Greater Manchester. The data set that will be studied in this work derives from the study that provided the data in Henderson et al. (2002). It contains 1221 pairs of points in Lancashire and Greater Manchester. Concretely, it contains the residential coordinates for the 233 cases of diagnosed chronic granulocytic leukemia registered between 1982 up to 1998 (inclusive), together with 988 controls. For the selection of controls, population counts in each of the 8131 census enumeration districts that make up the study-region, stratified by age and sex, were extracted from the 1991 census. The counts were then used to obtain a stratified random sample of two controls per case with coordinates given by their corresponding centroid coordinates (slightly jittered to avoid coincident points). In Figure 1.17, the contour of Lancashire and Greater Manchester and the samples of cases and controls are showed. It is very interesting problem to examine whether the distribution of this kind of cancer mirrored

Figure 1.17: Sub-regions of Lancashire and Greater Manchester on the North West of England (left), distribution of 233 cases of diagnosed leukaemia (center) and 988 controls on Lancashire and Greater Manchester (right) in the North West of England.

that of the population as a whole or whether there was evidence, as implied by concerned local residents, of clustering. This data set is available on the website of Prof. Peter J. Diggle, Lancaster University. We would like to thank him for the helpful explanatory provided on it.

1.4.2 Models for simulations

According to Section 1.3, the results of two simulations studies are presented in this work. First, existing methods for estimating level sets in dimension 1 will be compared in order to determinate which one is more competitive. A set of 18 one-dimensional densities will be used as test models. On the other hand, a new data-driven algorithm for estimating the parameter r for $r-$convex supports in general dimension d is proposed. In this case, three two-dimensional sets or supports which are $r-$convex will be proposed as models for the simulation results.

Densities for comparing methods of level set estimation. The set of 18 densities which will be considered includes the models proposed by Marron and Wand (1992) and three more characteristic densities, see Figure 1.18. They will be denoted by numbers between 1 and 18. The Marron and Wand's density functions goes from model number 1 until model 15. The models 16, 17 and 18 correspond to the marronite, caliper and matterhorn densities proposed in Berlinet and Devroye (1994).

The set of 18 densities considered is wide enough to analyze in depth the behavior of the algorithms to reconstructing level sets. Marron and Wand's densities include densities with different number and kind of modes with interesting properties. The

last three models was included because they are peculiar densities different from the Marron and Wand's densities. For instance, the model 16 present two asymmetric and separated modes. The model 17 has two jumps. It is a non continuous function. So, it is not derivable. About number of modes, it is bimodal. Finally, the model 18 was included because it is very special. It presents an non finite peak in the origin.

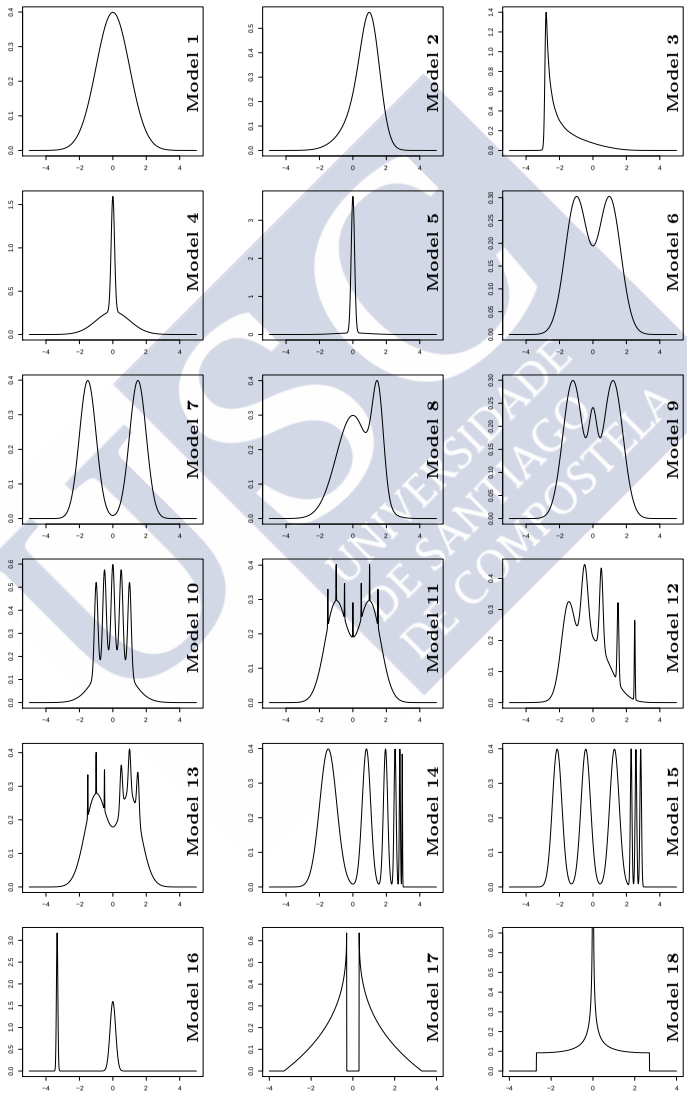


Figure 1.18: Density models for simulation study.

Supports for estimating the parameter r under $r-$convexity assumption. Three support models in $[0,1]^2 \subset \mathbb{R}^2$ with similar volumes will be used for analyzing the behavior of the new method to estimate the parameter $r > 0$. The first one is a circular ring and the other two models are two letters, **C** and **S**. In Figure 1.19, they are represented by including the representation on the horizontal and vertical axis to show the scale of these three sets. In addition, the biggest ball which rolls freely outside the set is represented.

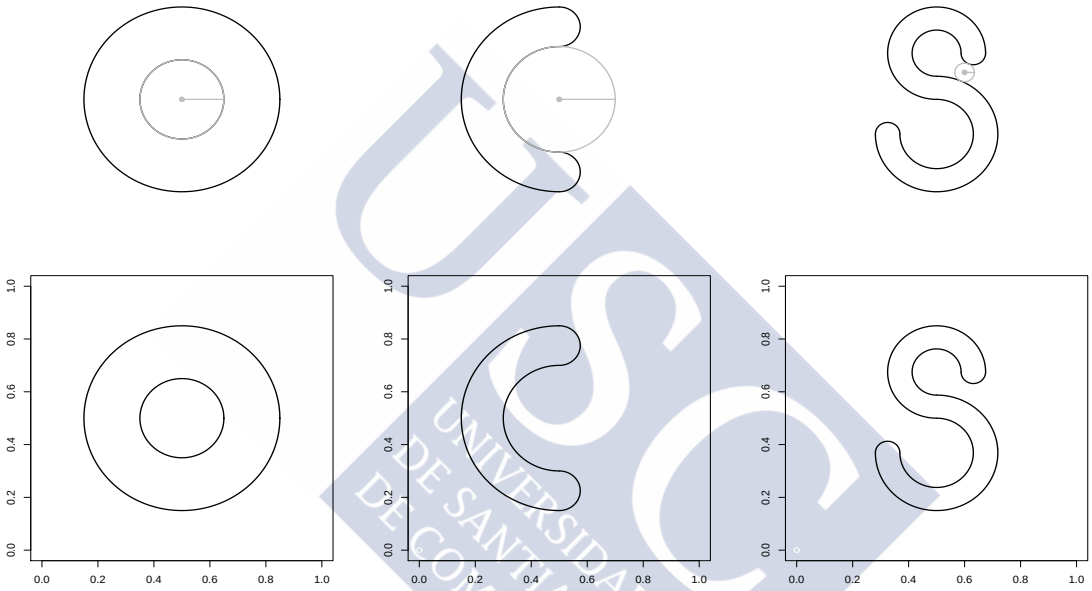


Figure 1.19: Support models in $\mathbb{R}^2$ for simulations. $S = B_{0.35}[(0.5,0.5)] \backslash B_{0.15}((0.5,0.5))$ (first column), $S = \mathbf{C}$ (second column) and $S = \mathbf{S}$ (third column).

According to the Definition 1.2.8, these three sets are $r-$ convex for different values of r. For the first one, $S = B_{0.35}[(0.5,0.5)] \backslash B_{0.15}((0.5,0.5))$, $r \leqslant 0.15$. For the second one, $S = \mathbf{C}$, $r \leqslant 0.2$. The last one set $S = \mathbf{S}$ is $r-$convex for $r \leqslant 0.0353$. Note that $C_r(S)$ and S could be very different. In particular, if $S = B_{0.35}[(0.5,0.5)] \backslash B_{0.15}((0.5,0.5))$ and $r > 0.15$ then $C_r(S) = B_{0.35}[(0.5,0.5)]$ overestimates S considerably. On the other hand, we can observe that the boundary of **S** is not as regular as the boundaries of the other two models.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 2

A revision on the existing data-driven methods for set estimation

Once distances between sets and some geometric properties of interest have been defined in Section 1.2, it is possible to review the literature on set estimation. According to the previous ideas, set estimation is the geometric counterpart of the classical theory of nonparametric functional estimation. In both theories the estimators typically depend on a sequence of smoothing parameters, the theoretical results make special emphasis on asymptotic properties, especially consistency and convergence rates.

This chapter is organized as follows. A brief outline of the classical support estimators available in the literature and their properties is given in Section 2.1. The most general case, when no assumption is made on the shape of the set S, will be first considered. Then, support estimation problem under the convexity and $r-$convexity assumptions is analyzed. In Section 2.2, the main theoretical results in literature about reconstructing density level sets are considered. In addition, the data-driven methods to estimate density level sets will be described in depth. Their behavior will be analyzed through an extensive simulation study and some useful conclusions for practitioners will be extracted in Section 2.3.

Two publications arising from the work compiled in this chapter, see Saavedra-Nieves et al. (2014) and Saavedra-Nieves et al. (Under second review).

2.1 Support estimation

According to the previous notation, $\mathcal{X}_n = \{X_1, ..., X_n\}$ denotes a random sample of i.i.d. observations from a random vector X with absolutely continuous probability distribution $\mathbb{P}_X$. One of the goals of this research work is to reconstruct in a data-driven way the support S. The most general case, when no assumptions are made on the shape of S, will be discussed. The problem changes substantially if some additional information on the support is a priori known. More sophisticated estimators can be considered in this case. Convexity and more flexible shape restrictions, such as $r-$convexity, will be taken into account.

2.1.1 The general case

As has been stated in Section 1.1, the support estimation problem is established as the problem of estimating the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X from independent and identically distributed observations, $\mathcal{X}_n = \{X_1, ..., X_n\}$, taken in it. The main problem is to reconstruct the set by using the available information.

In the most general case, no assumptions are made on the shape of S. Then, sample points $\mathcal{X}_n$ are the only source information and $\mathcal{X}_n$ is an d_H-consistent estimator. That is, with probability one, $d_H(S, \mathcal{X}_n) \to 0$ (it is understood that the limit as $n \to \infty$ is taken). However, with probability one, $d_\mu(S, \mathcal{X}_n) = \mu(S \triangle \mathcal{X}_n) = \mu(S \backslash \mathcal{X}_n) = \mu(S) > 0$ since that $\mathcal{X}_n$ is a finite size set.

Chevalier (1976) and Devroye and Wise (1980) proposed a more general estimator. They considered a smoothing version of the sample points, $\mathcal{X}_n$. Specifically,

$$S_n = \bigcup_{i=1}^{n} B_{\epsilon_n}[X_i], \tag{2.1}$$

where, remember, $B_{\epsilon_n}[X_i]$ denotes the closed ball with center X_i and radius ϵ_n which is assumed that depends only on n.

Figure 2.1 shows the behavior of the estimator for different values of ϵ_n for the same random sample considered at the beginning of this work, see Figure 1.1. If it is too small then the estimator could be split, see Figure 2.1 (left). However, $S \subset S_n$ for high values of ϵ_n, see (right). Devroye and Wise (1980) proved the $d_\mu-$consistent of this estimator in probability and almost surely. If $\epsilon_n \to 0$ and $n\epsilon_n^d \to \infty$ then $d_H(S_n, S) \to 0$ in probability. The assumptions on ϵ_n are identical to those imposed on the bandwidth parameter in nonparametric density estimation, to ensure the consistency. In addition, Devroye and Wise (1980) centered in a concrete testing problem regarding the detection

Figure 2.1: Devroye and Wise's estimator for $\epsilon_n = 5$ (left), $\epsilon_n = 10$ (center) and $\epsilon_n = 40$ (red) and $\mathcal{X}_{2000}$ (black) (right).

of the abnormal behavior of a system. Roughly, a machine is observed in normal operation through the sequence of independent observations $\mathcal{X}_n = \{X_1, ..., X_n\}$ drawn from the density f, and the complement S^c is considered as a danger area. Given a new and unique observation X_{n+1} with density g (possibly different from f), one has to decide whether or not the system behaves abnormally, in the sense that the distribution of X_{n+1} is different from f. A natural testing strategy then consists in rejecting the null hypothesis if X_{n+1} does not belong to S_n. In this context, the distance $d_{\mu_f}(S, S_n)$ has clear interpretation in terms of error of the first kind (or false alarm probability). See Definition 1.2.1 to remember the definition of this distance. Moreover, Korostelëv and Tsybakov (1993) obtained the convergence rates of the estimator defined in (2.1) under some piecewise Lipschitz conditions for the boundaries of S. Cuevas and Rodríguez-Casal (2004) focused on the estimation of ∂S with respect to the Hausdorff metric whereas the almost sure d_H-consistency of S_n can be straightforwardly obtained under the assumption that $\epsilon_n \to 0$, consistency results for $d_H(\partial S, \partial S_n) \to 0$ are not immediate. If $\epsilon_n \to 0$ almost surely together with $S \subset S_n$ then ∂S_n is an almost sure d_H-consistent estimator of ∂S. Another alternative to ensure the almost sure d_H-consistency of the boundary is by assuming certain shape restriction on S. In this context, it is natural to select ϵ_n such that the estimator ϵ_n fulfills the same shape restriction as S. Baíllo et al. (2000) examined the properties of the detection method proposed by Devroye and Wise (1980) and they obtained the convergence rate for the probability of false alarm. In addition, they show that the parameter ϵ_n can be used to incorporate some prior information on the shape of S. Baíllo and Cuevas (2001) assumed that S was star-shaped and they incorporated this additional information to

the Devroye and Wise's estimator selecting the smallest value of ϵ_n such that S_n is also star-shaped. A method of choosing ϵ_n from the sample was proposed by taken into account this idea. Almost sure d_H-consistenty was proved for the estimator of the boundary, ∂S_n. Biau et al. (2008) used the Devroye and Wise's estimator to reconstruct the support of a density function f. In this case, they calculated the exact convergence rates using a general distance d_g as a criterion of accuracy where g is again a density function. Under some mild analytic conditions on f and g, there exists an explicit non-negative constant c such that $\sqrt{n\epsilon_n^d}\mathbb{E}(d_g(S_n, S)) \to c$ as $n \to \infty$, provided $n\epsilon_n^d \to \infty$ and $n\epsilon_n^{d+2} \to 0$.

2.1.2 The convex case

More sophisticated estimators can be used if some additional information on the set is a priori given. Korostelëv and Tsybakov (1993) refers to Geffroy (1964), Rényi and Sulanke (1963), and Rényi and Sulanke (1964) as the first works on support estimation. Rényi and Sulanke (1963) and Rényi and Sulanke (1964) studied the case when $S \subset \mathbb{R}^2$ is a convex support. They proposed a natural estimator, the convex hull of the sample points,

$$H_n = \text{conv}(\mathcal{X}_n).$$

This is just the intersection of all convex sets including $\mathcal{X}_n$. In addition, the estimator fulfills the convexity shape restriction assumed on S. Figure 2.2 shows the convex hull for the same sample points considered previously, see Section 1.1.

Figure 2.2: $\mathcal{X}_{2000}$ on the Aral Sea (left). Convex hull of $\mathcal{X}_{2000}$ (right).

Korostelëv and Tsybakov (1993) proved that H_n is the maximum likelihood estimator in the family of all closed convex sets. Dümbgen and Walther (1996) studied

how closely is S approximated by the convex hull H_n of the sample points. The proximity between the set and the convex hull is studied in terms of the Hausdorff distance in an arbitrary dimension d. It is proved that $d_H(S, H_n) = O((\log n/n)^{1/d})$ almost surely. Furthermore, if ∂S is the boundary is under the conditions of Theorem 1.2.11, $d_H(S, H_n)$ is of order $(\log n/n)^{2/(d+1)}$.

There are some papers literature which studied some geometric characteristics of H_n such as the number of vertices, the number of facets, the volume or the surface area. Bräker and Hsing (1998) studied the asymptotic behaviour of the expected area and perimeter of H_n in the bidimensional case under more general conditions than those considered by Rényi and Sulanke (1963) and Rényi and Sulanke (1964). See Schneider (1988) for an extensive review of classical references in this line or Reitzner (2003) for more results.

2.1.3 A more flexible geometric condition

In practise, the convexity assumption may be too restrictive. Of course, the convex hull of the sample could be not the best option when S is not convex, see Figure 2.2 (right).

According to the ideas in Section 2.1.2, if it is assumed that S is $r-$convex then the natural estimator for the support S is the $r-$convex hull,

$$S_n = C_r(\mathcal{X}_n). \tag{2.2}$$

The estimator defined in (2.2) was first studied by Rodríguez-Casal (2007) under the shape restriction that the set S belongs to Serra's regular model. If S is $r-$convex, $d_H(S, C_r(\mathcal{X}_n)) = O\left((\log n/n)^{1/d}\right)$ almost surely, see Rodríguez-Casal (2007). Note that, although the family of $r-$convex sets is much wider than the family of convex sets, the convergence rates of $d_H(C_r(\mathcal{X}_n), S)$ and $d_H(H_n, S)$ are of the same order, see Dümbgen and Walther (1996). However, if S belongs to Serra's regular model then $d_H(S, C_r(\mathcal{X}_n)) = O\left((\log n/n)^{2/(d+1)}\right)$ almost surely. The same convergence rates are obtained for $d_H(\partial S, \partial C_r(\mathcal{X}_n))$ and $d_\mu(S, C_r(\mathcal{X}_n))$. Again, the order of convergence of $d_H(C_r(\mathcal{X}_n), S)$ is equal to that obtained for $d_H(H_n, S)$ when S is convex and satisfies the smoothness conditions of Theorem 1.2.11.

Although this estimator is well known in the computational geometry since Edelsbrunner et al. (1983) introduced an efficient algorithm to construct the $r-$convex hull of a bidimensional set of points and the $r-$convexity is a shape restriction more flexible than convexity, the estimator proposed in (2.2) has a big limitation. In practise, S is unknown and, as consequence, r too. Walther (1999) studied the influence of this smoothing parameter and he proved that, under some restrictions on a set A, $C_r(A)$

tends to the closure of A if r tends to zero. However, if r tends to infinity then $C_r(A)$ will tend to the convex hull of A. Mandal and Murthy (1997) proposed a new method to estimate r from the sample points $\mathcal{X}_n$ by using the concept of minimum spanning tree. However, this method is only valid for the bidimensional case.

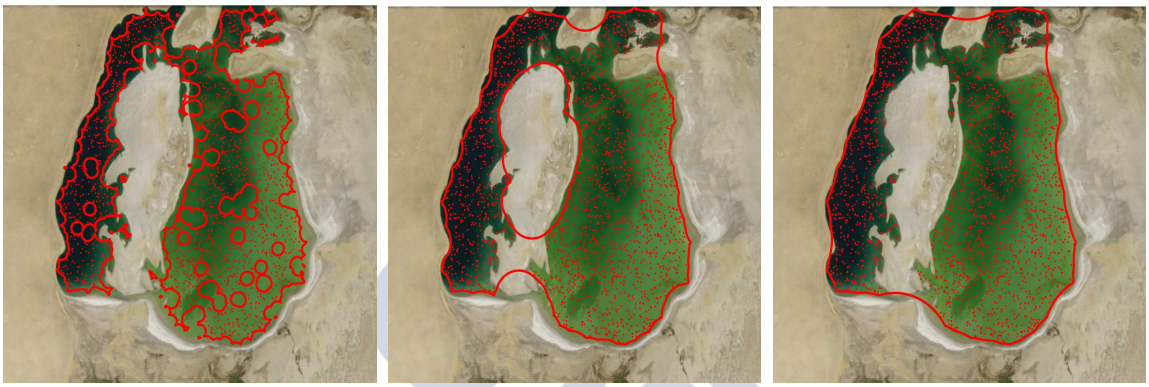


Figure 2.3: $C_{10}(\mathcal{X}_{2000})$ (left). $C_{40}(\mathcal{X}_{2000})$ (center). $C_{100}(\mathcal{X}_{2000})$ (right) is almost equal to the convex hull of $\mathcal{X}_{2000}$.

Figure 2.3 shows the influence of r by using our well-known sample of size $n = 2000$ on the Aral Sea, see Section 1.1. If r is closed enough to zero then $C_r(\mathcal{X}_n)$ coincides practically with the sample points or it is a very split estimator. However, if high values of r are considered then $C_r(\mathcal{X}_n)$ is almost equal to the convex hull of sample points. In addition, land areas will be contained in $C_r(\mathcal{X}_n)$ if r is too large. In this case, it is possible to find a big gap or spacing in the estimator which does not contain any sample point. Janson (1987) calibrated the size of the maximal spacings when the sample distribution is uniform. So, under uniformity restrictions, if a big enough gap can be found in $C_r(\mathcal{X}_n)$ then the chosen value of r will be too large and we should choose another smaller one. These are the basic ideas for the new automatic method which will be proposed in Chapter 3.

2.2 Density level set estimation

Density level sets play a crucial role in various scientific fields, and their estimation has received considerably interest in literature. Since Hartigan (1975) introduced a notion of populational cluster as the connected components of density level sets, many interesting applications have appeared. For more on this approach to clustering, see

for instance, Stuetzle and Nugent (2010). The idea behind the concept of cluster is quite related to the notion of mode and, in fact, some clustering algorithms are based on the estimation of modes, see Cuevas at al. (2000). An interesting application of this clustering approach to astronomical sky surveys was proposed by Jang (2006). Klemelä (2004, 2006) applied a similar point of view to develop methods for visualizing multivariate density estimates. Goldenshluger and Zeevi (2004) used level set estimation in the context of the Hough transform, which is a well-known computer vision algorithm. Some problems in flow cytometry involve the statistical problem of reconstructing a level set for the difference of two probability densities, see Roederer and Hardy (2001). In addition, interesting applications include detection of mine fields based on arial observations, the analysis of seismic data, as well as certain issues in image segmentation, see Huo and Lu (2004). Anomaly or novelty detection is another important application of level set estimation, see Gardner et al. (2006) or Markou and Singh (2003) for a review. An outlier can be defined as the observation that belongs to the set $\{f < f_\tau\}$. In other words, the outlier does not belong to the effective support determined by the level set $L(\tau)$. This approach follows the lines of the nonparametric set-based proposal in Devroye and Wise (1980) to decide if a manufacturing process is out of control. For quality control schemes see also Baíllo at al. (2000) or Baíllo and Cuevas (2006).

The broad scope of level set estimation clearly motivates the need to study in depth the practical performance of the existing methods. In general and according to the Section 1.1, the problem has been approached in the literature using three different nonparametric methodologies: Plug-in methods, excess mass methods and hybrid methods (for a Bayesian alternative, see Gayraud and Rousseau, 2005). The existing data-driven methods for each methodology will be reviewed separately in Sections 2.2.1, 2.2.2 and 2.2.3, respectively. By using the same notation, $\mathcal{X}_n = \{X_1, ..., X_n\}$ denotes a random sample of i.i.d. observations from an absolutely continuous random vector X with density function f.

2.2.1 Plug-in methodology

The simplest option to estimate level sets is the so-called plug-in methodology. It is based on replacing the unknown density f by a suitable nonparametric estimator f_n in (1.2), usually the kernel one. See definition of the kernel density estimator in (1.4). Therefore, this group of methods proposes

$$\hat{L}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geqslant \hat{f}_\tau\}, \tag{2.3}$$

as an estimator, where $\hat{f}_\tau$ estimates f_τ as follows:

1. Through numerical integration methods by solving the equation

$$\int_{\{f_n \geqslant t\}} f_n(x)dx = 1 - \tau$$

in t. This algorithm is consistent with the plug-in philosophy, as $\hat{f}_\tau$ is the corresponding threshold if f is replaced by f_n. However, this may be inefficient from a computational point of view. For consistency results, see Cadre (2006).

2. Hyndman (1996) proposed a method for estimating f_τ by calculating the quantile τ of the empirical distribution of $f_n(X_1), ..., f_n(X_n)$. The computational cost of this approach is lower than that of the previous method. In addition, it can be a useful method for general dimension. For consistency results, see Cadre et al. (2009).

The plug-in methodology is the most common approach, and has received considerable attention in the literature, e.g., Tsybakov (1997), Baíllo (2003), Mason and Polonik (2009), Rigollet and Vert (2009) or Mammen and Polonik (2013). However, it presents two important problems. On one hand, the estimator proposed in (2.3) is not useful if some geometric conditions are assumed a priori on the shape of $L(\tau)$ since it is not clear how to include them on the shape of the estimator. On the other hand, its performance is heavily dependent on the choice of the bandwidth matrix H for calculating f_n. In particular and for the one-dimensional case, H is a positive real number or, equivalently, $H = h^2$ where h denotes a positive real number called bandwidth parameter. If high values of h are chosen, an oversmoothing effect in the kernel estimator is obtained. However, too small values of it produce the opposite effect, see Wand and Jones (1995).

Most of the selection criteria for the smoothing parameter were designed to reconstruct the density function f. However, here, the goal is to reconstruct density level sets. At this point, three quite important questions must be made: Is a good bandwidth for estimating f the best alternative to reconstruct $L(\tau)$? Is the smoothing parameter influential for estimating $L(\tau)$? Does it play the same important role than in density estimation? In what follows, we will consider only the one-dimensional case.

Two standard bandwidths for density estimation have been taken as reference in order to give an answer for the first question. The first one, h_{ISE}, has been obtained by minimizing for a given sample the Integrated Squared Error

$$ISE(h) = \int (f_n(t) - f(t))^2 \, dt.$$

The second one, h_{μ_f}, has been obtained by minimizing for the given sample

$$d_{\mu_f}(L(\tau), \hat{L}(\tau)) = \int_{L(\tau)\triangle\hat{L}(\tau)} f(t)\ dt.$$

The smoothing parameter h_{ISE} minimizes the difference between f and f_n. However, h_{μ_f} is focused on the reconstruction of $L(\tau)$. In addition, it gives more weight to those regions in which sample points tend to be denser.

Figures 2.4, 2.5, 2.6 and 2.7 contain the boxplots and scatter plots for 1000 values of h_{ISE} and h_{μ_f}. They have been calculated by generating 1000 samples of size $n = 1600$ and different values of τ for the model densities 1, 4, 8 y 10, see Section 1.4.2. The two smoothing parameters present completely different behaviors, see Figures 2.4 or 2.5 for models 1 and 4, respectively. On the other hand, as expected, the bandwidth h_{μ_f} is strongly sensitive to the value of the parameter τ that determinates the threshold of the level set.

Figure 2.8 represents the different values of τ on the horizontal axis and the quotients $\frac{h_{\mu_f}}{h_{ISE}}$ for a fixed sample of size $n = 1600$ generated from the same previous models on the vertical axis. As it can be seen, these quotients can almost take any value, even for a fixed sample.

The effect of the bandwidth parameter in density level set estimation is studied graphically for different values of τ in order to give an answer for the second and third questions, see Figures 2.9, 2.10, 2.11 and 2.12. The first of the 1000 samples considered previously for each density model has been fixed. The set where f takes values is represented on the horizontal axis. The plug-in estimator has been calculated for the sequence of values of bandwidth parameters on the vertical axis. These estimators have been represented with gray color. The theoretical density level set is delimited with vertical dotted lines.

Important differences can be observed if the smoothing parameters h_{μ_f} and h_{ISE} are compared. In general, the last one provides clearly worse estimations. For instance, see model 1 in Figure 2.9 when $\tau = 0.8$, model 4 in Figure 2.10 when $\tau = 0.2$, model 8 in Figure 2.11 specially when $\tau = 0.5$ or model 10 in Figure 2.12 when $\tau = 0.8$. In some of these cases, the number of the connected components for the theoretical level set is overestimated.

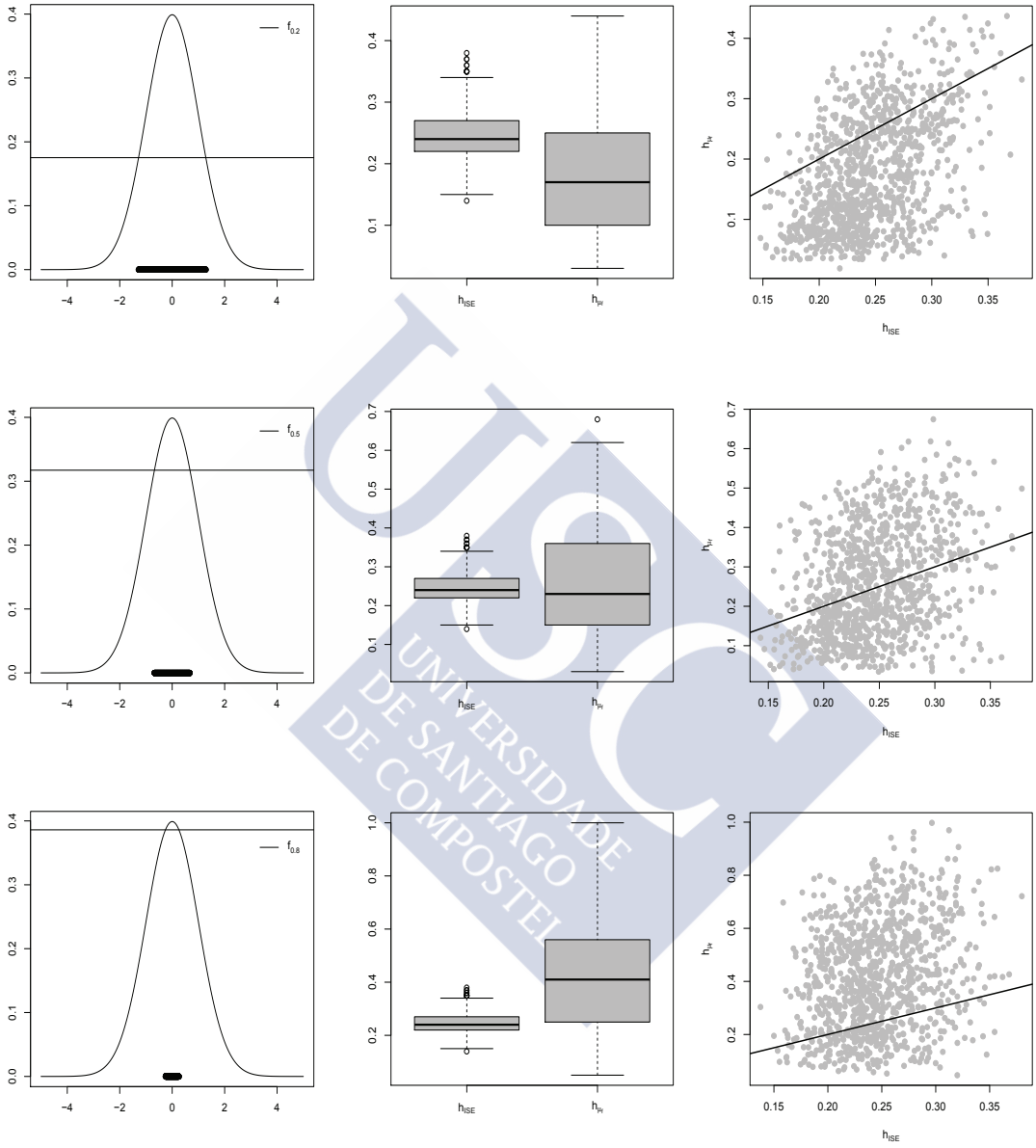


Figure 2.4: In the first column, the level sets for the model 1 are represented for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the boxplots for h_{ISE} (left) and h_{μ_f} (right) are showed. In the third column, the scatter plot with h_{ISE} on the horizontal axis and h_{μ_f} on the vertical one.

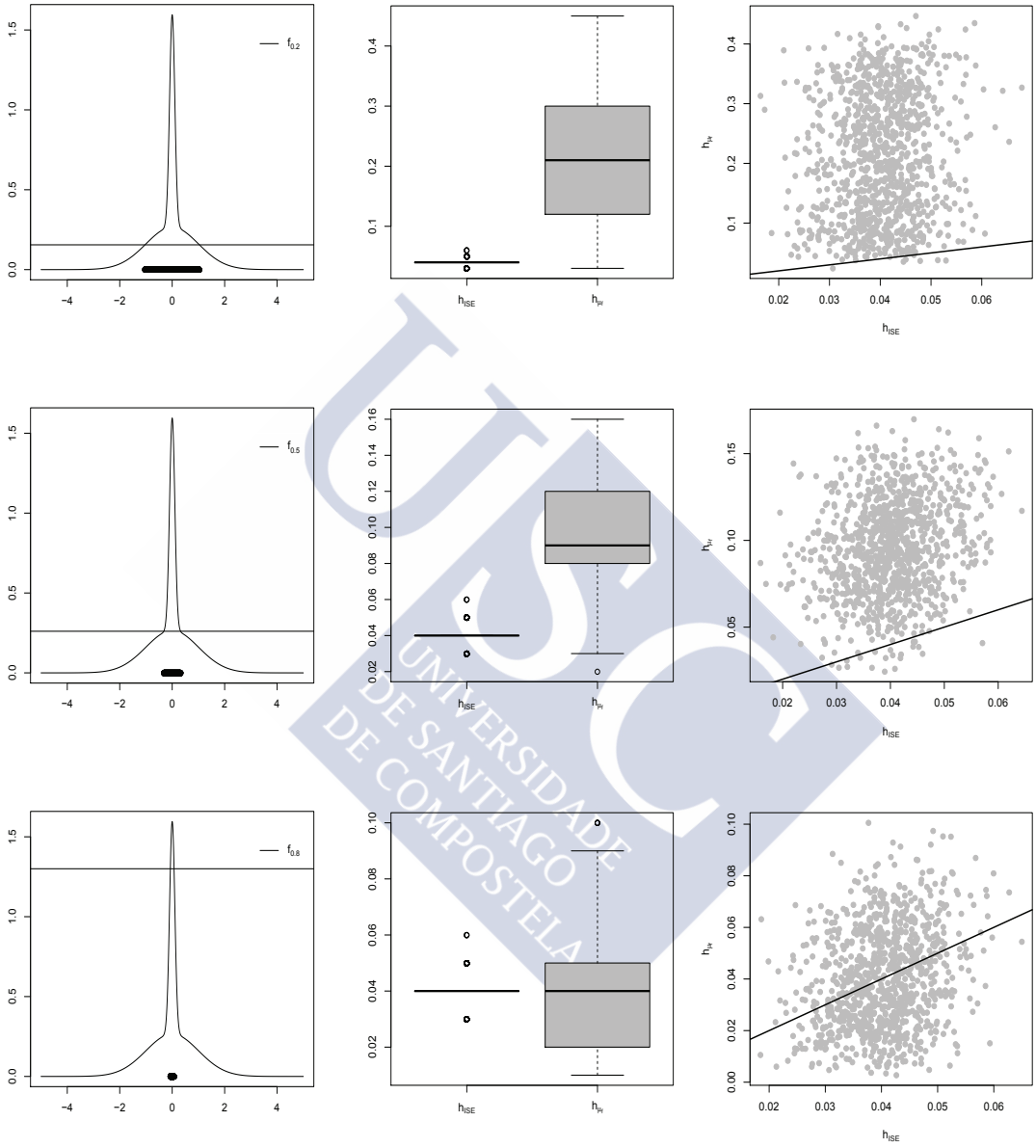


Figure 2.5: In the first column, the level sets for the model 4 are represented for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the boxplots for h_{ISE} (left) and h_{μ_f} (right) are showed. In the third column, the scatter plot with h_{ISE} on the horizontal axis and h_{μ_f} on the vertical one.

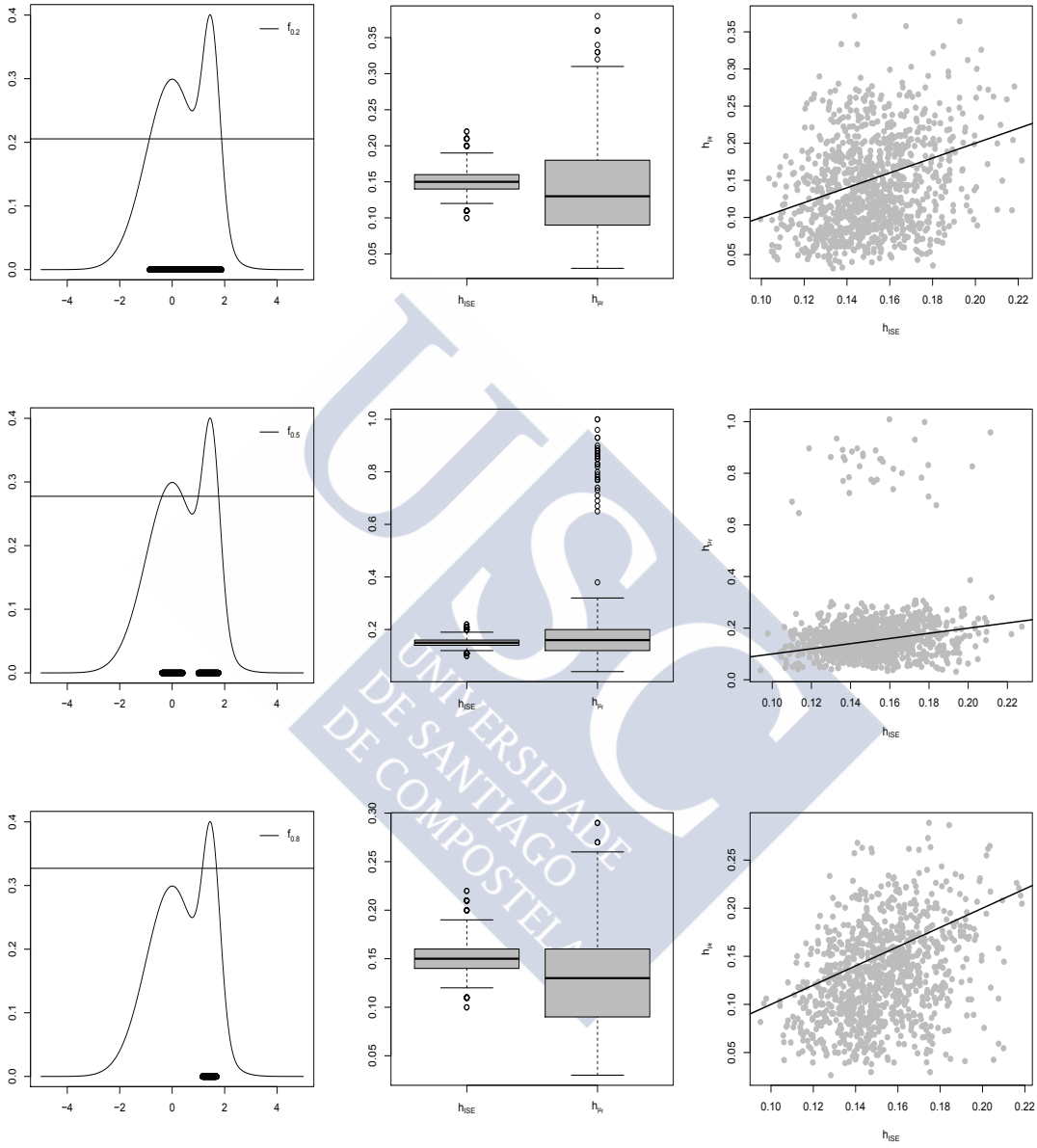


Figure 2.6: In the first column, the level sets for the model 8 are represented for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the boxplots for h_{ISE} (left) and h_{μ_f} (right) are showed. In the third column, the scatter plot with h_{ISE} on the horizontal axis and h_{μ_f} on the vertical one.

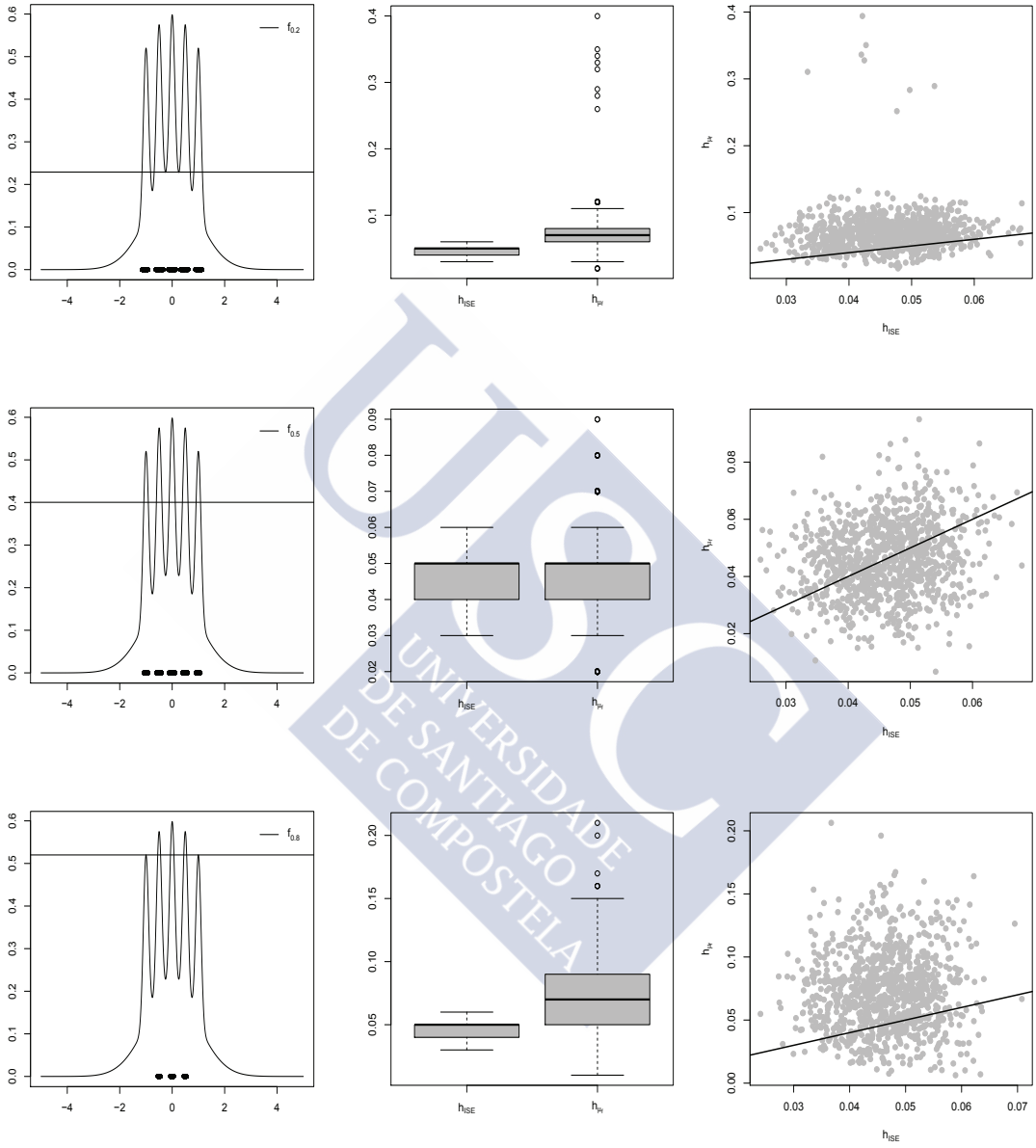


Figure 2.7: In the first column, the level sets for the model 10 are represented for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the boxplots for h_{ISE} (left) and h_{μ_f} (right) are showed. In the third column, the scatter plot with h_{ISE} on the horizontal axis and h_{μ_f} on the vertical one.

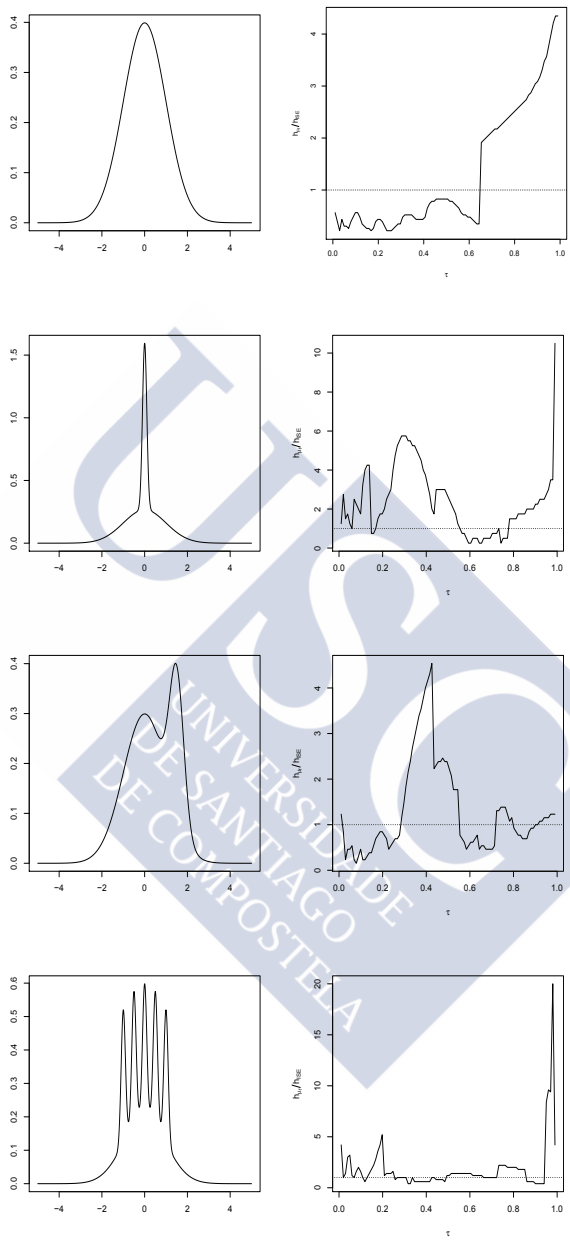


Figure 2.8: In the first column, models 1 (first row), 4 (second row), 8 (third row) y 10 (fourth row) are showed. In the second column, different values of τ are represented on the horizontal axis and $\frac{h_{\mu_f}}{h_{ISE}}$, on the vertical one for each of considered densities and a given sample.

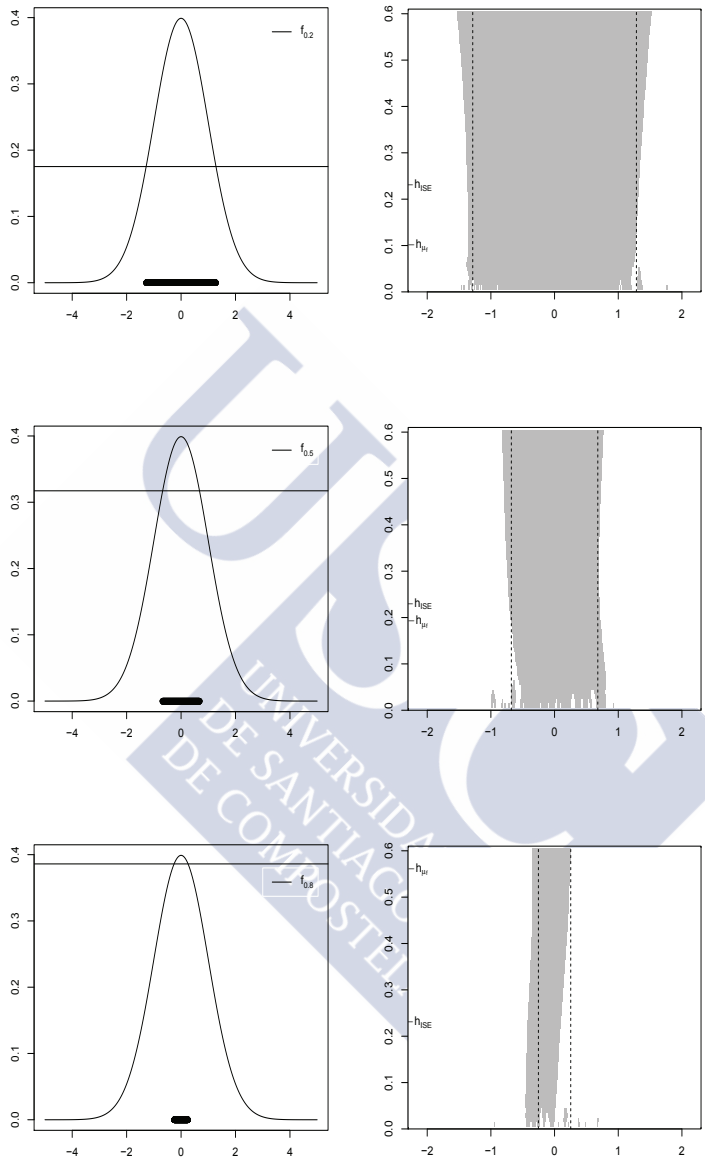


Figure 2.9: In the first column, level sets for model 1 are showed for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the theoretical level set is delimited with dotted lines and different values of the bandwidth parameter are represented on the vertical axis including h_{ISE} and h_{μ_f}. For this sequence of values, the plug-in estimator has been calculated and represented with gray color.

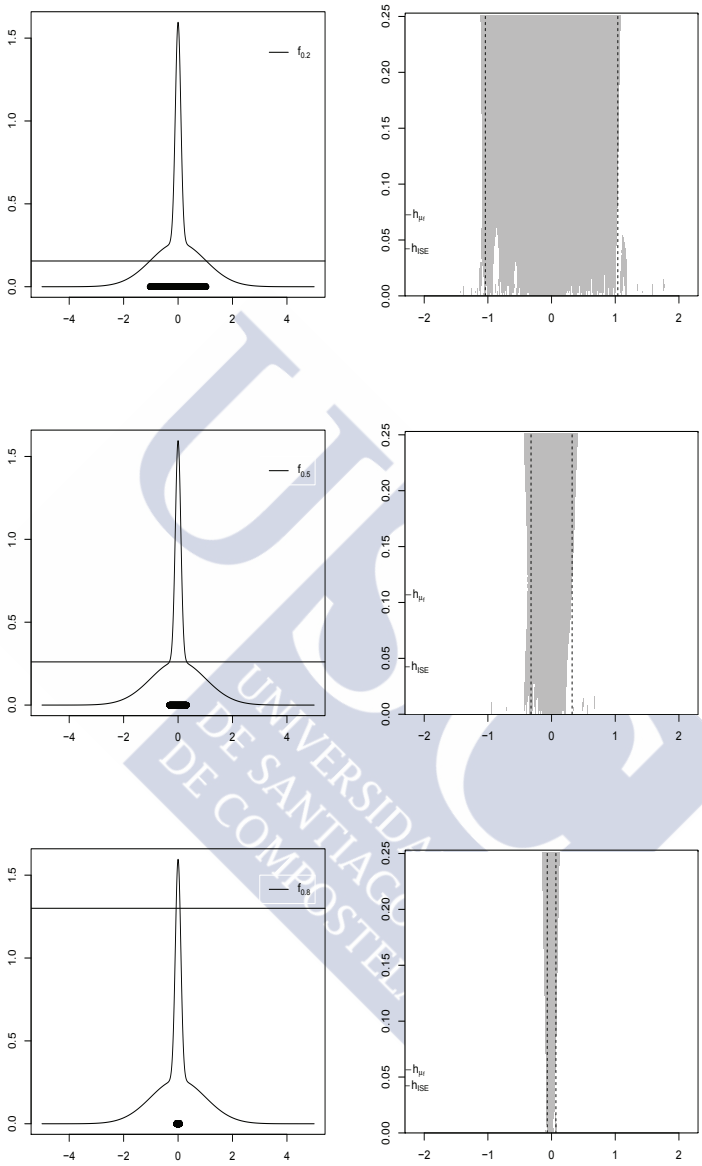


Figure 2.10: In the first column, level sets for model 4 are showed for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the theoretical level set is delimited with dotted lines and different values of the bandwidth parameter are represented on the vertical axis including h_{ISE} and h_{μ_f}. For this sequence of values, the plug-in estimator has been calculated and represented with gray color.

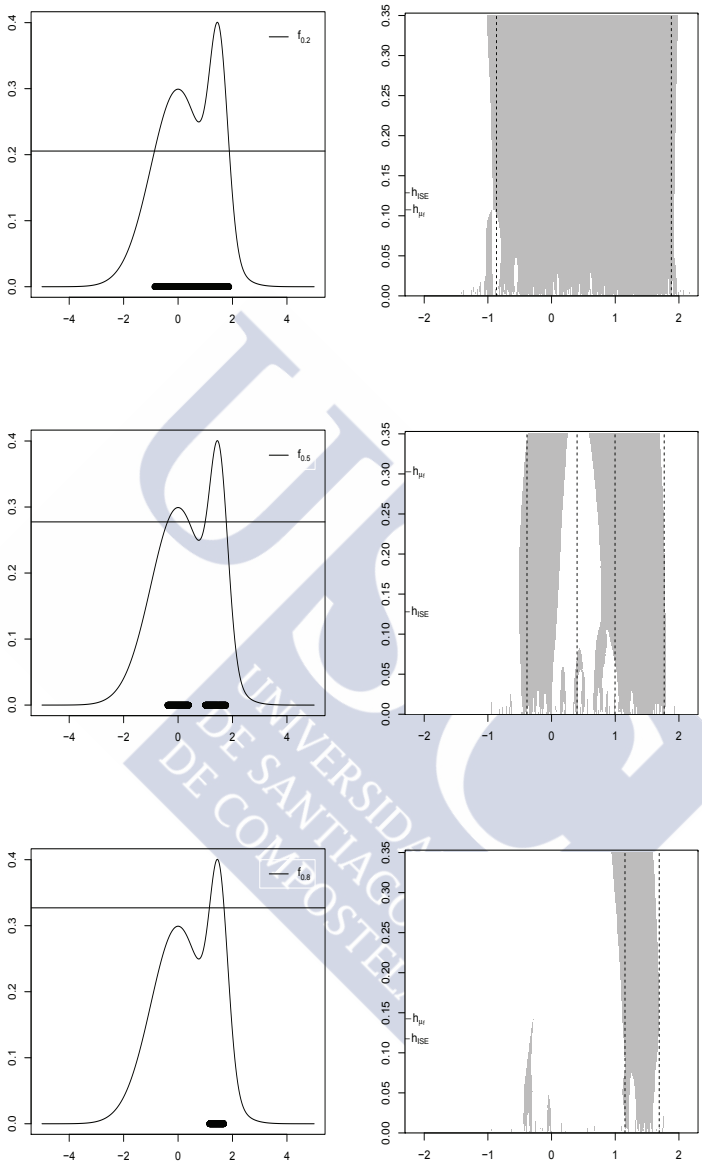


Figure 2.11: In the first column, level sets for model 8 are showed for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the theoretical level set is delimited with dotted lines and different values of the bandwidth parameter are represented on the vertical axis including h_{ISE} and h_{μ_f}. For this sequence of values, the plug-in estimator has been calculated and represented with gray color.

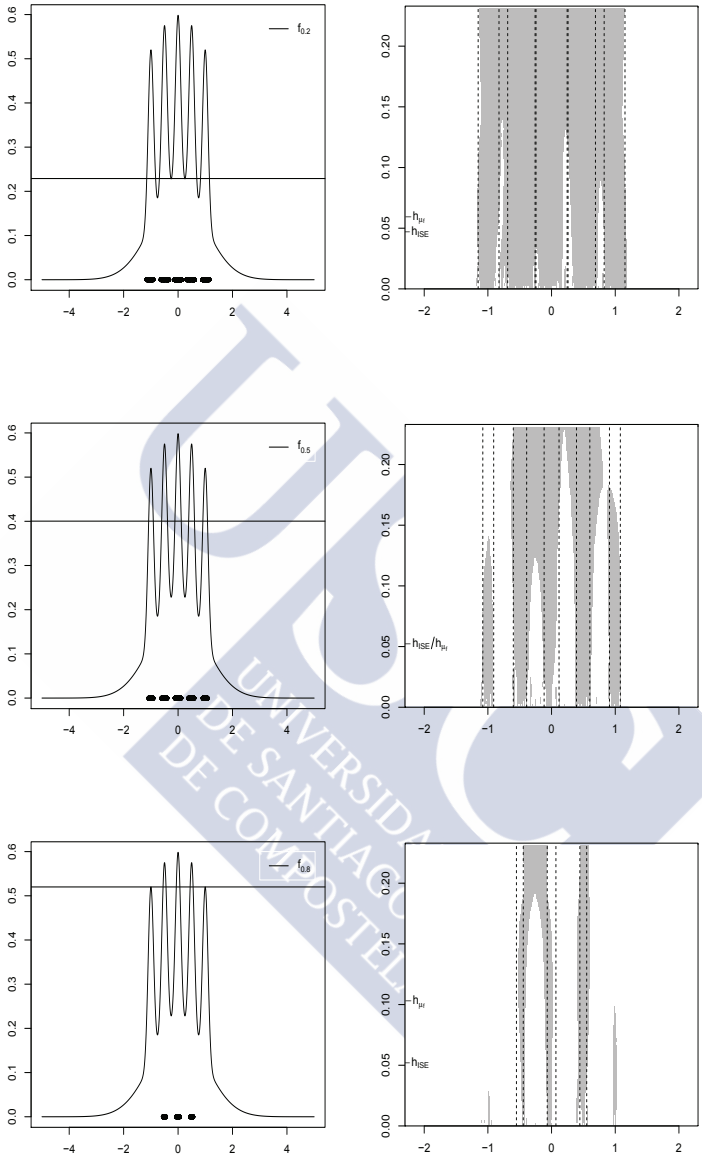


Figure 2.12: In the first column, level sets for model 10 are showed for $\tau = 0.2$ (first row), $\tau = 0.5$ (second row) and $\tau = 0.8$ (third row). In the second column, the theoretical level set is delimited with dotted lines and different values of the bandwidth parameter are represented on the vertical axis including h_{ISE} and h_{μ_f}. For this sequence of values, the plug-in estimator has been calculated and represented with gray color.

Next, the specific data-driven plug-in methods for density level set estimation will be presented in detail. Of course, selectors designed for estimating density functions such as cross validation or Sheather and Jones could be also considered, see Bowman (1984) and Sheather and Jones (1991), respectively.

2.2.1.1 Baíllo and Cuevas' method

Baíllo and Cuevas (2006) used quality control ideas for estimating level sets. In this context, $L(\tau)$ can be seen as a population tolerance region. Let us assume that a machine working in normal operation produces a sequence of independent observations $\mathcal{X}_n$ drawn from the density f. For example, X_i could be a value for a certain quality characteristic of a manufactured product. If the process starts to run out of control, the distribution of the samples will change. The aim is to detect a real change in this distribution as soon as possible, subject to the bound $\tau \in (0,1)$ on the rate of false alarms. The key idea is that, if there is no change in the distribution for a new observation, it is most likely to be within the tolerance limits of the region $L(\tau)$. In practice, we may determine that the $n+1-$ observation is a change-point in the distribution of the process, that is, a new observation X_{n+1} does not follow the distribution of $\mathcal{X}_n$ if X_{n+1} does not belong to the plug-in estimator of $L(\tau)$. Hence, choosing a good smoothing parameter to reconstruct a level set in the context of quality control is an interesting problem. Baíllo and Cuevas (2006) proposed a bandwidth selector by minimizing (over a $51-$point equally spaced grid with center h_{SJ} and width $1.2h_{SJ}$ where h_{SJ} denotes the classical selector by Sheather and Jones, 1991 for the one-dimensional case) a cross validation estimate of

$$|\mathbb{P}_{f_n}(h) - \tau| \text{ where } \mathbb{P}_{f_n}(h) = \int_{\{f_n < \hat{f}_\tau\}} f$$

denotes the probability of false alarm and $\hat{f}_\tau$, a estimator of f_τ. Specifically, $\mathbb{P}_{f_n}(h)$ is approximated by

$$\mathbb{P}_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{f_{n,-i}(X_i) < \hat{f}_{\tau,-i}\}},$$

where $f_{n,-i}$ denotes the kernel estimator with bandwidth h, constructed from $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$ and $\hat{f}_{\tau,-i}$ verifies

$$\int_{\{f_{n,-i} \geqslant \hat{f}_{\tau,-i}\}} f_{n,-i} = 1 - \tau.$$

2.2.1.2 Samworth and Wand's method

Samworth and Wand (2010) proposed a new automatic rule to select the smoothing parameter in the one-dimensional case for the plug-in estimation of density level sets $L(\tau)$. Their proposal is based on a uniform-in-bandwidth asymptotic approximation of the specific set estimation risk function, $\mathbb{E}\{d_{\mu_f}(L(\tau), \hat{L}(\tau))\}$, derived under the following regularity conditions:

SW1. f is uniformly continuous on $\mathbb{R}$. There exist finitely many points $x_1 < ... < x_{2r}$ such that

$$f(x_j) = f_\tau$$

for $j = 1, 2, ..., 2r$ and moreover there exists $\delta > 0$ such that f is twice continuously differentiable in

$$\cup_{j=1}^{r}[x_{2j-1} - \delta, x_{2j} + \delta]$$

with $f^{'}(x_{2j-1}) > 0$ and $f^{'}(x_{2j}) < 0$ for $j = 1, ..., r$.

SW2. Let $h^- = {h_n}^-$ and $h^+ = {h_n}^+$ be nonnegative sequences such that $h^- \leqslant h^+$, such that

$$n(h^-)^4/\sqrt{\log(1/h^-)} \to \infty \text{ and } h^+ \to 0$$

as $n \to \infty$. Then, $h = h_n$ is a sequence with ${h_n}^- \leqslant h_n \leqslant {h_n}^+$ for all n.

SW3. The kernel K is nonnegative, continuously differentiable, of bounded variation, and satisfies:

$$\int xK(x)\, dx = 0,\ \mu_2(K) \equiv \int x^2 K(x)\, dx < \infty \text{ and } \int K^{'}(x)^2\, dx < \infty.$$

The first restriction guarantees that f_τ is the only positive real number such that

$$\int_{\{f(t) \geqslant f_\tau\}} f(t)\, dt = 1 - \tau.$$

Under the previous assumptions, with probability one and for n large enough, $\hat{f}_n(x)$ presents an analogous property. In other words, $\hat{f}_\tau$ is the only positive real number such that

$$\int_{\{f_n(t) \geqslant \hat{f}_\tau\}} f_n(t)\, dt = 1 - \tau.$$

The authors analyzed the asymptotic behavior of $\mathbb{E}\{d_{\mu_f}(\hat{L}(\tau), L(\tau))\}$ under SW1, SW2 and SW3. Before detailing the conclusions, it is necessary to introduce some convenient notation. Let ϕ and Φ denote standard normal distribution function and density function, respectively, and write $R(K) = \int K^2(t)\, dt$. Define the quantities:

$$D_1 = \frac{1}{2}\mu_2(K)\left\{\sum_{j=1}^{2r}\frac{1}{\mid f^{'}(x_j)\mid}\right\}^{-1}\left[\sum_{j=1}^{2r}\frac{f^{''}(x_j)}{\mid f^{'}(x_j)\mid} + \frac{1}{f_\tau}\sum_{j=1}^{r}\{f^{'}(x_{2j}) - f^{'}(x_{2j-1})\}\right],$$

$$D_2 = R(K)f_\tau\left\{\sum_{j=1}^{2r}\frac{1}{\mid f^{'}(x_j)\mid}\right\}^{-2}\sum_{j=1}^{2r}\frac{1}{f^{'}(x_j)^2}$$

and

$$D_{3,j} = \frac{R(K)f_\tau}{\mid f^{'}(x_j)\mid}\left\{\sum_{k=1}^{2r}\frac{1}{\mid f^{'}(x_k)\mid}\right\}^{-1},\ j = 1, 2, ..., 2r.$$

At this point, Theorem 2.2.1 can be established. It is fundamental to propose the new data-driven method for selecting the bandwidth parameter.

Theorem 2.2.1. *Under conditions SW1, SW2 and SW3, it is verified*

$$\mathbb{E}\{d_{\mu_f}(\hat{L}(\tau), L(\tau))\} = \sum_{j=1}^{2r}\left[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2}) - 1\}\right] + o\left(\frac{1}{(nh)^{1/2}} + h^2\right)$$

as $n \to \infty$, *uniformly for* $h \in [h^-, h^+]$ *where*

$$B_{1,j} = 2f_\tau\frac{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}{\mid f^{'}(x_j)\mid},\ B_{2,j} = \frac{\mid 1/2\mu_2(K)f^{''}(x_j) - D_1\mid}{\{R(K)f_\tau - 2D_{3,j} + D_2\}^{1/2}}$$

and

$$B_{3,j} = f_\tau\frac{\mid 1/2\mu_2(K)f^{''}(x_j) - D_1\mid}{\mid f^{'}(x_j)\mid}.$$

Numerical assessment of risk approximation. Theorem 2.2.1 guarantees the next approximation for the asymptotic risk:

$$\mathbb{E}\{d_{\mu_f}(\hat{L}(\tau), L(\tau))\} \simeq \sum_{j=1}^{2r}\left[\frac{B_{1,j}\phi(B_{2,j}n^{1/2}h^{5/2})}{(nh)^{1/2}} + B_{3,j}h^2\{2\Phi(B_{2,j}n^{1/2}h^{5/2}) - 1\}\right]. \quad (2.4)$$

The right-hand side of (2.4) can be useful for selecting the bandwidth parameter from $\mathcal{X}_n$. However, first, it is prudent to assess the quality of this approximation to the risk.

For a given f, h and τ the risk $\mathbb{E}\{d_{\mu_f}(\hat{L}(\tau), L(\tau))\}$ is very difficult to obtain exactly. Instead, we work with a Monte Carlo approximation

$$\frac{1}{M}\sum_{i=1}^{M} d_{\mu_f}(\hat{L}_i(\tau), L(\tau)), \tag{2.5}$$

where $\hat{L}_1(\tau)$,...,$\hat{L}_M(\tau)$ are M simulated realizations of $\hat{L}(\tau)$.

Figure 2.13 shows the asymptotic error (broken line) and the Monte Carlo approximation for the exact risk (solid line) for the density models 1, 4, 8 and 10, $M = 500$ samples of size $n = 1000$, $\tau = 0.5$ and $\tau = 0.8$. For $\tau = 0.5$, the asymptotic error approximates is quite good for model 1. Although the model 4 has a single mode, the asymptotic error is a bad approximation of the exact one determinated by Monte Carlo. Specifically, the problem seems to be caused for the very large values of $|f''|$ at the crossing points of the threshold $f_{0.5}$. This level is very close to the rapid transition from shallow to steep gradient. The same conclusions can be extracted for models 8 and 2. On the other hand and from the obtained results, it is not easy to determinate the influence of the parameter τ in the error approximation.

The expression (2.4) present a unique minimum under some conditions, see Corollary 2.2.2.

Corolary 2.2.2. *Under conditions SW1 and SW3 and assume further that SW1 it is verified for $r = 1$ and the underlying density f is symmetric about some point on the real line. Then there exists a unique $c_{opt} \in (0, \infty)$ depending on f and K but not on n, such that any sequence of bandwidths h_{opt} that minimizes $\mathbb{E}\{d_{\mu_f}(\hat{L}(\tau), L(\tau))\}$ satisfies*

$$h_{opt} = c_{opt} n^{-1/5}\{1 + o(1)\}$$

as $n \to \infty$.

Selection of the bandwidth. Corollary 2.2.2 gives the desired result in a restricted scenario; however, the result in fact holds much more widely. From this point, it is assumed that it exists an only bandwidth h_{opt} that minimizes the asymptotic risk and

$$h_{opt} = c_{opt} n^{-1/5}\{1 + o(1)\},$$

where $c_{opt} \in (0, \infty)$ minimizes the asymptotic risk

$$AR(c) = \frac{1}{n^{2/5}}\sum_{j=1}^{2r}\left[\frac{B_{1,j}}{c^{1/2}}\phi(B_{2,j}c^{5/2}) + B_{3,j}c^2\{2\Phi(B_{2,j}c^{5/2}) - 1\}\right].$$

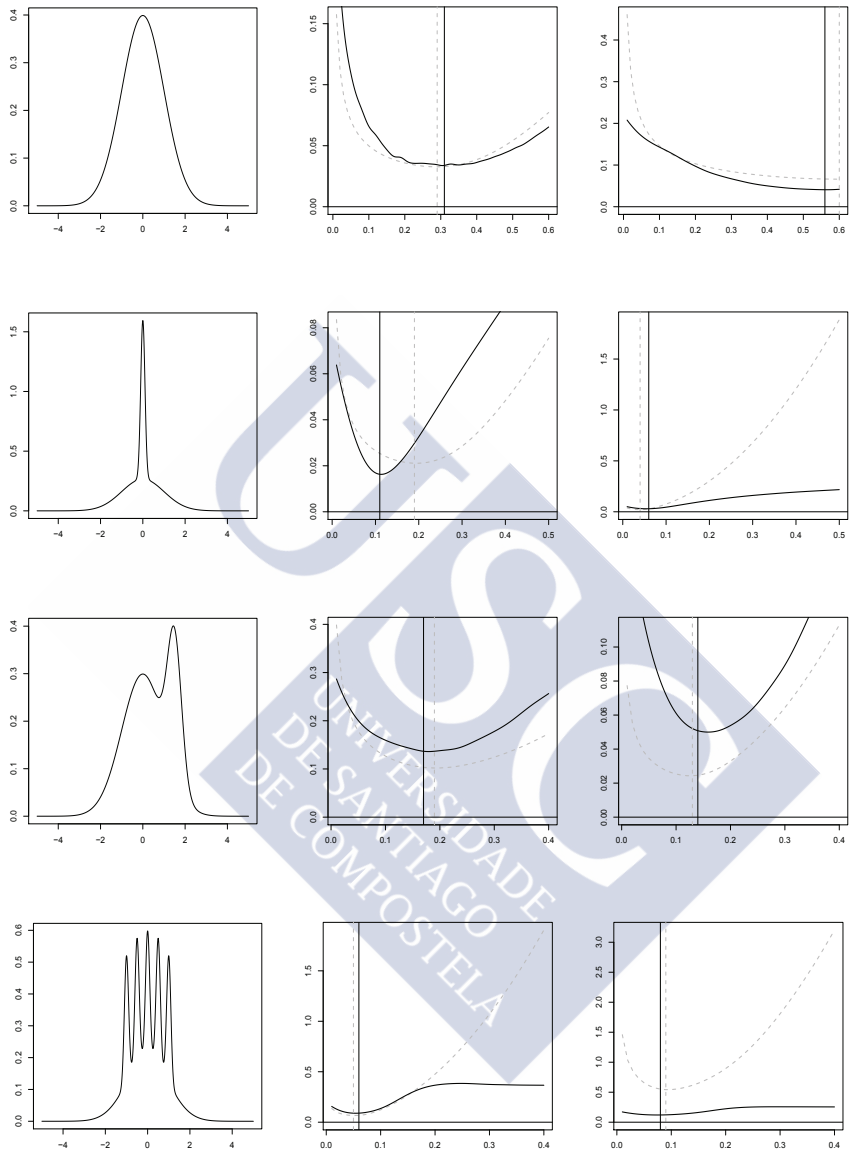


Figure 2.13: In the first column, the models 1 (first row), 4 (second row), 8 (third row) and 10 (fourth row) are showed. In the second columm, the asymptotic error (broken line) and the Monte Carlo approximation (solid line) with $n = 1000$, $M = 500$ and $\tau = 0.5$ are represented . In the third columm, the asymptotic error (broken line) and the Monte Carlo approximation (solid line) with $n = 1000$, $M = 500$ and $\tau = 0.8$.

In practise, it is necessary to obtain an estimator $\hat{c}_{opt}$ of c_{opt} in order to find a bandwidth selector. The most natural way consists in considering the estimators $\hat{D}_1$, $\hat{D}_2$ and $\hat{D}_{3,j}$ of D_1, D_2 and $D_{3,j}$, respectively in order to obtain plug-in estimators for $\hat{B}_{1,j}$, $\hat{B}_{2,j}$ and $\hat{B}_{3,j}$ of $B_{1,j}$, $B_{2,j}$ and $B_{3,j}$. Then,

$$\hat{c}_{opt} = \arg \min_{c\in(0,\infty)} \hat{AR}_n(c),$$

where

$$\hat{AR}_n(c) = \frac{1}{n^{2/5}} \sum_{j=1}^{2r} \left[\frac{\hat{B}_{1,j}}{c^{1/2}} \phi(\hat{B}_{2,j} c^{5/2}) + \hat{B}_{3,j} c^2 \{2\Phi(\hat{B}_{2,j} c^{5/2}) - 1\} \right].$$

This minimization problem has a unique solution under conditions of Corollary 2.2.2, with probability one and for n large enough. This solution could be approximated numerically. So, the final bandwidth would be obtained as

$$\hat{h}_\tau = \hat{c}_{opt} n^{-1/5}.$$

Plug-in estimators of f_τ, $f^{'}(x_j)$ and $f^{''}(x_j)$ for $j = 1, ..., 2r$ must be considered in order to construct the estimators $\hat{D}_1$, $\hat{D}_2$ and $\hat{D}_{3,j}$. If K is smooth then kernel estimators can be constructed for f and for $f^{'}$ and $f^{''}$ too. Concretely, the necessary estimations are $f_{n,\hat{h}_0}(\hat{x}_{j,\hat{h}_0})$, $f^{'}_{n,\hat{h}_1}(\hat{x}_{j,\hat{h}_0})$ and $f^{''}_{n,\hat{h}_2}(\hat{x}_{j,\hat{h}_0})$ of f_τ, $f^{'}(x_j)$ and $f^{''}(x_j)$, respectively, where $\hat{x}_{j,h_0}$ is an estimator of x_j and $\hat{h}_i$, $i = 0, 1, 2$ denote the different bandwidths for estimating f and their derivaties, $f^{'}$ and $f^{''}$.

Estimation of x_j, $j = 1, ..., 2r$ is easy. Once f is reconstructed nonparametrically by using f_{n,h_0} and fixed j, $\hat{x}_{j,h_0}$ can be calculated by solving the equation $f_{n,h_0} - \hat{f}_\tau$, where $\hat{f}_\tau$ denotes the estimator of f_τ calculated from $\mathcal{X}_n$.

Theorem 2.2.3 guarantees that the previous procedure is consistent.

Theorem 2.2.3. *Under regularity conditions in Theorem 3 by Samworth and Wand (2010), assume that c_{opt} is unique and that $AR^{''}(c_{opt}) > 0$. Then,*

$$\frac{\hat{h}_\tau}{h_{opt}} = 1 + O_P(n^{-2/9})$$

when $n \to \infty$. Moreover, recalling that $\hat{h}_\tau = \hat{c}_{opt} n^{-1/5}$, it is verified that

$$\frac{\hat{AR}_n(\hat{c}_{opt})}{AR(c_{opt})} = 1 + O_P(n^{-2/9}).$$

Bandwidth selector.

In practise, pilot bandwidths h_0, h_1 and h_2 are estimated using direct plug-in strategies with two levels of kernel functional estimation, see Samworth and Wand (2010) for precise details about their estimation. Next, the full algorithm is detailed for $K = \phi$:

1. The inputs are the sample $\mathcal{X}_n = \{X_1, ..., X_n\}$ and the parameter $0 < \tau < 1$.
2. Calculate the direct plug-in bandwidths $\hat{h}_0$, $\hat{h}_1$ and $\hat{h}_2$ for estimating f, $f^{'}$ and $f^{''}$, respectively, via Gaussian kernel.
3. Use the nonparametric estimation of f with bandwidth $\hat{h}_0$ for estimating f_τ, r and x_i, $i = 1, ..., 2r$.
4. Obtain the estimators $\hat{B}_{1,j}$, $\hat{B}_{2,j}$ and $\hat{B}_{3,j}$.
5. The selected bandwidth for estimating the density f with Gaussian kernel is

$$\hat{h}_\tau = \hat{c}_{opt} n^{-1/5}$$

with

$$\hat{c}_{opt} = \arg \min_{c \in (0,\infty)} \hat{AR}_n(c).$$

2.2.1.3 Modified Samworth and Wand's method

It has been detected that the algorithm proposed by Samworth and Wand (2010) can provide level set estimators equal to the empty set, mainly for large values of τ. The

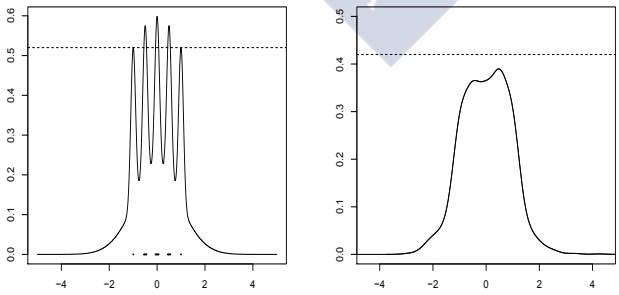


Figure 2.14: It is represented the level set for model 10 with $\tau = 0.8$ (left) and the kernel estimator for a sample of size $n = 1600$ with bandwidth $\hat{h}_\tau$ by estimating the threshold with $\hat{h}_0$ for $\tau = 0.8$ (right).

reason for this is simple. The estimation of the pilot bandwidth $\hat{h}_0$ that is used to estimate the threshold f_τ can be considerably greater than the final smoothing parameter $\hat{h}_\tau$ obtained by the original Samworth and Wand's algorithm. Then, the kernel density estimator calculated from $\hat{h}_\tau$ does not intersect $\hat{f}_\tau$, see Figure 2.14. To solve this problem, the threshold must be recalculated from $\hat{h}_\tau$.

According to the previous comments, adding a new step to the original method proposed in Samworth and Wand (2010) is necessary:

6. Use the nonparametric estimation of f with bandwidth $\hat{h}_\tau$ for estimating f_τ, r and x_i, $i = 1, ..., 2r$.

2.2.1.4 Singh, Scott and Nowak's method

Singh et al. (2009) presented a plug-in procedure for reconstructing density level sets that is based on an empirical density estimator, the regular histogram. This method considers a collection of cells A as a regular partition of $[0,1]^d$ into hypercubes $\mathcal{A}_j$ of dyadic sidelength 2^{-j}, where j is a nonnegative integer. The estimator of the level set $G(t)$, for a given value of $t > 0$, at a resolution level of j is defined as

$$\hat{G}(t) = \bigcup_{\{A \in \mathcal{A}_j : f_{n,H}(A) \geqslant t\}} A \text{ with } f_{n,H}(A) = \frac{\mathbb{P}_n(A)}{\mu(A)},$$

where $\mathbb{P}_n$ denotes the empirical probability induced by $\mathcal{X}_n$ and μ is the Lebesgue measure. $L(\tau)$ can be estimated by $\hat{G}(\hat{f}_\tau)$ where, to avoid the problem of bandwidth selection, $\hat{f}_\tau$ is computed using the empirical procedure proposed by Walther (1997). That is,

$$\hat{f}_\tau = \max\{t > 0 : \mathbb{P}_n(\hat{G}(t)) \geqslant 1 - \tau\},$$

where $\hat{G}(t)$ denotes the Singh, Scott and Nowak's estimator for $G(t)$. In this way, smoothing the sample data for estimating f_τ is avoided. The consideration of $[0,1]^d$ is not a real restriction because applications like translations or homothecies can be used.

The algorithm suggested in Singh et al. (2009) depends on the resolution level denoted by j. This parameter is selected using a data-driven procedure. The histogram resolution search is focused on regular partitions of dyadic sidelength 2^{-j}, $j \in \{0, 1, ..., J\}$. The choice of J is completely specified for the authors. Since the selected resolution needs to be adapted to the local regularity of the density around the level of interest, it is introduced:

$$\hat{\mathcal{V}}_{\hat{f}_\tau, j} = \min_{A \in A_j} \max_{A' \in A_{j'} \cap A} |\hat{f}_\tau - f_{n,H}(A')|$$

where $j^{'} = \lfloor j + \log_2(s_n) \rfloor$ and s_n is a slowly diverging monotone sequence, for example, $\log n$ or $\log\log n$ and $A_{j'} \cup A$ denotes the collection of subcells with sidelength $2^{-j^{'}} \in [2^{-j}/s_n, 2^{-j+1}/s_n)$ within the cell A. The empirical vernier $\hat{\mathcal{V}}_{\hat{f}_\tau, j}$ is balanced by a penalty term

$$\Psi_{j'} = \max A \in A_{j'} \sqrt{\frac{8\log(2^{j'(d+1)}16/\delta)}{n\mu(A)} \max f_{n,H}(A), \frac{8\log(2^{j'(d+1)}16/\delta)}{n\mu(A)}}$$

where $0 < \delta < 1$ is a confidence parameter, and $\mu(A) = 2^{-}j'd$. Notice that the penalty is computable from the given observations. Then,

$$\hat{j} = arg \min_{0 \leqslant j \leqslant J} \left\{ \hat{\mathcal{V}}_{\hat{f}_\tau, j+\Psi_{j'}} \right\}.$$

Observe that the value of the vernier decreases with increasing resolution as better approximations to the true level are available. On the other hand, the penalty is designed to increase with resolution to penalize high complexity estimates that might overfit the given sample of data. Thus, the above procedure chooses the appropriate resolution automatically by balancing these two terms. In Singh et al. (2009), it was proved that near-minimax optimal rates of convergence for a specific class of level sets are achieved for the resulting density level set estimator.

2.2.2 Excess mass methodology

Another possibility is to assume that the set of interest satisfies some geometric condition, such as convexity. In this case, the excess mass approach, first proposed by Hartigan (1987) and Müller and Sawitzki (1987), provides an alternative for the reconstruction of density level sets. Some previous contributions can be seen in Chernoff (1964) or Eddy and Hartigan (1977).

This group of algorithms utilizes the fact that the density level set $G(t)$, for a given value of $t > 0$, maximizes the functional

$$H_t(B) = \mathbb{P}(B) - t\mu(B),$$

where B is a Borel set, $\mathbb{P}$ denotes the probability measure induced by f and μ is the Lebesgue measure. Then, if $\mathcal{B}$ is a given class of sets, a natural estimator $\hat{G}(t)$ of $G(t)$ under the shape restriction $G(t) \in \mathcal{B}$ would be the maximizer, on $\mathcal{B}$, of the empirical excess mass

$$H_{t,n}(B) = \mathbb{P}_n(B) - t\mu(B),$$

where $\mathbb{P}_n$ denotes the empirical probability induced by the sample $\mathcal{X}_n$. This method incorporates geometrical conditions in a natural way on the estimator. If no shape

restriction is assumed the maximizer of $H_{t,n}$ would be the sample $\mathcal{X}_n$. Some interesting works can be found in literature. For instance, Hartigan (1987) and Grübel (1988) considered the case where $\mathcal{B}$ is the class of convex sets in the two and one-dimensional cases, respectively. Nolan (1991) proved the uniform convergence of the empirical excess mass functional to the theoretical one for the class of all closed ellipsoids in $\mathbb{R}^d$. Asymptotic results for the estimator for more general classes $\mathcal{B}$ were given in Polonik (1995).

If the goal is to reconstruct density level sets with a fixed probability content we can follow, again, the empirical procedure proposed by Walther (1997), $\hat{L}(\tau) = \hat{G}(\hat{f}_\tau)$, where

$$\hat{f}_\tau = \max\{t > 0 : \mathbb{P}_n(\hat{G}(t)) \geqslant 1 - \tau\}.$$

This methodology has been widely studied in the literature. Müller and Sawitzki (1991) proposed an efficient algorithm for estimating one-dimensional sets by assuming that the theoretical level set can be written as a finite union of M closed intervals. This algorithm assumes that M is known a priori. So, non convex sets could be estimated if $M > 1$. In Figure 2.15, one-dimensional levels sets are showed. They can be formed by a unique interval or the union or several ones. Next, Müller and Sawitzki's method is presented in a detail way in Section 2.2.2.1.

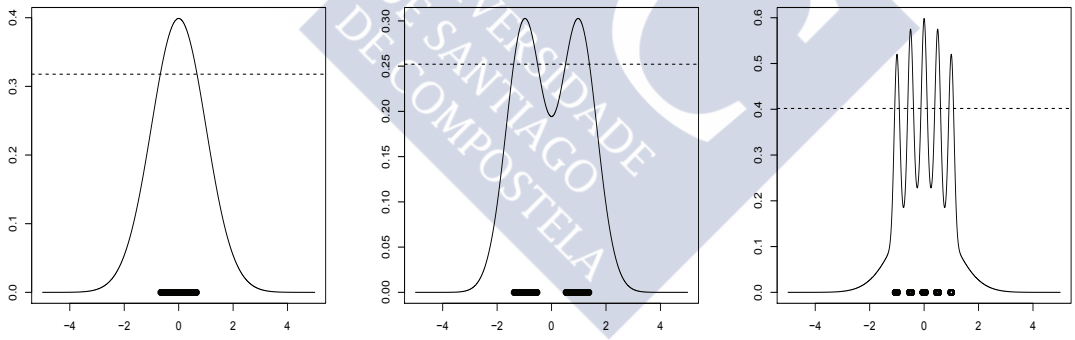


Figure 2.15: In each column, the level sets for models 1 (first column), 6 (second column) and 10 (third column) are represented for $\tau = 0.5$.

2.2.2.1 Müller and Sawitzki's method

Müller and Sawitzki (1991) studied the function $t \mapsto E(t)$ where $E(t) = \sup_B\{H_t(B)\}$ and the supreme is considered over the class of Borel sets for testing multimodality. From an analytical point of view, a usual definition relates a mode to a local maximum

of the density. However, with this idea, a distribution can have a mode at a point x while giving arbitrarily small probability to some neighborhood containing x. Müller and Sawitzki adopted a different definition: A mode is present where an excess of probability mass is concentrated, see Figure 2.16.

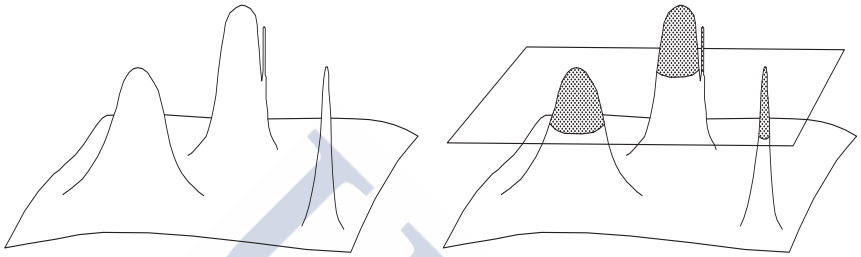


Figure 2.16: Modes in two-dimensional densities.

In this context, the connected components of $G(t)$ are called $t-$clusters. As t increases, the $t-$clusters concentrate on modes (local maxima of f). So, the usual notion of mode can be replaced by the concept of $t-$cluster.

If a density has exactly m $t-$clusters, the excess $E(t)$ can be expressed as

$$E(t) = \sup \sum_{j=1}^{m} \int_{C_j} (f(x) - t)\, dx,$$

where the supremum is taken over all families $\{C_j : j = 1, ..., m\}$ of pairwise disjoint connected sets. In general, it is defined

$$E_m(t) = \sup \sum_{j=1}^{m} \int_{C_j} (f(x) - t)\, dx.$$

So, it is possible to write:

$$E_m(t) = \sup \sum_{j=1}^{m} H_t(C_j).$$

From an empirical point of view, given a random sample $\mathcal{X}_n$, it is considered

$$E_{n,M}(t) = \sup \sum_{j=1,...,M} H_{t,n}(C_j) \tag{2.6}$$

where the parameter M is the maximum number of modes.

The test statistict proposed by Müller and Sawitzki for testing multimodality is defined as

$$\Delta_{n,M} = \max_t D_{n,M}(t)$$

where

$$D_{n,M}(t) = E_{n,M}(t) - E_{n,1}(t).$$

A large difference $D_{n,M}(t)$ indicates a violation of the hypothesis of unimodality.

Components C_j maximizing the sum in (2.6) will be called empirical $t-$clusters and denoted by empirical $t-$clusters. In the one-dimensional situation the empirical $t-$clusters are closed intervals with endpoints at data points, or empty. In the absence of flat parts of f they consistently estimate the real $t-$clusters, see Proposition 2 in Müller and Sawitzki (1991).

The Müller and Sawitzki's algorithm was designed for the one-dimensional case. It estimates the excess mass and it finds the empirical $t-$clusters. Therefore, it provides a method for estimating density level sets $G(t)$ when some information a priori about the number of modes of it is available. Then, $G(t)$ can be nonconvex.

2.2.3 Hybrid methodology

As the name suggests, hybrid methods assume a priori geometric restrictions on $L(\tau)$ and they also use a pilot nonparametric density estimator to define the sets $\mathcal{X}_{n,+}(\hat{f}_\tau) = \{X \in \mathcal{X}_n : f_n(X) \geqslant \hat{f}_\tau\}$ and $\mathcal{X}_{n,-}(\hat{f}_\tau) = \mathcal{X}_n \backslash \mathcal{X}_{n,+}(\hat{f}_\tau)$. In this chapter, two new hybrid methods to estimate convex and $r-$convex sets for some $r > 0$ are proposed. These two new algorithms are based on the convex hull and $r-$convex hull methods for estimating the support, see Korostelëv and Tsybakov (1993) and Rodríguez-Casal (2007), respectively. They are presented in Sections 2.2.3.1 and 2.2.3.2. Another classic hybrid method is the so-called the granulometric smoothing method, see Walther (1997). It assumes that the level set and its complementary are both $r-$convex. This method adapts the estimator for the support proposed by Devroye and Wise (1980) to the context of level set estimation. It is studied in depth in Section 2.2.3.3.

2.2.3.1 The convex hull method

The convex hull method estimate level sets $L(\tau)$ by assuming a priori that the level set is convex. This condition is not too restrictive for small values of τ. In Figure 2.17

several density models with convex level sets are showed. The convexity hypothesis is true even when the model is multimodal.

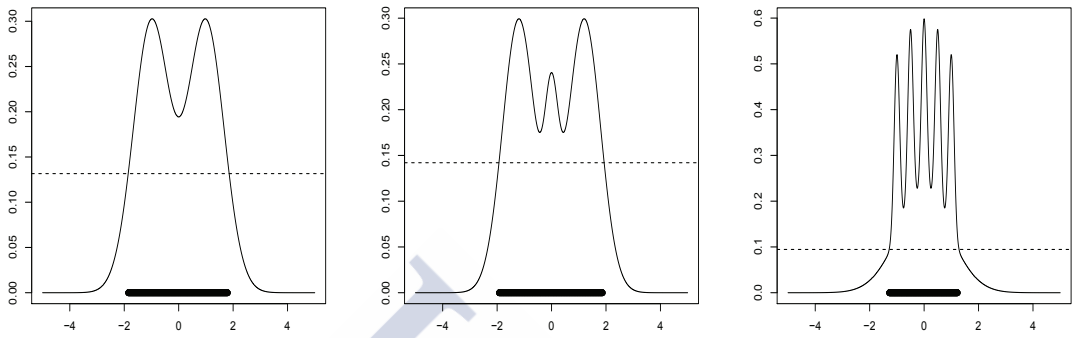


Figure 2.17: Level sets for models 6 (left), 9 (center) and 10 (right) with $\tau = 0.1$.

The estimator proposed adapts the ideas of convex hull for support estimation to this setting. Next, the steps of algorithm are detailed:

1. A nonparametric kernel estimator f_n is used for calculating the threshold estimator $\hat{f}_\tau$ by using the method by Hyndman (1996) or numerical integration process. Then, it is possible to define the set

$$\mathcal{X}_{n,+}(\hat{f}_\tau) = \{X \in \mathcal{X}_n : f_n(X) \geqslant \hat{f}_\tau\}.$$

2. The estimator of the level set $\hat{L}(\tau)$ is defined as the convex hull of $\mathcal{X}_{n,+}(\hat{f}_\tau)$, $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau))$. For dimension one, $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau)) = [\min(\mathcal{X}_{n,+}), \max(\mathcal{X}_{n,+})]$.

The threshold could be estimated by following the empirical process proposed in Walther (1997). Therefore, $\hat{f}_\tau = \max\{t : \mathbb{P}_n(\hat{G}(t)) \geqslant 1-\tau\}$ where $\hat{G}(t) = \text{conv}(\mathcal{X}_{n,+}(t))$.

For the data set presented in Section 1.1, corresponding to 322 cases of diagnosed of leukaemia on the North West of England, this estimator has been constructed in Figures 2.18 and 2.19. Convexity can be useful in some cases, see Figure 2.18 but it may be very restrictive in the most of the situations, see Figure 2.19.

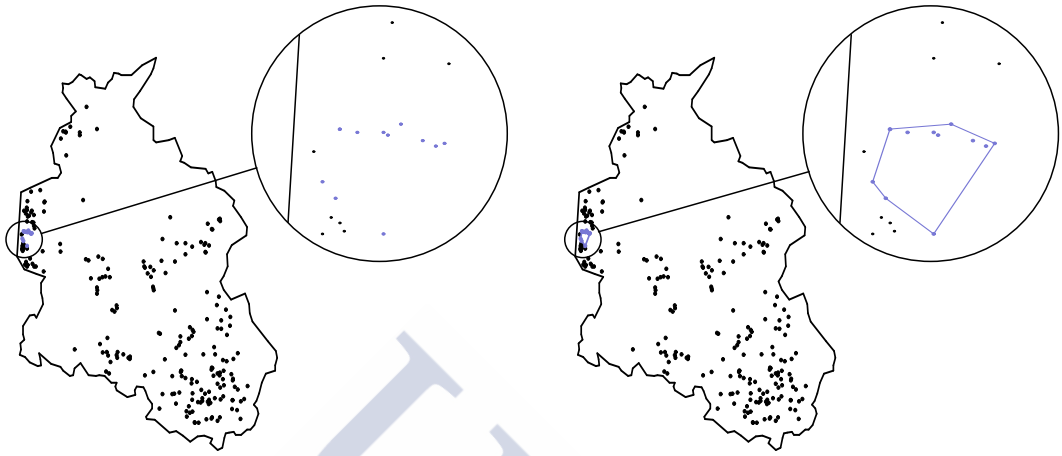


Figure 2.18: $\mathcal{X}_{n,+}(\hat{f}_{0.95})$ in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.95})$ in black (left) with $\mathcal{X}_n \subset \mathbb{R}^2$. $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_{0.95}))$ (right).

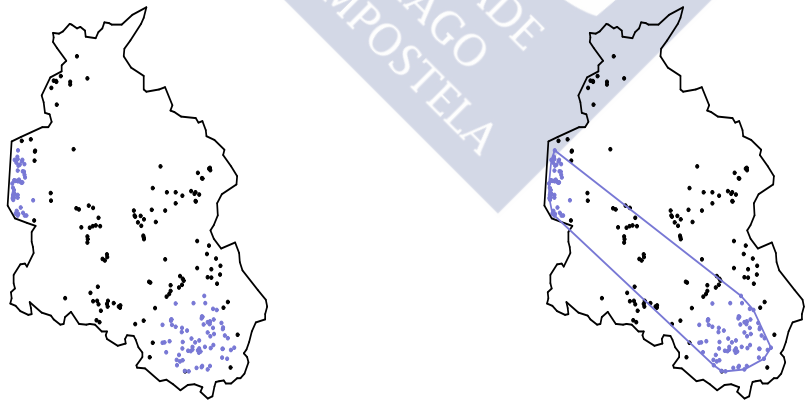


Figure 2.19: $\mathcal{X}_{n,+}(\hat{f}_{0.5})$ in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.5})$ in black (left) with $\mathcal{X}_n \subset \mathbb{R}^2$. $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_{0.5}))$ (right).

2.2.3.2 The $r-$convex hull method

According to the previous comments, convexity may be a very restrictive condition. The $r-$convex hull method estimates the level set $L(\tau)$ by assuming $r-$convexity as shape restriction, see Definition 1.2.8. Rodríguez-Casal (2007) studied by first time the $r-$convex hull as a support estimator. Next, this support estimator will be adapted for reconstructing level sets:

1. A nonparametric kernel estimator f_n is used for calculating the threshold estimator $\hat{f}_\tau$ by using the method by Hyndman (1996) or numerical integration process. It is possible to define the set

$$\mathcal{X}_{n,+}(\hat{f}_\tau) = \{X \in \mathcal{X}_n : f_n(X) \geqslant \hat{f}_\tau\}.$$

2. In this case, the estimator of the level set $\hat{L}(\tau)$ is defined as the $r-$convex hull of $\mathcal{X}_{n,+}(\hat{f}_\tau)$ with $r > 0$, $C_r(\mathcal{X}_{n,+}(\hat{f}_\tau))$.

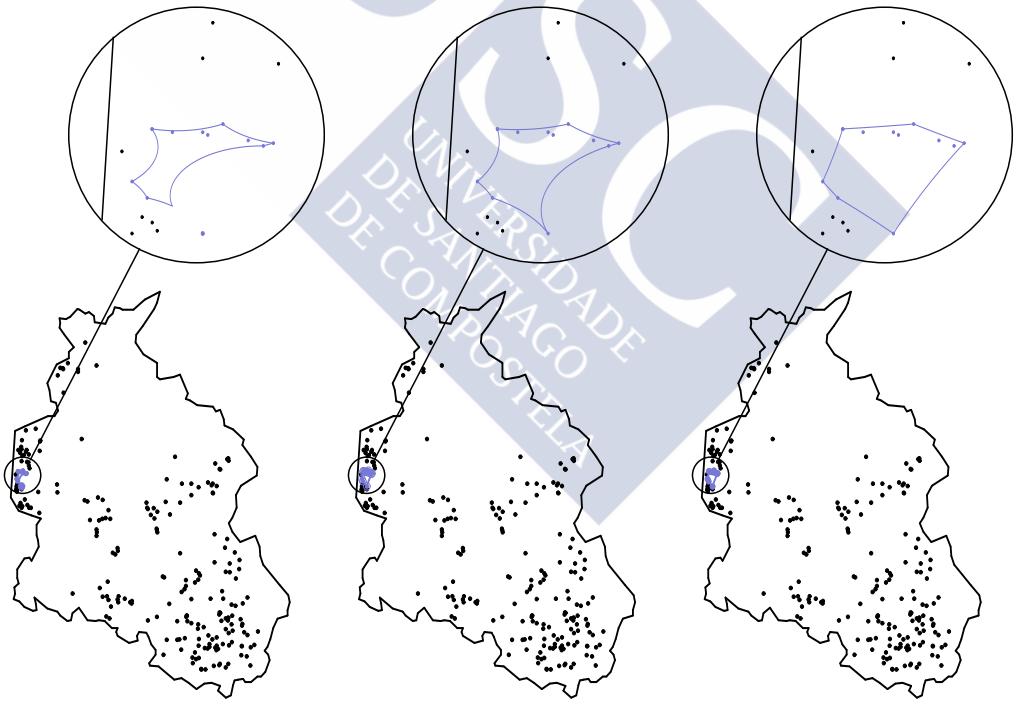


Figure 2.20: The set $\mathcal{X}_{n,+}(\hat{f}_{0.95})$ is represented in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.95})$, in black for $\mathcal{X}_n \subset \mathbb{R}^2$. $C_{0.02}(\mathcal{X}_{n,+}(\hat{f}_{0.95}))$ (center). $C_{0.03}(\mathcal{X}_{n,+}(\hat{f}_{0.95}))$ (center). $C_{0.3}(\mathcal{X}_{n,+}(\hat{f}_{0.95}))$ (right).

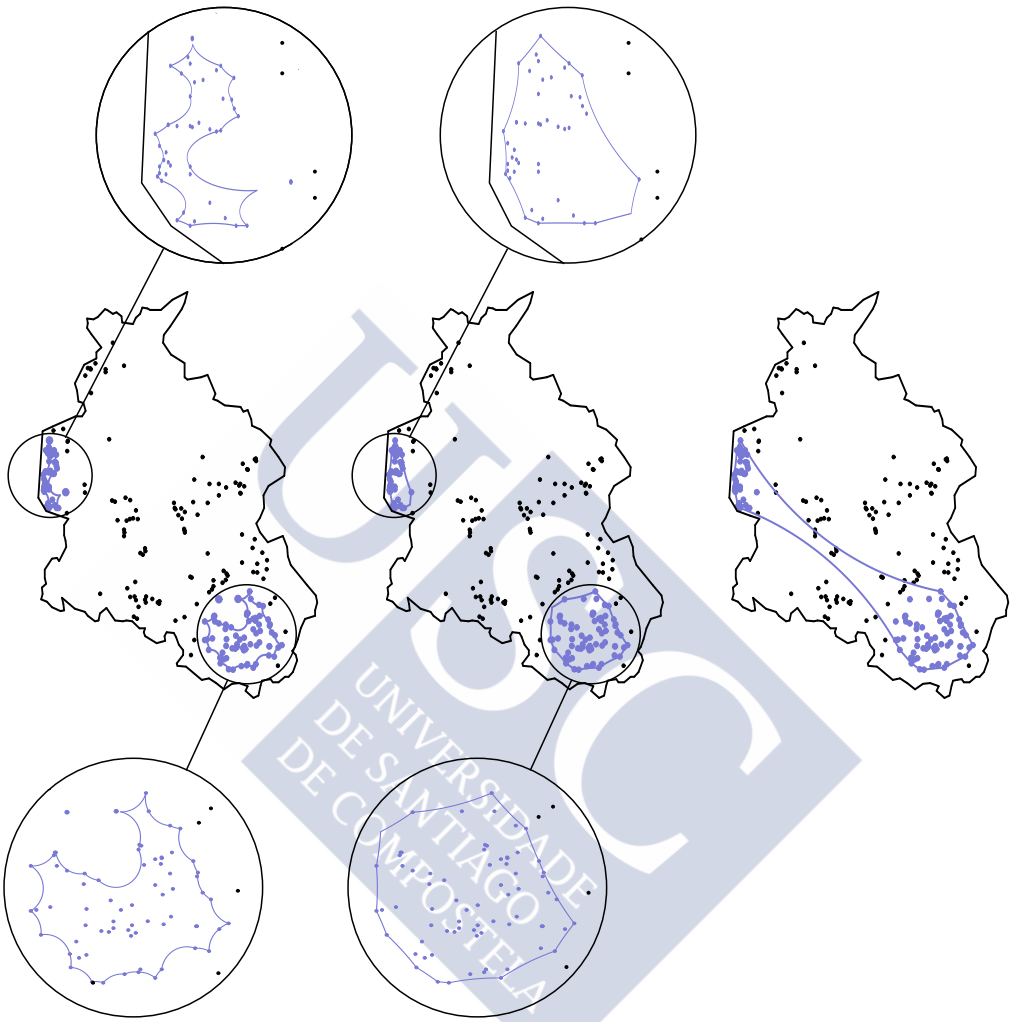


Figure 2.21: The set $\mathcal{X}_{n,+}(\hat{f}_{0.5})$ is represented in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.5})$, in black for $\mathcal{X}_n \subset \mathbb{R}^2$. $C_{0.03}(\mathcal{X}_{n,+}(\hat{f}_{0.5}))$ (center). $C_{0.3}(\mathcal{X}_{n,+}(\hat{f}_{0.5}))$ (center). $C_{0.9}(\mathcal{X}_{n,+}(\hat{f}_{0.5}))$ (right).

The threshold could be estimated by following the empirical process proposed in Walther (1997). Therefore, $\hat{f}_\tau = \max\{t : \mathbb{P}_n(\hat{G}(t)) \geqslant 1 - \tau\}$ where $\hat{G}(t) = C_r(\mathcal{X}_{n,+}(t))$.

The main disadvantage of this estimator is the dependence on the parameter $r > 0$. It is usually unknown. In Figures 2.20 and 2.21, this estimator has been constructed for the 322 cases of diagnosed of leukaemia on the North West of England presented in Section 1.1. If r is too small the estimator could be split; however, high values of

r provides estimators almost equal to the convex hull. In Figure 2.21, it can be seen that the estimator is very sensitive to the selection of r.

2.2.3.3 Granulometric smoothing method

The granulometric smoothing method is the classical hybrid. It was proposed by Walther (1997) and it is designed for estimating level sets $L(\tau)$. The estimator is a union of balls with radius r. So, it adapts the support estimator proposed by Devroye and Wise (1980).

This algorithm assumes as geometric restrictions that a ball of radius r rolls freely in the level set and in the closure of its complement, see Theorem 1.2.11. This is the case if the density f is smooth enough. Then, it is proposed the following algorithm:

1. A nonparametric kernel estimator f_n is used for calculating the threshold estimator $\hat{f}_\tau$ by using the method by Hyndman (1996) or numerical integration process. It is possible to define the sets

$$\mathcal{X}_{n,+}(\hat{f}_\tau) = \{X \in \mathcal{X}_n : f_n(X) \geqslant \hat{f}_\tau\},\ \mathcal{X}_{n,-}(t) = \mathcal{X}_n \backslash \mathcal{X}_{n,+}(t).$$

2. The estimators of $G(t)$ and $L(\tau)$ can be defined:

$$\hat{G}(t) = ((\mathcal{X}_{n,-} \oplus B_{r_n}[0])^c \cap \mathcal{X}_{n,+}) \oplus B_{r_n}[0],$$

$$\hat{L}(\tau) = \hat{G}(\hat{f}_\tau) \text{ where } \hat{f}_\tau = \max\{t : \mathbb{P}_n(\hat{G}(t)) \geqslant 1 - \tau\},$$

and r_n is sequence of smoothing parameters which represents the radius of balls rolling freely on the boundary of $L(\tau)$. In practise, for a fixed sample of size n, r_n is replaced by r.

More specifically, the estimator consists of the union of balls around those points in $\mathcal{X}_{n,+}$ that have a distance of at least r_n from each point in $\mathcal{X}_{n,-}$. So, unlike the previous hybrid methods, $\mathcal{X}_{n,+}$ is not necessarily contained in the estimator of $L(\tau)$.

The main disadvantage of this estimator is the dependence on the unknown parameter $r_n > 0$. Figure 2.23 shows this estimator for the data corresponding to 322 cases of diagnosed of leukaemia on the North West of England. Small values of r provide split estimators. However, if r is large enough this estimator could be equal to the empty set.

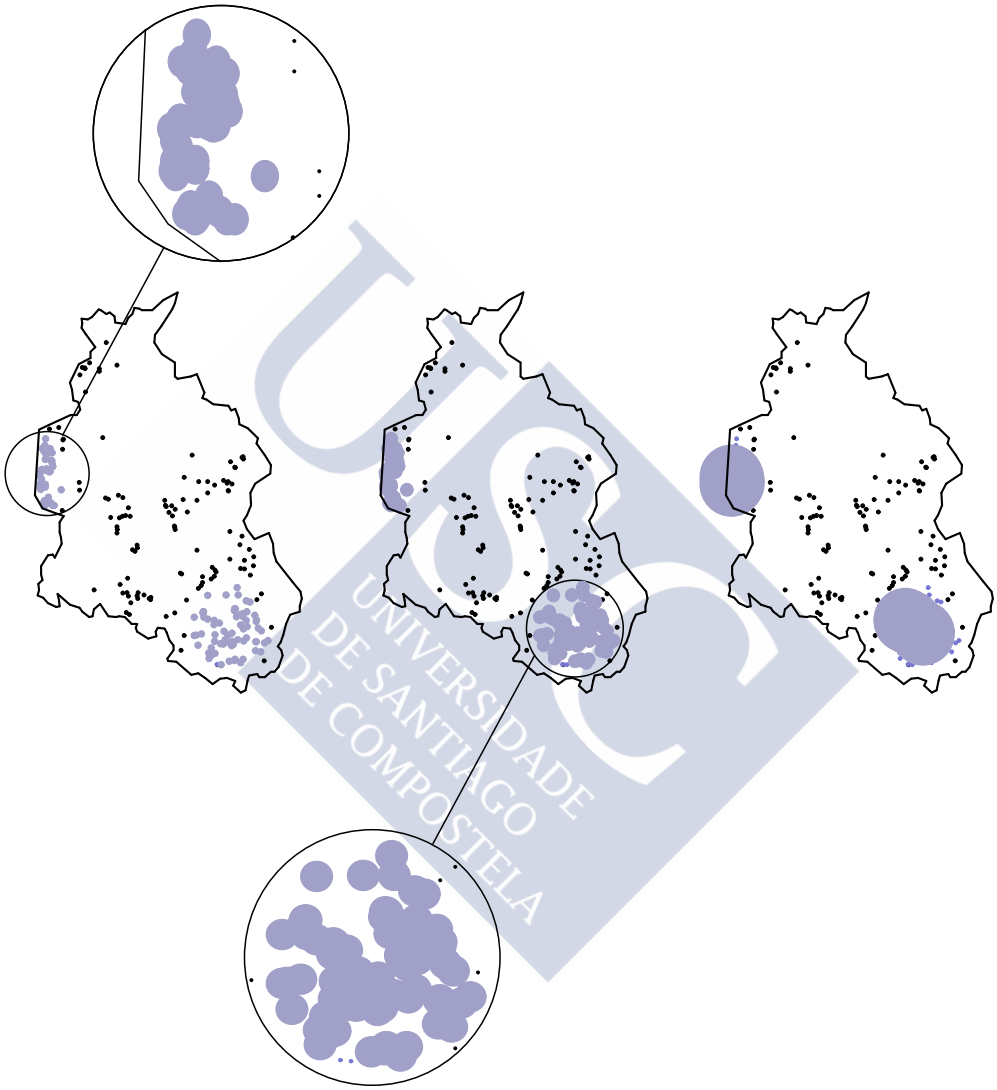


Figure 2.22: $\mathcal{X}_{n,+}(\hat{f}_{0.95})$ in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.95})$ in black for $\mathcal{X}_n \subset \mathbb{R}^2$. Walther's estimator for $r = 0.01$ (left). Walther's estimator for $r = 0.02$ (center). Walther's estimator for $r = 0.1$ is equal to the empty set (right).

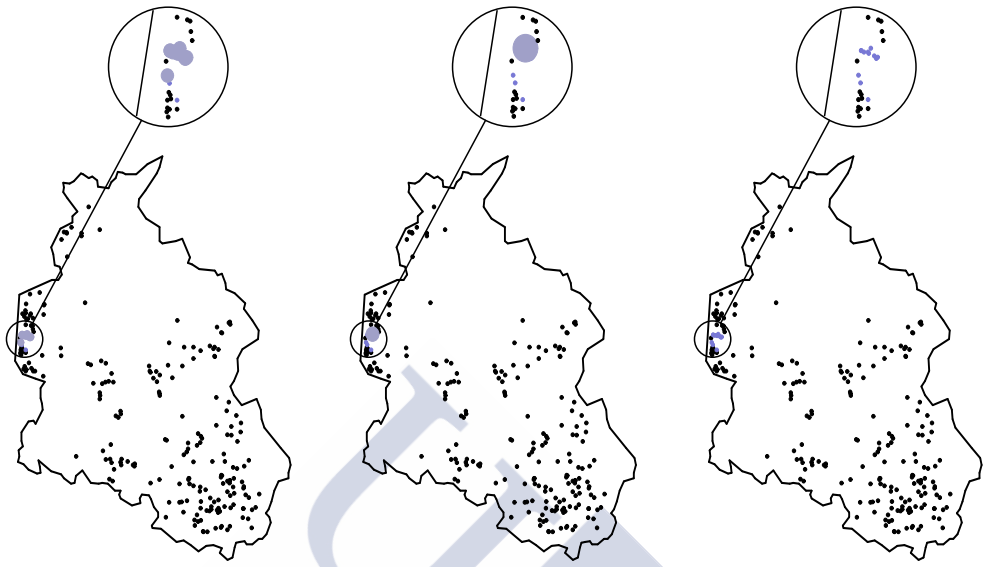


Figure 2.23: $\mathcal{X}_{n,+}(\hat{f}_{0.95})$ in blue and $\mathcal{X}_{n,-}(\hat{f}_{0.95})$ in black for $\mathcal{X}_n \subset \mathbb{R}^2$. Walther's estimator for $r = 0.01$ (left). Walther's estimator for $r = 0.02$ (center). Walther's estimator for $r = 0.03$ is equal to the empty set (right).

2.3 A comparative simulation study for density level sets

The existing data-driven methods for reconstructing density level sets presented in Section 2.2 will be compared through a detailed simulation study to analyze their performance. The notation for these algorithms in the following sections is showed in Table 2.1.

We focus on the one-dimensional setting in order to include some methods that do not have multidimensional counterparts, such as the plug-in algorithm proposed by Samworth and Wand (2010) or the excess mass method proposed by Müller and Sawitzki (1991). This comparison could be used, for instance, as a guide to identify which of these three methodologies is more promising in general dimension d. It could arise the need to generalize some one-dimensional methods to general dimension d.

We have generated 1000 samples of size n equal to 1600, 400 and 100 from 18 density models, see Section 1.4. Models from 1 to 15 are the densities proposed by Marron and Wand (1992). The models 16, 17 and 18 correspond to the marronite, caliper and matterhorn densities proposed in Berlinet and Devroye (1994). These models were added because they present some special properties. Specifically, they have two

asymmetric and separated modes, two jumps and a non finite peak, respectively.

Full name	Short name
Baíllo and Cuevas' method	BC
Samworth and Wand's method	SW
Sheather and Jones' method	SJ
Cross validation method	CV
Müller and Sawitzki's method with M modes	MS_M
Convex hull method	CH
$r-$convex hull method	CH_r
Walther's method with radius r	W_r

Table 2.1: Names for existing data-driven methods to be compared.

In addition, five values of M for the Müller y Sawitzki's method have been considered: $M = 1$, $M = 2$, $M = 3$, $M = 4$ and $M = 5$. Other five values of r were taken into account for Walther's method and $r-$convex hull method: $r_1 = 0.01$, $r_2 = 0.05$, $r_3 = 0.1$, $r_4 = 0.2$ and $r_5 = 0.3$.

Three values for the parameter τ have been considered too: $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$. The threshold f_τ was estimated for plug-in methods by using the two procedures described in Section 2.2.1. Although the comparison is not shown here, the two algorithms for estimating f_τ provide similar results. Therefore, in the following, only the results for Hyndman's method will be considered. The bandwidth given by Sheather and Jones has been used as the pilot selector to calculate f_n for the estimation of f_τ. Although it is possible to estimate the threshold by using Hyndman's method for Singh, Scott and Nowak's algorithm, Müller and Sawitzki's method and hybrid algorithms, we propose to imitate the empirical procedure studied in Walther (1997) in order to estimate the threshold which satisfies the fixed probability content. In this way, it is not necessary to use a kernel density estimator. We will show the results for the last one alternative. We have checked that these two algorithms offer very similar results.

For each fixed random sample and each method (plug-in, excess mass or hybrid), we have calculated the estimator $\hat{L}(\tau)$ of $L(\tau)$ and the errors in the estimation $d_\mu(\hat{L}(\tau) \triangle L(\tau))$, $d_{\mu_f}(\hat{L}(\tau) \triangle L(\tau))$ and $d_H(\hat{L}(\tau), L(\tau))$. Although the results are not showed here, the correlations between these three error criteria were analyzed using the Spearman coefficient. The error criteria d_μ and d_{μ_f} often allows to obtain the same conclusions except for the density models that take values close to zero in their domains or that

present considerable jumps, see models 3, 15 or 17. The d_H-errors must be analyzed separately. Thus, 1000 errors were obtained for each algorithm, model, value of τ and error criteria.

Sections 2.3.1, 2.3.2 and 2.3.3 show the comparison of plug-in methods, excess mass methods and hybrid methods, respectively. This chapter is closed with the comparison of most competitive algorithms in each methodology. Some interesting conclusions are exposed in Section 2.3.4.

2.3.1 Simulation results for plug-in methododology

In this section, we will compare Baíllo and Cuevas' (BC), Samworth and Wand's (SW), Sheather and Jones' (SJ) and cross validation (CV) methods. The first two one are specific bandwidth selectors to estimate level sets. The last two algorithms are general selectors to estimate density functions. For instance, the behaviour of these general methods was studied in the literature in Cao et al. (1994). Singh, Scott and Nowak's algorithm is not included in this comparison because, for the sample sizes considered in the study, its behavior was not satisfactory.

To facilitate the presentation of the results, the following figures are divided into rectangles of different colours according to the method (vertical axis) and density model (horizontal axis), where light colours correspond to small errors, and vice versa. This representation allows us to detect the most competitive algorithm for each fixed value of τ. Given a density, the empirical means of the 1000 errors have been ordered by testing whether the mean errors of the compared methods are equal. If the null hypothesis of equality between two methods is rejected, then each algorithm is painted by using a different colour (darker or lighter according to the mean error). Otherwise, both algorithms are represented using the same colour. This approach is used in the rest of this chapter.

Figure 2.24 shows the comparison for plug-in methods for $d_{\mu_f}-$errors, $n = 400$ and $\tau = 0.5$. Figures 2.25, 2.26 and 2.27 was elaborated for $d_{\mu_f}-$errors but, in this case, $n = 1600$ and $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$, respectively. Figures 2.28 and 2.29 show the results for a different error criteria, d_H, for $n = 1600$ with $\tau = 0.5$ and $\tau = 0.8$, respectively.

If the the error criteria is d_{μ_f} then the most competitive methods for selecting the bandwidth parameter are cross validation and Sheather and Jones, see Figures 2.24, 2.25, 2.26 and 2.27. In general, these algorithms provide similar results. However, the value of the parameter τ is very decisive. For $\tau = 0.8$, specific methods for choosing the smoothing parameter (BC and SW) provide the best results for the unimodal densities 1, 2, 3 and 5 (see Figure 2.27). They are very simple level sets. For lower values of τ, cross validation and Sheather and Jones present a better global behavior. If we want

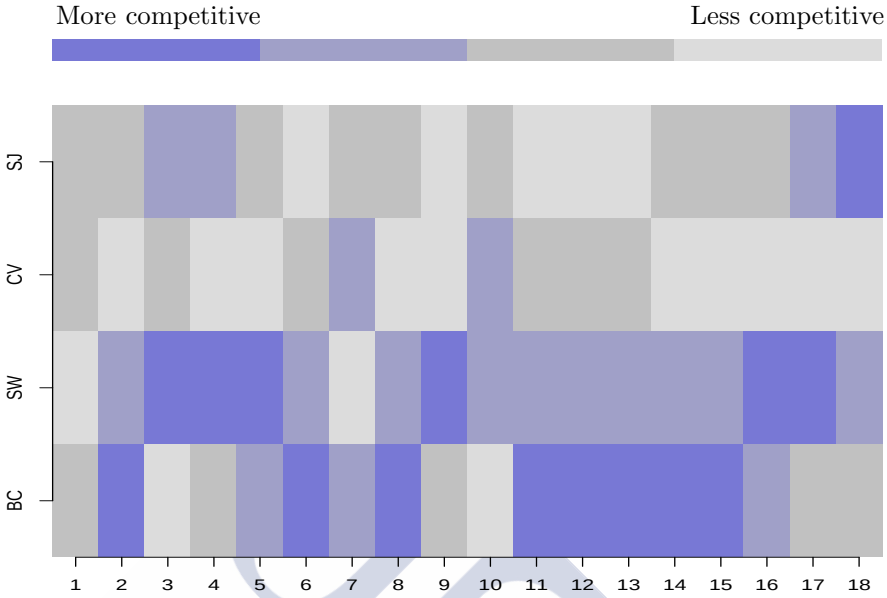


Figure 2.24: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 400$.

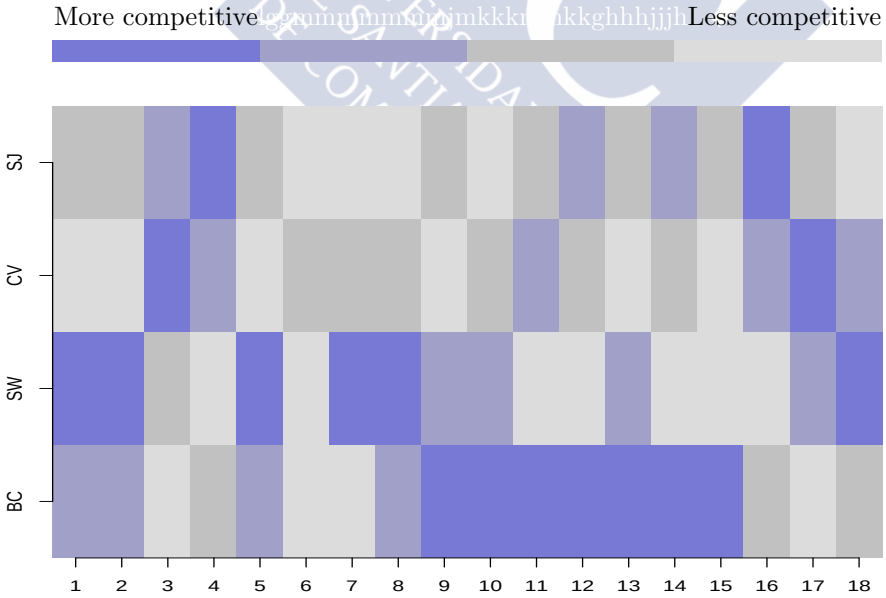


Figure 2.25: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.2$ and $n = 1600$.

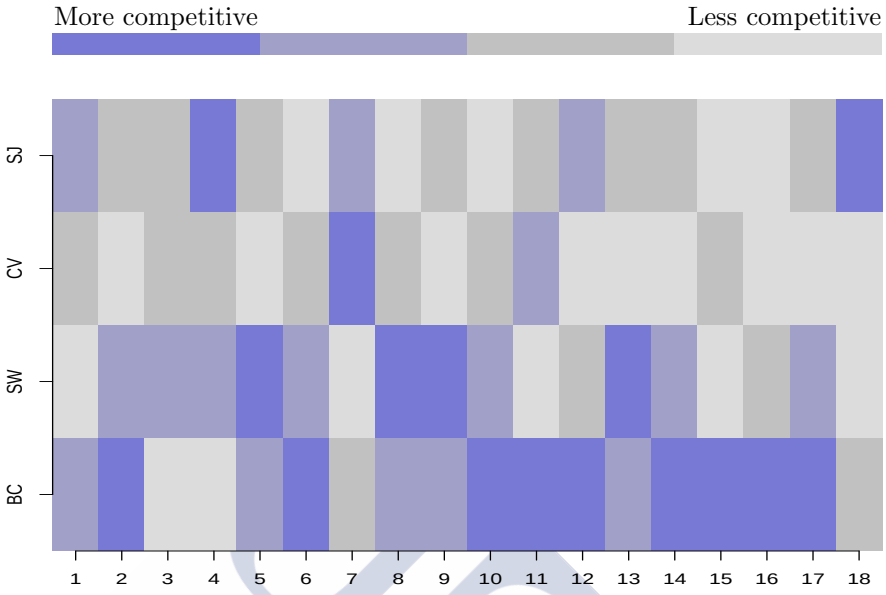


Figure 2.26: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

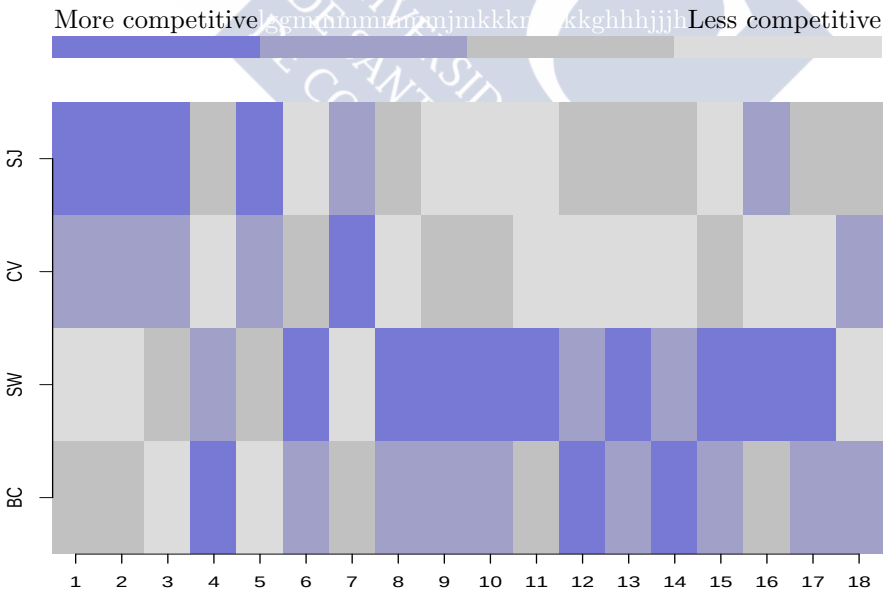


Figure 2.27: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.8$ and $n = 1600$.

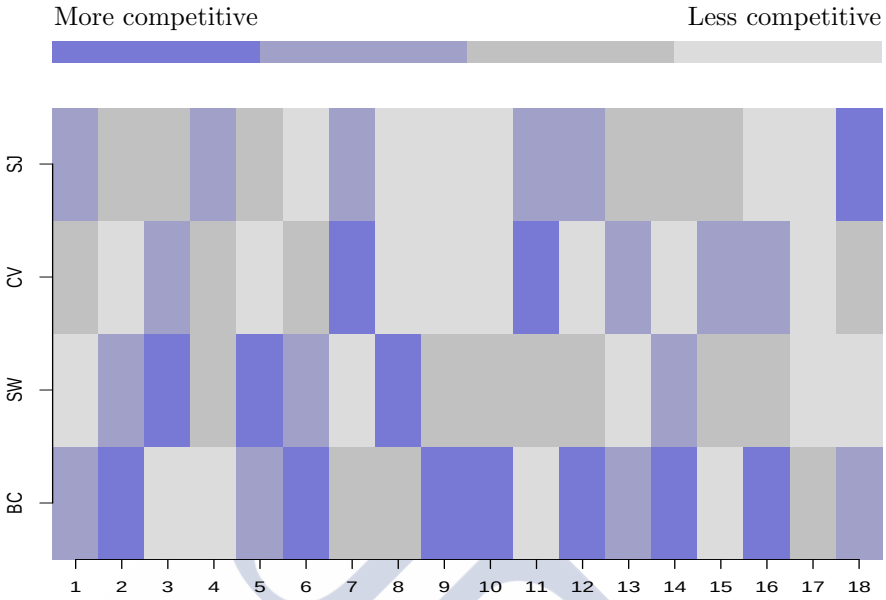


Figure 2.28: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_H as error criteria, $\tau = 0.5$ and $n = 1600$.

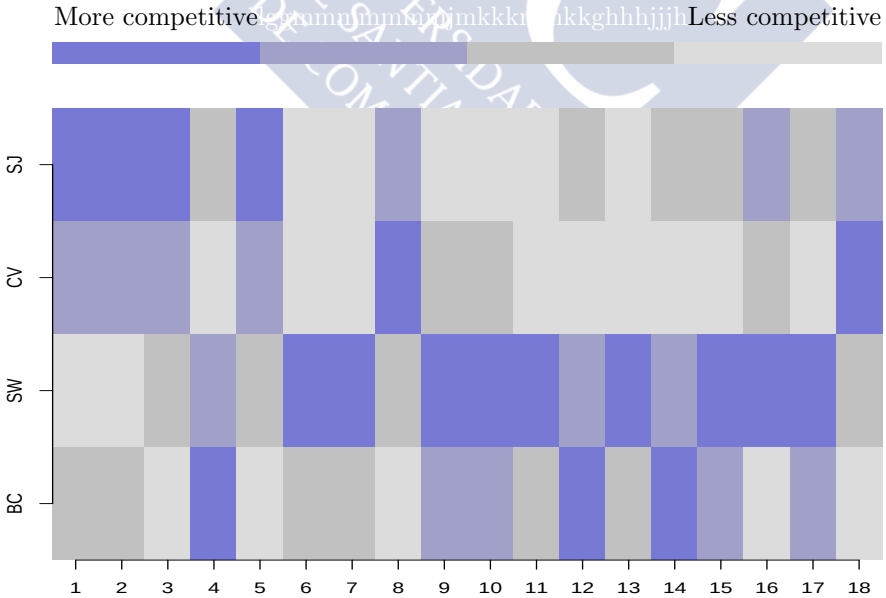


Figure 2.29: Comparison of plug-in methods (vertical axis) for the 18 model densities (horizontal axis) considering d_H as error criteria, $\tau = 0.8$ and $n = 1600$.

to study the effect of sample size then Figures 2.24 and 2.25 must be analyzed. Specific bandwidth selectors for density level sets have even a worse performance for $n = 400$.

The conclusions are similar if the error criteria d_μ is considered. Again, the most competitive algorithms are cross validation and Sheather and Jones. So, these results are not showed.

Figure 2.28 summarizes the results for d_H-errors for $n = 1600$ with $\tau = 0.5$. Baíllo y Cuevas' method is the less competitive for $\tau = 0.5$, see models 1, 2, 5, 6, 9, 10, 12, 13, 14 or 16. Samworth y Wand's methods presents bad results for simple densities with one or two modes, see models 2, 3, 5, 6 or 8. However, it improves its results if models from 9 until 18 are considered. Although specific density selectors have the best good global results, they are not the most competitive methods for complicated models like 11, 13, 15 or 16. Figure 2.29 shows the comparison for d_H-errors for $n = 1600$ with $\tau = 0.8$. In this case, Samworth y Wand's algorithm presents the worst behavior, see models 6, 7, 9, 10, 11, 13, 15, 16 and 17. However, specific level set estimation methods have the best results for densities with a unique mode like model 1, 2, 3, 5 or 8 and for the model 18 with a non finite peak. For these densities, cross validation and Sheather and Jones are not competitive

As a conclusion, specific methods to estimate level sets do not improve the results of the classic bandwidth selection rules. In addition, cross validation and Sheather and Jones methods often provide similar results and they present the best global behavior.

2.3.2 Simulation results for excess mass methododology

Müller and Sawitzki's method depends on an unknown parameter M. This is the main disadvantage of this algorithm. We have considered five values for the number of clusters, $M = 1,\ 2,\ 3,\ 4$ and 5 and we have denoted the Müller and Sawitzki's method with M modes by MS_M. Next, the influence of this parameter will be analyzed by using Figures 2.30 and 2.31. They have been elaborated with identical criterial considered in Section 2.3.1. In this case, different colors are assigned to the five values of M fixed ($M = 1$, $M = 2$, $M = 3$, $M = 4$ and $M = 5$ in the vertical axis) for each density model (horizontal axis). In addition, we have written the real number of modes for each density on the vertical axis too.

The influence of parameter M for $\tau = 0.2$ and $n = 1600$ can be analyzed from Figure 2.30. The models from 1 to 5 have one mode and $M = 1$ is the most competitive option. The same situation is repeated for models 8 and 18. The densities 7, 13 and 16 are bimodal. Again, $M = 2$ provides the best results. For models 14 and 15 with 6 modes, $M = 5$ presents the best results. It is the closest to the real number of modes. However, models like 11 or 12 have some non significant modes so the most competitive results

are provided by values of M smaller than the real one.

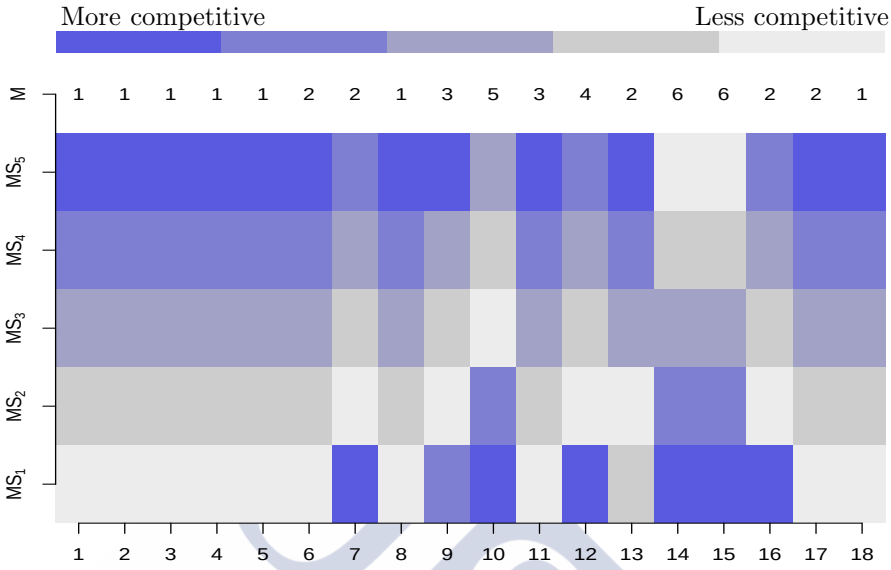


Figure 2.30: Comparison of Müller and Sawitzki's method for different values of M (vertical axis) and the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.2$ and $n = 1600$.

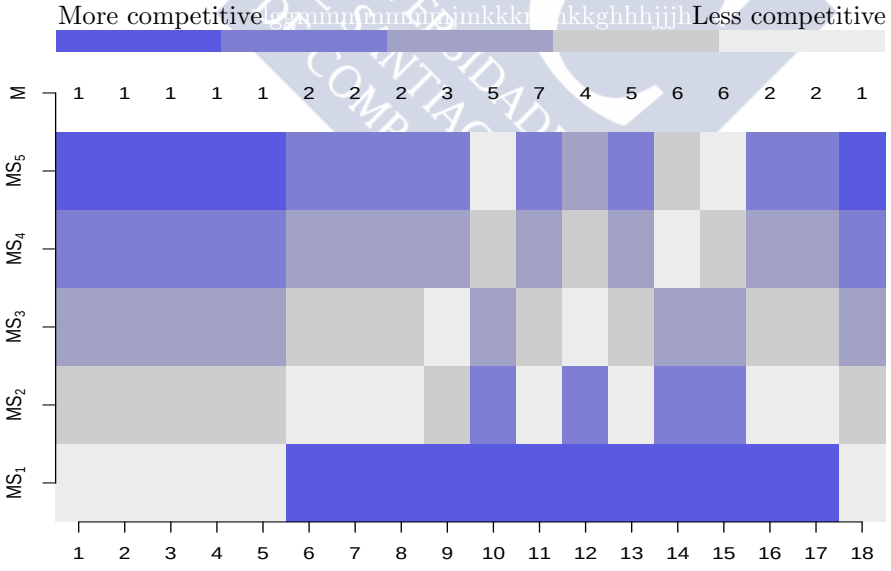


Figure 2.31: Comparison of Müller and Sawitzki's method for different values of M (vertical axis) and the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

We show the results for $\tau = 0.5$ and $n = 1600$ in Figure 2.31. The densities 1, 2, 3, 4 and 5 are unimodal and $M = 1$ provides the best results. Densities 6, 7, 8, 16 and 17 have two modes and, in this case, the best value of M is $M = 2$. The models 9 and 10 have three and five modes. In this case, $M = 3$ and $M = 5$ provide the best results, respectively. For model 15 with six modes, $M = 5$ is the most competitive alternative since that it is the closest value to the real number of modes. However, the best value of M for the Müller and Sawitzki's method is not equal to the real value of M for the models 11, 12, 13 and 14 because some of their modes are not significant. It can be seen in Figure 2.32 where the boxplots of $d_{\mu_f}-$errors are showed for some of these models.

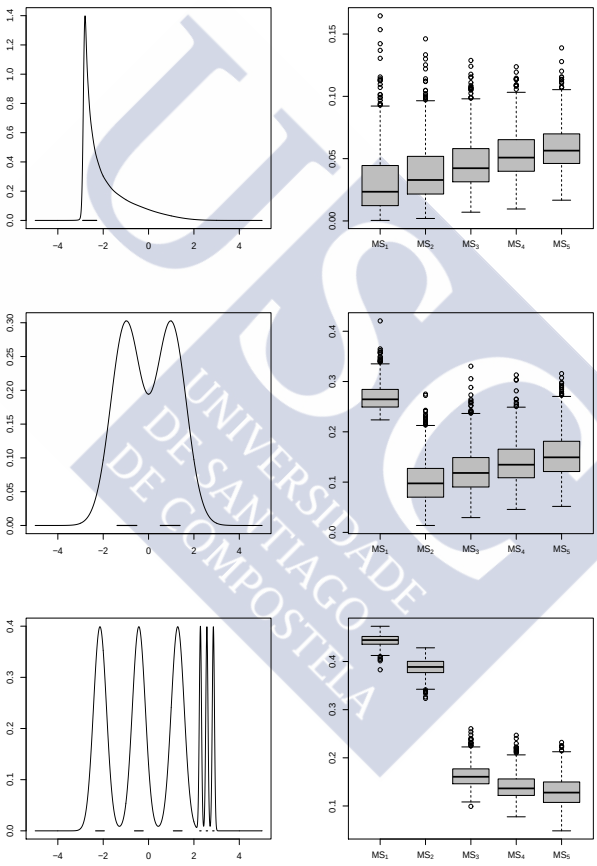


Figure 2.32: In each row, the boxplots for d_{μ_f} errors have been considered for the Müller and Sawitzki method with $M = 1$, $M = 2$, $M = 3$, $M = 4$ and $M = 5$ for densities 3 (first row), 6 (second row) and 15 (fourth row) with $n = 1600$ and $\tau = 0.5$.

Similar conclusions are obtained for $\tau = 0.8$. So, the results are not showed. It is clear that Müller and Sawitzki's method is very sensitive to the parameter M from Figures 2.30 and 2.31. In addition, if misspecification of M occurs it can be seen that big values of M are better than

a small values because the means of errors are lower, see for example model 7 in Figures 2.30 and 2.31. The conclusions for the rest of error criterias are similar.

2.3.3 Simulation results for hybrid methododology

Granulometric smoothing method and $r-$convex hull method depend on an unknown parameter r. This is the main disadvantage of these algorithms. In this work, we have considered five values for the radius of balls, r: $r_1 = 0.01$, $r_2 = 0.05$, $r_3 = 0.1$, $r_4 = 0.2$ and $r_5 = 0.3$. Next, we will study the influence of this parameter for these two algorithms and then, they and convex hull method will be compared.

2.3.3.1 Influence of parameter r for the $r-$convex hull method

According to the previous ideas, the $r-$convex hull algorithm depends on an unknown parameter r. From the five values considered, its influence will be studied in Figure 2.33 for $\tau = 0.5$ and $n = 1600$, in Figure 2.34 for $\tau = 0.8$ and $n = 1600$ and in Figure 2.35 for $\tau = 0.5$ and $n = 100$. The parameter r is represented on the vertical axis and the 18 density models on the horizontal axis.

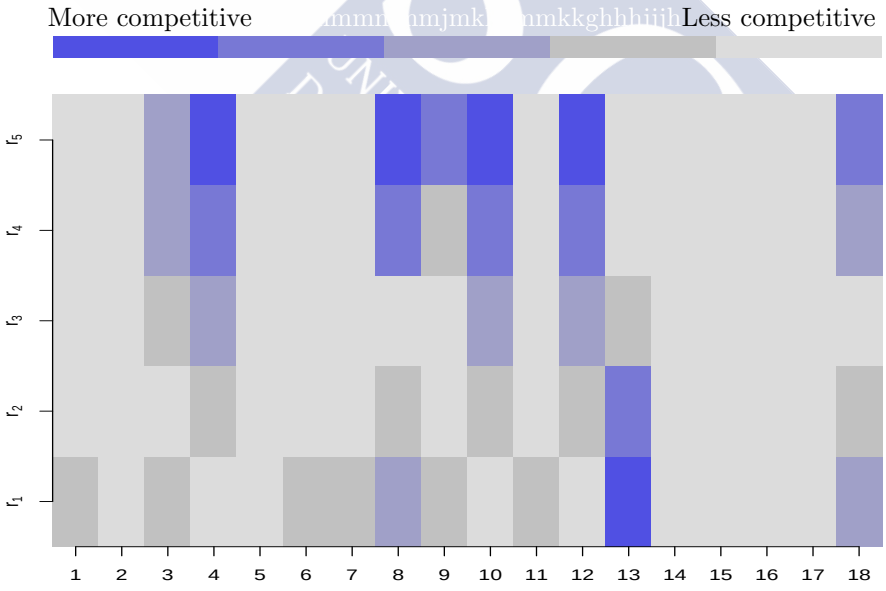


Figure 2.33: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

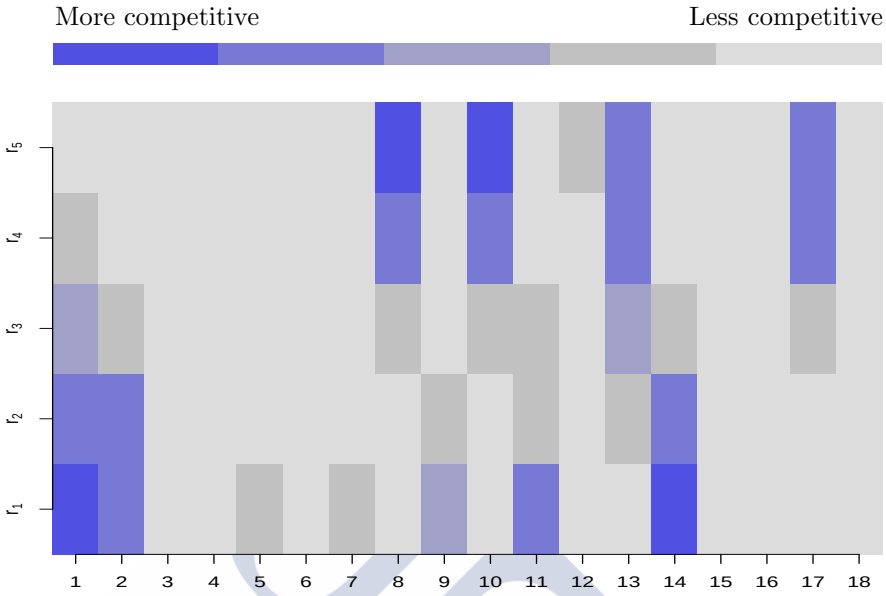


Figure 2.34: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.8$ and $n = 1600$.

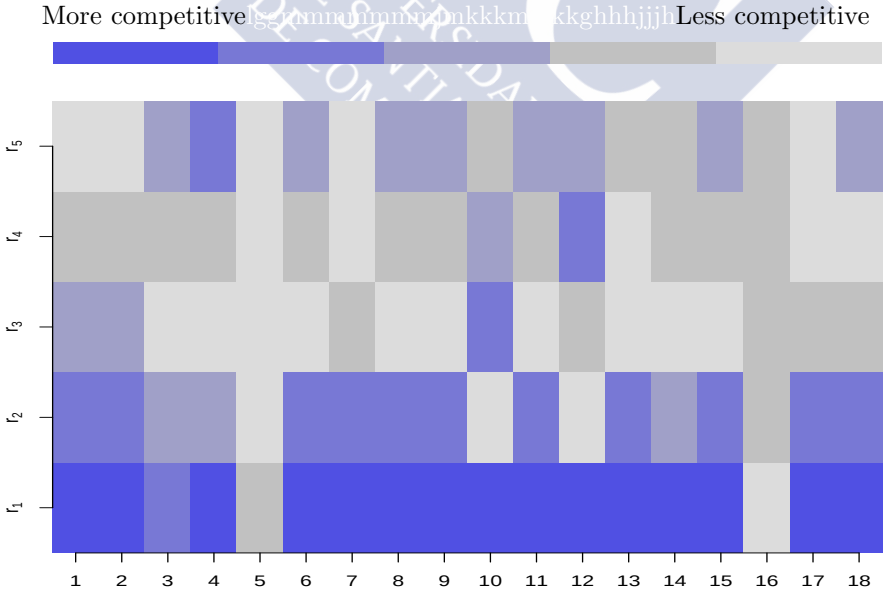


Figure 2.35: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 100$.

Figures 2.33 and 2.34 show the results for $n = 1600$ with $\tau = 0.5$ and $\tau = 0.8$, respectively. From this information, the $r-$ convex hull method depend on the parameter r in a weak way, see densities 1, 2, 5, 6, 7, 11, 14, 15, 16 and 17 for $\tau = 0.5$ and models 3, 4, 5, 6, 7, 9, 12, 15, 16 and 18 for $\tau = 0.8$. For $\tau = 0.2$, similar results are obtained. On the other hand, if low values of n are considered the importance of selecting r gets stronger, compare Figures 2.33 and 2.35. In the last one, we show the results for $n = 100$ and $\tau = 0.5$.

2.3.3.2 Influence of parameter r for the Walther's method

As we have told before, Walther's methods requires the specification of the parameter r which represents the radius of closed balls rolling freely in complementary of the level set. Figures 2.36, 2.37 and 2.38 will be used in order to analyze its influence. The five values of the parameter r are represented on the vertical axis and the 18 denity models on the horizontal axis.

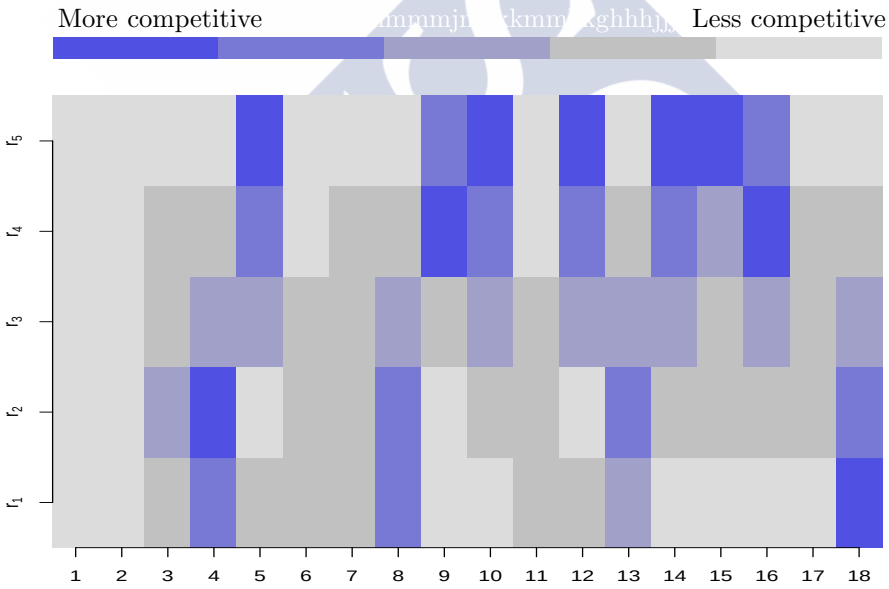


Figure 2.36: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

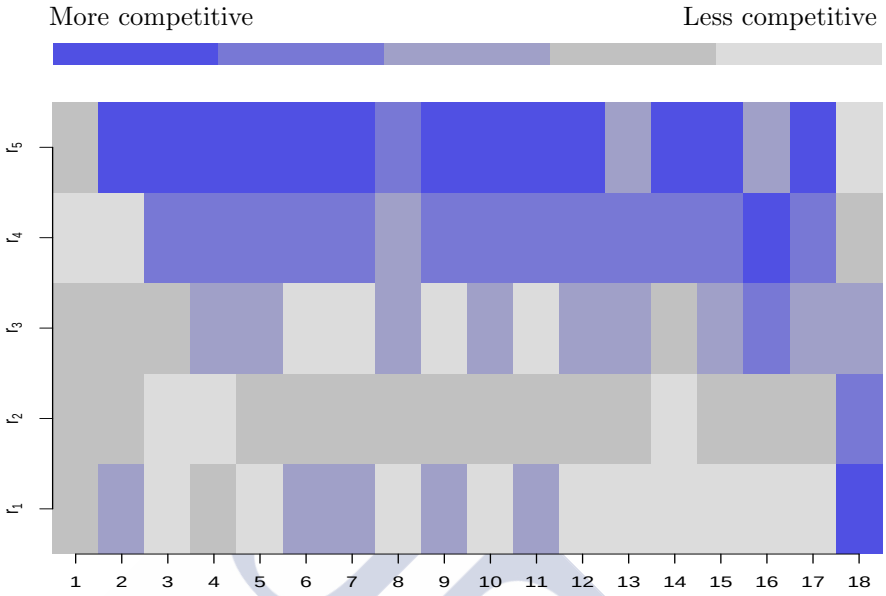


Figure 2.37: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.8$ and $n = 1600$.

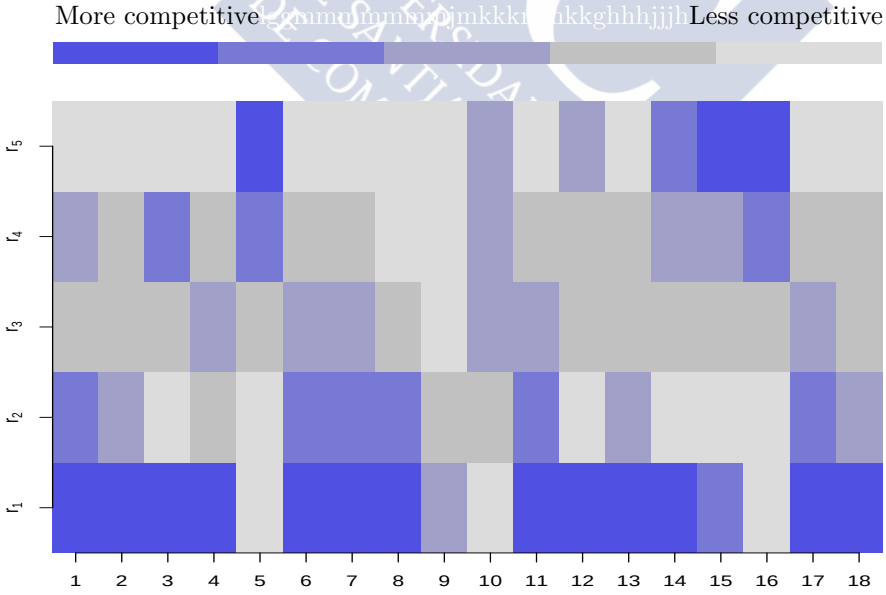


Figure 2.38: Influence of the parameter r (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 100$.

Figures 2.36 and 2.37 show the results for $n = 1600$ with $\tau = 0.5$ and $\tau = 0.8$, respectively. Figure for $\tau = 0.2$ is not showed but the conclusions are similar to the case $\tau = 0.5$. From this information, Walther's method depend heavily on the parameter r. In addition, this dependence is stronger if $\tau = 0.8$. In this case, small values of r provide better results, see densities from 2 to 12 in Figure 2.37.

On the other hand, if low values of n are considered the importance of selecting r gets even stronger, compare Figures 2.36 and 2.38. In the last one, we show the results for $n = 100$ and $\tau = 0.5$

2.3.3.3 Comparison of hybrid methods

The convex hull method, the r_3-convex hull method and the Walther's method with $r = r_3$ will compared. The real value of the parameter r is unknown so we have fixed an intermediate value for it. It should be note that according to the Sections 2.3.3.1 and 2.3.3.2, $r-$convex hull method is less sensitive to the selection to the parameter r.

Figures 2.39, 2.40 and 2.41 show the results obtained for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$, respectively, when $n = 1600$. Each method is represented on the vertical axis and each density model on the horizontal axis.

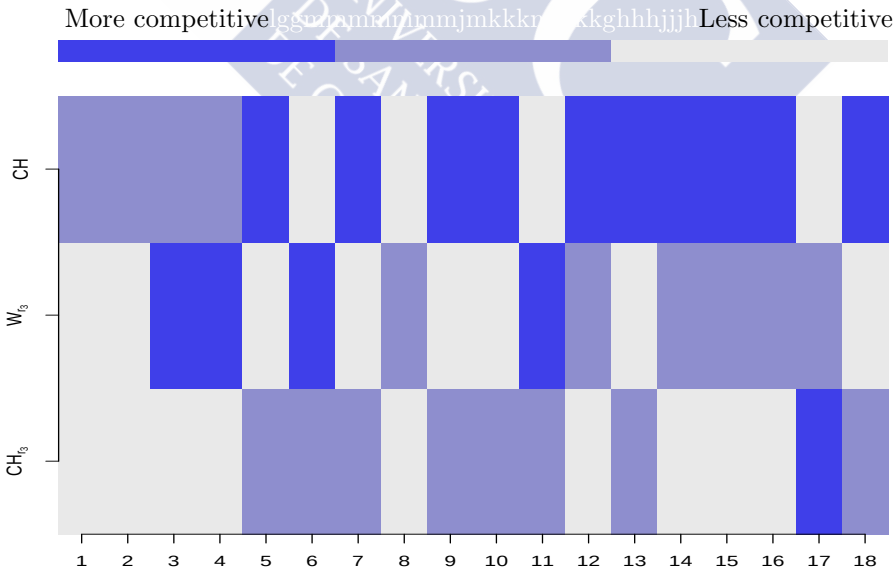


Figure 2.39: Comparison of hybrid methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.2$ and $n = 1600$.

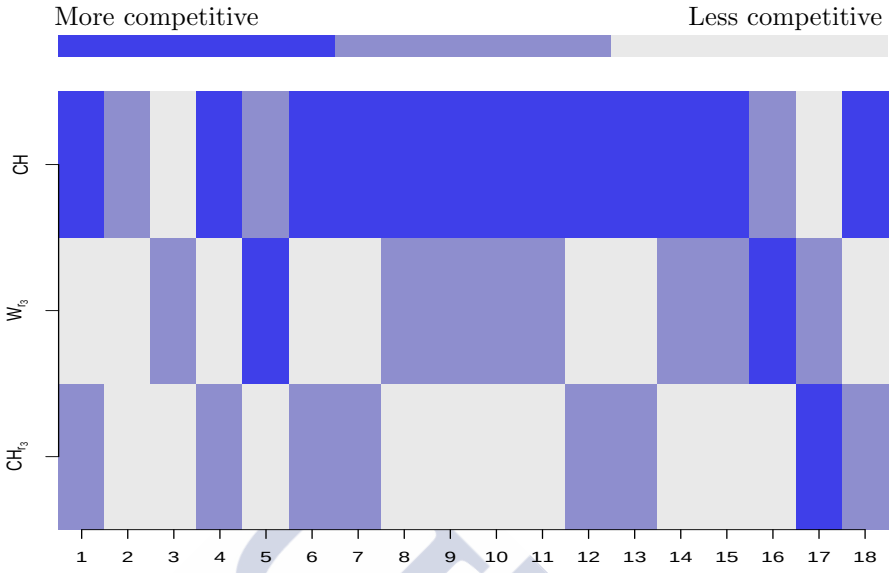


Figure 2.40: Comparison of hybrid methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

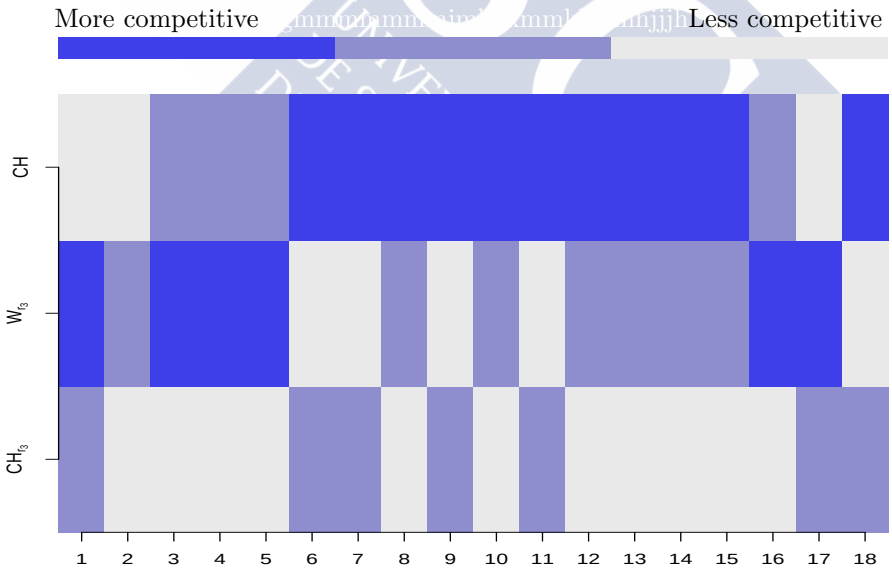


Figure 2.41: Comparison of hybrid methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.8$ and $n = 1600$.

Some of density models present convex level sets for $\tau = 0.2$ although they are not unimodal (see for example densities 6, 8 or 11 in Figure 2.39). In this cases, when the

convexity assumption is true, convex hull method can be very competitive. However, models 1, 2, 3 and 4 have convex level sets for some value of τ and r_3- convex hull method is the most competitive for them. In addition, sometimes convexity hypothesis can be very restrictive (see models 7 or 10, for example) and then, r_3- convex hull or granulometric smoothing methods provide better or similar results although the first one is slightly better for high values of τ.

If $\tau = 0.5$ or $\tau = 0.8$ then convexity is a very restrictive shape condition for most of models (see densities 6, 7, 8, 9, 10, 11, 12, 13, 14 or 15 in Figure 2.40 or models 6, 7, 9, 10, 11, 12, 13, 14 or 15 in Figure 2.41). So, granulometric smoothing and r_3- convex hull methods have better behavior than convex hull method. However, the r_3- convex hull method is the most competive for $\tau = 0.8$, see models 1, 3, 4, 5, 16 and 17.

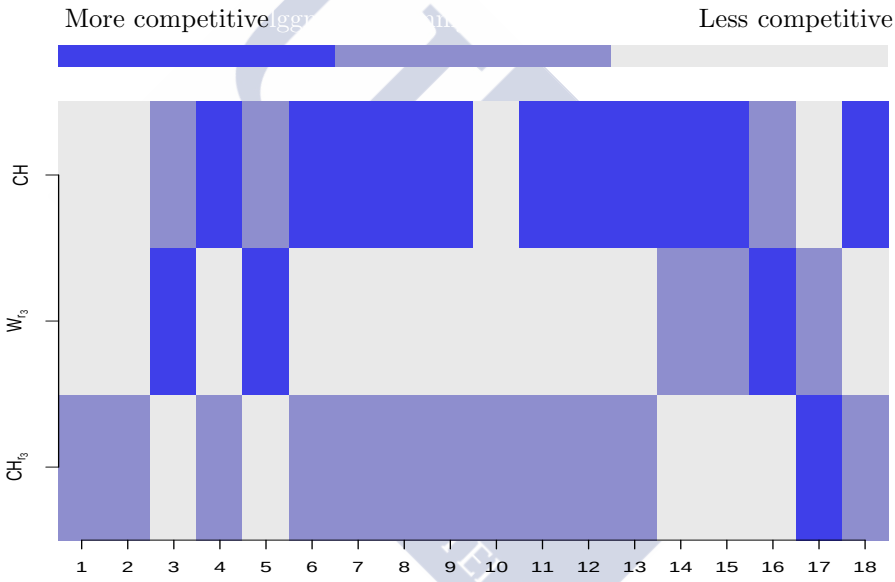


Figure 2.42: Comparison of hybrid methods (vertical axis) for the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 400$.

If smaller sample sizes are considered, the r_3-convex hull algorithm gets worse its performance. Figure 2.42 shows the simulation results when $n = 400$ and $\tau = 0.5$. In this case, the r_3-convex hull method is the most competitive alternative only for densities 3, 5, 14, 15 and 16. However, granulometric smoothing algorithm presents the best behaviour for models 1, 2, 4, 6-13 and 18. Nonetheless, note that the r_3-convex hull method does not provide the worst result for most of the models considered. This is quite promising, because the parameter r is not estimated from the data, and the optimal value was expected that this would change from model to model.

2.3.4 Final comparison

Finally, we will compare the most competitive methods in each group. So, we will consider cross validation method, Müller and Sawitzki's method, granulometric smoothing method, $r-$convex hull method and convex hull method. It is necessary to specify a value for the parameters M and r for Müller and Sawitzki's method and granulometric smoothing method or $r-$convex hull method. We have fixed $M = 3$, $M = 2$ and $r = r_3$.

Figure 2.43, Figure 2.44 and Figure 2.45 show the results for the error criteria d_{μ_f} with $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$, respectively. When $\tau = 0.2$, Müller and Sawitzki's method is more competitive with $M = 1$ than with $M = 2$ for unimodal densities 1–5, as shown in Figure 2.43. The same conclusion can be extracted for models 6, 8, 11, and 18, because, for this value of τ, their level sets have only one interval. Müller and Sawitzki's method with $M = 2$ does not exhibit very good results, because most of the models are not bimodal. In spite of this, it is the most competitive algorithm for models 10, 12, and 16. In this case, cross-validation provides quite good results, except for densities 3, 4, 6, 7, 9, 11, and 17, where one of the two hybrid methods presents the best behaviour. However, these two algorithms have a very important disadvantage in that they depend on an unknown parameter.

When $\tau = 0.5$, the cross-validation selector provides quite competitive results (see models 1, 2, 4, 5, 6, 10, 12, 14, 16, 17, and 18 in Figure 2.44). Granulometric smoothing and the r_3-convex hull method produce the best results for densities 6, 7, and 8 and 3, 8, 9, 11, 15, and 16, respectively. In this case, Müller and Sawitzki's method is not particularly competitive, especially when $M = 1$ for models 6–17. All of these densities have level sets with more than one interval.

Cross-validation exhibits the most competitive behaviour for unimodal densities when $\tau = 0.8$ and d_{μ_f} is the error criteria considered (see models 1, 2, 4, 5, 8, and density 16 in Figure 2.45). Granulometric smoothing presents its worst performance for densities 3, 4, 5, and 16. However, although the r_3-convex hull method is not the most competitive for many of the models, it presents very regular behaviour. Müller and Sawitzki's method with $M = 1$ does not provide very good results, e.g., densities 6, 7, 9, 10, 11, 12, 14, and 15. None of these models are unimodal. If we consider $M = 2$, this method is the most competitive for densities 12, 13, and 17.

Figure 2.46, Figure 2.47 and Figure 2.48 show the results for the error criteria d_H with $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$, respectively. When $\tau = 0.2$, MS_1 and MS_2 are not too competitive. MS_1 offers the best results only for density 10 and MS_2, only for models 12, 16 and 18. Although CH_{r_3} exhibits better results than the excess mass method, CV and W_{r_3} are the algorithms with the best regular behavior, see models 1, 2, 5, 7, 8, 13, 15 and 16 or densities 4, 6, 7, 8, 9, 13 and 14, respectively.

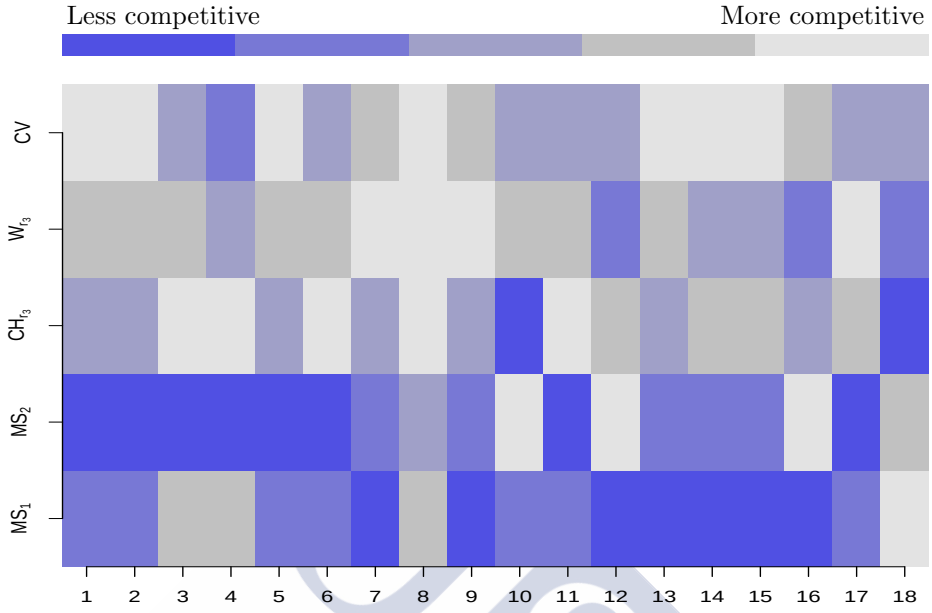


Figure 2.43: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.2$ and $n = 1600$.

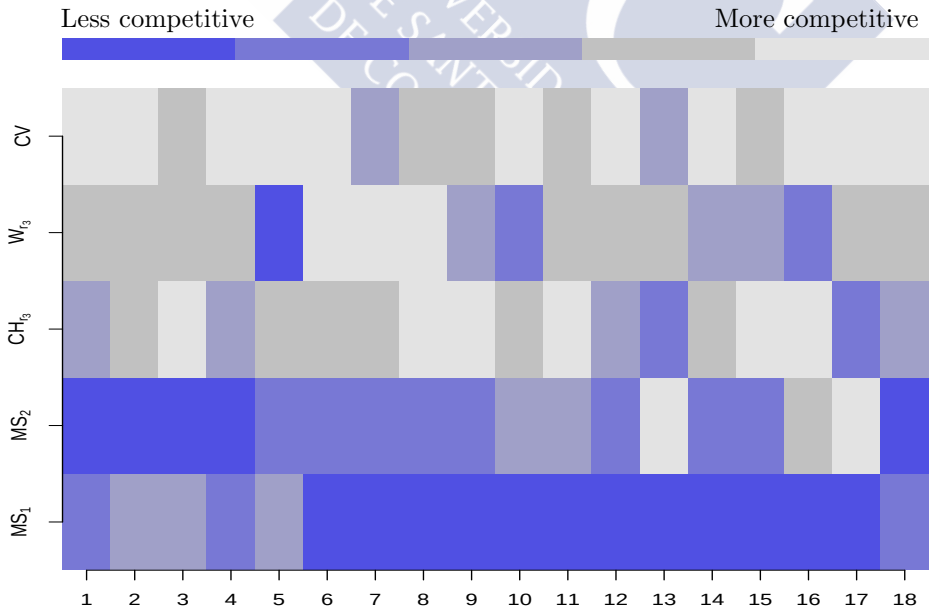


Figure 2.44: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.5$ and $n = 1600$.

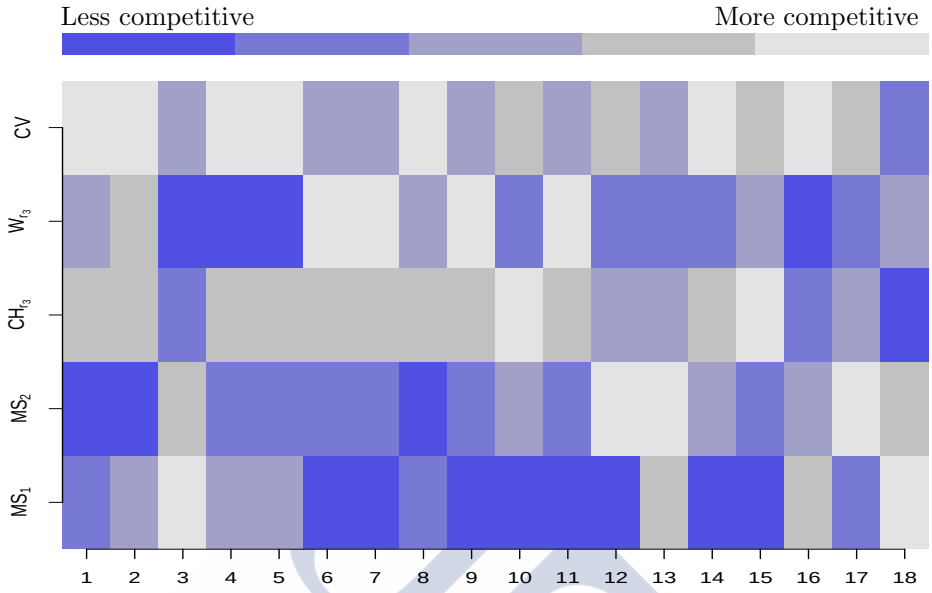


Figure 2.45: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis) considering d_{μ_f} as error criteria, $\tau = 0.8$ and $n = 1600$.

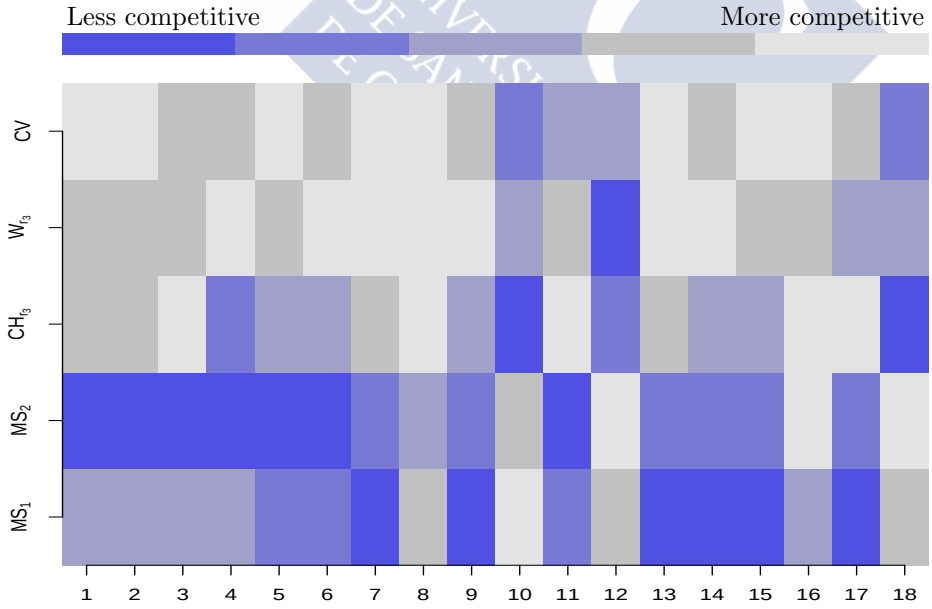


Figure 2.46: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis), considering d_H as error criteria, $\tau = 0.2$ and $n = 1600$.

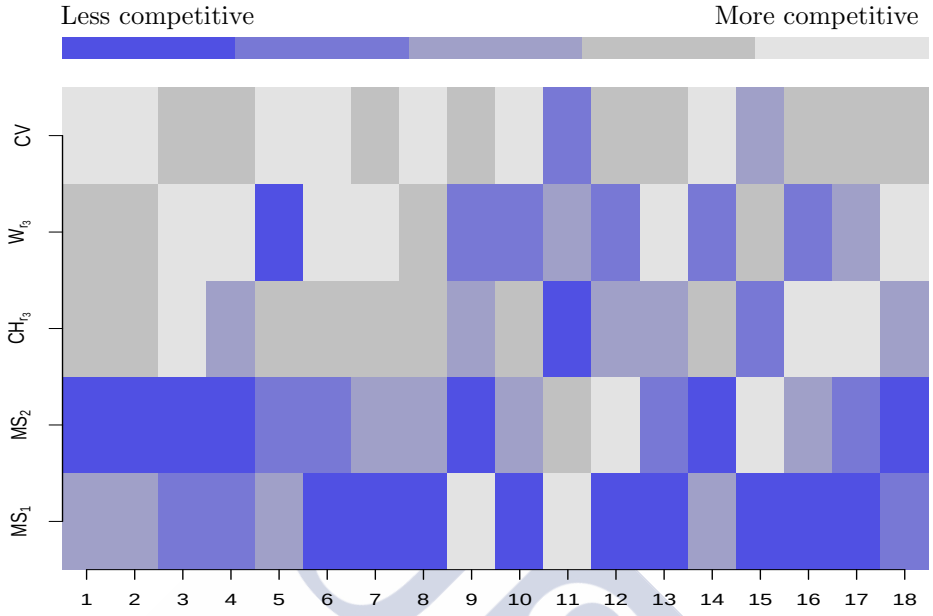


Figure 2.47: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis) considering d_H as error criteria, $\tau = 0.5$ and $n = 1600$.

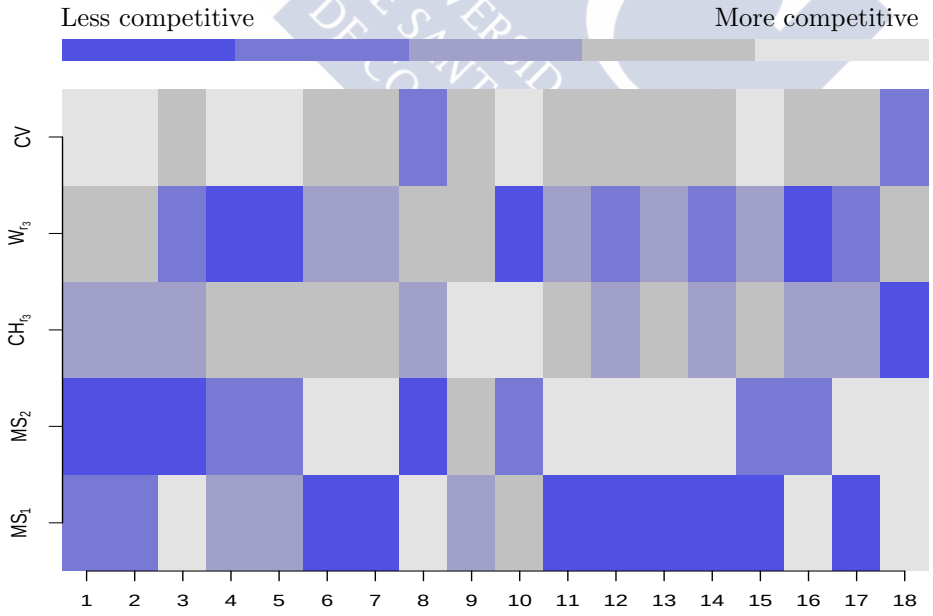


Figure 2.48: Comparison of CV, W_{r_3}, CH_{r_3}, MS_2, and MS_1 (vertical axis) with the 18 model densities (horizontal axis) considering d_H as error criteria, $\tau = 0.8$ and $n = 1600$.

Cross-validation exhibits the most competitive behavior for models 1, 2, 5, 6, 9, 10 and 14 when $\tau = 0.5$ and d_H is the error criteria considered. For the rest of densities, it offers quite good results except for the density 11 where MS_1 is the most competitive method. However, MS_1 and MS_2 are not too competitive in general. They present the worst results for models 1-4, 9, 14 or 18 (although the densities 1-4 and 18 are unimodal) and 6-8, 10, 12, 13 or 15-17, respectively. On the other hand, CH_{r_3} and W_{r_3} present their best performances for models 3, 16 or 17 and 3, 4, 6, 7, 13 or 18, respectively.

When $\tau = 0.8$ and it is considered d_H as error criteria, MS_2 improves its results considerably, see bimodal models 6, 7 or 17 and densities 11-14 and 18. Some of the modes for models 11-14 are not too significant. However, MS_1 is only the most competitive algorithm for the unimodal densities 3 and 18 or for the models 8 and 16 whose levels sets have an only connected component for $\tau = 0.8$. In this case, CH_{r_3} offers better results than W_{r_3}, see densities 3-7 and 9-17. Cross validation is again the most regular method with the best performance for models 1, 2, 4, 5, 10 or 15.

Although the results for lower values of the sample size are not showed here, similar conclusions can be extracted.

2.3.5 Conclusions of the simulation study

As has been stated previously, if no assumption is made on the shape of the density level set to be estimated, then plug-in methods provide good results. In general, cross-validation or even Sheather and Jones' method are good alternatives for reconstructing the level set. The results showed in this chapter suggest that specific bandwidth selectors for density level sets present worse general behaviour for the sample size considered.

In contrast, excess mass and hybrid methods are useful for incorporating the shape restrictions of the density level set into the estimators. In particular, Müller and Sawitzki's algorithm assumes that some information about the number of clusters M is given a priori. We therefore fixed two values, $M = 1$ and $M = 2$. Although most model densities satisfy one of these two conditions, this algorithm did not provide very competitive results in general. However, one of the main advantages of the excess mass methodology is that it does not need to smooth the data to reconstruct a density level set with a fixed probability content.

If, however, some geometric properties of the level set are known, then hybrid methods present a competitive alternative. For instance, if τ is small, the convex hull method was shown to provide good results. Most of these densities have convex level sets for sufficiently small values of τ. Under more flexible shape restrictions, the r-convex hull method and granulometric smoothing could be used. However, their main disadvantage is the dependence on r, an unknown parameter. These approaches are

very promising, because they remain quite competitive even when the value of r is fixed. Selecting r automatically from the sample points would significantly improve their practical performance. We will explain in depth this possibility in Chapter 4.

Chapter 3

A new data-driven method for estimating the support

Having reviewed the basics of support estimation, we now turn our attention to this problem under the assumption of $r-$convexity. The main goal of this work is to present a data-driven and stochastic method for estimating the unknown parameter r. As consequence, an algorithm for reconstructing the shape of a point cloud will be provided. This problem, for the bidimensional case, has already been studied in literature by Mandal and Murthy (1997). They proposed a selector for r based on the concept of minimum spanning but convergence rates for the resulting support estimator were not provided.

This chapter is organized as follows. In Section 3.1, the $r-$convex hull of the sample points is analyzed as an estimator for a $r-$convex support. The main disadvantage of this estimator is the selection of the smoothing parameter r. Its real value is unknown since that the support S is unknown too. The value of the optimal parameter r is established in Section 3.2. A new geometric condition will be necessary to obtain some interesting theoretical results. It is discussed in Section 3.2.1. The new data-driven and stochastic algorithm for selecting the smoothing parameter of $C_r(\mathcal{X}_n)$ is presented in Section 3.3. It is based on the theory of maximal spacings, see Janson (1987). Consistency of this estimator is established in Section 3.3.1. In addition, the resulting support estimator obtained from this smoothing parameter is able to achieve the same convergence rates as the convex hull for estimating convex sets. This will be proved in Section 3.4. The numerical questions involving the practical application of the algorithm are analyzed in Section 3.5. In Section 3.6, the performances of the new selector and Mandal and Murthy (1997)'s method will be analyzed through a simulation study. Finally, an application of the new algorithm is presented in Section 3.7. It is tested if the water area is decreasing in the Aral Sea.

A publication arising from the work compiled in this chapter, see Rodríguez-Casal and Saavedra-Nieves (2014).

3.1 Preliminaries

Support estimation deals with the problem of reconstructing the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X assuming that a random sample $\mathcal{X}_n = \{X_1, ..., X_n\}$ from X is given. According to Section 2.1.3, if it is assumed that S is $r-$convex then a natural estimator for the support is the $r-$convex hull of $\mathcal{X}_n$, $S_n = C_r(\mathcal{X}_n)$. This estimator is well known in the computational geometry literature for producing good global reconstructions if the sample points are (approximately) uniformly distributed on the set S. See Edelsbrunner (2014) for a survey on the subject.

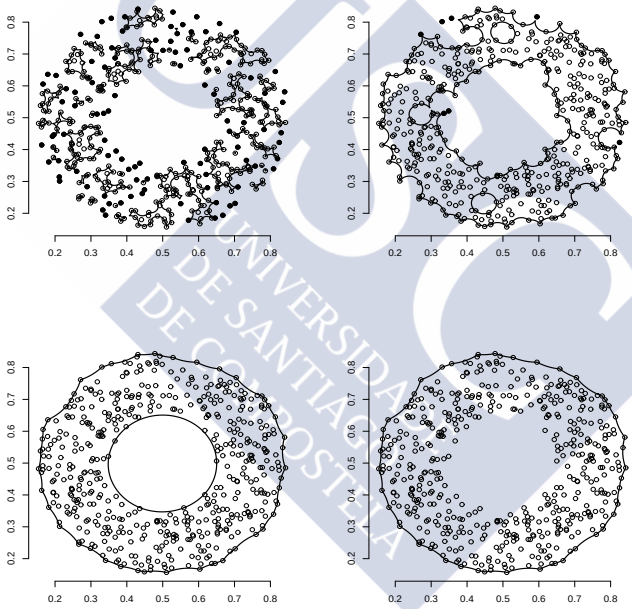


Figure 3.1: In the first row, $C_r(\mathcal{X}_{500})$ for $r = 0.02$ (left) and $r = 0.03$ (right). In the second row, $C_r(\mathcal{X}_{500})$ for $r = 0.15$ (left) and $r = 0.155$ (right).

However, this estimator depends on an unknown parameter r. For the particular problem of reconstructing the Aral Sea, we have analyzed the importance of selecting this smoothing parameter correctly. Next, its influence will be showed again in Figure 3.1. We have represented $C_r(\mathcal{X}_{500})$ for different values of r from a uniform sample of size $n = 500$ on the support $S = B_{0.35}[(0.5, 0.5)] \backslash B_{0.15}((0.5, 0.5))$. Small values of r provide split estimators. However, if the value of r is too large then $B_{0.15}((0.5, 0.5))$

will be contained in the estimator. This gap which does not intersect the sample points does not belong to the theoretical support. This intuitive idea will be fundamental to propose the automatic selection criterion to estimate r. The problem of selecting r can be also analyzed by using graphical descriptive methods. For the sample considered previously, we have calculated the area of $C_r(\mathcal{X}_n)$ for a sequence of values from 0 to 0.3. In Figure 3.2 (left) the relationship between the parameter r and the area of the estimator (on vertical axis) and r (on horizontal axis) is showed. Similar conclusions could be obtained if the length of the boundary is considered. Other alternatives are considered in Figure 3.2 (center and right). In Figure 3.2 (center), we have represented the circles with area equal to the area of $C_r(\mathcal{X}_n)$ for a sequence of values of r using a gradient of colours. A jump can be observed for values of r greater than 0.15. In this case, the closed ball $B_{0.15}[(0.5, 0.5)]$ is contained in $C_r(\mathcal{X}_n)$. In Figure 3.2 (right), we have represented with blue the circumference with radius equal to the value of r which provides this change for the area behaviour. From Figure 3.2, it is easy to observe the existence of a region which is included in the estimator but it does not belong to the support. Although these tools can be useful for a first approximation, they depend strongly on the sequence or grid of radius considered.

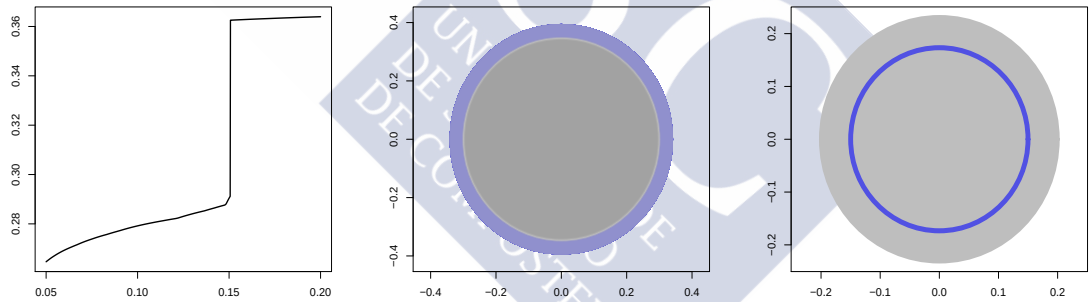


Figure 3.2: Graphical methods for selecting the parameter r.

The goal of this research work is to present an automatic method to select the smoothing parameter of the $r-$convex hull from a random and uniform sample of points. The theory of maximal spacings will be used, see Janson (1987).

3.2 Defining the optimal parameter

The problem of reconstructing a $r-$convex support S in an automatic way can be solved if the parameter r is estimated from a random sample of points $\mathcal{X}_n$ taken in S. Next, it will be presented an algorithm to do this. The first step is to determine

precisely the optimal value of r. We will take into account a very simple property: If S is $r-$convex for $r > 0$ then it is r^*-convex for all $0 < r^* \leqslant r$, see Section 1.2.2. So, it seems reasonable to estimate the highest value of r which verifies that S is $r-$convex, see Definition 3.2.1. However, if S is convex then it is easy to prove that S is $r-$convex for all $r > 0$. Consequently and for simplicity in the exposition, it is assumed that S is not convex. If S is a $r-$convex set that is not convex then $\{\gamma > 0 : C_\gamma(S) = S\}$ is a nonemptyset and upper bounded. Therefore, it is possible to present Definition 3.2.1.

Definition 3.2.1. *Let $S \subset \mathbb{R}^d$ be a nonempty, compact, non convex and $r-$convex set for some $r > 0$. It is defined*

$$r_0 = \sup\{\gamma > 0 : C_\gamma(S) = S\}. \tag{3.1}$$

Of course, if S is convex r_0 would be infinity. In addition, if r_0 is a maximum of the set $\{\gamma > 0 : C_\gamma(S) = S\}$ then it is possible to guarantee that S is r_0-convex. In this case, it is clear that $C_r(\mathcal{X}_n)$, for $r < r_0$, is a non admisible estimator since it is always outperformed by $C_{r_0}(\mathcal{X}_n)$. This is because $C_r(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$ and hence, $d_\mu(C_{r_0}(\mathcal{X}_n), S) \leqslant d_\mu(C_r(\mathcal{X}_n), S)$ (the same holds for the Hausdorff distance). So, the only admisible parameter would be $r = r_0$. On the other hand, if $r > r_0$ then $C_r(\mathcal{X}_n)$ can considerably overestimate S specially if S has a big hole inside, see Figure 3.1. In Proposition 3.2.11, it is proved that the supreme defined in (3.1) is a maximum. However, it is not enough to assume that S is $r-$convex for obtaining the proof. It was necessary to suppose that S satisfies a new geometric property slightly stronger than $r-$convexity:

(R_λ^r) A closed ball of radius $\lambda > 0$ rolls freely in S and a closed ball of radius $r > 0$ rolls freely in $\overline{S^c}$.

In Definition 1.2.10 was analyzed the intuitive concept of rolling freely. In Proposition 3.2.11 it is proved that, under (R_λ^r), if $\{r_n\}$ converges to r_0 and $C_{r_n}(S) = S$ then it is verified that $C_{r_0}(S) = S$. The idea of the proof is as follows. If $C_{r_n}(S) = S$ then S would be r_n-convex, for all r_n. Then, Lemma 3.2.2 would guarantee that a closed ball of radius r_n rolls freely in $\overline{S^c}$, for all r_n. Therefore, and according to Lemma 3.2.9, a closed ball of radius r_0 rolls freely in $\overline{S^c}$. That is, the rolling property is preserved in the limit. However, it will be showed next that the rolling property with radius r_0 is not enough to guarantee that S is r_0-convex. Although the rolling property is preserved if the limit is considered (see Lemma 3.2.9), we have not been able to prove the analogous result for $r-$convexity. However, under (R_λ^r), the equivalence between r_0-convexity and rolling property in $\overline{S^c}$ for radius r_0 can be obtained, see Lemma 3.2.2 and Proposition 3.2.10. Before presenting these results in Section 3.2.2, it is necessary

to study the relationship between some classical geometric notions and (R^r_λ) in Section 3.2.1.

3.2.1 A new flexible geometric condition

Sets fulfilling condition (R^r_λ) have a number of desirable properties which make them easier to handle. Walther (1997, 1999) did not explicitly consider the case where the radius λ of balls rolling in S can be different from r, the radius of balls rolling freely in $\overline{S}^c$, see Figure 3.3. But this is important for defining the parameter r_0.

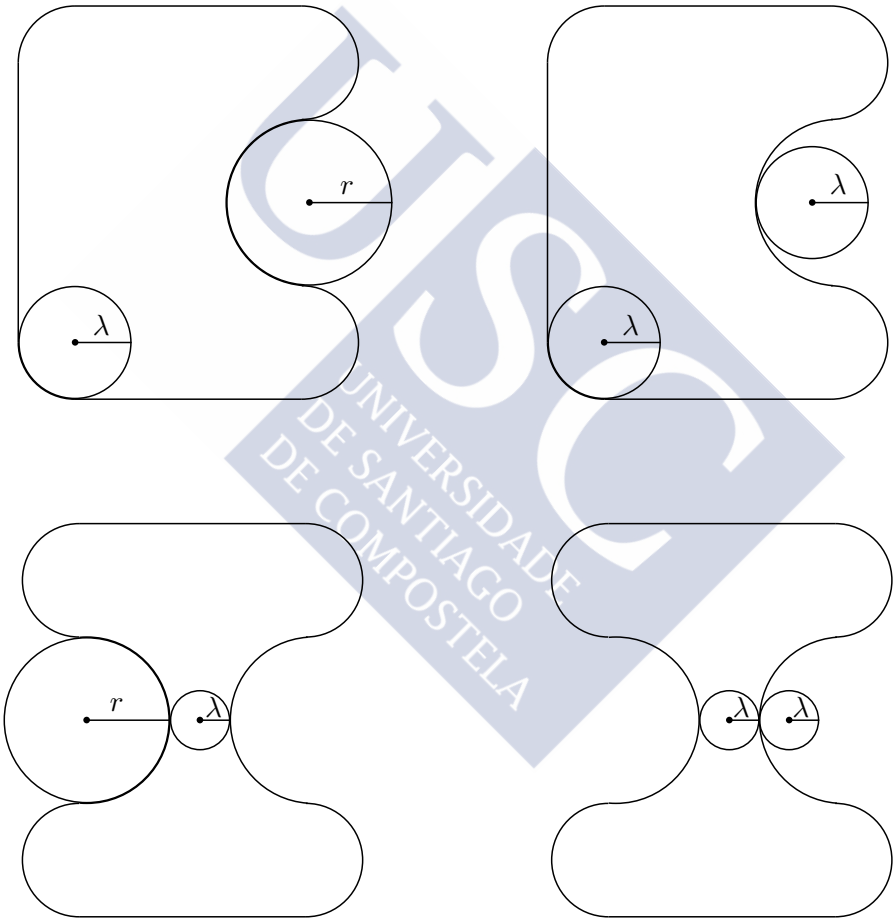


Figure 3.3: (R^r_λ) is a more general condition.

The $\lambda-$rolling property ensures that S is smooth but we are not imposing any relationship between r and λ. According to Figure 3.4, (R^r_λ) includes even the case with $r = \infty$ and λ very close to zero.

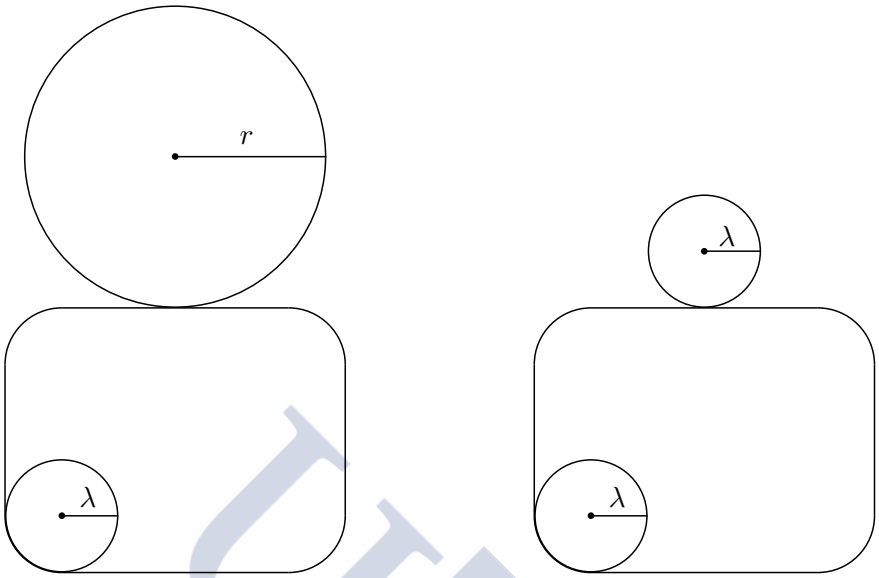


Figure 3.4: $r > 0$ can be very large and $\lambda > 0$ can be very close to zero.

Walther (1997, 1999) allows us to prove the equivalence between these three properties for two closed sets A and $\overline{A^c}$: A and $\overline{A^c}$ are $r-$convex, A and $\overline{A^c}$ belong to the Serra's model and A satisfies the property (R_r^r). In particular, under (R_λ^r), it is possible to check that both S and $\overline{S^c}$ are $m-$convex for $m = \min\{\lambda, r\}$, see Figure 3.5.

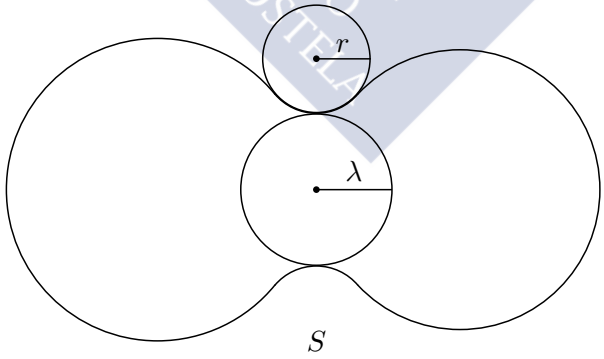


Figure 3.5: S is $r-$convex since that $\min\{\lambda, r\} = r$.

Cuevas et al. (2012) studied the relationship between these geometric notions too. The authors prove that they are not equivalent.

Lemma 3.2.2. *(Cuevas et al., 2012) Let $A \subset \mathbb{R}^d$ be a compact and $\gamma-$convex set for some $\gamma > 0$. Then, a closed ball of radius γ rolls freely in $\overline{A^c}$.*

Remark 3.2.3. *The outside rolling property for a set A established in Cuevas et al. (2012) is slightly different from our rolling approach for $\overline{A^c}$. However and since $\partial\overline{A^c} \subset \partial A$, Lemma 3.2.2 holds.*

According to Figure 3.6, the reciprocal is not true in general.

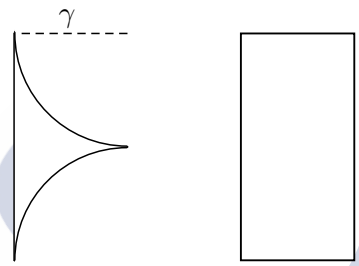


Figure 3.6: $\gamma-$rolling in $\overline{A^c} \not\Rightarrow \gamma-$convexity.

In Proposition 3.2.10, it will be proved that, under (R_λ^r) for any $\lambda > 0$, S is $r-$convex too. So, (R_λ^r) is a sufficient condition for guaranteeing $r-$convexity of the support S; however, (R_λ^r) is not a necessary condition. Figure 3.7 shows three $r-$convex sets which do not satisfy (R_λ^r) for any $\lambda > 0$.

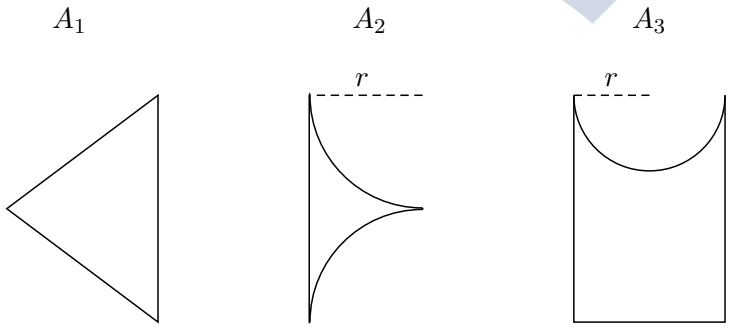


Figure 3.7: A_1 is convex and, so, $r-$convex for all $r > 0$. A_2 and A_3 are $r-$convex.

Next results provide the necessary tools in order to prove Lemma 3.2.9 and Proposition 3.2.10. The first one guarantee the existence of an outward pointing unit vector

for each point belonging to the boundary of the set.

Lemma 3.2.4. *Let $A \subset \mathbb{R}^d$ be a closed and nonempty set such that a ball of radius λ rolls freely in A. Then, for all $a \in \partial A$ exists $\eta(a)$ (not necessarily unique) such that $\|\eta(a)\| = 1$ and $B_\lambda[a - \lambda\eta(a)] \subset A$.*

Proof. According to the property of rolling freely for a given $a \in \partial A$ exists $x \in A$ such that $a \in B_\lambda[x] \subset A$ verifying that $\|x - a\| = \lambda$. If $\|x - a\| < \lambda$ then $a \in B_{\lambda - \|x-a\|}[x] \subset \mathrm{Int}(A)$ which is a contradiction since that $a \in \partial A$. Then, it is possible to define

$$\eta(a) = \frac{(a - x)}{\|a - x\|}.$$

It is verified that $x = a - \lambda\eta(a)$. So,

$$B_\lambda[a - \lambda\eta(a)] \subset A. \qquad \square$$

The vector $\eta(a)$ (see Figure 4.7) is not unique necessarily. Lemma 3.2.5 relates the uniqueness of this unit vector and the existence of some $x \notin A$ such that a coincides with the metric projection of x onto A.

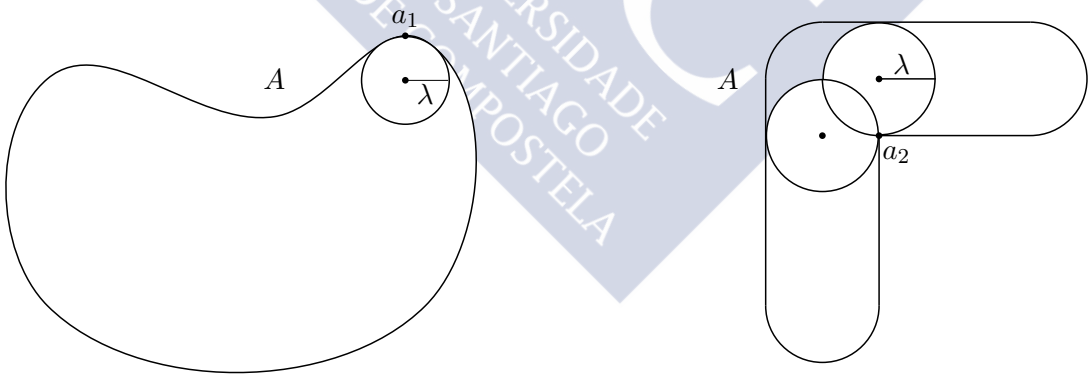


Figure 3.8: A ball of radius λ rolls freely in A. For $a_1 \in \partial A$ exists a unique $x \in A$ such that $a_1 \in B_\lambda[x] \subset A$. For $a_2 \in \partial A$, $a_2 \in B_\lambda[x]$ for a non finite number of points $x \in A$.

Lemma 3.2.5. *Let $A \subset \mathbb{R}^d$ be a nonempty and closed set and $a \in \partial A$. Let us assume that there exists $x \notin A$ such that*

$$\rho = \|x - a\| = d(x, A),$$

that is, a is a metric projection of x onto A. If exists $\lambda > 0$ and a unit vector $\eta(a)$ such that $B_\lambda[a - \lambda\eta(a)] \subset A$, then

$$x = a + \rho\eta(a).$$

Proof. To see this suppose the contrary, that is, let us suppose that exists x verifying the required conditions with $x \neq a + \rho\eta(a)$. Then, x, a and $a - \lambda\eta(a)$ can not lie on the same line and hence,

$$\|a - \lambda\eta(a) - x\| < \|a - \lambda\eta(a) - a\| + \|a - x\| = \lambda + \rho. \tag{3.2}$$

Let $z \in \partial B_\lambda[a - \lambda\eta(a)] \cap [x, a - \lambda\eta(a)]$, where $[x, a - \lambda\eta(a)]$ denotes the line segment with endpoints x and $a - \lambda\eta(a)$ (see Figure 3.9). Then,

$$\|a - \lambda\eta(a) - x\| = |a - \lambda\eta(a) - z\| + \|z - x\| = \lambda + \|z - x\|.$$

According to (3.2),

$$\|z - x\| = \|a - \lambda\eta(a) - x\| - \lambda < \lambda - \rho - \lambda = \rho,$$

which is a contradiction since $z \in A$ and $\rho = d(x, A)$. □

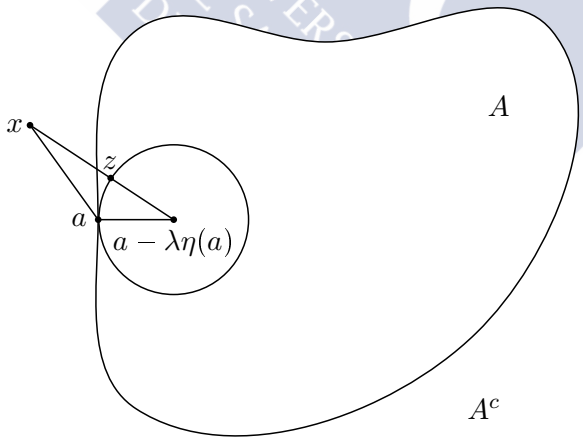


Figure 3.9: Elements of Lemma 3.2.5.

Remark 3.2.6. *According to Lemma 3.2.5, the unit vector $\eta(a)$ is unique, whenever $a \in \partial A$ is the metric projection of some $x \notin A$ onto A. Alternatively, if there exists more than one ball such that $a \in B_\lambda[x] \subset A$ then a can not be the metric projection of any point $x \notin A$, see Figure 3.10.*

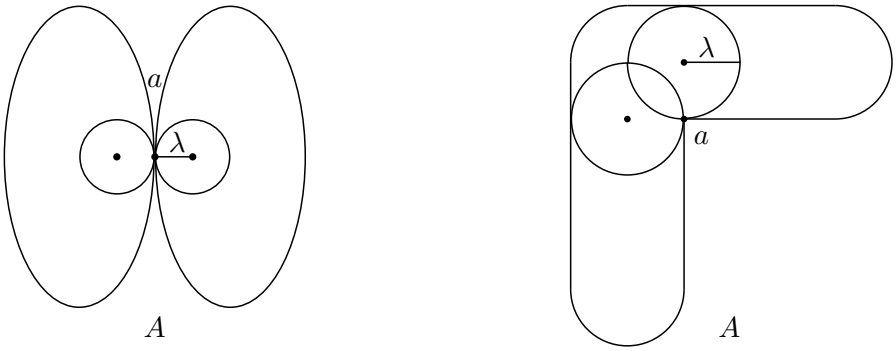


Figure 3.10: $a \in \partial A$ is not a metric projection of $x \notin A$ on A because there exist two or more unit vectors $\eta(a)$ such that $B_\lambda[a - \lambda\eta(a)] \subset A$.

Lemma 3.2.7 guarantees a reasonable topological behaviour of sets under rolling freely condition, see Figure 3.11. For these type of sets, it can be proved easily that the outside rolling property of a set A defined in Cuevas et al. (2012) is totally equivalent to our rolling definition for $\overline{A^c}$. In particular, it holds under (R^r_λ).

Lemma 3.2.7. *Let $A \subset \mathbb{R}^d$ be a nonempty and closed set. If a ball of radius λ rolls freely in A then*

$$\mathrm{Int}(\overline{A^c}) = A^c \text{ and } \partial A = \partial\overline{A^c}.$$

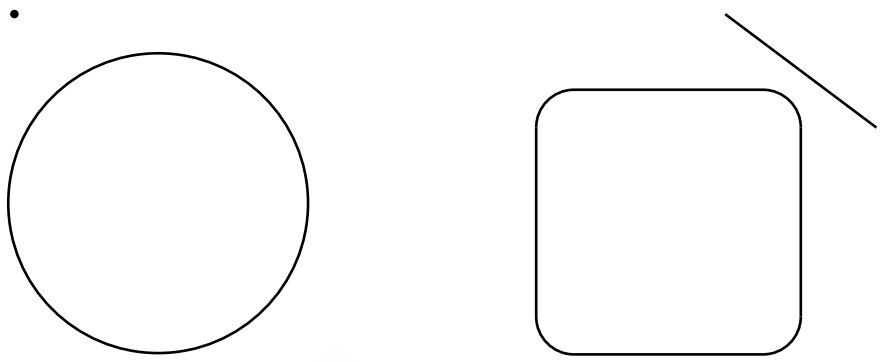
Proof. First, we will prove that $\mathrm{Int}(\overline{A^c}) = A^c$. Since that A^c is open and $A^c \subset \overline{A^c}$ then $A^c \subset \mathrm{Int}(\overline{A^c})$. Next, it will be proved that $\mathrm{Int}(\overline{A^c}) \subset A^c$. Let us suppose the contrary, that is, there exists $x \in \mathrm{Int}(\overline{A^c})$ such that $x \in A$. Then, $x \in A \cap \overline{A^c} = \partial A$. Rolling freely in A guarantees that there exists $p \in A$ such that $x \in B_\lambda[p] \subset A$ with $\|x - p\| = \lambda$. Since that $x \in \mathrm{Int}(\overline{A^c})$, there exists $\epsilon > 0$ such that $B_\epsilon[x] \subset \overline{A^c}$. Let us assume that $\epsilon < \lambda$ and let us consider the point

$$y_\tau = x + \tau\frac{p - x}{\|p - x\|}, \ \tau \in (0, \epsilon).$$

Then, $y_\tau \in B_\lambda(p) \subset \mathrm{Int}(A)$. So, a contradiction is obtained since $y_\tau \in B_\epsilon[x] \subset \overline{A^c}$. Proving $\partial A = \partial\overline{A^c}$ is easy because the boundary of a set can be written as the closure minus the interior. In addition, A is closed and $\mathrm{Int}(\overline{A^c}) = A^c$. So,

$$\partial\overline{A^c} = \overline{\overline{A^c}} \setminus \mathrm{Int}(\overline{A^c}) = \overline{A^c} \setminus A^c = \overline{A^c} \setminus \mathrm{Int}(A^c) = \partial A^c = \partial A. \qquad \square$$

Lemma 3.2.8 guarantees that, under (R^r_λ), the outward pointing unit vector established in Lemma 3.2.4 is unique.



A_1 A_2

Figure 3.11: A_1 and A_2 are not under the conditions imposed in Lemma 3.2.7.

Lemma 3.2.8. *Let $A \subset \mathbb{R}^d$ be a closed set verifying (R_λ^r). Then, for any point $a \in \partial A$ there exists a unique unit vector $\eta(a)$ such that*

$$B_\lambda(a - \lambda\eta(a)) \subset A \text{ and } B_r(a + r\eta(a)) \subset \overline{A^c}.$$

Proof. Let $a \in \partial A$. Under (R_λ^r), a ball of radius λ rolls freely in A. Then,

$$\exists x \in A \text{ such that } B_\lambda(x) \subset A.$$

In addition, it is possible to write (see Lemma 3.2.4)

$$x = a - \lambda\eta(a) \text{ with } \eta(a) = \frac{a - x}{\|a - x\|}.$$

According to Lemma 3.2.7, $\partial A = \partial\overline{A^c}$ and, so, $a \in \partial\overline{A^c}$. Under (R_λ^r), it is verified that a ball of radius r rolls freely in $\overline{A^c}$. Then,

$$\exists y \in \overline{A^c} \text{ such that } B_r(y) \subset \overline{A^c}$$

verifying that $\|y - a\| = d(y, A) = r$. So, a is metric projection of a point $y \notin A$. According to the Lemma 3.2.5,

$$y = a + r\eta(a) \text{ and then } B_r(a + r\eta(a)) \subset \overline{A^c}. \qquad \square$$

Lemma 3.2.9 shows that the rolling freely property present a continuous behavior.

Lemma 3.2.9. *Let $A \subset \mathbb{R}^d$ be a closed set. Let $\{r_n\}$ be a sequence of positive terms converging to $\overline{r}$. If a ball of radius r_n rolls freely in $\overline{A^c}$, for all n, then a ball of radius $\overline{r}$ will roll freely in $\overline{A^c}$.*

Proof. Let us suppose that $r_n \neq \overline{r}$ for all n since that, otherwise, the proof would be trivial. It is verified that

$$\forall a \in \partial A \text{ and } \forall n \in \mathbb{N} \ \exists x_n \text{ such that } a \in B_{r_n}[x_n] \subset \overline{A^c}.$$

For each $a \in \partial A$, let us consider the sequence of closed balls $\{B_{r_n}[x_n]\} \subset \overline{A^c}$. It is not restrictive to assume that $\{r_n\}$ is an monotone increasing sequence. In another case, it would be possible to consider a monotone subsequence of $\{r_n\}$ denoted by $\{r_n\}$ again converging to $\overline{r}$. If a decreasing subsequence was considered, the proof would be trivial. Then, only the increasing case will be considered. Then, $\{r_n\}$ converges to $\overline{r}$ and $\{x_n\}$ converges to x_a since $\{x_n\}$ is bounded and it contains a convergent subsequence which we denote by $\{x_n\}$ again. Two steps are necessary to get the proof.

Step 1: It will be proved that for any $a \in \partial A$ it is verified that $B_{\overline{r}}[x_a] \subset \overline{A^c}$. To see this suppose the contrary, that is, let us suppose that $a \in \partial A$ such that $B_{\overline{r}}[x_a] \not\subset \overline{A^c}$ with $\overline{A^c} = A^c \cup \partial A$. Then,

$$\exists \overline{a} \in \text{Int}(A) \text{ such that } \overline{a} \in B_{\overline{r}}[x_a] \text{ and, so, } \overline{a} \notin \overline{A^c}.$$

Without loss of generality, we can assume that $\|\overline{a} - x_a\| < \overline{r}$. If $\|\overline{a} - x_a\| = \overline{r}$ it is enough to consider a new point on the segment $[x_a, \overline{a}]$. Since $\overline{a} \in \text{Int}(A)$ then $d(\overline{a}, \partial A) > 0$. So, there will exist $\widetilde{a} \in \text{Int}(A) \cap [x_a, \overline{a}]$ such that $\|\widetilde{a} - x_a\| < \overline{r}$. In this case, $\overline{a} = \widetilde{a}$ would be taken, see Figure 3.13.

Since $\{r_n\} \uparrow \overline{r}$,

$$\exists n_0 \in \mathbb{N} \text{ such that } \|\overline{a} - x_a\| < r_n < \overline{r}, \ \forall n \geqslant n_0.$$

So,

$$\forall n \geqslant n_0, \ \overline{a} \in B_{r_n}(x_a) \subset B_{r_n}[x_a]. \tag{3.3}$$

Let us define for all $n \geqslant n_0$,

$$d_n = d(\overline{a}, \partial B_{r_n}[x_a]).$$

In addition, $\{r_n\}$ is an increasing sequence. Then, it is verified that

$$B_{r_1}[x_a] \subset B_{r_2}[x_a] \subset ...$$

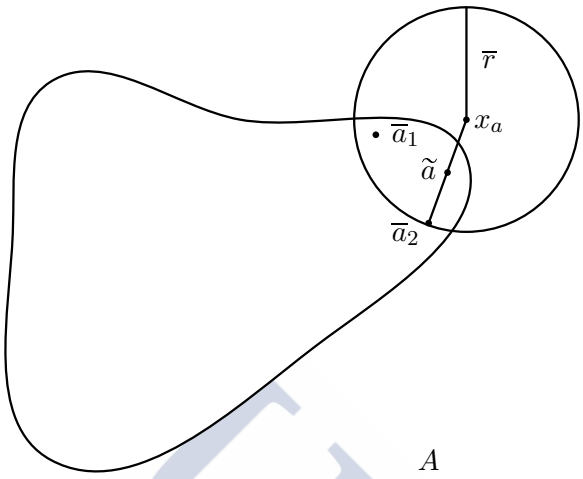


Figure 3.12: Elements of Step 1 in Lemma 3.2.9 with $\|\overline{a}_1 - x_a\| < \overline{r}$ and $\|\overline{a}_2 - x_a\| = \overline{r}$.

and, as consequence and taking (3.3) into account,

$$0 < d_{n_0} \leqslant d_{n_1} \leqslant d_{n_2} \leqslant ...$$

Let us consider $d_{n_0}/2$, since $\{x_n\}$ converges to x_a,

$$\exists n_1 \in \mathbb{N} \text{ such that } \|x_a - x_n\| < d_{n_0}/2, \ \forall n \geqslant n_1.$$

So,

$$\overline{a} \in B_{r_n}[x_n], \ \forall n \geqslant n_2 = \max\{n_0, n_1\}.$$

To see this, notice that, if $n \geqslant n_2$ then

$$\|\overline{a} - x_n\| \leqslant \|\overline{a} - x_a\| + \|x_a - x_n\| < r_n - d_n + \frac{d_{n_0}}{2} < r_n - d_n + \frac{d_n}{2} < r_n.$$

This fact is a contradiction since

$$B_{r_n}[x_n] \subset \overline{A^c}, \ \forall n$$

because $\overline{a} \in \text{Int}(A)$.

Step 2: It will be proved that $a \in B_{\overline{r}}[x_a]$. We will assume that $a \notin B_{\overline{r}}[x_a]$ and we will

show that this is impossible under the assumptions we have done. If $a \notin B_{\overline{r}}[x_a]$ then $\|a - x_a\| > \overline{r}$ and it is possible to define $\epsilon = \|a - x_a\| - \overline{r} > 0$. Since $\{x_n\}$ converges to x_a, there exists $n_0 \in \mathbb{N}$ such that $\|x_n - x_a\| < \epsilon$. For all $n \geqslant n_0$,

$$\|x_n - a\| \geqslant \|a - x_a\| - \|x_a - x_n\| > \|a - x_a\| - \epsilon = \overline{r}.$$

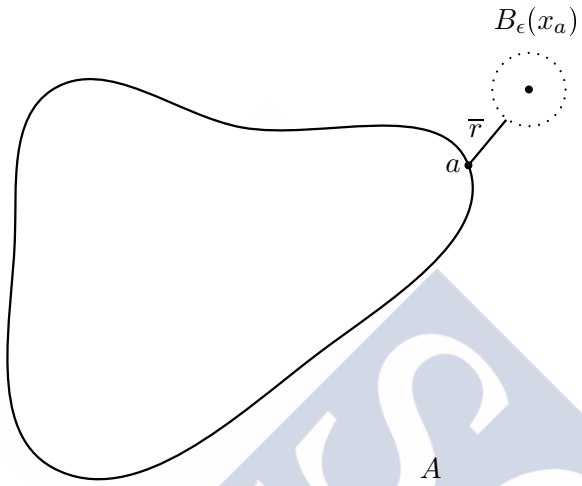


Figure 3.13: Elements of Step 2 in Lemma 3.2.9: $a \in \partial A$, x_a and $B_\epsilon(x_a)$.

Since $\{r_n\}$ is an monotone increasing sequence converging to $\overline{r}$, $a \notin B_{r_n}[x_n]$. This is a contradiction since we are assuming that $a \in B_{r_n}[x_n]$ for all n. □

Proposition 3.2.10 guarantees, under (R^r_λ), the $r-$convexity of the support S. This result shows that the reciprocal of Lemma 3.2.2 is true if (R^r_λ) is assumed.

Proposition 3.2.10. *Let $S \subset \mathbb{R}^d$ be a nonempty, compact support verifying (R^r_λ). Then, S is $r-$ convex.*

Proof. Let us prove that $S = C_r(S)$. Since $S \subset C_r(S)$ for any $r > 0$, it is enough to check if $C_r(S) \subset S$. Equivalently, it will be checked that for all $x \in S^c$ there exists an open ball of radius r containing x that does not intersect S. Let us fix $x \notin S$. If $d(x, S) \geqslant r$ then

$$x \in B_r(x) \text{ and } B_r(x) \cap S = \varnothing.$$

Otherwise, if $d(x, S) < r$, let s be a projection of x on S and let us define $\rho = d(x, S) = \|x - s\|$. According to Lemmas 3.2.4 and 3.2.5,

$$B_\lambda[s - \lambda\eta(s)] \subset S,$$

where $\eta(s) = (s - x)/\|s - x\|$ and $x = s + \rho\eta(s)$. In addition, $s \in \partial S = \partial\overline{S^c}$ and, according to the imposed conditions, a ball of radius r rolls freely in $\overline{S^c}$. So,

$$\exists c \in \mathbb{R}^d \text{ such that } s \in B_r[c] \text{ and } B_r[c] \subset \overline{S^c}.$$

It is verified that $B_r(c) \cap S = \varnothing$ and, according to Lemma 3.2.5,

$$c = s + r\eta(s).$$

since s is projection of c on S. We are supposing that $\rho < r$. So,

$$\|x - c\| = \|(\rho - r)\eta(s)\| = r - \rho < r.$$

Then, $x \notin C_r(S)$ since $x \in B_r(c)$ and $B_r(c) \cap S = \varnothing$. □

Figure 3.14 shows the elements used in the proof of Proposition 3.2.10.

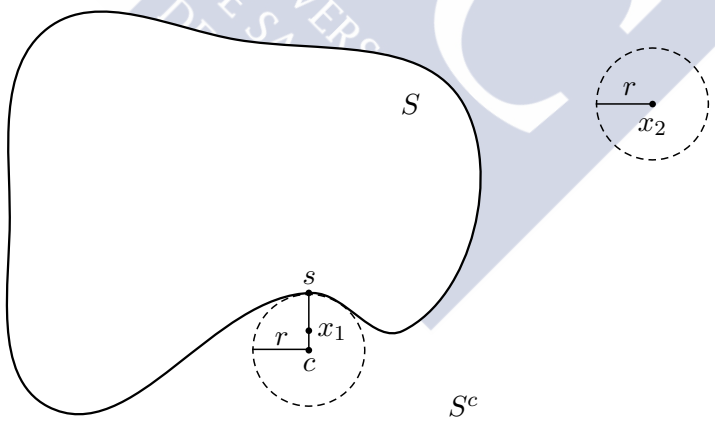


Figure 3.14: Elements of Proposition 3.2.10 with $d(x_1, S) < r$ and $d(x_2, S) > r$.

3.2.2 Studying the smoothing parameter

Having presented the relationships between the different geometric conditions, we are now ready to prove that the supreme defined in (3.1) is, in fact, a maximum.

Proposition 3.2.11. *Let $S \subset \mathbb{R}^d$ be a nonempty, compact and nonconvex set verifying (R_λ^r) and let r_0 be the parameter defined in (3.1). Then,*

$$C_{r_0}(S) = S.$$

As consequence, a ball of radius r_0 rolls freely in $\overline{S^c}$.

Proof. It will be proved that $r_0 \in \{\gamma > 0 : C_\gamma(S) = S\}$. According to the properties of the supreme,

$$r_0 \in \overline{\{\gamma > 0 : C_\gamma(S) = S\}}$$

and, so, there exists $\{r_n\}$ converging to r_0 such that $\{r_n\} \subset \{\gamma > 0 : C_\gamma(S) = S\}$. Then,

$$C_{r_n}(S) = S, \ \forall n \in \mathbb{N}.$$

According to Lemma 3.2.2, a ball of radius r_n rolls freely in $\overline{S^c}$ for all n. Then, a ball of radius r_0 rolls freely in $\overline{S^c}$, see Lemma 3.2.9. Since a ball of radius $\lambda > 0$ rolls freely in the interior of S, it is possible to guarantee that S is under $(R_\lambda^{r_0})$. According to Proposition 3.2.10, S is a r_0-convex set. □

3.3 The new data-driven method

According to the previous comments, if S is compact, nonempty, nonconvex and it is under (R_λ^r) then the existence of the optimal parameter r_0 is guaranteed. In addition, it is satisfied that $C_{r_0}(S) = S$. The uniformity test proposed by Berrendero et al. (2012) will be used to estimate r_0 in a data-driven way from $\mathcal{X}_n$. This test is based on the multivariate spacings theory studied by Janson (1987). In the univariate case, the spacings defined by a random sample of points $\mathcal{X}_n$ in a support interval $S = [a, b]$ are defined as the gap lengths left by the sample points in the interval. They are calculated in a simple way in terms of differences between consecutive order statistics.

If $d > 1$ the definition of spacings is not so straightforward. However, there still is a natural way to define the largest (or maximal) spacing with some valuable properties derived for it. In this work, we will consider spherical spacings. The maximal spacing for S is defined in a formal way as

$$\Delta_n(S) = \sup\{\gamma : \exists x \text{ with } B_\gamma[x] \subset S \backslash \mathcal{X}_n\}. \tag{3.4}$$

Obviously, the maximal spacing depends on S and, of course, on the sample points. The Lebesgue measure (volume) of the balls with radius $\Delta_n(S)$ is denoted by $V_n(S)$.

Theorem 3.3.1 will be essential in order to present the uniformity test.

Theorem 3.3.1. *(Janson (1987)) Let $\mathcal{X}_n$ be an i.i.d and uniform sample on S, with $\mu(S) = 1$ and $\mu(\partial S) = 0$. Then, the following weak convergence holds*

$$nV_n - \log n - (d-1)\log\log n - \log\beta \xrightarrow{d} U,$$

where V_n denotes the volume associated with the largest spacing Δ_n defined in (3.4), β is a known constant, $\xrightarrow{d}$ denotes the convergence in distribution and U is a random variable with distribution $\mathbb{P}(U \leqslant u) = \exp(-\exp(-u))$, para $u \in \mathbb{R}$.

Remark 3.3.2. *Janson (1987) gave explicitly the value of the constant β for the most of cases. In particular, if $d = 2$ then $\beta = 1$.*

Berrendero et al. (2012) use Theorem 3.3.1 in order to propose a test of uniformity on the support S establishing as null hypothesis:

$$H_0: \ X \text{ is uniform on } S.$$

With significance level α, H_0 would be rejected whenever

$$V_n(S) > \frac{a(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n}, \tag{3.5}$$

where $a = \mu(S)$ and μ denotes the Lebesgue measure.

However, the main goal of their work is to present a uniformity test for a more general case in which the support S is unknown but it satisfies some geometric restriction. It was assumed that S verified (R_ν^ν) for some $\nu > 0$ and $S_n = C_\nu(\mathcal{X}_n)$ was considered as an estimator of S since, under (R_ν^ν), S is $\nu-$convex. However, ν is again unknown and, in addition, its influence is very strong. No data-driven method was provided for selecting ν. Under H_0, the null hypothesis of uniformity could be rejected on $C_\nu(\mathcal{X}_n)$ if high values of ν are considered, see Figure 3.15. In this case, the maximal spacing is estimated as

$$\hat{\Delta}_n = \sup\{\gamma : \exists x \text{ with } B_\gamma[x] \subset S_n \backslash \mathcal{X}_n\}, \tag{3.6}$$

and the critical region (3.5) is replaced by

$$\hat{V}_n > \frac{a_n(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n},$$

where $a_n = \mu(S_n)$ and $\hat{V}_n$ denotes the volume of the ball of radius $\hat{\Delta}_n$ given in (3.6). In practise, the authors considered an alternative critical region

$$\hat{V}_n > \frac{a_n^*(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n},$$

where $a_n^* = \frac{n}{(n-v_n)}\mu(S_n)$ and v_n denotes the number of vertices of S_n.

Figure 3.15 shows the maximal spacing for a sample of size $n = 500$ with uniform distribution on the circular ring $B_{0.35}[(0.5, 0.5)]\backslash B_{0.15}((0.5, 0.5))$ and the maximal spacing when it is assumed that the support is known, see (left), or unknown, see (center) and (right). In this last case, $S_n = C_{0.15}(\mathcal{X}_{500})$ and $S_n = C_{0.155}(\mathcal{X}_{500})$ are considered as support estimators for the circular ring. A bad choice of the smoothing paremeter (right) allows us to detect a very big gap or spacing, clearly incompatible with the uniformity hypothesis.

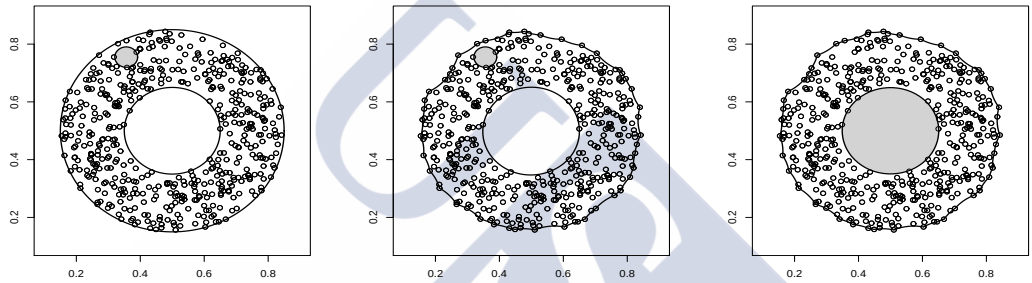


Figure 3.15: Maximal spacing with support $S = B_{0.35}[(0.5, 0.5)]\backslash B_{0.15}((0.5, 0.5))$ (left), $S_n = C_{0.15}(\mathcal{X}_{500})$ (center) and $S_n = C_{0.155}(\mathcal{X}_{500})$ (right).

Similarly, Figure 3.16 shows the maximal spacings for the estimators of the Aral Sea considered in Figure 2.3. A bad choice (a big value) of the smoothing parameter allows to detect a large gap in Figure 3.16 (right), again incompatible with uniformity assumption. The test will reject that $\mathcal{X}_n$ is uniform on the support considered (known or unknown) if it contains a big spacing. Then, null hypothesis of uniformity should be rejected on $S_n = C_{100}(\mathcal{X}_{2000})$ with level α. However, $\mathcal{X}_{2000}$ is a uniform sample on the Aral Sea. This means that the estimator contains a large spacing which is not contained in the Aral Sea. Then, the smoothing parameter have been chosen in a wrong way because the sample is uniform on the original support. It must be selected smaller than 100.

From an intuitive point of view, the test will reject that a sample $\mathcal{X}_n$ is uniform on the support considered (known or unknown) if this one contains a gap with a relatively large area and that does not intersect the sample points. The algorithm that we propose for estimating r_0 is based in the case of the unknown support following the somewhat opposite approach by Berrendero et al. (2012). We will assume that $\mathcal{X}_n$ follows a uniform distribution on S. If a large spacing is found in $C_r(\mathcal{X}_n)$ then we should reduce the value of r. As an example in Figure 3.15 (right), the null hypothesis of uniformity

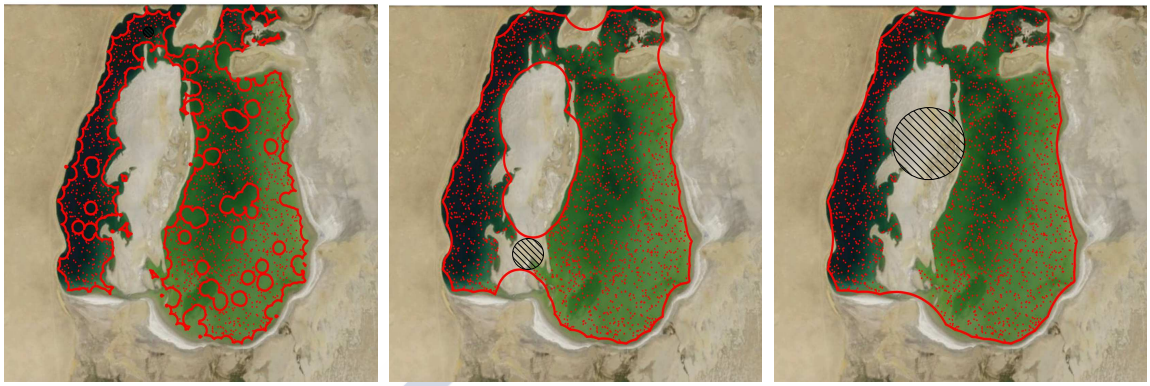


Figure 3.16: $C_{10}(\mathcal{X}_{2000})$ (left). $C_{40}(\mathcal{X}_{2000})$ (center). $C_{100}(\mathcal{X}_{2000})$ (right) is almost equal to the convex hull of $\mathcal{X}_{2000}$.

is rejected on $S_n = C_{0.155}(\mathcal{X}_{500})$ with level α. However, $\mathcal{X}_{500}$ is, by construction, a uniform sample on S. This means that the estimator contains a big enough gap that is not compatible with the uniformity of the sample points on S_n. This is because S_n is not contained in S. Then, it is possible to deduce that the choice of the smoothing parameter is wrong and this one is smaller than 0.155. According to the Definition 3.2.1, we suggest to estimate r_0 as

$$\hat{r}_0 = \sup\{\gamma > 0 : H_0 \text{ is accepted on } C_\gamma(\mathcal{X}_n)\}. \tag{3.7}$$

In the next section some technical aspects will be considered. For instance, the existence of the supreme defined in (3.7) will be guaranteed, with probability one, for n large enough.

3.3.1 Consistency for the estimator of the optimal parameter

Next theoretical auxiliary results will be useful for guaranteeing the consistency of the estimator proposed in (3.7). First we will prove that, with probability increasing to one, $\hat{r}_0$ is at least as big as r_0.

Theorem 3.3.3. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (3.1) and $\{\alpha_n\} \subset (0,1)$ a sequence converging to zero that denotes the significance level of*

the tests performed in (3.7). Then,

$$\lim_{n\to\infty} \mathbb{P}(\hat{r}_0 \geqslant r_0) = 1.$$

Proof. From the definition of $\hat{r}_0$, see (3.7), it is clear that

$$\mathbb{P}(\hat{r}_0 \geqslant r_0) \geqslant \mathbb{P}(\hat{V}_{n,r_0} \leqslant \hat{c}_{n,\alpha_n,r_0}),$$

where $\hat{V}_{n,r_0}$ denotes the volume of the maximal spacing in $C_{r_0}(\mathcal{X}_n)$,

$$\hat{c}_{n,\alpha_n,r_0} = \frac{\mu(C_{r_0}(\mathcal{X}_n))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta)}{n}$$

and u_{α_n} satisfies $\mathbb{P}(U \leqslant u_{\alpha_n}) = 1 - \alpha_n$ and U is the random variable defined in Theorem 3.3.1. Since, with probability one, $C_{r_0}(\mathcal{X}_n) \subset S$, we have $\hat{V}_{n,r_0} \leqslant V_n(S)$ where, remember, $V_n(S)$ denotes the volume of a ball with radius the maximal spacing of S. Hence,

$$\mathbb{P}(\hat{r}_0 \geqslant r_0) \geqslant \mathbb{P}(V_n(S) \leqslant \hat{c}_{n,\alpha_n,r_0}) = \mathbb{P}\left(\frac{u_{\alpha_n}}{A_n} U_n \leqslant u_{\alpha_n}\right),$$

where

$$U_n = \frac{nV_n(S)}{\mu(S)} - \log n - (d-1)\log\log n - \log\beta$$

and

$$A_n = \frac{n\hat{c}_{n,\alpha_n,r_0}}{\mu(S)} - \log n - (d-1)\log\log n - \log\beta.$$

According to the Janson (1987)'s Theorem, $U_n \xrightarrow{d} U$. Next, it will be proved easily that

$$\frac{u_{\alpha_n}}{A_n} \xrightarrow{\mathbb{P}} 1.$$

To see this, notice that it is possible to write

$$\frac{u_{\alpha_n}}{A_n} = \frac{u_{\alpha_n}}{\frac{n\hat{c}_{n,\alpha_n,r_0}}{\mu(S)} - \log n - (d-1)\log\log n - \log\beta}$$

$$= \frac{u_{\alpha_n}}{\frac{n}{\mu(S)}\frac{\mu(C_{r_0}(\mathcal{X}_n))(u_{\alpha_n}+\log n+(d-1)\log\log n+\log\beta)}{n} - \log n - (d-1)\log\log n - \log\beta}$$

$$= \frac{u_{\alpha_n}}{\frac{\mu(C_{r_0}(\mathcal{X}_n))}{\mu(S)}(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta) - \log n - (d-1)\log\log n - \log\beta}.$$

In addition,

$$\mu(C_{r_0}(\mathcal{X}_n))/\mu(S) = 1 + O_P((\log(n)/n)^{2/(d+1)}),$$

see Theorem 3 in Rodríguez-Casal (2007). Therefore,

$$\frac{u_{\alpha_n}}{A_n} \xrightarrow{\mathbb{P}} 1.$$

According to the Slutsky's Lemma and since that

$$\frac{u_{\alpha_n}}{A_n} \xrightarrow{\mathbb{P}} 1 \text{ and } U_n \xrightarrow{d} U \text{ then } \frac{u_{\alpha_n}}{A_n} U_n \xrightarrow{d} U.$$

Notice that U has a continuous distribution, so convergence in distribution implies that

$$\sup_u |\mathbb{P}\left((u_{\alpha_n}/A_n)U_n \leqslant u\right) - \mathbb{P}(U \leqslant u)| \to 0.$$

Since $\mathbb{P}(U \leqslant u_{\alpha_n}) = 1 - \alpha_n$ and $\alpha_n \to 0$, this ensures that

$$\mathbb{P}\left((u_{\alpha_n}/A_n)U_n \leqslant u_{\alpha_n}\right) \to 1.$$

Therefore, $\mathbb{P}(\hat{r}_0 \geqslant r_0) \to 1$. □

Next, it will be proved that $\hat{r}_0$ cannot be arbitrarily larger that r_0. The following proposition ensures that, for a given $\gamma > r_0$, there exists an open ball contained in $C_\gamma(S)$ which does not meet S.

Proposition 3.3.4. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and let be $\gamma > 0$ such that $S \subsetneq C_\gamma(S)$. Then, there exists $\epsilon > 0$ and $x \in C_\gamma(S)$ such that $B_\epsilon(x) \subset C_\gamma(S)$ and $B_\epsilon(x) \cap S = \varnothing$.*

Proof. Let us assume, for a moment, that we can find $s \in \partial S$ such that $s \in \text{Int}(C_\gamma(S))$. In this case, there exists $\rho > 0$ satisfying that $B_\rho(s) \subset C_\gamma(S)$, see Figure 3.17. On the other hand, by assumption, S is r_0-convex which implies, using Lemma 3.2.2, that a ball of radius r_0 rolls freely in $\overline{S^c}$. Lemma 3.2.7 guarantees that $s \in \partial S = \partial\overline{S^c}$. So, there exists a ball $y \in \mathbb{R}^d$ such that $s \in B_{r_0}[y] \subset \overline{S^c}$. Therefore, $B_{r_0}(y) \subset \overline{S^c}$. As consequence of Lemma 3.2.7, $B_{r_0}(y) \subset \text{Int}(\overline{S^c}) = S^c$. Then, $B_{r_0}(y) \cap S = \varnothing$. It is clear that we can find an open ball $B_\epsilon(x)$ such that $B_\epsilon(x) \subset B_{r_0}(y) \cap B_\rho(s)$, see Figure 3.17. By construction $B_\epsilon(x) \subset B_{r_0}(y)$ and, hence, $B_\epsilon(x) \cap S = \varnothing$. Finally, $B_\epsilon(x) \subset B_\rho(s)$ and, therefore, $B_\epsilon(x) \subset C_\gamma(S)$. This would finished the proof in this case.

It remains to prove what happens if $\partial S \subset \partial C_\gamma(S)$. We will show that this is impossible under the assumptions we have done. First, the hypothesis $\partial S \subset \partial C_\gamma(S)$ imply that a ball of radius γ rolls freely in $\overline{S^c}$. This is a straightforward consequence of Lemma 3.2.2 since $C_\gamma(S)$ is γ-convex. So, S would satisfy the (R^γ_λ) shape restriction that implies, see Proposition 3.2.10, γ-convexity. This is a contradiction since we are assuming that $S \subsetneq C_\gamma(S)$. □

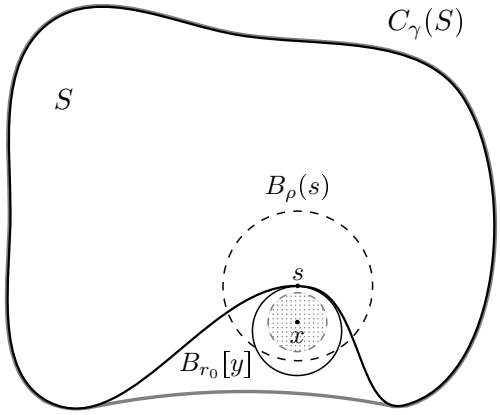


Figure 3.17: Elements of proof in Lemma 3.3.4. ∂S in black, $\partial C_\gamma(S)$ in gray, $B_\rho(s)$, $B_{r_0}[y]$ and $B_\epsilon(x)$ in gray.

Lemma 3.3.5. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (3.1). Then, for all $r > r_0$, there exists an open ball $B_\rho(x)$ such that $B_\rho(x) \cap S = \varnothing$ and*

$$\mathbb{P}\left(B_\rho(x) \subset C_r(\mathcal{X}_n),\ \textit{eventually}\right) = 1.$$

Proof. Let be r^* such that $r > r^* > r_0$. Since $C_{r_0}(S) = S \subsetneq C_{r^*}(S)$, according to Lemma 3.3.4,

$$\exists B_\epsilon(x) \text{ such that } B_\epsilon(x) \subset C_{r^*}(S) \text{ and } B_\epsilon(x) \cap S = \varnothing.$$

It can be assumed, without loss of generality, that $r \leqslant \frac{\epsilon}{2} + r^*$. If this is not the case then it would be possible to replace r^* by $r^{**} > r^*$ satisfying $r^{**} < r \leqslant \frac{\epsilon}{2} + r^{**}$. For this r^{**},

$$B_\epsilon(x) \subset C_{r^*}(S) \subset C_{r^{**}}(S) \text{ and } B_\epsilon(x) \cap S = \varnothing.$$

Now, we apply Proposition B.0.5 in Appendix B in order to ensure that

$$\mathbb{P}\left(S \oplus B_{r^*}[0] \subset \mathcal{X}_n \oplus B_r[0],\ \text{eventually}\right) = 1. \tag{3.8}$$

Since $S \in \mathcal{G}_S(r_0)$, for $0 < \epsilon^* < r_0$, it is verified that, see Appendix B,

$$\mathbb{P}(S \oplus B_{r^*}[0] \nsubseteq [(S \cap \mathcal{X}_n) \oplus (r^* + 2\epsilon^*)B_1[0]])$$

$$\leqslant D(\epsilon^*, S \oplus (r^* + 2\epsilon^*)B_1[0]) \exp\left\{-nab \min\{r^* + \epsilon^*, r_0\}^{\frac{d-1}{2}} {\epsilon^*}^{\frac{d+1}{2}}\right\},$$

where $D(\epsilon^*, B) = \max\{card\ N : N \subset B,\ |x-y| > \epsilon^* \text{ for different } x, y \in N\}$, a is a dimensional constant and $b = 1/\mu(S)$. Therefore,

$$\mathbb{P}(S \oplus B_{r^*}[0] \not\subset [(S \cap \mathcal{X}_n) \oplus B_{r^*+2\epsilon^*}[0]]) \leqslant A \exp\{-nW\}$$

where A and W denote the corresponding constants. Since

$$\sum_{i=1}^{\infty} \exp\{-nW\} < \infty.$$

We have, using the Borel Cantelli Lemmas, that

$$\mathbb{P}(S \oplus B_{r^*}[0] \not\subset [(S \cap \mathcal{X}_n) \oplus B_{r^*+2\epsilon^*}[0]]),\ \text{infinitely often}) = 0.$$

Then, with probability one and for n large enough,

$$S \oplus B_{r^*}[0] \subset (S \cap \mathcal{X}_n) \oplus B_{r^*+2\epsilon^*}[0]).$$

Then, since $S \in \mathcal{G}_C(r_0)$ and, with probability one, $\mathcal{X}_n \subset S$, it is verified for n large enough that

$$S \oplus B_{r^*}[0] \subset \mathcal{X}_n \oplus B_{r^*+2\epsilon^*}[0].$$

In addition, assuming that $\epsilon^* < (r - r^*)/2$, with probability one and for n large enough,

$$S \oplus B_{r^*}[0] \subset \mathcal{X}_n \oplus B_{r^*+2\epsilon^*}[0] \subset \mathcal{X}_n \oplus B_r[0].$$

According to (3.8), if $S \oplus B_{r^*}[0] \subset \mathcal{X}_n \oplus B_r[0]$ then $(S \oplus B_{r^*}[0]) \ominus B_{r^*}[0] \subset (\mathcal{X}_n \oplus B_r[0]) \ominus B_{r^*}[0]$, that is, $C_{r^*}(S) \subset (\mathcal{X}_n \oplus B_r[0]) \ominus B_{r^*}[0]$. This imply that

$$C_{r^*}(S) \ominus B_{r-r^*}[0] \subset ((\mathcal{X}_n \oplus B_r[0]) \ominus B_{r^*}[0]) \ominus B_{r-r^*}[0].$$

In addition,

$$((\mathcal{X}_n \oplus B_r[0]) \ominus B_{r^*}[0]) \ominus B_{r-r^*}[0] = (\mathcal{X}_n \oplus B_r[0]) \ominus B_r[0] = C_r(\mathcal{X}_n),$$

where we have used that, for sets A, C and D, $(A \ominus C) \ominus D = A \ominus (C \oplus D)$. Finally, since $B_\epsilon(x) \subset C_{r^*}(S)$ and $\epsilon/2 \geqslant (r - r^*)$, we have $B_{\epsilon/2}(x) \subset C_{r^*}(S) \ominus B_{\epsilon/2}[0] \subset C_{r^*}(S) \ominus B_{r-r^*}[0] \subset C_r(\mathcal{X}_n)$. This concludes the proof of the lemma by taking $\rho = \epsilon/2$. □

Proposition 3.3.6. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (3.1) and $\{\alpha_n\} \subset (0,1)$ a sequence of significance levels converging to zero such that $\log(\alpha_n)/n \to 0$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\hat{r}_0 \leqslant r_0 + \epsilon,\ \textit{eventually}\right) = 1$$

Proof. Given $\epsilon > 0$ let be $r = r_0 + \epsilon$. According to Lemma 3.3.5, there exists $x \in \mathbb{R}^d$ and $\rho > 0$ such that $B_\rho(x) \cap S = \varnothing$ and

$$\mathbb{P}\left(B_\rho(x) \subset C_r(\mathcal{X}_n), \text{ eventually}\right) = 1.$$

Since, with probability one, $\mathcal{X}_n \subset S$ we have $B_\rho(x) \cap \mathcal{X}_n = \varnothing$. Hence, if $B_\rho(x) \subset C_r(\mathcal{X}_n)$, we have $\hat{V}_{n,r} \geqslant \mu(B_\rho(x)) = c_\rho > 0$. Similarly, $\hat{V}_{n,r'} \geqslant \hat{V}_{n,r} \geqslant c_\rho$ for all $r' \geqslant r$. On the other hand, since $-u_{\alpha_n}/\log(\alpha_n) = \log(-\log(1-\alpha_n))/\log(\alpha_n) \to 1$, we have, with probability one,

$$\sup_{r'} \hat{c}_{n,\alpha_n,r'} \leqslant \frac{\mu(\text{conv}(S))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta)}{n},$$

and

$$\frac{\mu(\text{conv}(S))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta)}{n} \to 0$$

where $\text{conv}(S)$ denotes the convex hull of S. This means that, with probability one, there is n_0 such that if $n \geqslant n_0$ we have $\sup_{r'} \hat{c}_{n,\alpha_n,r'} < c_\rho$. Therefore, if $B_\rho(x) \subset C_r(\mathcal{X}_n)$, we get $\hat{r}_0 \leqslant r$. This last statement follows from $\hat{V}_{n,r'} > \hat{c}_{n,\alpha_n,r'}$ for all $r' \geqslant r$ and the definition of $\hat{r}_0$, see (3.7). □

Theorem 3.3.7 shows that $\hat{r}_0$ is finite and it estimates r_0 consistently. We assume that S is not convex only for simplicity in the exposition considering the case $r_0 = \infty$. If S is convex it can be shown that $\hat{r}_0$ goes to infinity (which is the value of r_0 in this case) because, with high probability, the test is not rejected for all values of r.

Theorem 3.3.7. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (3.1) and $\hat{r}_0$ defined in (3.7). Let $\{\alpha_n\} \subset (0,1)$ be a sequence of significance levels converging to zero verifying $\lim_{n\to\infty} \log(\alpha_n)/n = 0$. Then, $\hat{r}_0$ converges to r_0 in probability.*

Proof. The proof is a straightforward consequence of Theorem 3.3.3 and Proposition 3.3.6. □

3.4 Consistency for the resulting estimator for the support

Once the consistency of $\hat{r}_0$ as an estimator of the parameter r_0 has been established, it is necessary to analyze the quality of $C_{\hat{r}_0}(\mathcal{X}_n)$ as an estimator for the support S. The distances between sets allow us to calculate the error when the support is estimated by measuring the distance between S and its estimations. In this work, two usual metrics between sets are often considered in order to assess the performance of the supports estimators, Hausdorff distance and distance in measure. In Theorem 3.4.1, some conditions for guaranteeing the consistency of the estimator are analyzed.

Theorem 3.4.1. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (3.1) and $\hat{r}_0$ defined in (3.7). Let $\{\alpha_n\} \subset (0,1)$ be a sequence of significance levels converging to zero verifying $\lim_{n\to\infty} \log(\alpha_n)/n = 0$. If $\lim_{\overline{r}\to r_0^+} d_H(S, C_{\overline{r}}(S)) = 0$ then $d_H(S, C_{\hat{r}_0}(\mathcal{X}_n)) \to 0$, in probability. The same holds for $d_\mu(S, C_{\hat{r}_0}(\mathcal{X}_n)) = 0$.*

Proof. For the uniform distribution on S, Theorem 3 of Rodríguez-Casal (2007) ensures that, under $(R^{\widetilde{r}}_{\widetilde{r}})$, then $\mathbb{P}(\mathcal{E}_n) \to 1$, where

$$\mathcal{E}_n = \left\{ d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leqslant A \left(\frac{\log n}{n} \right)^{2/(d+1)} \right\},$$

and A is some constant. Under the hypothesis of Theorem 3.4.1 this holds for any $\widetilde{r} \leqslant \min\{r, \lambda\}$. Fix one $\widetilde{r} \leqslant \min\{r, \lambda\} \leqslant r_0$ and observe that $C_{\widetilde{r}}(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$. If the event $\mathcal{E}_n$ holds we have

$$d_H(S, C_{r_0}(\mathcal{X}_n)) \leqslant d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leqslant A \left(\frac{\log n}{n} \right)^{2/(d+1)}. \tag{3.9}$$

On the other hand, since $\lim_{\overline{r}\to r_0^+} d_H(S, C_{\overline{r}}(S)) = 0$, given $\epsilon > 0$,

$$\exists \overline{r}_\epsilon \text{ such that } d_H(S, C_r(S)) < \epsilon \text{ for all } r \text{ verifying } r_0 < r < \overline{r}_\epsilon. \tag{3.10}$$

According to the Theorem 3.3.7, $\hat{r}_0$ converges to r_0 in probability. In addition, with probability increasing to one, $\hat{r}_0$ is at least as big as r_0, see Theorem 3.3.3. Then, if $\mathcal{R}_n = \{r_0 \leqslant \hat{r}_0 \leqslant \overline{r}_\epsilon\}$, it is verified that $\mathbb{P}(\mathcal{R}_n) \to 1$. Therefore, we have

$$C_{r_0}(\mathcal{X}_n) \subset C_{\hat{r}_0}(\mathcal{X}_n) \subset C_{\overline{r}_\epsilon}(S)$$

and, as consequence,

$$d_H(C_{\hat{r}_0}(\mathcal{X}_n), S) \leqslant \max\{d_H(C_{r_0}(\mathcal{X}_n), S), d_H(C_{\overline{r}_\epsilon}(S), S)\}.$$

Then, if the events $\mathcal{E}_n$ and $\mathcal{R}_n$ hold (notice that $\mathbb{P}(\mathcal{E}_n \cap \mathcal{R}_n) \to 1$) it is enough to take (3.9) and (3.10) into account in order to finish the proof. □

According to the Theorem 3.4.1, $C_{\hat{r}_0}(\mathcal{X}_n)$ is a consistent estimator for the support if Hausdorff or Lebesgue measures are considered when $\lim_{\overline{r}\to r_0^+} d_H(S, C_{\overline{r}}(S)) = 0$. What happens if this continuity of the Hausdorff distance does not hold? We have proved that, with probability increasing to one, $\hat{r}_0 \geqslant r_0$. For instance, if we consider as a support the set represented in Figure 3.18 then, for a large but finite value of n, it is satisfied that $d_H(S, C_r(\mathcal{X}_n)) \geqslant r_0/2 > 0$ whenever $r > r_0$. Of course, this does not imply that $d_H(S, C_{\hat{r}_0}(\mathcal{X}_n)) \geqslant r_0/2$ but the condition $\hat{r}_0 \geqslant r_0$ complicates the proof. This problem can be easily solved by considering $C_{r_n}(\mathcal{X}_n)$ as an estimator with $r_n = \nu \hat{r}_0$ with a fixed $\nu \in (0,1)$, see Theorem 3.4.2.

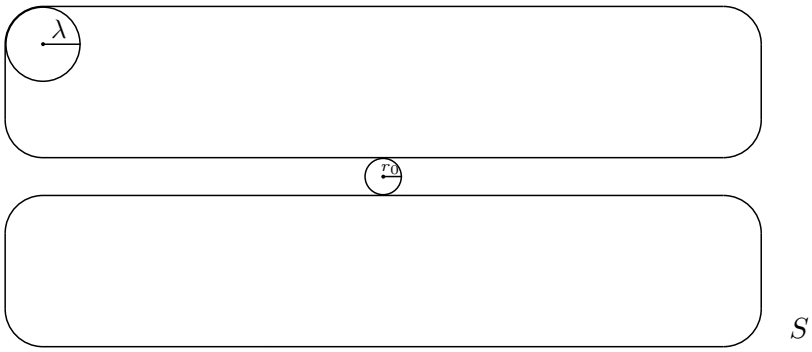


Figure 3.18: A ball of radius r_0 rolls freely in $\overline{S^c}$ and a ball of radius λ rolls freely in S.

Theorem 3.4.2. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in the (3.1) and $\hat{r}_0$ defined in (3.7). Let $\{\alpha_n\} \subset (0,1)$ be a sequence of significance levels converging to zero under the conditions of Theorem 3.3.7. Let be $\nu \in (0,1)$ and $r_n = \nu\hat{r}_0$. Then,*

$$d_H(S, C_{r_n}(\mathcal{X}_n)) = O_P\left(\frac{\log n}{n}\right)^{\frac{2}{d+1}},$$

$$d_H(\partial S, \partial C_{r_n}(\mathcal{X}_n)) = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right)$$

and

$$\mu(S \triangle C_{r_n}(\mathcal{X}_n)) = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right).$$

Proof. For the uniform distribution on S, Theorem 3 of Rodríguez-Casal (2007) ensures that, under $(R^{\widetilde{r}}_{\widetilde{r}})$, then $\mathbb{P}(\mathcal{E}_n) \to 1$, where

$$\mathcal{E}_n = \left\{ d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leqslant A\left(\frac{\log n}{n}\right)^{2/(d+1)} \right\},$$

and A is some constant. Under the hypothesis of Theorem 3.4.2 this holds for any $\widetilde{r} \leqslant \min\{r, \lambda\}$. Fix one $\widetilde{r} \leqslant \min\{r, \lambda\}$ such that $\widetilde{r} < \nu r_0$ and define $\mathcal{R}_n = \{\widetilde{r} \leqslant r_n \leqslant r_0\}$. Since, by Theorem 3.3.7, $r_n = \nu\hat{r}_0$ converges in probability to νr_0 and $\widetilde{r} < \nu r_0 < r_0$, we have that $\mathbb{P}(\mathcal{R}_n) \to 1$. If the events $\mathcal{E}_n$ and $\mathcal{R}_n$ hold (notice that $\mathbb{P}(\mathcal{E}_n \cap \mathcal{R}_n) \to 1$) we have $C_{\widetilde{r}}(\mathcal{X}_n) \subset C_{r_n}(\mathcal{X}_n) \subset S$ and, therefore,

$$d_H(S, C_{r_n}(\mathcal{X}_n)) \leqslant d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leqslant A\left(\frac{\log n}{n}\right)^{2/(d+1)}.$$

This completes the proof of the first statement of Theorem 3.4.2. Similarly, we can proof the result for the other error criteria considered in Theorem 3.4.2. □

Remark 3.4.3. *The selector proposed by Mandal and Murthy (1997), r_n^{MM}, goes to zero in probability. In Pateiro-López and Rodríguez-Casal (2013) it is proved that, for a deterministic sequence of parameters d_n ($d_n \leqslant r_0$ and $d_n^2 n/\log(n) \to \infty$), the convergence rate (in probability) for the distance in measure is, for the bidimensional case, $d_n^{-1/3} n^{-2/3}$. This is the convergence rate of the new proposal plus a penalizing term $d_n^{-1/3}$ which goes to infinity if $d_n \to 0$. It is expected that this penalizing factor, $(r_n^{MM})^{-1/3}$ also appears for the the Mandal and Murthy's proposal.*

3.5 Numerical aspects of the algorithm

The practical implementation of this method requires considering some numerical aspects in order to detail it completely.

For n large enough, the existence of the estimator defined in (3.7) is guaranteed under the hypothesis of Theorem 3.3.7. However, in practise, it could not to exist for a specific sample $\mathcal{X}_n$ and a given value of the significance level α. Therefore, the influence of α must be taken into account. The null hypothesis will be (incorrectly) rejected on $C_r(\mathcal{X}_n)$ for $0 < r \leqslant r_0$ with probability α approximately. This is not important from the theoretical point of view, since we are assuming that $\alpha = \alpha_n$ goes to zero as the sample size increases. But, what to do if, for a given sample, we reject H_0 for *all* r (or at least *all* reasonable values of r)? In order to fix a minimum acceptable value of r, it is assumed that S (and, hence, the estimator) will have no more than C cycles. Too split estimators will not be considered even in the case that we reject H_0 for all r. The minimum value that ensures a number of cycles not greater than C will be taken in this latter case, see below.

Dichotomy algorithms can be used to compute $\hat{r}_0$. The practitioner must select a maximum number of iterations I and two initial points r_m and r_M with $r_m < r_M$ such that the null hypothesis of uniformity is rejected and accepted on $C_{r_M}(\mathcal{X}_n)$ and $C_{r_m}(\mathcal{X}_n)$, respectively. According to the previous comments, it is assumed that the number of cycles of $C_{r_m}(\mathcal{X}_n)$ must not be greater than C. Choosing a value close enough to zero is usually sufficient to select r_m. However, if selecting this r_m is not possible because, for very low and positive values of r, the hypothesis of uniformity is still rejected on $C_r(\mathcal{X}_n)$ then r_0 is estimated as the positive closest value to zero r such that the number of cycles of $C_r(\mathcal{X}_n)$ is smaller than or equal to C. On the other hand, if the hypothesis of uniformity is accepted even on $\text{conv}(\mathcal{X}_n)$ then we propose $\text{conv}(\mathcal{X}_n)$ as the estimator for the support.

To sum up, the next inputs should be given: the significance level $\alpha \in (0,1)$, a maximum number of iterations I, a maximum number of cycles C and two initial values r_m and r_M. Given these parameters $\hat{r}_0$ will be computed as follows:

1. In each iteration and while the number of them is smaller than I:
   (a) $r = (r_m + r_M)/2$.
   (b) If the null hypothesis is not rejected on $C_r(\mathcal{X}_n)$ then $r_m = r$.
   (c) Otherwise, $r_M = r$.
2. Then, $\hat{r}_0 = r_m$.

According to the correction of the bias proposed by Ripley and Rasson (1977) for the convex hull estimator, Berrendero et al. (2012) suggested rejecting the null hypothesis of uniformity when

$$\hat{V}_{n,r} > \frac{\mu(S_n)(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n - v_n},$$

where v_n denotes the number of vertices of $S_n = C_r(\mathcal{X}_n)$ (points of $\mathcal{X}_n$ that belong to ∂S_n). In this work, it is proposed to redefine the critical region as

$$\hat{V}_{n,r} > \hat{c}^*_{n,\alpha,r},$$

where $\hat{c}^*_{n,\alpha,r}$ is equal to

$$\frac{\mu(S_n)(u_\alpha + \log(n - v_n) + (d-1)\log\log(n - v_n) + \log\beta)}{n - v_n},$$

that is, we suggest to replace n by $n - v_n$ in the definition of $\hat{c}_{n,\alpha,r}$ elsewhere not only in the denominator. Although the main theoretical results in Section 3.4 are established in terms of $\hat{c}_{n,\alpha,r}$ instead of $\hat{c}^*_{n,\alpha,r}$, the proofs are completely analogous in both cases since v_n is negligible with respect to n, see the upper bound for the expected number of vertices in Theorem 3 by Pateiro-López and Rodríguez-Casal (2013) in the two-dimensional case or Pateiro-López (2008) for general dimension.

Some technical aspects related to the computation of the maximal spacings must be also considered. Testing the null hypothesis of uniformity is a procedure repeated I times in this algorithm. This may seem to be very computing intensive since the test involves calculating the maximal spacing. Berrendero et al. (2012) found the maximal spacing in two stages. First, based on the Voronoi diagram and Delaunay triangulation of the sample, an initial radius is determined, stored as a candidate to be the maximal spacing. Then, by enlarging this initial value iteratively, they define an

increasing sequence of radius and check whether any of them satisfies the conditions to define the maximal spacing. However, it is not necessary to know the exact value of the maximal spacing. In fact, in order to perform the test, it is only necessary to check if, for a fixed r, $C_r(\mathcal{X}_n)$ contains an open ball, that does not intersect the sample points with volume greater than the test's critical value $\hat{c}^*_{n,\alpha,r}$. For instance, for the two-dimensional case, we will simply check if an open ball of radius equal to $\sqrt{\hat{c}^*_{n,\alpha,r}/\pi}$ and center x is contained in $C_r(\mathcal{X}_n)\backslash\mathcal{X}_n$. If this disc exists then $x \notin B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$ where

$$B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n) = \bigcup_{X_i \in \mathcal{X}_n} B_{\hat{c}^*_{n,\alpha,r}}(X_i)$$

is the dilation of radius $\hat{c}^*_{n,\alpha,r}$ of the sample. Therefore, the centers of the possible maximal balls necessarily lie outside $B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$. Following Berrendero et al. (2012), to check if the null hypothesis of uniformity is rejected on $C_r(\mathcal{X}_n)$, we will follow the following steps:

1. Determine the set $D(r) = C_r(\mathcal{X}_n) \cap \partial B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$. Notice that, if $x \in D(r)$ then $B_{\hat{c}^*_{n,\alpha,r}}(x) \cap \mathcal{X}_n = \varnothing$.

2. Calculate $M(r) = \max\{d(x, \partial C_r(\mathcal{X}_n) : x \in D(r)\}$.

3. If $M(r) \leqslant \hat{c}^*_{n,\alpha,r}$ then the null hypothesis of uniformity is not rejected.

It should be noted that $\partial C_r(\mathcal{X}_n)$ and $\partial B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$ can be easily computed (at least for the bidimensional case), see Pateiro-López and Rodríguez-Casal (2010).

3.6 A comparative simulation study

The performances of the algorithm proposed in this paper and Mandal and Murthy (1997)'s method will be analyzed in this section. They will be denoted by RS and MM, respectively. A total of 1000 uniform samples of four different sample sizes n have been generated on three support models in the Euclidean space $\mathbb{R}^2$, see Section 1.4.2.

The first set, $S = B_{0.35}[(0.5, 0.5)]\backslash B_{0.15}((0.5, 0.5))$, is a circular ring with $r_0 = 0.15$. The other two ones are two interesting sets, $S = \mathbf{C}$ and $S = \mathbf{S}$ with $r_0 = 0.2$ and $r_0 = 0.0353$, respectively. The values of n considered are $n = 100$, $n = 500$, $n = 1000$ and $n = 1500$. In addition, four values for α have been taken into account, $\alpha_1 = 10^{-1}$, $\alpha_2 = 10^{-2}$, $\alpha_3 = 10^{-3}$ and $\alpha_4 = 10^{-4}$. The maximum number of cycles C was fixed arbitrarily in all the simulations equal to 4.

For each fixed random sample, both estimators of the smoothing parameter of the $r-$convex hull have been calculated. So, one thousand estimations have been obtained

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.1592	0.1456	0.1438	0.1410
	$\alpha_2 = 10^{-2}$	0.1592	0.1509	0.1499	0.1495
	$\alpha_3 = 10^{-3}$	0.1592	0.1516	0.1507	0.1503
	$\alpha_4 = 10^{-4}$	0.1592	0.1517	0.1507	0.1504
MM		0.1969	0.1295	0.1084	0.0977

Table 3.1: Empirical means of 1000 RS and MM estimations for the smoothing parameter of the $r-$convex hull with $S = B_{0.35}[(0.5, 0.5)] \backslash B_{0.15}((0.5, 0.5))$. In this case, $r_0 = 0.15$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.2724	0.2007	0.1903	0.1888
	$\alpha_2 = 10^{-2}$	0.2929	0.2150	0.2056	0.2032
	$\alpha_3 = 10^{-3}$	0.2982	0.2188	0.2089	0.2055
	$\alpha_4 = 10^{-4}$	0.2988	0.2226	0.2105	0.2068
MM		0.1636	0.1072	0.0897	0.0809

Table 3.2: Empirical means of 1000 RS and MM estimations for the smoothing parameter of the $r-$convex hull with $S = \mathbf{C}$. In this case, $r_0 = 0.2$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.0954	0.0833	0.0637	0.0548
	$\alpha_2 = 10^{-2}$	0.0954	0.0878	0.0695	0.0602
	$\alpha_3 = 10^{-3}$	0.0958	0.0886	0.0736	0.0631
	$\alpha_4 = 10^{-4}$	0.1077	0.0887	0.0778	0.0659
MM		0.1644	0.1055	0.088	0.0792

Table 3.3: Empirical means of 1000 RS and MM estimations for the smoothing parameter of the $r-$convex hull with $S = \mathbf{S}$. In this case, $r_0 = 0.0353$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.9288	0.3293	0.2085	0.1623
	$\alpha_2 = 10^{-2}$	0.9288	0.3143	0.1970	0.1492
	$\alpha_3 = 10^{-3}$	0.9294	0.3123	0.1957	0.1484
	$\alpha_4 = 10^{-4}$	0.9288	0.3122	0.1957	0.1483
MM		1.4165	0.3378	0.2316	0.1837
		0.9337	0.2956	0.1819	0.1364

Table 3.4: Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = B_{0.35}[(0.5, 0.5)] \backslash B_{0.15}((0.5, 0.5))$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.6041	0.1472	0.0920	0.0712
	$\alpha_2 = 10^{-2}$	0.6677	0.1589	0.0833	0.0640
	$\alpha_3 = 10^{-3}$	0.6820	0.1953	0.0832	0.0631
	$\alpha_4 = 10^{-4}$	0.6837	0.2440	0.0865	0.0626
MM		0.4145	0.1681	0.1125	0.0885
		0.3727	0.1277	0.0800	0.0606

Table 3.5: Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = \mathbf{C}$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.6389	0.2591	0.1842	0.1485
	$\alpha_2 = 10^{-2}$	0.6389	0.2537	0.1821	0.1455
	$\alpha_3 = 10^{-3}$	0.6411	0.2530	0.1821	0.1464
	$\alpha_4 = 10^{-4}$	0.6797	0.2529	0.1816	0.1476
MM		1.2319	0.4851	0.2445	0.1514
		1.0794	0.3320	0.2038	0.1541

Table 3.6: Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = \mathbf{S}$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

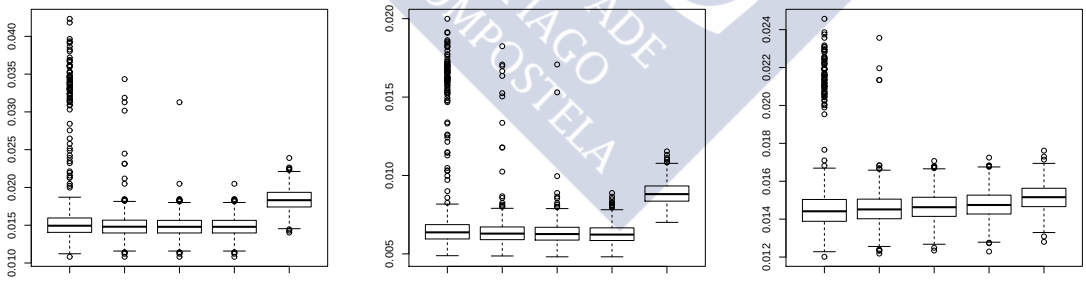


Figure 3.19: Boxplots of the estimations for the distance in measure for RS and MM methods when $n = 1500$ for $S = B_{0.35}[(0.5, 0.5)] \backslash B_{0.15}((0.5, 0.5))$ (right), $S = \mathbf{C}$ (center) and $S = \mathbf{S}$ (left). From left to right, RS considering α_1, α_2, α_3 and α_4 and MM.

for each algorithm, fixed a model and the values of n and α. The empirical means of these one thousand estimations are showed in Tables 3.1, 3.2 and 3.3 for the RS and MM methods. We should mention that MM method is included in these table only for illustrative purposes. The results of these two algorithms are not directly comparable since the goal of MM is not to estimate the parameter r_0 defined in (3.1). On the other hand, Tables 3.4, 3.5 and 3.6 contain the empirical means of one thousand Monte Carlo estimations for the distance in measure between the RS and MM support estimators and the corresponding theoretical models, respectively. In addition, we have also estimated the distance in measure between the r_0-convex hull of each sample and its corresponding support model for the different sample sizes. The means of these estimations can be considered as the benchmark. They are showed (multiplied by 10) in the last row of Tables 3.4, 3.5 and 3.6. A grid of 334^2 points was considered in the unit square for estimating the distance in measure. The parameter ν was fixed equal to 0.95 for calculating the RS support estimator.

In order to asses the variability on the estimation, Figure 3.19 contains the boxplots for the distance in measure between the resulting support estimators for the RS and MM methods when $n = 1500$.

Conclusions. According to the results showed in Tables 3.1, 3.2 and 3.3, RS presents a good global behavior for estimating the smoothing parameter r_0. Only when $S = \mathbf{C}$ and $n = 100$, MM provides better results, see Table 3.2. In this particular case, the estimations of RS are specially greater than 0.2, the real value of parameter r_0. In general, MM provides too small estimations, mainly for high values of the sample size, see Tables 3.1 and 3.2.

The role of the level of significance α must be also discussed. Taking low values of α reduces the number of outliers considerably for the three support models presented. In addition, if the model considered is not too complex then small values of α provide slightly better results for n large enough reducing the risk of rejecting the null hypothesis of uniformity when it is satisfied, see for instance $S = B_{0.35}[(0.5, 0.5)]\backslash B_{0.15}((0.5, 0.5))$ or $S = \mathbf{C}$ in Tables 3.1 and 3.2. Therefore, excessively low values of r will not be selected. However, if the support model is not so simple then choosing large values of α provides better estimations for the smoothing parameter, see Table 3.3 for $S = \mathbf{S}$. Anyway, for moderate and large values of the sample size the dependence on α of RS method is small.

Finally and according to the Tables 3.4, 3.5 and 3.6, RS always provides the smallest estimation errors for the criteria considered except when $S = \mathbf{C}$ with $n = 100$ or even $n = 500$ if the value of α is too small, see Table 3.5. Therefore, RS support estimator is more competitive than MM algorithm. According to the previous comments, it can

be seen that the number of outliers for RS increases if large values of α are considered for the three support models, see Figure 3.19.

3.7 A real example

In order to assess the applicability of our estimation method to real examples, we have analyzed the two Aral Sea's images presented in Section 1.4. These two photographs show the Aral Sea in 2011 and 2000, respectively.

In the example considered the goal is to test if water area is decreasing in the two regions in Figure 1.16 (center). The alternative hypothesis, H_1, is Aral Sea's water is decreasing. So, if A_{2000} and A_{2011} denotes the water areas in 2000 and 2011, respectively then we can write:

$$H_0 : A_{2000} \leqslant A_{2011} \text{ versus } H_1 : A_{2000} > A_{2011}.$$

Or equivalently,

$$H_0 : A_{2000}/A_{2011} \leqslant 1 \text{ versus } H_1 : A_{2000}/A_{2011} > 1.$$

By using the discrimination method, we have constructed two uniform samples of size n on the two water regions in Figure 1.16 (center) denoted by $\mathcal{X}_{n,2000}$ and $\mathcal{X}_{n,2011}$, respectively. For each one of these two samples, the method proposed in this work is used to estimate r_0. The values obtained are denoted by $\hat{r}_{0,2000}$ and $\hat{r}_{0,2011}$, respectively. We have measure the difference between the areas of $C_{\hat{r}_{0,2000}}(\mathcal{X}_{n,2000})$ and $C_{\hat{r}_{0,2011}}(\mathcal{X}_{n,2011})$,

$$T = \mu(C_{\hat{r}_{0,2000}}(\mathcal{X}_{n,2000}))/\mu(C_{\hat{r}_{0,2011}}(\mathcal{X}_{n,2011})).$$

To decide if the null hypothesis is or not rejected, the considered test statistic T should be calibrated and estimated the critical value, CV, under the null hypothesis. The next procedure is proposed:

1. First, a new sample $\mathcal{X}_{2n}$ is defined as the union of $\mathcal{X}_{n,2000}$ and $\mathcal{X}_{n,2011}$.
2. The algorithm presented in this work is used for estimating $\hat{r}_{0,2n}$ from $\mathcal{X}_{2n}$. Then, $S^* = C_{\hat{r}_{0,2n}}(\mathcal{X}_{2n})$ will be an auxiliary support.
3. This point must be repeated $G = 1000$ times:
   (a) Generate two uniform samples on S^* of size n, $\mathcal{X}^*_{n,1}$ and $\mathcal{X}^*_{n,2}$.
   (b) Estimate r_0 by using the proposed method from $\mathcal{X}^*_{n,1}$ and $\mathcal{X}^*_{n,2}$. We denote the estimations by $\hat{r}^*_{0,1}$ and $\hat{r}^*_{0,2}$, respectively.
   (c) Compute $T^* = \mu(C_{\hat{r}^*_{0,1}}(\mathcal{X}^*_{n,1}))/\mu(C_{\hat{r}^*_{0,2}}(\mathcal{X}^*_{n,2}))$.

4. If T is greater than CV (95%-quantile of G values T^* calculated in the previous step) then the null hypothesis will be rejected. The $P-$value could be approximate by calculating the average of bootstrap values which are greater than T.

Tables 3.7, 3.8 and 3.9 contain the results obtained by considering the previous algorithm for three different values of n: $n = 1000$, $n = 2000$ and $n = 3000$.

T	CV	$p-$value
2.91	1.02	0

Table 3.7: Values of T, critical value estimated and the $p-$value approximated for Aral Sea with $n = 1000$.

T	CV	$p-$value
2.85	1.01	0

Table 3.8: Values of T, critical value estimated and the $p-$value approximated for Aral Sea with $n = 2000$.

T	CV	$p-$value
2.94	1.01	0

Table 3.9: Values of T, critical value estimated and the $p-$value approximated for Aral Sea with $n = 3000$.

According to the previous results, in these three cases, the null hypothesis is rejected with level of significance $\alpha = 0.05$. So, the Aral Sea is losing water. In fact, the value of T suggests that the water area in 2000 is around three times greater than the water area in 2011.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

A new data-driven method for estimating density level sets

Once proposed a new automatic algorithm for reconstructing the support under the assumption of $r-$convexity, we now turn our attention to the problem of estimating $r-$convex density level sets. Just as the support case, unknowing the value of the parameter r will be the main problem. In this chapter an algorithm to estimate it will be presented. As consequence, an estimator for density level sets will be proposed too. In this way, the proposal in Section 2.2.3.2 will be improved by introducing slight modifications. This problem, as far as we know, have not been previously pointed out in the literature.

This chapter is organized as follows. The problem of reconstructing a $r-$convex density level set is reviewed in Section 4.1. As we have told before, the main disadvantage of using the $r-$convex hull as an estimator is the selection of the smoothing parameter r. The optimal parameter is defined and an estimator for it is established in Sections 4.2 and 4.3, respectively. The consistency of this new estimator is proved in Section 4.4. In Section 4.5, the resulting density level set estimator is presented. Its consistency and convergence rates will be showed. Here, a serie of theoretical results in Walther (1997) contained in Appendix B will be used. These sections only consider the theoretical results for estimating level sets $G(t)$ defined in (1.1). However, from a practical point of view, estimating level sets $L(\tau)$ defined in (1.2) can be more interesting. Therefore, a practical implementation of the method proposed has been designed for reconstructing $L(\tau)$. It requires considering some numerical aspects detailed in Section 4.6. This chapter closes with a real data example in Section 4.7. The performance of the new algorithm will be illustrated by comparing the distribution of controls and cases in the leukaemia data set described in Section 1.4.1.

4.1 Preliminaries

Level set estimation theory deals with the problem of reconstructing an unknown set of type $G(t)$ or $L(\tau)$ defined in (1.1) and (1.2), respectively given a random sample of points $\mathcal{X}_n$ generated from a distribution with density function f. A new algorithm for estimating levels sets under $r-$convexity assumption was proposed in Section 2.2.3.2. It provides quite good results, see Sections 2.3.3 or 2.3.4. This method divides the sample points into two subsamples using the information of the kernel estimator f_n defined in (1.4). One of these subsamples contains the sample points such that f_n evaluated on them is equal to or greater than the threshold. These points likely belong to the theoretical level set $L(\tau)$. Therefore, the resulting estimator for the level set is constructed as the $r-$convex hull of this subset of sample points. The main disadvantage of this algorithm is the unknowing of the parameter r and, as we have showed in Section 2.2.3.2, its influence may be significant. The main goal of this chapter is to present a data-driven method to estimate the smoothing parameter. As consequence, a new density level set estimator is proposed under the assumption of $r-$convexity. Some slight modifications on the construction of the two subsamples on the original method of $r-$convex hull have been introduced in order to get theoretical guarantees on the performance of the method.

4.2 Defining the optimal parameter

In the same way that support estimation, the first step is to determinate the optimal value of the smoothing parameter to be estimated. Again, we are interested in estimating the greatest value of $r > 0$ such that $G(t)$ is $r-$convex. Its optimality can be ensured using the same reasonings considered in Section 3.2 for the support. For simplicity in the exposition, we will assume that $G(t)$ is not convex in order to guarantee that the set $\{\gamma > 0 : C_\gamma(G(t)) = G(t)\}$ is upper bounded. Notice that, in this case, the parameter depends on the level $t > 0$ considered, see Figure 4.1.

Definition 4.2.1. *Let $G(t)$ be a compact, nonempty, nonconvex and $r-$convex level set for some $r > 0$. It is defined*

$$r_0(t) = \sup\{\gamma > 0 : C_\gamma(G(t)) = G(t)\}. \tag{4.1}$$

The following geometric property has been assumed on the level set $G(t)$:

(R_λ^r) A closed ball of radius $\lambda > 0$ rolls freely in $G(t)$ and a closed ball of radius $r > 0$ rolls freely in $\overline{G(t)^c}$.

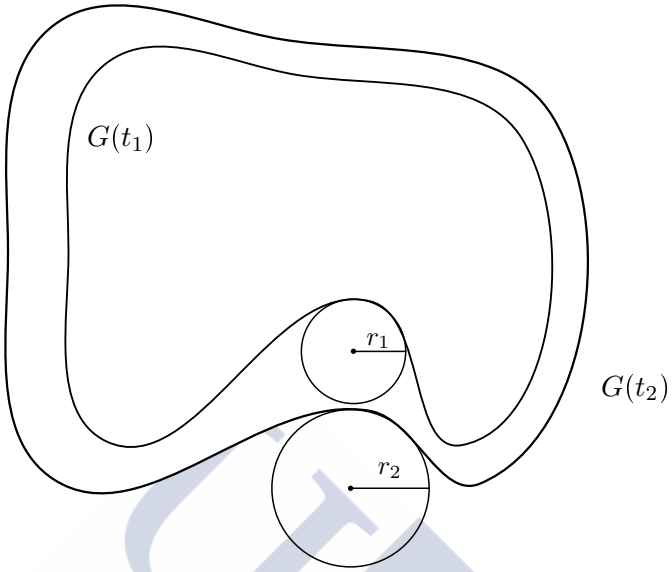


Figure 4.1: Level sets $G(t_i)$, with $i = 1, 2$ and t_1 greater than t_2. In addition, $C_{r_i}(G(t_i)) = G(t_i)$ with r_i denoting $r_0(t_i)$ for $i = 1, 2$. In this case, $r_2 > r_1$.

The rolling freely property was analyzed in depth in Definition 1.2.10. Satisfying the shape condition (R^r_λ) is a quite natural general property for level sets of densities. In fact, in Theorem 2 by Walther (1997) was proved that, under some assumptions on the density f, its level sets satisfy the conditions in Theorem 1.2.11 for $r = \lambda = m/k$, see below. Then, according to Theorem 2 in Walther (1997), the following assumptions are considered on f:

A.
1. The threshold t of $G(t)$ belongs to (l, u) with $-\infty < l \leqslant u < \sup(f)$.
2. $f \in \mathcal{C}^p(U)$, $p \geqslant 1$ where U is a bounded open set containing $\overline{G(l-\zeta)} \setminus \operatorname{Int}(G(u+\zeta))$ for some $\zeta > 0$ where $G(u+\zeta)$ is bounded, see Figure 4.2.
3. The gradient of f, ∇f, satisfies $|\nabla f| \geqslant m > 0$ as well as Lipschitz condition on U:
$$|\nabla f(x) - \nabla f(y)| \leqslant k|x - y| \text{ for } x, y \in U.$$

Under (A), it is verified that $r_0(t) \geqslant m/k$. In addition, the consideration of the shape condition (R^r_λ) has allowed us to adapt some useful and necessary properties of the support to the context of density level set estimation. For instance, the $r-$convexity of the level set $G(t)$ is guaranteed under (R^r_λ), see Proposition 4.2.2. On the other hand, the balls of radius r and radius λ that roll freely in $\overline{G(t)^c}$ and $G(t)$, respectively, under

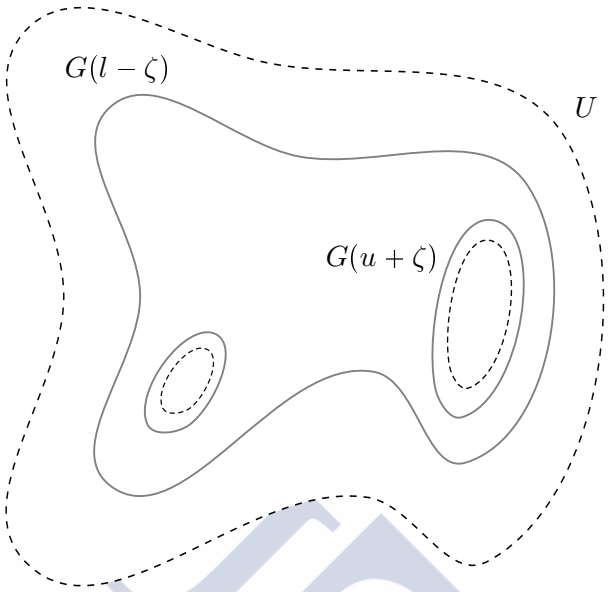


Figure 4.2: $G(u+\zeta)$ and $G(l-\zeta)$ in gray. The open set U in dashed line.

(R_λ^r) have been characterized, see Lemma 4.2.3. The proofs of these two results are not showed here since they are a straightforward consequence of Proposition 3.2.10 and Lema 3.2.8 in Chapter 3, respectively. The proof of Proposition 4.2.4 is similar to that Proposition 3.3.4 and we will skip the details here.

Proposition 4.2.2. *Let $G(t) \subset \mathbb{R}^d$ be a compact and nonempty level set verifying (R_λ^r). Then, $G(t)$ is $r-$convex.*

Lemma 4.2.3. *Let $G(t) \subset \mathbb{R}^d$ be a closed level set verifying (R_λ^r). Then, for each $x_t \in \partial G(t)$ there exists a unique unit vector $\eta(x_t)$ such that*

$$B_\lambda(x_t - \lambda\eta(x_t)) \subset G(t) \text{ and } B_r(x_t + r\eta(x_t)) \subset \overline{G(t)^c}.$$

Proposition 4.2.4. *Let $G(t) \subset \mathbb{R}^d$ be a compact and nonempty level set verifying (R_λ^r) and $\gamma > 0$ such that $G(t) \subsetneq C_\gamma(G(t))$. Then, there exists $x_t \in \mathrm{Int}(C_\gamma(G(t))) \cap \partial G(t)$.*

4.3 Defining the estimator for the smoothing parameter

According to the previous comments, the method of the $r-$convex hull proposed in Section 2.2.3.2 divides the original sample $\mathcal{X}_n$ into two subsamples. The estimator for the density level set is constructed from the sample points where the density estimator is greater than or equal to the threshold. Therefore, it takes into account the information contained only in one of the two subsamples. Then, the information about the complement of the level set $G(t)$ is not taken advantage. Our proposal here will solve this problem by modifying slightly the original algorithm in Section 2.2.3. First, an estimator for the parameter defined in (4.1) will be proposed. Its definition depends on a sequence D_n satisfying the assumption:

D. D_n is equal to $M(\log n/n)^{p/(d+2p)}$ for a big enough value of the constant $M > 0$.

Definition 4.3.1. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A) and (D), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f. An estimator for the parameter defined in (4.1) can be defined as*

$$\hat{r}_0(t) = \sup\{\gamma > 0 : C_\gamma(\mathcal{X}_n^+(t)) \cap \mathcal{X}_n^-(t) = \varnothing\}, \tag{4.2}$$

where

$$\mathcal{X}_n^+(t) = \{X \in \mathcal{X}_n : f_n(X) \geqslant t + D_n\} \text{ and } \mathcal{X}_n^-(t) = \{X \in \mathcal{X}_n : f_n(X) < t - D_n\}.$$

The original sample $\mathcal{X}_n$ is divided into three subsamples, $\mathcal{X}_n^+(t)$, $\mathcal{X}_n^-(t)$ and $\mathcal{X}_n \backslash (\mathcal{X}_n^+(t) \cup \mathcal{X}_n^-(t))$. From an intuitive point of view, $\mathcal{X}_n^+(t)$ and $\mathcal{X}_n^-(t)$ should be contained in $G(t)$ and its complementary, respectively. This property is proved in Lemma 4.3.2, even for convex sets. According to Definition 4.3.1, we have assumed that $G(t)$ is not convex only for simplicity in the exposition. If $G(t)$ is convex then $\hat{r}_0(t) = \infty$ and, therefore, the convex hull of sample points, $\text{conv}(\mathcal{X}_n^+(t))$, would reconstruct the level set $G(t)$. In addition, Lemma 4.3.3 ensures that $\mathcal{X}_n^+(t) \neq \varnothing$. If $G(t)$ is nonconvex then it can be seen that, with probability one and for n large enough, the set $\{\gamma > 0 : C_\gamma(\mathcal{X}_n^+(t)) \cap \mathcal{X}_n^-(t) = \varnothing\}$ is nonempty and upper bounded. So, the estimator proposed in (4.2) is well-defined. In order to guarantee that the estimator satisfies these interesting and natural properties, two conditions on the kernel estimator f_n of f must be considered, see again Walther (1997) for more details:

K.
1. The kernel function K is a continuous kernel of order at least p with bounded support and finite variation.
2. The bandwidth parameter is of the order $(\log n/n)^{1/(d+2p)}$.

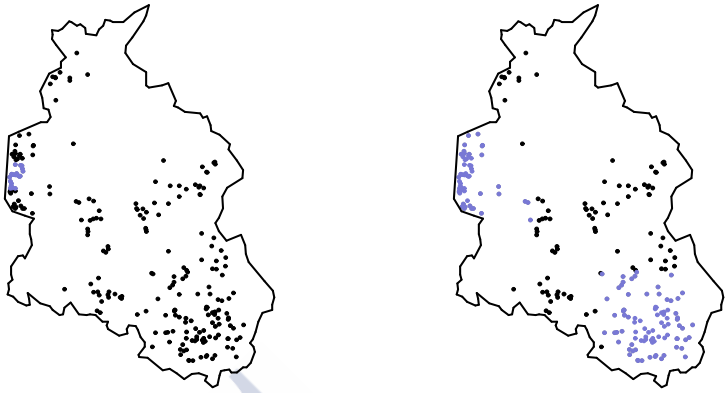


Figure 4.3: The set $\mathcal{X}_n^+(t_i)$ is represented in blue for $t = t_i$, $i = 1, 2$ with t_1 (left) greater than t_2 (right).

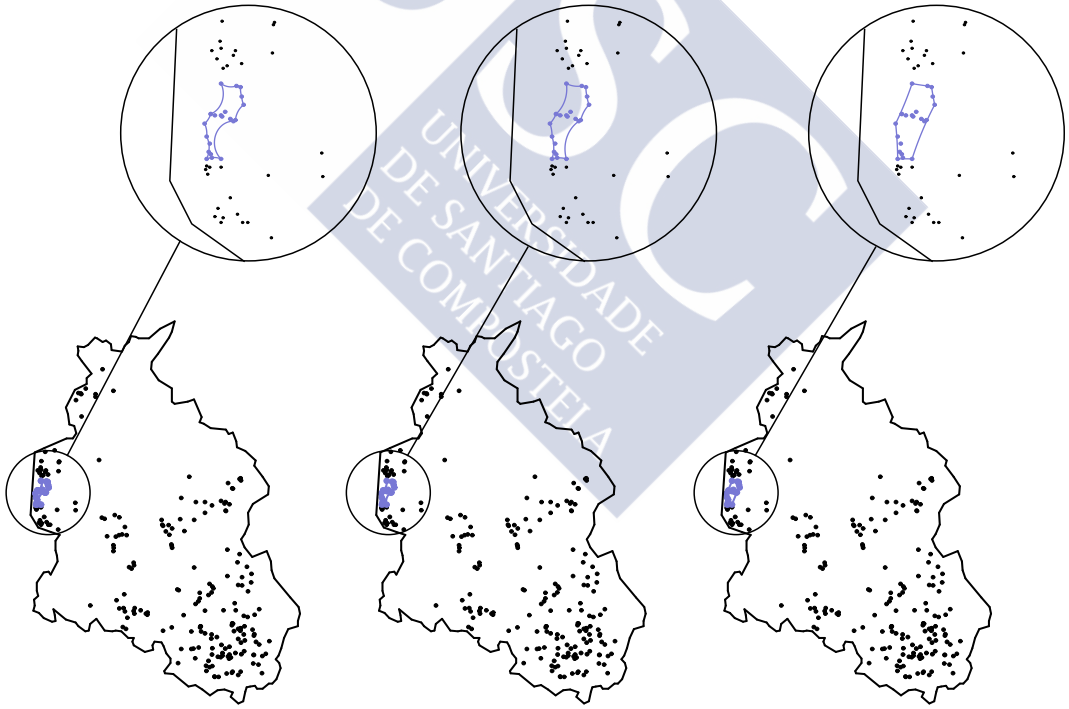


Figure 4.4: The set $\mathcal{X}_n^+(t_1)$ is represented in blue. $C_{0.02}(\mathcal{X}_n^+(t_1))$ (left), $C_{0.03}(\mathcal{X}_n^+(t_1))$ (center) and $C_{0.3}(\mathcal{X}_n^+(t_1))$ (right).

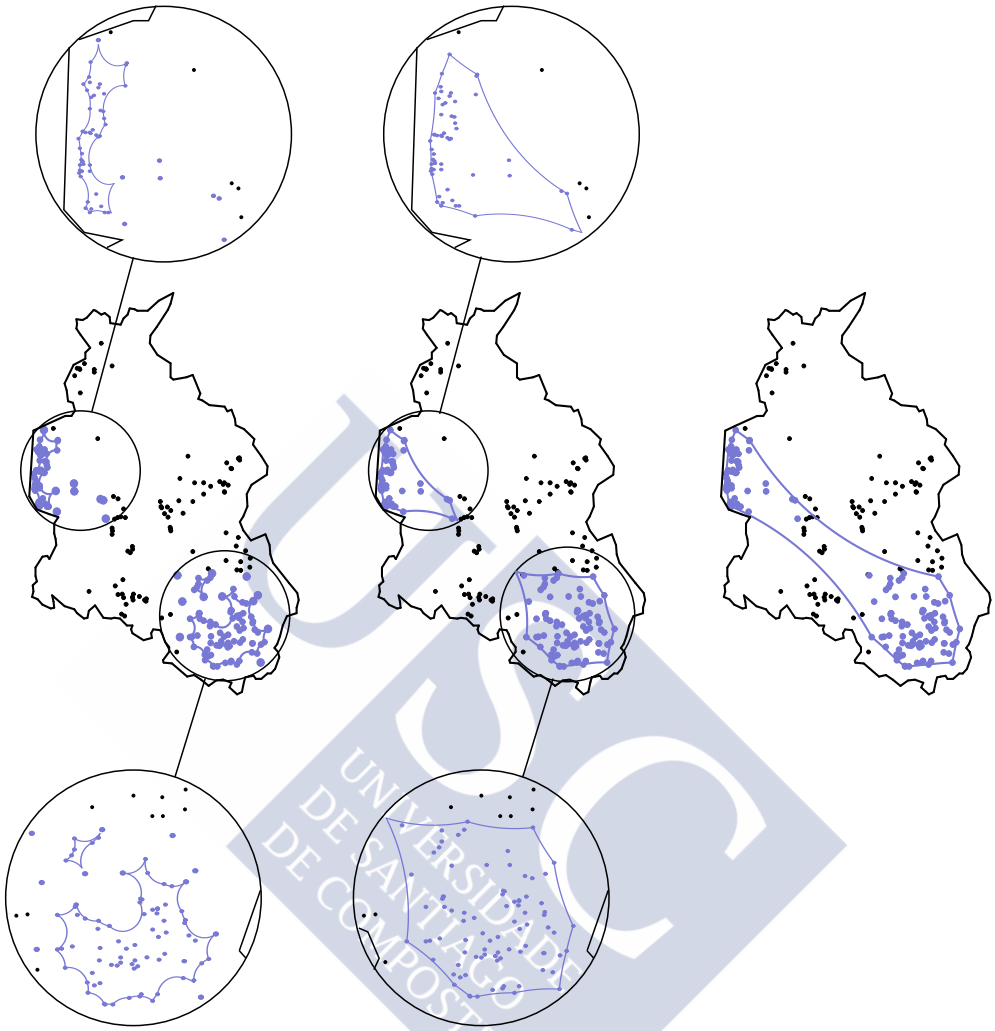


Figure 4.5: The set $\mathcal{X}_n^+(t_2)$ is represented in blue. $C_{0.03}(\mathcal{X}_n^+(t_2))$ (left), $C_{0.3}(\mathcal{X}_n^+(t_2))$ (center) and $C_{0.9}(\mathcal{X}_n^+(t_2))$ (right).

In Figure 4.3, we show $\mathcal{X}_n^+(t)$ in blue for the data corresponding to 322 cases of diagnosed of leukaemia on the North West of England by considering two different values of the parameter $t > 0$. In practise, the role of the sequence D_n must be taken into account. The procedure used for calculating D_n and hence $\mathcal{X}_n^+(t)$ and $\mathcal{X}_n^-(t)$ is explained in depth in Section 4.6. In these two cases, the $r-$convex hulls of $\mathcal{X}_n^+(t)$ are represented for different values of the parameter r in Figures 4.4 and 4.5, respectively. It is clear, see Figure 4.5, that the influence of the parameter r is important for estimating $G(t)$. Reconstructing it in a data-driven way is necessary.

Lemma 4.3.2. *Let $G(t)$ be a compact and nonempty level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f and $\mathcal{X}_n^+(t)$ and $\mathcal{X}_n^-(t)$ established in Definition 4.3.1. Then,*

$$\mathbb{P}\left(\mathcal{X}_n^+(t) \subset G(t),\ \mathcal{X}_n^-(t) \subset G(t)^c,\ \textit{eventually}\right) = 1.$$

Proof. First, we will prove that,

$$\mathbb{P}(\mathcal{X}_n^+(t) \subset G(t),\ \text{eventually}) = 1.$$

For this, it is enough to prove

$$\mathbb{P}\left(\sup_{z \in G(t)^c} f_n(z) < t + M\left(\frac{\log n}{n}\right)^{p/(d+2p)},\ \text{eventually}\right) = 1. \tag{4.3}$$

Then, let $z \in G(t)^c$ and C be the compact set defined in Proposition B.0.2. Two cases are considered: $z \in C$ or $z \in C^c$.

1. Let $z \in C^c$. Since $z \in G(t)^c$ then $z \notin G(l)$ because $G(l) \setminus \operatorname{Int}(G(u)) \subset C$. Therefore, according to Proposition B.0.3, with probability one and for n large enough,

   $$f_n(z) \leqslant \sup_{y \in G(l)^c \cap C^c} f_n(y) < l - \frac{w}{2} < l,$$

   where w denotes a positive constant. Therefore,

   $$\mathbb{P}\left(\sup_{z \in G(t)^c \cap C^c} f_n(z) < l,\ \text{eventually}\right) = 1,$$

   and since $l < t + D_n$ for all $t \in (l, u)$,

   $$\mathbb{P}\left(\sup_{z \in G(t)^c \cap C^c} f_n(z) < t + D_n,\ \text{eventually}\right) = 1.$$

2. Let $z \in C$. According to Proposition B.0.2 we can guarantee that,

   $$\sup_C |f_n - f| = O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right),\ \text{almost surely.}$$

   So, there exists $N > 0$ such that

   $$\sup_C |f_n - f| \leqslant N\left(\frac{\log n}{n}\right)^{p/(d+2p)},\ \text{almost surely.} \tag{4.4}$$

Since $z \notin G(t)$ then $f(z) < t$. Taking into account (4.4), for n large enough, it is verified that

$$f_n(z) \leqslant |f_n(z) - f(z)| + f(z) < |f_n(z) - f(z)| + t \leqslant N \left(\frac{\log n}{n}\right)^{p/(d+2p)} + t.$$

If $M \geqslant N$,

$$f_n(z) < t + M \left(\frac{\log n}{n}\right)^{p/(d+2p)} = t + D_n, \text{ almost surely.}$$

This concludes the proof of (4.3).
Similarly, we will prove that,

$$\mathbb{P}(\mathcal{X}_n^-(t) \subset G(t)^c, \text{ eventually}) = 1.$$

For this, it is enough to prove

$$\mathbb{P}\left(\inf_{z \in G(t)} f_n(z) \geqslant t - M \left(\frac{\log n}{n}\right)^{p/(d+2p)}, \text{ eventually}\right) = 1. \tag{4.5}$$

Let $z \in G(t)$. Again, two cases are considered: $z \in C$ or $z \in C^c$.

1. Let $z \in C^c$. Then, $z \notin (G(l) \setminus \operatorname{Int}(G(u)))$. But $z \in G(t) \subset G(l)$. Therefore, $z \in \operatorname{Int}(G(u))$ and, as consequence, $f(z) > u$. According to Proposition B.0.3, with probability one,

$$f_n(z) \geqslant \inf_{y \in G(u) \cap C^c} f_n(y) > u + \frac{w}{2} > u,$$

where w denotes a positive constant. Therefore,

$$\mathbb{P}\left(\inf_{z \in G(t) \cap C^c} f_n(z) > u, \text{ eventually}\right) = 1,$$

and since $t - D_n < u$ for all $t \in (l, u)$,

$$\mathbb{P}\left(\inf_{z \in G(t) \cap C^c} f_n(z) > t - D_n, \text{ eventually}\right) = 1.$$

2. Let $z \in C$. Since $z \in G(t)$ then $f(z) \geqslant t$. Taking into account (4.4),

$$f_n(z) \geqslant f(z) - |f_n(z) - f(z)| \geqslant t - |f_n(z) - f(z)|$$
$$\geqslant t - N \left(\frac{\log n}{n}\right)^{p/(d+2p)}, \text{ almost surely.}$$

If $M \geqslant N$,

$$f_n(z) \geqslant t - M \left(\frac{\log n}{n}\right)^{p/(d+2p)} = t + D_n, \text{ almost surely.}$$

This concludes the proof of (4.5). The lemma is a straightforward consequence of (4.3) and (4.5). □

Lemma 4.3.3 bounds the Euclidian distance between $G(t)$ and $\mathcal{X}_n^+(t)$ guaranteeing, in particular, that the set $\mathcal{X}_n^+(t)$ is nonempty eventually, see Figure 4.6.

Lemma 4.3.3. *Let $G(t)$ be a compact and nonempty level set. Under assumptions (A) and (D), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f and $\mathcal{X}_n^+(t)$ established in Definition 4.3.1. Then, for all $\epsilon > 0$ it is verified that*

$$\mathbb{P}\left(\sup_{x \in G(t)} d(x, \mathcal{X}_n^+(t)) \leqslant \epsilon, \textit{eventually}\right) = 1.$$

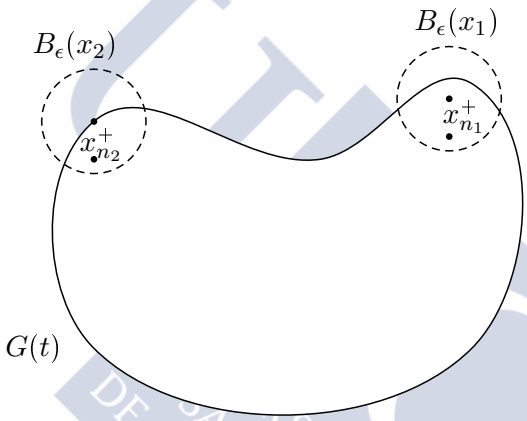


Figure 4.6: $x_1,\ x_2 \in G(t)$ and $x_{n_i}^+ \in B_\epsilon(x_i),\ i \in \{1, 2\}$.

Proof. Let $\epsilon > 0$. It is clear that it is enough to show the result for a value of ϵ small enough. The followings steps complete the proof:

1. Let $x \in G(t)$. Under (A), a ball of radius m/k rolls freely in $G(t)$ and $\overline{G(t)^c}$. According to Lemma 1 in Arias-Castro and Rodríguez-Casal (2014), if $\epsilon \leqslant m/k$,

$$\exists B_{\frac{\epsilon}{2}}(y) \subset B_\epsilon(x) \text{ such that } B_{\frac{\epsilon}{2}}(y) \subset G(t).$$

We define $B_t^x = B_{\epsilon/4}(y)$. Obviously, $B_t^x \subset G(t)$. In addition, it verifies that

$$B_t^x \subset G(t) \ominus \frac{\epsilon}{4} B_1[0]$$

since, for all $z \in B_t^x$, $z + (\epsilon/4)B_1[0] \subset B_{\epsilon/2}(y) \subset G(t)$. On the other hand, considering Proposition B.0.4 for ϵ small enough and $T = \epsilon m/8$,

$$G(t) \ominus \frac{\epsilon}{4} B_1[0] \subset G(t + T).$$

Therefore, see Figure 4.7,

$$B_t^x \subset G(t) \ominus \frac{\epsilon}{4} B_1[0] \subset G(t+T).$$

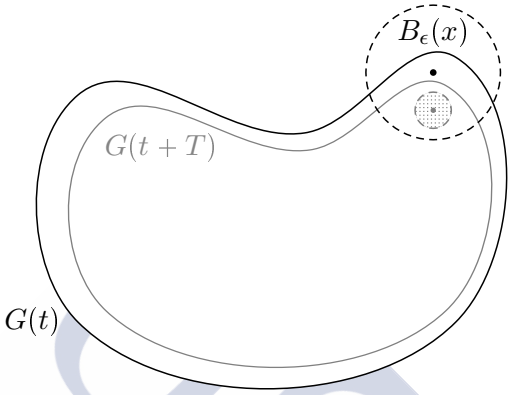


Figure 4.7: Elements in proof of Lemma 4.3.3. $G(t)$ in black, $G(t+T)$ in gray, $B_\epsilon(x)$ in black and B_t^x in gray.

2. Let $\mathcal{F} = \{B_t^x : \ x \in G(t)\}$. Under (A), the level set $G(t)$ is bounded since $G(u+\zeta)$ is bounded and $\overline{G(l-\zeta)} \setminus \text{Int}(G(u+\zeta)) \subset U$ where U is a bounded set too. As consequence, $\overline{G(l-\zeta)}$ is bounded and, therefore, $G(t) \subset G(l) \subset \overline{G(l-\zeta)}$ too. Then, there exists a finite cover for $G(t)$ of balls of radius, for instance, $\epsilon/10$. Therefore, there exists $z_1, \cdots, z_s \in G(t)$ such that

$$G(t) \subset \bigcup_{i=1}^{s} B_{\frac{\epsilon}{10}}(z_i).$$

Then, for all $B_t^x = B_{\epsilon/4}(y) \in \mathcal{F}$ where $y \in G(t)$,

$$\exists z_j \in \{z_1, \cdots, z_s\} \text{ such that } \|z_j - y\| < \frac{\epsilon}{10}.$$

Next, we will prove that the ball $B_{\epsilon/10}(z_j) \subset B_t^x$. Let $z \in B_{\epsilon/10}(z_j)$,

$$\|z - y\| \leqslant \|z - z_j\| + \|z_j - y\| < \frac{\epsilon}{10} + \frac{\epsilon}{10} = \frac{\epsilon}{5} < \frac{\epsilon}{4}.$$

As consequence, if a ball in $\mathcal{F}$ does not meet $\mathcal{X}_n$ then there exists a ball $B_{\frac{\epsilon}{10}}(z_i)$ with $z_i \in \{z_1, \cdots, z_s\}$ such that $B_{\frac{\epsilon}{10}}(z_i) \cap \mathcal{X}_n = \varnothing$. So,

$$\mathbb{P}(\exists x \in G(t) : \mathcal{X}_n \cap B_t^x = \varnothing) \leqslant \sum_{i=1}^{s} \mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing\right). \tag{4.6}$$

In addition, if ϵ is small enough then

$$f(z) > l - \zeta \text{ for all } z \in B_{\frac{\epsilon}{10}}(z_i),\ i = 1, \cdots, s. \tag{4.7}$$

Let $z \in B_{\frac{\epsilon}{10}}(z_i)$ for some $i \in \{1, \cdots, s\}$. Since $z_i \in G(t)$ then $f(z_i) \geqslant t > l$. In addition, f is continuous in U. Then, two cases must be considered:

(a) If $z_i \in U$ then given $\zeta > 0$, see assumption (A) for more details,

$$\exists \delta_i > 0 \text{ such that } \forall w \in B_{\delta_i}(z_i) \text{ it is verified that } \|f(w) - f(z_i)\| < \zeta.$$

Then,

$$\exists \delta_i > 0 \text{ such that } \forall w \in B_{\delta_i}(z_i) \text{ it is verified that } f(w) > f(z_i) - \zeta \geqslant l - \zeta.$$

(b) If $z_i \notin U$ then $z_i \in \text{Int}(G(u + \zeta))$ since $z_i \in G(l) \cap U^c$. Therefore,

$$\exists \delta_i > 0 \text{ such that } B_{\delta_i}(z_i) \subset \text{Int}(G(u + \zeta)).$$

Then, for all $w \in B_{\delta_i}(z_i)$, $f(w) > u + \zeta > l - \zeta$.

In order to guarantee (4.7), it is enough to take $\epsilon < 10 \min\{\delta_i : i = 1, \cdots, s\}$.

3. Next, using (4.6), we will prove that

$$\mathbb{P}(\exists x \in G(t) : \mathcal{X}_n \cap B_t^x = \varnothing, \text{ infinitely often}) = 0.$$

Using the same reasoning as in the Step 2, it is enough to analyze if for each fixed $i \in \{1, \cdots, s\}$

$$\mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing, \text{ infinitely often}\right) = 0.$$

According to Borel-Cantelli's Lemmas, it is enough to show that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing\right) < \infty.$$

Since the observations are independent and identically distributed, we can write

$$\mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing\right) = \mathbb{P}\left(\forall i \in \{1, ..., n\},\ X_i \notin B_{\frac{\epsilon}{10}}(z_i)\right)$$

$$= \prod_{i=1}^{n} \mathbb{P}\left(X_i \notin B_{\frac{\epsilon}{10}}(z_i)\right) = \left[\mathbb{P}\left(X_1 \notin B_{\frac{\epsilon}{10}}(z_i)\right)\right]^n = \left[1 - \mathbb{P}\left(X_1 \in B_{\frac{\epsilon}{10}}(z_i)\right)\right]^n$$

$$\leqslant e^{-n\mathbb{P}\left(X_1 \in B_{\frac{\epsilon}{10}}(z_i)\right)}.$$

According to (4.7),

$$\mathbb{P}\left(X_1 \in B_{\frac{\epsilon}{10}}(z_i)\right) = \int_{B_{\frac{\epsilon}{10}}(z_i)} f(x)\, d\mu \geqslant \int_{B_{\frac{\epsilon}{10}}(z_i)} (l-\zeta)\, d\mu = (l-\zeta)\mu\left(B_{\frac{\epsilon}{10}}(z_i)\right) = \rho > 0$$

and

$$\mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing\right) \leqslant e^{-n\mathbb{P}\left(X_1 \in B_{\frac{\epsilon}{10}}(z_i)\right)} = e^{-n\rho}.$$

Then,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\mathcal{X}_n \cap B_{\frac{\epsilon}{10}}(z_i) = \varnothing\right) \leqslant \sum_{n=1}^{\infty} e^{-n\rho} < \infty.$$

4. According to Step 3, with probability one, there exists n_0 such that for all $x \in G(t)$,

$$\mathcal{X}_n \cap B_t^x \neq \varnothing,\ \forall n \geqslant n_0.$$

Then, there exists n_0 such that for all $x \in G(t)$,

$$\exists X_{i_x} \in \mathcal{X}_n \cap B_t^x \subset \mathcal{X}_n \cap B_\epsilon(x),\ \forall n \geqslant n_0.$$

Therefore, it only remains to prove that $X_{i_x} \in \mathcal{X}_n^+(t)$. According to Proposition B.0.2, it is possible to guarantee that

$$\sup_C |f_n - f| = O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right),\ \text{almost surely}$$

where $C \subset U$ is under conditions of Proposition B.0.2. Therefore, there exists $N > 0$ such that, with probability one,

$$\sup_C |f_n - f| \leqslant N\left(\frac{\log n}{n}\right)^{p/(d+2p)}.$$

Two cases are considered: $X_{i_x} \in C$ and $X_{i_x} \notin C$.

(a) If $X_{i_x} \in C$ and $D_n = M\left(\frac{\log n}{n}\right)^{p/(d+2p)}$ with $M \geqslant N$ then $\lim_{n\to\infty} D_n = 0$. So, fixed $T/2 > 0$ (see Step 1 in this proof),

$$\exists n_1 \in \mathbb{N} \text{ such that } D_n < T/2, \forall n \geqslant n_1.$$

Then,

$$|f_n(X_{i_x}) - f(X_{i_x})| \leqslant \sup_C |f_n - f| \leqslant D_n < T/2,\ \forall n \geqslant \{n_0, n_1\}.$$

Therefore, since $X_{i_x} \in B_t^x \subset G(t+T)$,

$$f_n(X_{i_x}) \geqslant f(X_{i_x}) - D_n \geqslant t + T - D_n > t + T - \frac{T}{2} = t + \frac{T}{2} \geqslant t + D_n.$$

(b) If $X_{i_x} \notin C$ then, since $X_{i_x} \in B_l^x \subset G(t+T)$, it is verified that $f(X_{i_x}) \geqslant t+T > t \geqslant l$. So, $X_{i_x} \in \text{Int}(G(l))$. Then, $X_{i_x} \in G(u) \cap C^c$. According to Proposition B.0.3 for a certain $w > 0$, with probability one,

$$\exists n_2 \text{ such that } f_n(z) \geqslant u + \frac{w}{2}, \ \forall z \in G(u) \cap C^c \text{ and } \forall n \geqslant n_2.$$

For D_n fixed previously, $\lim_{n\to\infty} D_n = 0$. So, given $w/2 > 0$,

$$\exists n_3 \in \mathbb{N} \text{ such that } D_n < w/2, \forall n \geqslant n_3.$$

Therefore, since $t \leqslant u$,

$$f_n(X_{i_x}) \geqslant u + \frac{w}{2} \geqslant t + D_n, \ \forall n \geqslant \max\{n_0, n_2, n_3\}. \qquad \square$$

Corollary 4.3.4 shows, in particular, that $\mathcal{X}_n^+(t)$ is a consistent estimator for $G(t)$ in Hausdorff distance. At this point, it is important to remember a similar property for the support S. The set of sample points $\mathcal{X}_n$ is a Hausdorff consistent estimator for S too.

Corolary 4.3.4. *Let $G(t)$ be a compact and nonempty level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f and $\mathcal{X}_n^+(t)$ established in Definition 4.3.1. Then, for all $\epsilon > 0$ it is verified that*

$$\mathbb{P}\left(d_H(G(t), \mathcal{X}_n^+(t)) \leqslant \epsilon, \ \textit{eventually}\right) = 1.$$

Proof. The proof is a straightforward consequence of Lemmas 4.3.2 and 4.3.3. $\square$

4.4 Consistency for the estimator of the optimal parameter

Lemma 4.4.1 is a useful and auxiliary tool for guaranteeing the consistency for the estimator established in Definition 4.3.1. It ensures the existence of points in $\mathcal{X}_n^-(t)$ inside any open ball contained in $G(t)^c$. A straightforward consequence is that, with probability one and for n large enough, $\mathcal{X}_n^-(t)$ is not empty.

Lemma 4.4.1. *Let $G(t)$ be a compact, nonempty and nonconvex. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f and $\mathcal{X}_n^-(t)$ established in Definition 4.3.1. Let $B_\epsilon(x)$ such that $B_\epsilon(x) \subset \text{Int}(G(l-\zeta))$ and $B_\epsilon(x) \cap G(t) = \varnothing$. Then,*

$$\mathbb{P}\left(\mathcal{X}_n^-(t) \cap B_\epsilon(x) \neq \varnothing, \ \textit{eventually}\right) = 1.$$

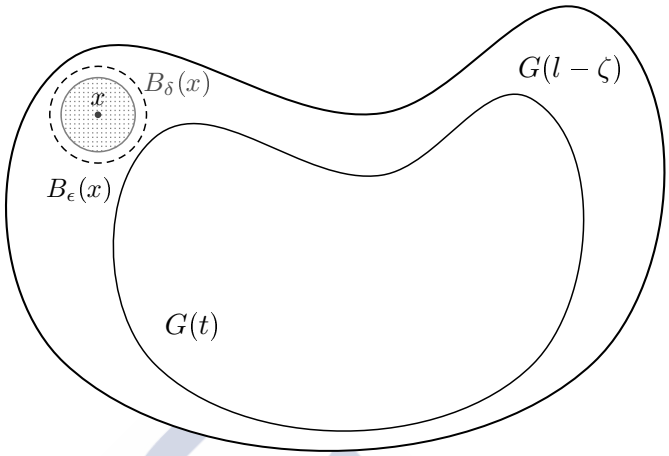


Figure 4.8: Elements of Lemma 4.4.1. $B_\delta(x) \subset B_\epsilon(x) \subset \text{Int}(G(l-\zeta)) \cap G(t)^c$.

Proof. Since $x \in G(t)^c \cap \text{Int}(G(l-\zeta))$, it is verified that $l-\zeta < f(x) < t$. The following steps complete the proof:

1. Under (A), f is continuous in x. Therefore, given $K = \frac{t-f(x)}{2} > 0$,

$$\exists \delta_1 > 0 \text{ such that } \forall y \in B_{\delta_1}(x) \text{ it is verified that } \|f(x) - f(y)\| < K.$$

   Since $f(x) = t - 2K$,

$$\forall y \in B_{\delta_1}(x) \text{ it is verified that } f(y) < t - K.$$

   In addition, $B_\epsilon(x) \subset \text{Int}(G(t-\zeta))$. Therefore,

$$\forall y \in B_\epsilon(x) \text{ it is verified that } f(y) > t - \zeta > 0.$$

   If $\delta = \min\{\delta_1, \epsilon\}$ then it is verified that $l - \zeta < f(y) < t - K$ for all $y \in B_\delta(x) \subset B_\epsilon(x)$. See Figure 4.8 for more details.

2. Next, we will prove that, with probability one and for n large enough, there exists $X_{i_x} \in \mathcal{X}_n \cap B_\delta(x)$. That is, we will prove that $\mathbb{P}(\mathcal{X}_n \cap B_\delta(x) \neq \varnothing, \text{eventually}) = 1$. According to the Borel-Cantelli's Lemmas, it is enough to prove that $\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{X}_n \cap B_\delta(x) = \varnothing) < \infty$. Since the observations are independent and identically distributed, we can write

$$\mathbb{P}(\mathcal{X}_n \cap B_\delta(x) = \varnothing) = \mathbb{P}(\forall i \in \{1, ..., n\},\ X_i \notin B_\delta(x))$$

$$= \prod_{i=1}^{n} \mathbb{P}(X_i \notin B_\delta(x)) = [\mathbb{P}(X_1 \notin B_\delta(x))]^n = [1 - \mathbb{P}(X_1 \in B_\delta(x))]^n$$

$$\leqslant e^{-n\mathbb{P}(X_1 \in B_\delta(x))}.$$

According to the previous step, $\forall y \in B_\delta(x)$ it is verified that $f(y) > t - \zeta > 0$. Therefore,

$$\mathbb{P}(X_1 \in B_\delta(x)) = \int_{B_\delta(x)} f(x)\, d\mu \geqslant \int_{B_\delta(x)} l - \zeta\, d\mu$$
$$= (l - \zeta)\mu(B_\delta(x)).$$

So,

$$\mathbb{P}(\mathcal{X}_n \cap B_\delta(x) = \varnothing) \leqslant e^{-n\mathbb{P}(X_1 \in B_\delta(x))} \leqslant e^{-n(l-\zeta)\mu(B_\delta(x))} > 0.$$

Then,

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{X}_n \cap B_\delta(x) = \varnothing) \leqslant \sum_{n=1}^{\infty} e^{-n(l-\zeta)\mu(B_\delta(x))} < \infty.$$

In addition, for all $X_{i_x} \in \mathcal{X}_n \cap B_\delta(x)$ it is satisfied that $f(X_{i_x}) < t - K$ con $K > 0$ and $f(X_{i_x}) > l - \zeta$ (see Step 1 of this proof). It remains to show that $X_{i_x} \in \mathcal{X}_n^-(t)$.

3. According to the previous step, with probability one, there exists n_0 such that

$$\mathcal{X}_n \cap B_\delta(x) \neq \varnothing,\ \forall n \geqslant n_0.$$

According to Proposition B.0.2,

$$\sup_C |f_n - f| = O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right),$$

where $C \subset U$ is under conditions of Proposition B.0.2. Therefore, with probability one and for n large enough,

$$\exists N > 0 \text{ such that } \sup_C |f_n - f| \leqslant N\left(\frac{\log n}{n}\right)^{p/(d+2p)}.$$

Two situations are considered: $X_{i_x} \in C$ and $X_{i_x} \notin C$.

(a) If $X_{i_x} \in C$ and $D_n = M\left(\frac{\log n}{n}\right)^{p/(d+2p)}$ with $M \geqslant N$ then $\lim_{n\to\infty} D_n = 0$. So, fixed $K/2 > 0$ (see Step 1 in this proof),

$$\exists n_1 \in \mathbb{N} \text{ such that } D_n < K/2, \forall n \geqslant n_1.$$

Then, with probability one,

$$|f_n(X_{i_x}) - f(X_{i_x})| \leqslant \sup_C |f_n - f| \leqslant D_n < K/2, \ \forall n \geqslant \max\{n_0, n_1\}.$$

Therefore, for all $n \geqslant \max\{n_0, n_1\}$,

$$f_n(X_{i_x}) \leqslant f(X_{i_x}) + D_n < t - K + D_n < t - K + \frac{K}{2} = t - \frac{K}{2} < t - D_n.$$

(b) If $X_{i_x} \notin C$ then, since $f(x) < t - K < t \leqslant u$, it is verified that $x \in G(u)^c$. Without losing generality, we can assume that $B_\delta(x) \subset G(u)^c$ since $G(u)^c$ is open and x is a interior point. In another case, it is enough to reduce the radius of the ball. So, $X_{i_x} \in G(l)^c \cap C^c$. According to Proposition B.0.3 for some $w > 0$, with probability one,

$$\exists n_1 \text{ such that } f_n(z) \leqslant l - \frac{w}{2}, \ \forall z \in G(l)^c \cap C^c \text{ and } \forall n \geqslant n_1.$$

For D_n previously fixed, $\lim_{n\to\infty} D_n = 0$. Therefore, fixed $w/2 > 0$,

$$\exists n_2 \in \mathbb{N} \text{ such that } D_n < w/2, \forall n \geqslant n_2.$$

Therefore, since $l \leqslant t$ and $X_{i_x} \in G(l)^c \cap C^c$,

$$f_n(X_{i_x}) \leqslant l - \frac{w}{2} \leqslant t - D_n, \ \forall n \geqslant \max\{n_0, n_1, n_2\}. \qquad \square$$

Proposition 4.4.2 proves that, with probability one and for n large enough, the estimator $\hat{r}_0(t)$ is greater than or equal to $r_0(t)$.

Proposition 4.4.2. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, $r_0(t)$ and $\hat{r}_0(t)$ established in Definitions 4.2.1 and 4.3.1, respectively. Then,*

$$\mathbb{P}(\hat{r}_0(t) \geqslant r_0(t), \ \textit{eventually}) = 1.$$

Proof. According to Lemma 4.3.2,

$$\exists n_1 \in \mathbb{N} \text{ such that } \mathcal{X}_n^+(t) \subset G(t) \text{ and } \mathcal{X}_n^-(t) \subset G(t)^c, \ \forall n \geqslant n_1.$$

Since $G(t)$ is $r_0(t)-$convex, it is verified that

$$C_{r_0(t)}(\mathcal{X}_n^+(t)) \subset C_{r_0(t)}(G(t)) = G(t).$$

Therefore, since $\mathcal{X}_n^-(t) \subset G(t)^c$,

$$\hat{r}_0(t) = \sup\{\gamma > 0 : C_\gamma(\mathcal{X}_n^+(t)) \cap \mathcal{X}_n^-(t) = \varnothing\} \geqslant r_0(t), \ \forall n \geqslant n_1. \qquad \square$$

It remains to prove that $\hat{r}_0(t)$ can not be arbitrarily larger than $r_0(t)$. This is established in Theorem 4.4.3. In order to prove consistency, see Theorem 4.4.4.

Theorem 4.4.3. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, $r_0(t)$ and $\hat{r}_0(t)$ established in Definitions 4.2.1 and 4.3.1, respectively. Then, for any $\epsilon > 0$*

$$\mathbb{P}(\hat{r}_0(t) \leqslant r_0(t) + \epsilon, \ \textit{eventually}) = 1.$$

Proof. Given $\epsilon > 0$, let be $r = r_0(t) + \epsilon > r_0(t)$. Let be $r^{'}$ such that $r > r^{'} > r_0(t)$. The proof is split in several steps:

1. First, we will prove that there exists $B_\gamma(x)$ verifying that

$$B_\gamma(x) \subset C_{r^{'}}(G(t)) \cap G(l-\zeta) \text{ and } B_\gamma(x) \cap G(t) = \varnothing. \tag{4.8}$$

   According to Proposition 4.2.4, there exists $x_t \in \text{Int}(C_{r^{'}}(G(t))) \cap \partial G(t)$:

   (a) Then, there exists $\gamma_1 > 0$ such that $B_{\gamma_1}(x_t) \subset C_{r^{'}}(G(t))$.

   (b) Since $x_t \in \partial G(t)$, $f(x_t) = t > l > l - \zeta$. Therefore, $x_t \in \text{Int}(G(l-\zeta))$. As consequence, there exists $\gamma_2 > 0$ such that $B_{\gamma_2}(x_t) \subset \text{Int}(G(l-\zeta))$.

   (c) In addition, a ball of radius m/k rolls freely in $\overline{G(t)^c}$. Then, there exists $y \in G(t)^c$ such that $x_t \in B_{m/k}[y]$ with $B_{m/k}(y) \cap G(t) = \varnothing$.

   We fixed $0 < \gamma \leqslant \min\{\gamma_1, \ \gamma_2, \ m/k\}/2$ and $x = x_t + \gamma\eta(x_t)$, see Figure 4.9 and Lemma 4.2.3 for remember details about the vector $\eta(x_t)$. For this γ, $B_\gamma(x)$ satisfies (4.8). In addition, notice that we can assume that, without loss of generality, $r \leqslant r^{'} + \gamma/2$. Otherwise, if $r - r^{'} > \gamma/2$, we could select $r^* = r^{'} + \gamma/2 < r$ verifying $r^* > r^{'} > r_0(t)$. For this r^*, (4.8) is still satisfied.

2. According to Lemma 4.3.3, with probability one and for n large enough,

$$G(t) \subset \mathcal{X}_n^+(t) \oplus B_{r-r^{'}}[0].$$

   Then, with probability one and for n large enough, it is verified that

$$G(t) \oplus B_{r^{'}}[0] \subset \left(\mathcal{X}_n^+(t) \oplus B_{r-r^{'}}[0]\right) \oplus B_{r^{'}}[0].$$

   Therefore, with probability one and for n large enough,

$$C_{r^{'}}(G(t)) = (G(t) \oplus B_{r^{'}}[0]) \ominus B_{r^{'}}[0] \subset \left[\left(\mathcal{X}_n^+(t) \oplus B_{r-r^{'}}[0]\right) \oplus B_{r^{'}}[0]\right] \ominus B_{r^{'}}[0]$$

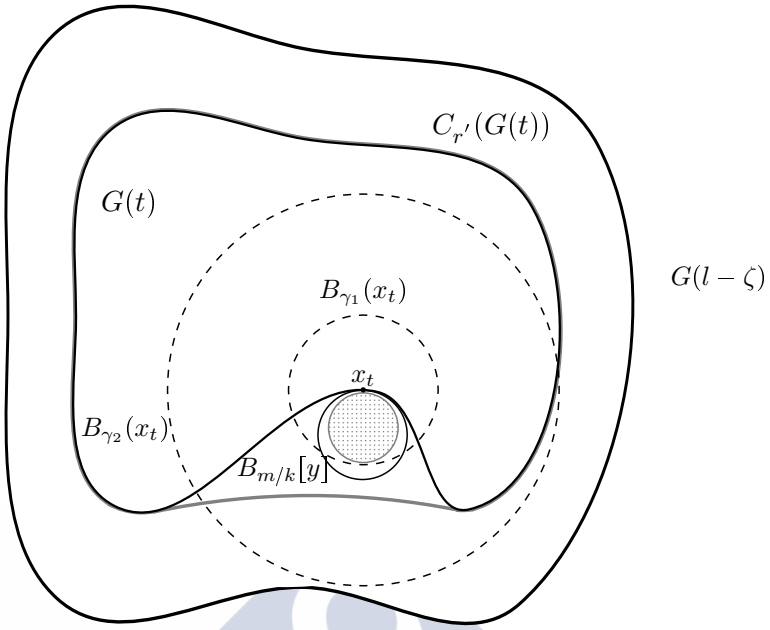


Figure 4.9: Elements of proof in Theorem 4.4.3. $B_\gamma[x]$ in gray color.

or equivalently,

$$C_{r'}(G(t)) = (G(t) \oplus B_{r'}[0]) \ominus B_{r'}[0] \subset \left[\left(\mathcal{X}_n^+(t) \oplus B_r[0]\right)\right] \ominus B_{r'}[0].$$

Then,

$$C_{r'}(G(t)) \ominus B_{r-r'}[0] \subset \left[\left(\mathcal{X}_n^+(t) \oplus B_r[0]\right) \ominus B_{r'}[0]\right] \ominus B_{r-r'}[0] = C_r(\mathcal{X}_n^+(t)).$$

Since $B_\gamma(x) \subset C_{r'}(G(t))$, if $r - r^{'} \leqslant \gamma/2$ then

$$B_{\frac{\gamma}{2}}(x) \subset C_{r'}(G(t)) \ominus B_{r-r'}[0] \subset C_r(\mathcal{X}_n^+(t)).$$

3. According to Lemma 4.4.1, with probability one and for n large enough, $\mathcal{X}_n^-(t) \cap B_{\gamma/2}(x) \neq \varnothing$ and, hence, $\mathcal{X}_n^-(t) \cap C_r(\mathcal{X}_n^+(t)) \neq \varnothing$. Therefore, we can conclude that $\hat{r}_0(t) \leqslant r$. □

Theorem 4.4.4. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, $r_0(t)$ and $\hat{r}_0(t)$ established in Definitions 4.2.1 and 4.3.1, respectively. Then, $\hat{r}_0(t)$ converges to r_0, almost surely.*

Proof. The proof is a straightforward consequence of Proposition 4.4.2 and Theorem 4.4.3. □

4.5 Consistency for the resulting estimator of the level set

Once the consistency for the estimator of the smoothing parameter $\hat{r}_0(t)$ defined in (4.2) was studied, it is natural to consider $C_{\hat{r}_0(t)}(\mathcal{X}_n^+(t))$ as an estimator for the level set $G(t)$. However and taking into account the support estimator presented in Chapter 3, we will propose $C_{r_n(t)}(\mathcal{X}_n^+(t))$ as the estimator of the level set $G(t)$ where $r_n(t) = \nu\hat{r}_0(t)$ for a fixed value $\nu \in (0,1)$. This estimator provides a consistent reconstruction of the theoretical level set and the convergence rates are provided in Theorem 4.5.8. Before exposing these key results, it is necessary to present some auxiliary proofs. For instance, Proposition 4.5.1 establishes that the estimator $C_{r_n(t)}(\mathcal{X}_n^+(t))$ is contained in the theoretical level set with probability one and for n large enough.

Proposition 4.5.1. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, $r_0(t)$ and $\hat{r}_0(t)$ established in Definitions 4.2.1 and 4.3.1. Let $\nu \in (0,1)$ be a fixed number and $r_n(t) = \nu\hat{r}_0(t)$. Then,*

$$\mathbb{P}(C_{r_n(t)}(\mathcal{X}_n^+(t)) \subset G(t), \ \textit{eventually}) = 1.$$

Proof. According to Lemma 4.3.2, with probability one,

$$\exists n_1 \in \mathbb{N} \text{ such that } \mathcal{X}_n^+(t) \subset G(t), \ \forall n \geqslant n_1.$$

Since $r_n(t)$ converges to $\nu r_0(t)$, almost surely, we have that, with probability one,

$$\exists n_2 \in \mathbb{N} \text{ such that } r_n(t) \leqslant r_0(t), \ \forall n \geqslant n_2.$$

If $n \geqslant \max\{n_1, n_2\}$,

$$C_{r_n(t)}(\mathcal{X}_n^+(t)) \subset C_{r_0(t)}(\mathcal{X}_n^+(t)) \subset G(t). \qquad \square$$

At this point, it is necessary to introduce some auxiliary sets in order to obtain the convergence rates of the resulting estimator for the level set, see Definitions 4.5.2, 4.5.3 and Figure 4.10. Really, these new sets are subsets of the original level set $G(t)$ and the sample $\mathcal{X}_n$, respectively. Notice that both are defined from the theoretical density function f. The kernel estimator f_n is not considered. On the other hand and although they depend on some parameters like n, this fact is not reflected in their names for simplicity in the exposition.

Definition 4.5.2. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A) and (D), the set $G^+(t) \subset \mathbb{R}^d$ is defined as the level set with threshold equal to $t + 2D_n$. That is, $G^+(t) = G(t + 2D_n)$.*

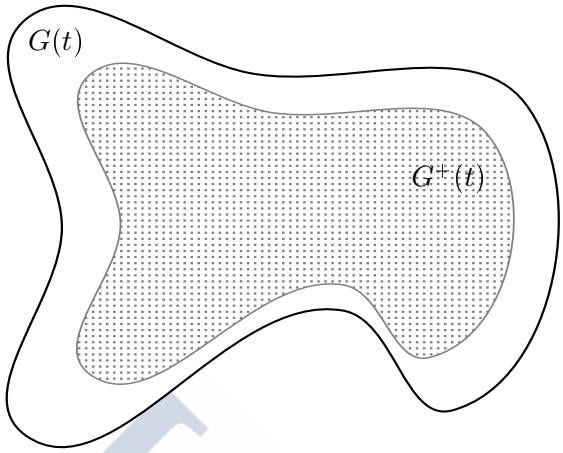


Figure 4.10: Sets $G^+(t)$ and $G(t)$ in Definition 4.5.2.

Definition 4.5.3. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A) and (D), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f and let $G^+(t) \subset \mathbb{R}^d$ be the level set established in Definition 4.5.2. The set $\mathcal{X}_n^{G^+}$ is defined by $\mathcal{X}_n \cap G^+(t)$. Therefore, it can be written as $\mathcal{X}_n^{G^+} = \{X_i \in \mathcal{X}_n : f(X_i) \geqslant t + 2D_n\}$.*

A new class of sets is presented in Definition 4.5.4. This family was already considered in Walther (1997).

Definition 4.5.4. *Let $A \subset \mathbb{R}^d$ be a set and $\gamma > 0$. Then, $\mathcal{G}_A(\gamma)$ denotes all sets B that verify (R_γ^γ) satisfying $B \subset A$.*

The smoothing parameter established in Definition 4.2.1 is studied in Lemma 4.5.5 for the sets $G^+(t)$.

Lemma 4.5.5. *Under assumptions (A) and (D), let $r_0(t)$ established in Definition 4.2.1. It is verified that*

$$\exists n_0 \in \mathbb{N} \text{ such that } r_0(t + 2D_n) \geqslant m/k, \ \forall n \geqslant n_0.$$

Proof. Since $\lim_{n\to\infty} D_n = 0$,

$$\exists n_0 \in \mathbb{N} \text{ such that } 2D_n < u - t, \ \forall n \geqslant n_0.$$

Therefore,

$$l < t + 2D_n < u, \ \forall n \geqslant n_0.$$

Under (A), $G^+(t)$ verifies that a ball of radius m/k rolls freely in $G^+(t)$ and $G^+(t)^c$ for $n \geqslant n_0$. Therefore,

$$0 < m/k \leqslant r_0(t + 2D_n), \ \forall n \geqslant n_0. \qquad \square$$

Next, it will be proved that $G^+(t) \in \mathcal{G}_{G(l)}(r_\nu)$ for n large enough and $r_\nu > 0$, see Lemma 4.5.6 for details about the positive constant r_ν.

Lemma 4.5.6. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Let $G^+(t)$ be the set established in Definition 4.5.2. Under assumptions (A), (D) and (K), let $\hat{r}_0(t)$ established in Definition 4.3.1, $\nu \in (0,1)$ be a fixed number and $r_n(t) = \nu \hat{r}_0(t)$. Then, there exists $0 < r_\nu < m/k$ such that*

$$\mathbb{P}(r_n(t) > r_\nu, \ \textit{eventually}) = 1.$$

Further,

$$\exists n_0 \in \mathbb{N} \textit{ such that } G^+(t) \in \mathcal{G}_{G(t)}(r_\nu), \ \forall n \geqslant n_0$$

and, therefore,

$$G^+(t) \in \mathcal{G}_{G(l)}(r_\nu), \ \forall n \geqslant n_0$$

for $\mathcal{G}_{G(t)}(r_\nu)$, $\mathcal{G}_{G(l)}(r_\nu)$ and $G^+(t)$ established in Definitions 4.5.4 and 4.5.2, respectively.

Proof. Let be $r_\nu > 0$ verifying that $r_\nu < \nu(m/k) < m/k$. It is easy to prove that $\mathbb{P}(r_n(t) > r_\nu, \text{ eventually}) = 1$ taking into account that $r_n(t)$ converges to $\nu r_0(t)$, almost surely. According to Lemma 4.5.5, it is verified that

$$\exists n_0 \in \mathbb{N} \text{ such that } r_0(t + 2D_n) \geqslant m/k, \ \forall n \geqslant n_0.$$

Since $0 < r_\nu < m/k$,

$$r_0(t + 2D_n) \geqslant m/k > r_\nu, \ \forall n \geqslant n_0.$$

Then, since $G^+(t) = G(t + 2D_n) \subset G(t)$ for all n,

$$G^+(t) \in \mathcal{G}_{G(t)}(r_\nu), \ \forall n \geqslant n_0. \qquad \square$$

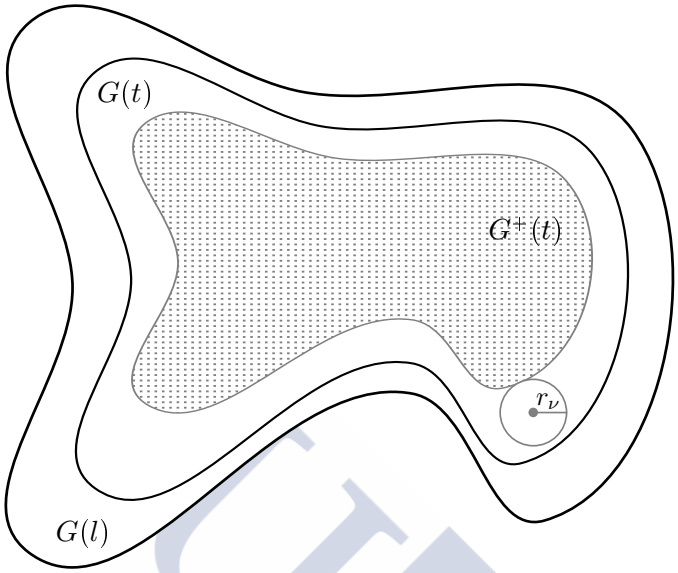


Figure 4.11: Elements in Lemma 4.5.6. $G(t)$, $G(l)$ and $G^+(t)$. A ball of radius r_ν (gray color) rolls freely in $\overline{G^+(t)^c}$.

In Lemma 4.5.7, it will proved that, given the threshold t, the set $\mathcal{X}_n^{G^+}$ is eventually contained in $\mathcal{X}_n^+(t)$.

Lemma 4.5.7. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, let $\mathcal{X}_n^+(t)$ be established in Definition 4.3.1 and let $\mathcal{X}_n^{G^+}$ be the subsample defined in Definition 4.5.3. Then,*

$$\mathbb{P}(\mathcal{X}_n^{G^+} \subset \mathcal{X}_n^+(t), \ \textit{eventually}) = 1.$$

Proof. Let $X_i \in \mathcal{X}_n^{G^+}$. Therefore, $f(X_i) \geqslant t + 2D_n$. According to Proposition B.0.2,

$$\sup_C |f_n - f| = O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right), \text{ almost surely.}$$

where $C \subset U$ is under conditions of Proposition B.0.2. So, with probability one and for n large enough,

$$\exists N > 0 \text{ such that } \sup_C |f_n - f| \leqslant N\left(\frac{\log n}{n}\right)^{p/(d+2p)}. \tag{4.9}$$

Two cases are considered: X_i belongs to C or X_i does not belong to C.

1. Let $X_i \in C$. According to (4.9), if $M \geqslant N$,

$$|f_n(X_i) - f(X_i)| \leqslant D_n.$$

Therefore,

$$f_n(X_i) \geqslant f(X_i) - D_n \geqslant t + 2D_n - D_n = t + D_n.$$

2. If $X_i \notin C$ then $X_i \in G(u) \cap C^c$ since $X_i \in G(l)$ and $G(l) \backslash \operatorname{Int}(G(u)) \subset C$. According to Proposition B.0.3, with probability one and for n large enough, $f_n(X_i) \geqslant u + v/2$ for some $v > 0$. In addition, since D_n converges to zero,

$$\exists n_0 \in \mathbb{N} \text{ such that } 2D_n < \frac{v}{2}, \ \forall n \geqslant n_0.$$

Then, with probability one and for n large enough,

$$f_n(X_i) \geqslant u + \frac{v}{2} \geqslant t + \frac{v}{2} \geqslant t + D_n. \qquad \square$$

According to Lemma 4.5.7, it is verified that $C_{r_n(t)}(\mathcal{X}_n^{G^+}) \subset C_{r_n(t)}(\mathcal{X}_n^+(t))$. That is, $\mathcal{X}_n^+(t)$ is at least as good as $\mathcal{X}_n^{G^+}$ in order to estimate $G(t)$. Remember that $\mathcal{X}_n^{G^+}$ is constructed from f. It does not depend on the kernel estimator f_n. In addition, $\mathcal{X}_n^{G^+}$ would be the natural sample for estimating $G^+(t)$. Theorem 4.5.8 uses these ideas for obtaining the convergence rates of the level set estimator proposed.

Theorem 4.5.8. *Let $G(t)$ be a compact, nonempty and nonconvex level set. Under assumptions (A), (D) and (K), let $\mathcal{X}_n$ be a random sample generated from a distribution with density function f, let $\mathcal{X}_n^+(t)$ be established in Definition 4.3.1 and let $r_n(t) = \nu \hat{r}_0(t)$ where $\nu \in (0,1)$ is a fixed number and $\hat{r}_0(t)$ defined in (4.2). Then,*

$$d_H(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{p/(d+2p)}, \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right\}\right), \text{ almost surely.}$$

The same convergence order holds for $d_\mu(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t))$.

Proof. Let r_ν be a positive constant under the conditions in Lemma 4.5.6 and let $r > 0$ such that $0 < r \leqslant r_\nu$. Let $\epsilon_n = \left(\frac{C \log n}{n}\right)^{\frac{2}{d+1}}$ where $C > 0$ denotes a big enough constant to be established later. Since $\lim_{n\to\infty} \epsilon_n = 0$,

$$\exists n_0 \in \mathbb{N} \text{ such that } 0 < \epsilon_n < \max\left\{\frac{r}{3}, 1\right\}, \ \forall n \geqslant n_0.$$

On the other hand,

$$\exists n_1 \in \mathbb{N} \text{ such that } r - 2\epsilon_n \geqslant r/2, \ \forall n \geqslant n_1.$$

According to Proposition B.0.5, if $f \geqslant b > 0$,

$$\mathbb{P}(A \oplus B_{r-3\epsilon_n}[0] \not\subset [(A \cap \mathcal{X}_n) \oplus B_r[0]] \text{ for some } A \in \mathcal{G}_{G(l)}(r_\nu))$$

$$\leqslant D(\epsilon_n, G(l) \oplus B_r[0]) D\left(\frac{\epsilon_n}{10r}, S^{d-1}\right) \exp\left\{-nab(r-2\epsilon_n)^{\frac{d-1}{2}}(\epsilon_n/2)^{\frac{d+1}{2}}\right\},$$

where $D(\epsilon, B) = \max\{card\ V : V \subset B,\ |x-y| > \epsilon \text{ for different } x, y \in V\}$, S^{d-1} denotes the unit sphere in $\mathbb{R}^d$ and a is a dimensional constant. Therefore, if $n \geqslant \max\{n_0, n_1\}$ then $r - 2\epsilon_n \geqslant r/2$ and

$$\mathbb{P}(A \oplus B_{r-3\epsilon_n}[0] \not\subset [(A \cap \mathcal{X}_n) \oplus B_r[0]] \text{ for some } A \in \mathcal{G}_{G(l)}(r_\nu))$$

$$\leqslant Q\epsilon_n^{-d}\epsilon_n^{-(d-1)} \exp\left\{-nab\left(\frac{r}{2}\right)^{\frac{d-1}{2}}\left(\frac{C\log n}{2^{(d+1)/2}n}\right)\right\} = Q\epsilon_n^{(-2d+1)} \exp\{-W\log n\}$$

with Q is a constant depending on r and the dimension d and $W = \frac{ab}{2^{(d+1)/2}}\left(\frac{r}{2}\right)^{\frac{d-1}{2}} C$. If C tends to infinite then W tends to it too. Then, given $Q > 0$

$$\exists n_2 \in \mathbb{N} \text{ such that } \exp\{-W\log n\} \leqslant Q,\ \forall n \geqslant n_2.$$

Therefore,

$$\mathbb{P}(A \oplus (r-3\epsilon_n)B_1[0] \not\subset [(A \cap \mathcal{X}_n) \oplus B_r[0]] \text{ for some } A \in \mathcal{G}_{G(l)}(r_\nu))$$

$$\leqslant Q^2\epsilon_n^{-(2d-1)} \leqslant Q^2\left(\frac{n}{\log n}\right)^{\frac{(2d-1)(d+1)}{2}} n^{-M},\ \forall n \geqslant \max\{n_0, n_1, n_2\}.$$

If $W > \frac{(2d-1)(d+1)}{2}$ it is verified that

$$\sum_{i=1}^{\infty}\left(\frac{n}{\log n}\right)^{\frac{(2d-1)(d+1)}{2}} n^{-W} < \infty.$$

So,

$$\mathbb{P}(A \oplus B_{r-3\epsilon_n}[0] \not\subset [(A \cap \mathcal{X}_n) \oplus B_r[0]] \text{ for some } A \in \mathcal{G}_{G(l)}(r_\nu), \text{ infinitely often}) = 0.$$

Then, with probability one,

$$\exists n_3 \in \mathbb{N} \text{ such that } A \oplus B_{r-3\epsilon_n}[0] \subset (A \cap \mathcal{X}_n) \oplus B_r[0],\ \forall A \in \mathcal{G}_{G(l)}(r_\nu) \text{ and } \forall n \geqslant n_3.$$

According to Lemma 4.5.6, for n large enough, $G^+(t) \in \mathcal{G}_{G(l)}(r_\nu)$. So, for n large enough, with probability one,

$$(G^+(t) \oplus B_{r-3\epsilon_n}[0]) \ominus B_r[0] \subset (\mathcal{X}_n^{G^+} \oplus B_r[0]) \ominus B_r[0] = C_r(\mathcal{X}_n^{G^+}).$$

Since $G^+(t)$ is $(r-3\epsilon_n)-$convex because $r-3\epsilon_n \leqslant r_\nu$, it is satisfied that

$$(G^+(t) \oplus B_{r-3\epsilon_n}[0]) \ominus B_r[0] =$$
$$= (G^+(t) \oplus B_{r-3\epsilon_n}[0]) \ominus (B_{r-3\epsilon_n}[0] \oplus B_{3\epsilon_n}[0])$$
$$= (G^+(t) \oplus B_{r-3\epsilon_n}[0] \ominus B_{r-3\epsilon_n}[0]) \ominus B_{3\epsilon_n}[0] = G^+(t) \ominus B_{3\epsilon_n}[0].$$

Therefore, since $r_\nu > r > 0$,

$$\exists n_4 \in \mathbb{N} \text{ such that } G^+(t) \ominus B_{3\epsilon_n}[0] \subset C_r(\mathcal{X}_n^{G^+}) \subset C_{r_\nu}(\mathcal{X}_n^{G^+}),\ \forall n \geqslant n_4.$$

According the Lemma 4.5.6,

$$\exists n_5 \in \mathbb{N} \text{ such that } r_n(t) \geqslant r_\nu,\ \forall n \geqslant n_5.$$

Then,

$$G^+(t) \ominus B_{3\epsilon_n}[0] \subset C_{r_n(t)}(\mathcal{X}_n^{G^+}),\ \forall n \geqslant \max\{n_4, n_5\}.$$

Therefore, since $\mathcal{X}_n^{G^+} \subset \mathcal{X}_n^+(t) \subset G(t)$ and $r_n(t) \leqslant r_0(t)$, it is verified

$$G^+(t) \ominus B_{3\epsilon_n}[0] \subset C_{r_n(t)}(\mathcal{X}_n^+(t)) \subset C_{r_0(t)}(G(t)) = G(t),\ \forall n \geqslant \max\{n_4, n_5\}. \quad (4.10)$$

Using (4.10), with probability one and for n large enough,

$$d_H(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)) \leqslant d_H(G^+(t) \ominus B_{3\epsilon_n}[0], G(t)).$$

By the triangle inequality,

$$d_H(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)) \leqslant d_H(G^+(t), G(t)) + d_H(G^+(t), G^+(t) \ominus B_{3\epsilon_n}[0]). \quad (4.11)$$

Since $\lim_{n\to\infty} D_n = 0$,

$$\exists n_6 \in \mathbb{N} \text{ such that } 2D_n < \min\left\{(m/2)c, \frac{\zeta}{2}\right\},\ \forall n \geqslant n_6.$$

According to Proposition B.0.4,

$$G(t) \subset G^+(t) \oplus B_{\frac{4}{m}D_n}[0],\ \forall n \geqslant n_6.$$

Since $G^+(t) \subset G(t)$, $d_H(G^+(t), G(t)) = O(D_n)$. On the other hand,

$$G^+(t) \ominus B_{3\epsilon_n}[0] \subset G^+(t) \subset G^+(t) \oplus B_{3\epsilon_n}[0].$$

Therefore, $d_H(G^+(t), G^+(t) \ominus B_{3\epsilon_n}[0]) = O(\epsilon_n)$. As consequence, using (4.11)

$$d_H(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)) = O(\max\{D_n, \epsilon_n\}),\ \text{almost surely.} \quad \square$$

Remark 4.5.9. *If the smoothing parameter is unknown for the granulometric smoothing method then it provides the same convergence rates than the algorithm proposed but it incurs a penalty, see Theorem 3 in Walther (1997). The rates obtained in Theorem 4.5.8 do no depend on any penalty term because, although $r_0(t)$ is a priori unknown, it is estimated in a data-driven way from $\mathcal{X}_n$.*

4.6 Numerical aspects of the algorithm

From a practical point of view, reconstructing level sets $L(\tau)$ may be more interesting than estimating $G(t)$. In this work, the algorithm for calculating the estimator for the smoothing parameter defined in (4.2) is detailed next for this particular case. Of course, it could be easily adapted if level sets $G(t)$ must be reconstructed.

Once the value of $\tau \in (0,1)$ is given by the practitioner, it should be natural, as first step, to estimate the threshold f_τ and, then, determinate the sets $\mathcal{X}_n^+(\hat{f}_\tau)$ and $\mathcal{X}_n^-(\hat{f}_\tau)$. However, these two previous sets depend on the sequence D_n that tends to zero when the sample size tends to infinity, see Definition 4.3.1. This sequence does not rely on $\mathcal{X}_n$. However, in practise and for a fixed value of n, we think that establishing some relationship between them could be really useful. For this, a bootstrap procedure will be proposed in order to estimate two values of two probability content verifying that $\hat{\tau}^- \leqslant \tau \leqslant \hat{\tau}^+$. In addition, it is assumed that $\hat{\tau}^+$ and $\hat{\tau}^-$ could not to be symmetric around the τ. From these two values, two thresholds $\hat{f}_\tau^+$ and $\hat{f}_\tau^-$ can be determinated. Therefore, it would be possible to calculate the subsets $\mathcal{X}_{n,+}(\hat{f}_\tau^+)$ and $\mathcal{X}_{n,-}(\hat{f}_\tau^-)$, see Notation for remembering their definitions. Notice that, in most of cases, $\mathcal{X}_n \neq \mathcal{X}_{n,+}(\hat{f}_\tau^+) \cup \mathcal{X}_{n,-}(\hat{f}_\tau^-)$. Therefore, the information contained in $\mathcal{X}_n \backslash (\mathcal{X}_{n,+}(\hat{f}_\tau^+) \cup \mathcal{X}_{n,-}(\hat{f}_\tau^-))$ is not taken in advantage, see first column in Figure 4.12. To solve this, we propose to use $k-$nearest neighbors considering $\mathcal{X}_{n,+}(\hat{f}_\tau^+)$ and $\mathcal{X}_{n,-}(\hat{f}_\tau^-)$ as training samples for classifying the full sample $\mathcal{X}_n$. In particular, the set $\mathcal{X}_n \backslash (\mathcal{X}_{n,+}(\hat{f}_\tau^+) \cup \mathcal{X}_{n,-}(\hat{f}_\tau^-))$ will be classified, see second column in Figure 4.12. Therefore, a value $\hat{k} \geqslant 1$ for the nearest neighbors must selected too. Below, the bootstrap procedure considered for calculating $\hat{f}_\tau^+$, $\hat{f}_\tau^-$ and $\hat{k}$ will be explained in detail. Before, the algorithm for estimating $r_0(f_\tau)$ will be exposed. For simplicity in the exposition, the estimator will be denoted by $\hat{r}_0(\hat{f}_\tau)$. Dichotomy algorithms will be considered. Therefore, a maximum number of iterations J and two initial points r_m and r_M with $r_m < r_M$ must be selected. In practise, it is necessary to guarantee that $C_{r_m}(\mathcal{X}_{n,+}(\hat{f}_\tau^+)) \cap \mathcal{X}_{n,-}(\hat{f}_\tau^-) = \varnothing$ and $C_{r_M}(\mathcal{X}_{n,+}(\hat{f}_\tau^+)) \cap \mathcal{X}_{n,-}(\hat{f}_\tau^-) \neq \varnothing$, respectively. Then, a value close enough to zero must be chosen for r_m and r_M should be big enough for guaranteeing that $C_{r_M}(\mathcal{X}_{n,+}(\hat{f}_\tau^+))$ coincides or is almost equal to $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau^+))$. Of course, if $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau^+)) \cap \mathcal{X}_{n,-}(\hat{f}_\tau^-) = \varnothing$ then $\hat{r}_0(\hat{f}_\tau) = \infty$ and, therefore, $\hat{L}(\tau) = \text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau^+))$. Taking the previous comments under consideration, $\hat{r}_0(\hat{f}_\tau)$ will be computed as follows:

1. Use $\hat{k}-$nearest neighbors algorithm considering $\mathcal{X}_{n,+}(\hat{f}_\tau^+)$ and $\mathcal{X}_{n,-}(\hat{f}_\tau^-)$ as a trainning sample for classifying the original full sample $\mathcal{X}_n$. The two resulting sets are denoted, for simplicity in the exposition, by the name of the original sets.

2. In each iteration and while the number of them is smaller than J:

(a) $r = (r_m + r_M)/2$.

(b) If $\mathcal{X}_{n,-}(\hat{f}_\tau^-) \cap C_r(\mathcal{X}_{n,+}(\hat{f}_\tau^+)) \neq \varnothing$ then $r_M = r$.

(c) Otherwise, $r_m = r$.

Then, $\hat{r}_0(\hat{f}_\tau) = r_m$ and $\hat{L}(\tau) = C_{\hat{r}_0(\hat{f}_\tau)}(\mathcal{X}_{n,+}(\hat{f}_\tau^+))$.

As we told in Chapter 3, it should be noted that the $r-$convex and convex hulls of a sample points can be easily computed (at least for the bidimensional case), see Pateiro-López and Rodríguez-Casal (2010) and Renka (1996), respectively.

Once the algorithm for estimating the smoothing parameter was exposed, it only remains to detail the procedure in order to calculate the two thresholds, $\hat{f}_\tau^+$ and $\hat{f}_\tau^-$, and $\hat{k}$. A bootstrap method is proposed for selecting them by minimizing an error criteria between sets, the distance in measure. In an analogous way, other distances between sets could be considered.

To sum up, the next inputs should be given: the probability content $\tau \in (0,1)$, a big enough sample size M, a step Δ and a positive integer I for defining the vectors $\tau^+ = (\tau, \tau+\Delta, ..., \tau+I\Delta)$ and $\tau^- = (\tau, \tau-\Delta, ..., \tau-I\Delta)$ verifying $\tau + I\Delta \leqslant (n-1)/n$ and $\tau - I\Delta \geqslant 1/n$ in order to avoid empty sets, a number of bootstrap iterations B, a vector k of length K containing the number of nearest neighbors to be considered and, as before, a maximum number of iterations J for the dichotomy algorithm. On the other hand, the selector for the bandwidth parameter of Bowman (1984) and Rudemo (1982) could be considered for density estimation in the univariate case since its generalization for the multivariate case is computed easily, see Duong (2007).

1. Estimate by Monte Carlo approach the threshold f_τ^* in the bootstrap world:

   (a) Draw a bootstrap sample of size M from f_n where f_n denotes the kernel estimator with bandwidth H obtained from $\mathcal{X}_n$. It is denoted by $\mathcal{X}_M^*$.

   (b) Obtain f_τ^* determinating the quantile τ of the empirical distribution of $f_n(\mathcal{X}_M^*)$. Therefore, $L^*(\tau) = \{f_n \geqslant f_\tau^*\}$ represents the theoretical level set in the bootstrap world.

2. This step must be repeated B times:

   (a) Draw a bootstrap sample of size n from f_n. It will be denoted by $\mathcal{X}_n^*$.

   (b) Calculate $f_n(\mathcal{X}_n^*)$ and $f_n^*(\mathcal{X}_n^*)$ where f_n^* is the kernel estimator calculated from $\mathcal{X}_n^*$ with bandwidth H^*.

   (c) In each iteration, while j_1 and j_2 are smaller or equal than $I+1$ and while j_3 is smaller than K:

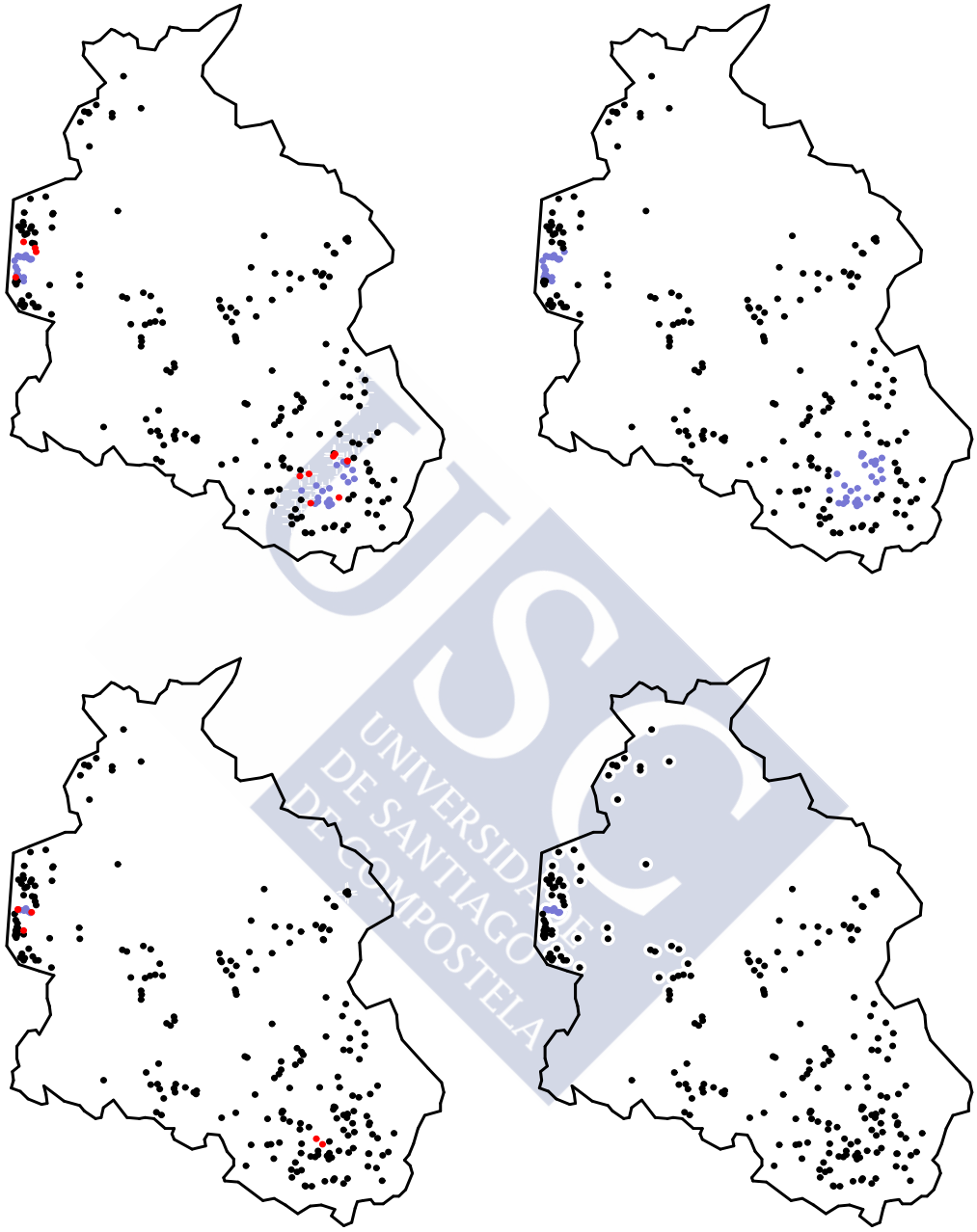


Figure 4.12: In the first column, $\mathcal{X}_{n,+}(\hat{f}_\tau^+)$ (blue), $\mathcal{X}_{n,-}(\hat{f}_\tau^-)$ (black) and $\mathcal{X}_n \backslash (\mathcal{X}_{n,+}(\hat{f}_\tau^+) \cup \mathcal{X}_{n,-}(\hat{f}_\tau^-))$ (red) for τ_1 (top) smaller than τ_2 (bottom). In the second column, $\mathcal{X}_{n,+}(\hat{f}_\tau^+)$ (blue) and $\mathcal{X}_{n,-}(\hat{f}_\tau^-)$ (black) after classification for τ_1 (top) smaller than τ_2 (bottom).

i. Obtain $\hat{f}^+_{\tau,*}$ and $\hat{f}^-_{\tau,*}$ determinating the quantiles $\tau^+(j_1)$ and $\tau^-(j_2)$ of the empirical distribution of $f^*_n(\mathcal{X}^*_n)$, respectively.

ii. Calculate $\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*})$, $\mathcal{X}^*_{n,-}(\hat{f}^-_{\tau,*})$ and $\mathcal{X}^*_n \backslash (\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*}) \cup \mathcal{X}^*_{n,-}(\hat{f}^-_{\tau,*}))$.

iii. Use $k(j_3)-$nearest neighbors algorithm considering $\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*})$ and $\mathcal{X}^*_{n,-}(\hat{f}^-_{\tau,*})$ as a trainning sample for classifying the full sample $\mathcal{X}^*_n$. The two resulting sets are denoted, for simplicity in the exposition again by $\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*})$ and $\mathcal{X}^*_{n,-}(\hat{f}^-_{\tau,*})$.

iv. Estimate the smoothing parameter from $\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*})$ and $\mathcal{X}^*_{n,-}(\hat{f}^-_{\tau,*})$ using the previous dichotomy algorithm. It will be denoted by $\hat{r}^*_0(\hat{f}_{\tau,*})$.

v. Estimate the error $d_\mu\left(L^*(\tau), C_{\hat{r}^*_0(\hat{f}_{\tau,*})}(\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*}))\right)$ induced by f_n as follows:

   A. Draw another bootstrap sample $\mathcal{Y}^*_M$ from f_n of size M.

   B. Determinate which points in $\mathcal{Y}^*_M$ are and are not in $L^*(\tau)$:

$$\mathcal{Y}^*_{M,+}(f^*_\tau) = \{Y \in \mathcal{Y}^*_M : f_n(Y) \geqslant f^*_\tau\}$$

   and

$$\mathcal{Y}^*_{M,-}(f^*_\tau) = \{Y \in \mathcal{Y}^*_M : f_n(Y) < f^*_\tau\}.$$

   C. Calculate the cardinal of the set

$$\left\{\mathcal{Y}^*_{M,-}(f^*_\tau) \cap C_{\hat{r}^*_0(\hat{f}_{\tau,*})}(\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*}))\right\} \cup \left\{\mathcal{Y}^*_{M,+}(f^*_\tau) \cap C_{\hat{r}^*_0(\hat{f}_{\tau,*})}(\mathcal{X}^*_{n,+}(\hat{f}^+_{\tau,*}))^c\right\}$$

   and divide the result obtained by M.

3. Select the values in τ^+, τ^- and k which provides the lowest empirical means of the B errors calculated. They will be denoted by $\hat{\tau}^+$, $\hat{\tau}^-$ and $\hat{k}$, respectively.

4. Obtain $\hat{f}^+_\tau$ and $\hat{f}^-_\tau$ determinating the quantiles $\hat{\tau}^+$ and $\hat{\tau}^-$ of the empirical distribution of $f_n(\mathcal{X}_n)$, respectively.

4.7 A real example

The question of whether the geographical incidence of disease shows any tendency towards clustering in geographical space has a long and rich history. For instance, do cases of disease tend to occur in proximity to other cases? The problem has become more urgent in recent years in the light of concerns raised about possible links between disease incidence and potential sources of environmental contamination, such as nuclear installations. Evidence of clustering might also lend support to other theories of disease incidence, such as a viral aetiology. For example, exposure to a common, persistent viral infection, either during gestation or as a young child with an immune system

that had been protected at a very early age, might provide clues to explaining possible leukaemia clustering.

A priori we may expect to observe a certain amount of clustering due to natural background variation in the population from which events arise. For example, cases of cancer will always cluster because of the distribution of population at risk. In such instances, we are more interested in detecting evidence of clustering over and above this underlying environmental heterogeneity; in other words, in discovering whether the distribution of one type of event clusters relative to that of another.

In order to assess the applicability of the estimation method presented in Section 4.6 and considering the data set presented in Section 1.4.1, the evidence for clustering of the cases of leukaemia in the North West of England will be studied. Analyzing whether the distribution of leukaemia mirrored that of the population as a whole or whether there was evidence, as implied by concerned local residents, of clustering. For this, it could help identify the peaks or the modes of the density estimation in the resulting surface allowing to visualize easily an excess of case intensity over that of population.

	τ	$\hat{\tau}^+$	$\hat{\tau}^-$	$\hat{k}$	$\hat{r}_0(\hat{f}_\tau)$
Cases	0.7	0.7	0.66	5	0.529
	0.75	0.75	0.69	5	0.571
	0.8	0.84	0.70	1	0.382
	0.85	0.85	0.8	3	0.434
	0.9	0.908	0.812	1	0.544
	0.95	0.975	0.95	5	∞
Controls	0.7	0.74	0.62	1	0.060
	0.75	0.78	0.66	1	0.265
	0.8	0.8	0.8	1	0.178
	0.85	0.93	0.85	1	∞
	0.9	0.967	0.822	1	∞
	0.95	0.973	0.907	1	∞

Table 4.1: Estimators of k, τ^+, τ^- and $r_0(f_\tau)$ for the samples of cases and controls with different values of τ.

Then, we have estimated the level sets $L(\tau)$ from the samples of cases and controls for relatively high values of the probability content τ. More specifically, the values of τ considered are 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95. In addition, we have fixed $I = 10$, $\Delta = \min\{(1 - \tau - 3/n)/I, (\tau - 3/n)/I, 0.01\}$ where n denotes the sample size of cases or controls depending on the situation, $k = (1, 3, 5)$, $M = 3000$, and $B = 500$.

Following the algorithm detailed in Section 4.6, Table 4.1 shows the values obtained for $\hat{k}$, $\hat{\tau}^+$, $\hat{\tau}^-$ and $\hat{r}_0(\hat{f}_\tau)$ for the samples of cases and controls with the different values of τ considered. According to the results obtained for $\hat{r}_0(\hat{f}_\tau)$, $r-$convexity property plays an interesting role, mainly for the sample of cases. Only for τ equal to 0.95 the level set estimator is convex. The level set estimators for the sample of controls are convex for the three largest values of τ considered, 0.85, 0.9 and 0.95. In addition, $\hat{\tau}^+$ and $\hat{\tau}^-$ are usually different. Only for the controls with $\tau = 0.8$, it is verified that $\hat{\tau}^+ = \hat{\tau}^- = 0.8$. The performance of the estimations for the parameter k is not too clear. In particular, for the cases, $\hat{k}$ takes the values 1, 3 and 5. However, it is always equal to 1 for the sample of controls.

The resulting level sets are showed for the two samples on North West of England in Figures 4.13, 4.14 and 4.15 for different values of the probability content τ. It is possible to observe an excess of case intensity over that of population. Greater Manchester is a metropolitan county in North West England that encompasses one of the largest metropolitan areas in the United Kingdom. However, Lancashire is a non-metropolitan county that emerged during the Industrial Revolution as a major commercial and industrial region. Therefore, there is evidence of clustering and the leukaemia cases could be related to environmental and industrial factors. Similar studies have been already considered in literature. For instance, see Cuzick and Edwards (1990) for the childhood leukaemia in Humberside, Diggle et al. (1990) for the lung and larynx cancers in Chorley-South Ribble, Kelsall and Diggle (1998) for the lung and stomach cancer in Walsall, Kelsall and Wakefield (2000) for the colorectal cancer in Birminghan or Henderson et al. (2002) for acute myeloid leukemia in North West of England.

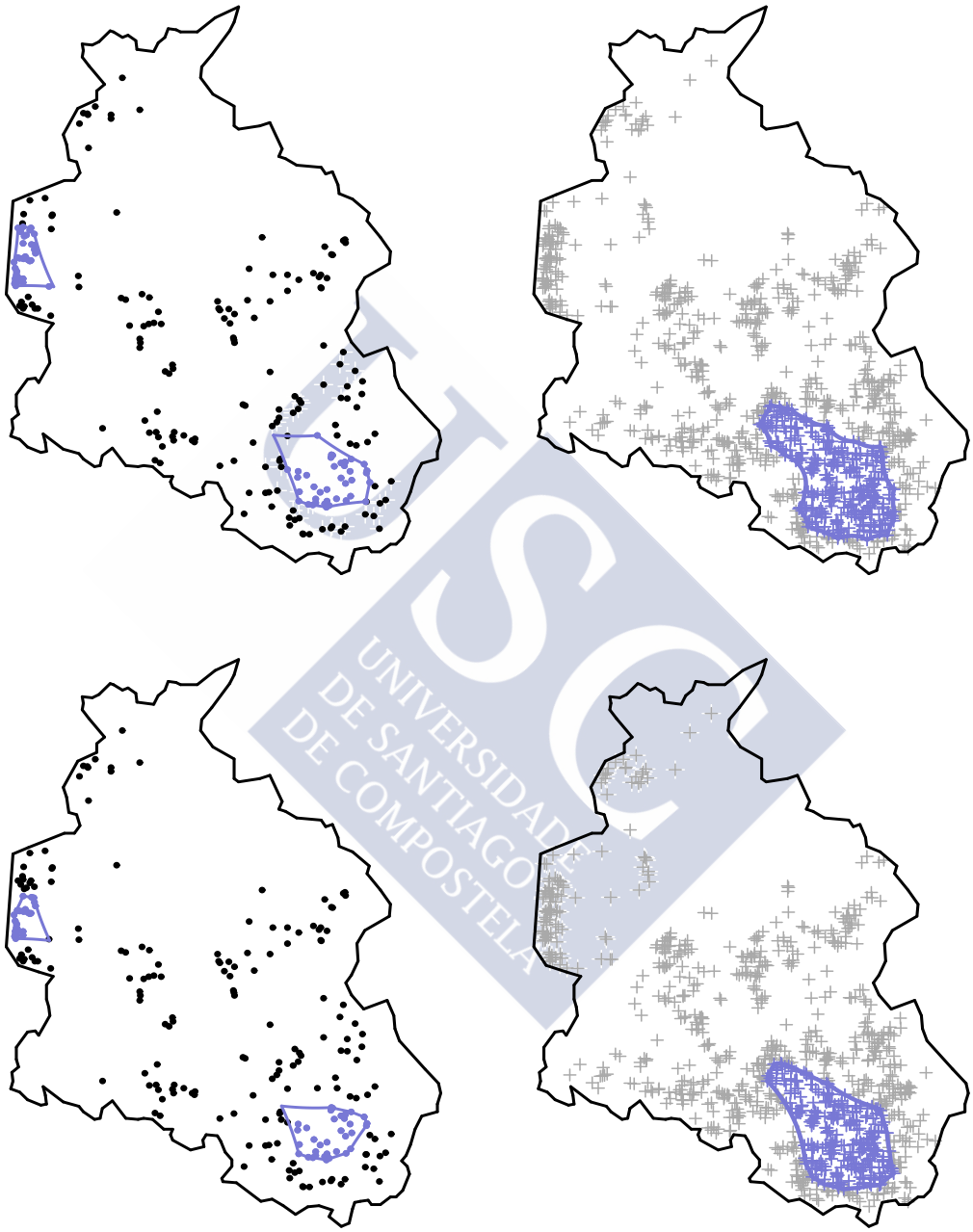


Figure 4.13: In the first column, estimated level sets for the sample of 322 cases diagnosed of leukaemia on the North West of England with $\tau = 0.7$ (top) and $\tau = 0.75$ (bottom). In the second column, estimated level sets for the sample of 988 controls of leukaemia on the North West of England with $\tau = 0.7$ (top) and $\tau = 0.75$ (bottom).

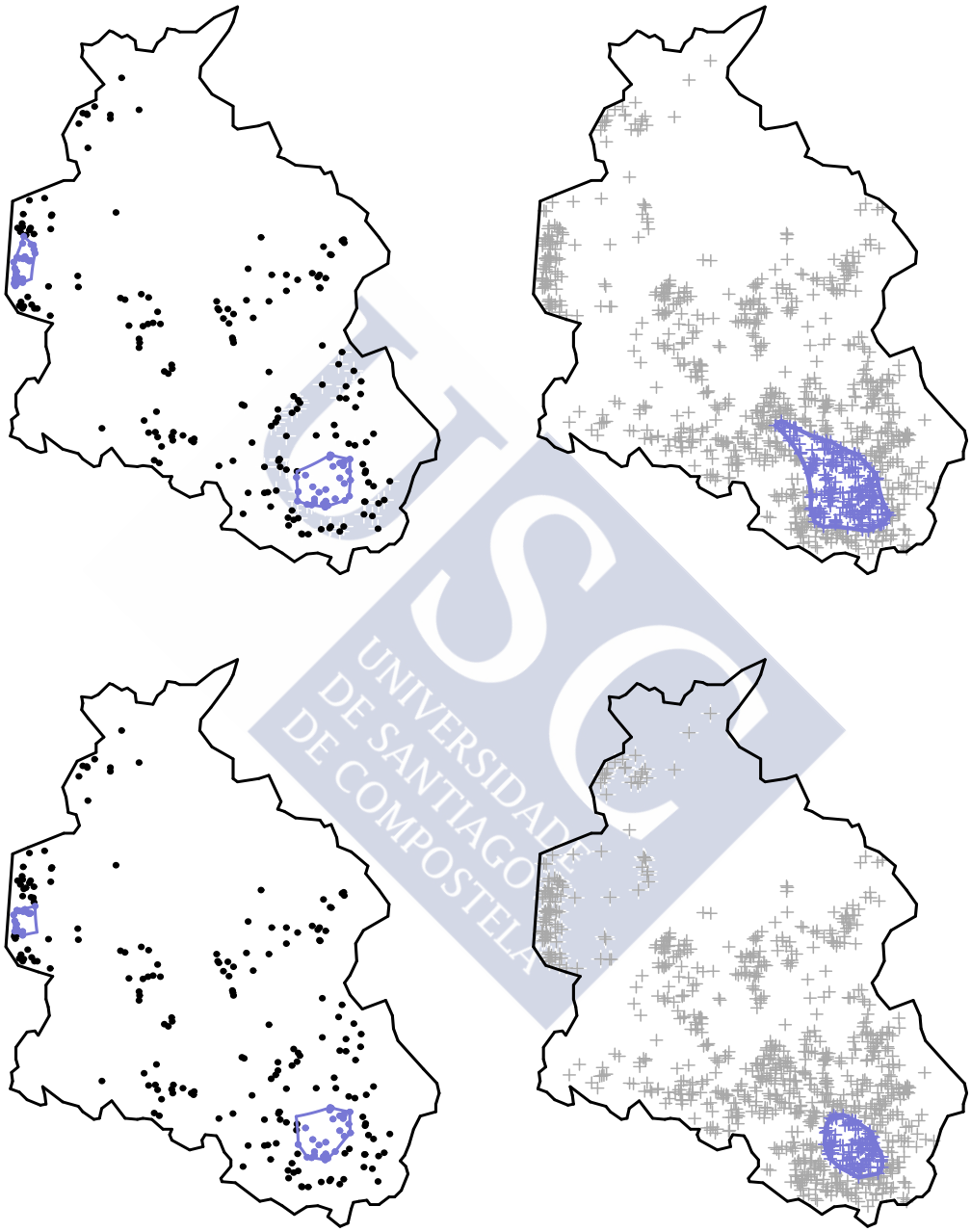


Figure 4.14: In the first column, estimated level sets for the sample of 322 cases diagnosed of leukaemia on the North West of England with $\tau = 0.8$ (top) and $\tau = 0.85$ (bottom). In the second column, estimated level sets for the sample of 988 controls of leukaemia on the North West of England with $\tau = 0.8$ (top) and $\tau = 0.85$ (bottom).

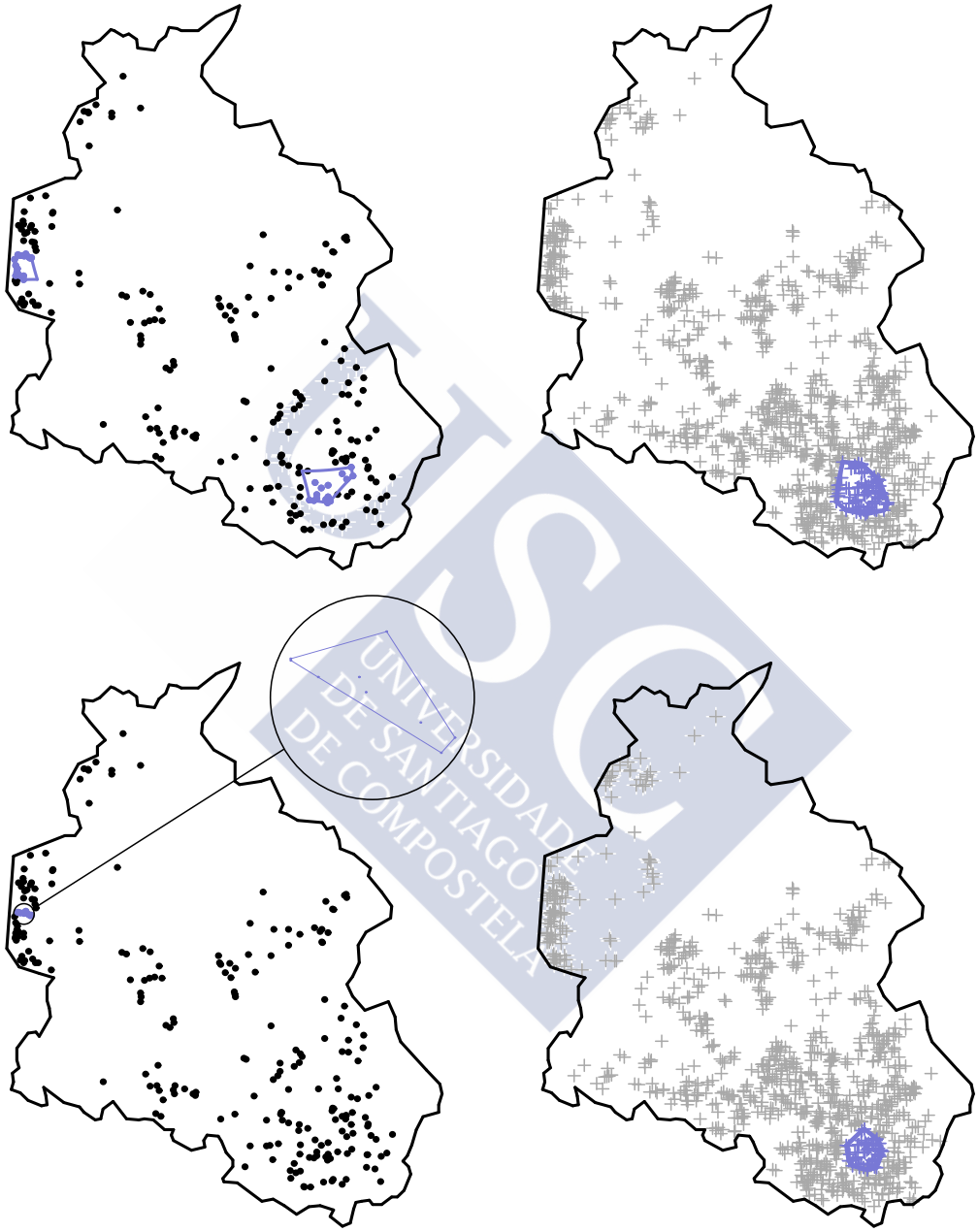


Figure 4.15: In the first column, estimated level sets for the sample of 322 cases diagnosed of leukaemia on the North West of England with $\tau = 0.9$ (top) and $\tau = 0.95$ (bottom). In the second column, estimated level sets for the sample of 988 controls of leukaemia on the North West of England with $\tau = 0.9$ (top) and $\tau = 0.95$ (bottom).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix A

Formulas of the density models for estimating level sets

In Chapter 1, the density models for studying the behavior of methods for estimating level sets have been presented. Only densities 17 and 18 are not normal mixtures. Then, they can be written as $f(x) = \omega_1 N(\mu_1, \sigma_1^2) + ... + \omega_n N(\mu_n, \sigma_n^2)$ where $\omega_1 + ... + \omega_n = 1$, μ_i and σ_i, $\omega_i \in \mathbb{R}^+$, $i = 1, ..., n$. Next, formulas for normal mixtures are exposed in Tables A.1 and A.2. The two last models can be seen in Table A.3.

Model	$\omega_1 N(\mu_1, \sigma_1^2) + ... + \omega_k N(\mu_k, \sigma_k^2)$
1 Gaussian	$N(0,1)$
2 Skewed	$\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
3 Strongly Skew	$\sum_{i=0}^{7} \frac{1}{8}N(3\{\left(\frac{2}{3}\right)^l - 1\}, \left(\frac{2}{3}\right)^{2l})$
4 Kurtotic	$\frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \left(\frac{1}{10}\right)^2\right)$
5 Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N\left(0, \left(\frac{1}{10}\right)^2\right)$

Table A.1: Parameters for normal mixtures.

Model	$\omega_1 N(\mu_1, \sigma_1^2) + ... + \omega_k N(\mu_k, \sigma_k^2)$
6 Bimodal	$\frac{1}{2}N(-1, \left(\frac{2}{3}\right)^2) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$
7 Bimodal separated	$\frac{1}{2}N\left(\frac{-3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
8 Asymmetric Bimodal	$\frac{3}{4}N\left(0, 1\right) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
9 Trimodal	$\frac{9}{20}N\left(\frac{-6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{9}{20}N\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10}N\left(0, \left(\frac{1}{4}\right)^2\right)$
10 Claw	$\frac{1}{2}N\left(0, 1\right) + \sum_{l=0}^{4} \frac{1}{10}N(l/2 - 1, \left(\frac{1}{10}\right)^2)$
11 Double Claw	$\frac{49}{100}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{49}{100}N\left(1, \left(\frac{2}{3}\right)^2\right)$ $+ \sum_{l=0}^{6} \frac{1}{350}N((l-3)/2, \left(\frac{1}{100}\right)^2)$
12 Asymmetric Claw	$\frac{1}{2}N\left(0, 1\right) + \sum_{l=-2}^{2}(2^{1-l}/31)N\left(l + \frac{1}{2}, (2^{-l}/10)^2\right)$
13 Asymmetric Double Claw	$\sum_{l=0}^{1} \frac{46}{100}N\left(2l - 1, \left(\frac{2}{3}\right)^2\right) + \sum_{l=1}^{3} \frac{1}{300}N\left(-l/2, \left(\frac{1}{100}\right)^2\right)$ $+ \sum_{l=1}^{3} \frac{7}{100}N\left(l/2, \left(\frac{7}{100}\right)^2\right)$
14 Smooth Comb	$\sum_{l=0}^{5}(2^{5^l}/63)N\left(\{65 - 96\left(\frac{1}{2}\right)^l\}/21, \left(\frac{32}{63}\right)^2/2^{2l}\right)$
15 Discrete Comb	$\sum_{l=0}^{2} \frac{2}{7}N\left((12l - 15)/7, \left(\frac{2}{7}\right)^2\right) + \sum_{l=8}^{10} \frac{1}{21}N\left(2l/7, \left(\frac{1}{21}\right)^2\right)$
16 Marronite	$\frac{1}{3}N\left(-\frac{20}{6}, \left(\frac{1}{24}\right)^2\right) + \frac{2}{3}N(0, \left(\frac{1}{6}\right)^2)$

Table A.2: Parameters for normal mixtures.

Modelo	$f(x)$
17 Caliper	$f(x) = \frac{2}{3}(1 - (\frac{x}{3} - \frac{1}{10})^{1/3})\mathbb{I}_{\{\frac{1}{10} \leqslant \frac{x}{3} \leqslant \frac{11}{10}\}}$
	$+2(1 - (-\frac{x}{3} - \frac{1}{10})^{1/3})\mathbb{I}_{\{-\frac{11}{10} \leqslant \frac{x}{3} \leqslant -\frac{1}{10}\}}$
18 Matterhorn	$f(x) = \frac{1}{20}\frac{1}{\lvert\frac{x}{20}\rvert\log(\lvert\frac{x}{20}\rvert)^2}\mathbb{I}_{\{-\frac{1}{e^2} \leqslant \frac{x}{20} \leqslant \frac{1}{e^2}\}}$

Table A.3: Density models.

Appendix B

Auxiliary results for set estimation

Many proofs in Chapters 3 and 4 take into account mathematical aspects considered in Walther (1997). Next, we will summarize these theoretical results. In particular, Proposition B.0.1 analyzes the behavior of dilation and erosion operators for the Lebesgue measure.

Proposition B.0.1. *Let $K \subset \mathbb{R}^d$ be a compact set and let $\mathcal{G}_K(r)$ be the family of sets defined in Definition 4.5.4 for some $r > 0$. If the sequence ϵ_n converges to 0^+ then it is verified that*

$$\mu(A \oplus B_{\epsilon_n}[0]) = \mu(A) + O(\epsilon_n)$$

and

$$\mu(A \ominus B_{\epsilon_n}[0]) = \mu(A) + O(\epsilon_n),$$

uniformly in $A \in \mathcal{G}_K(r)$.

Proposition B.0.2 can be obtained directly from proof of Theorem 3 in Walther (1997). It guarantees the existence of a compact set C where the convergence rate for the density kernel estimator is established.

Proposition B.0.2. *Under assumptions (A) and (K) established in Chapter 4, there exists $\upsilon > 0$ and a compact set C verifying that*

$$G(l) \backslash \operatorname{Int}(G(u)) \oplus B_\upsilon[0] \subset U$$

and

$$G(l) \backslash \operatorname{Int}(G(u)) \oplus B_{\frac{\upsilon}{2}}[0] \subset C.$$

In addition,

$$\sup_C |f_n - f| = O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right), \text{ almost surely.}$$

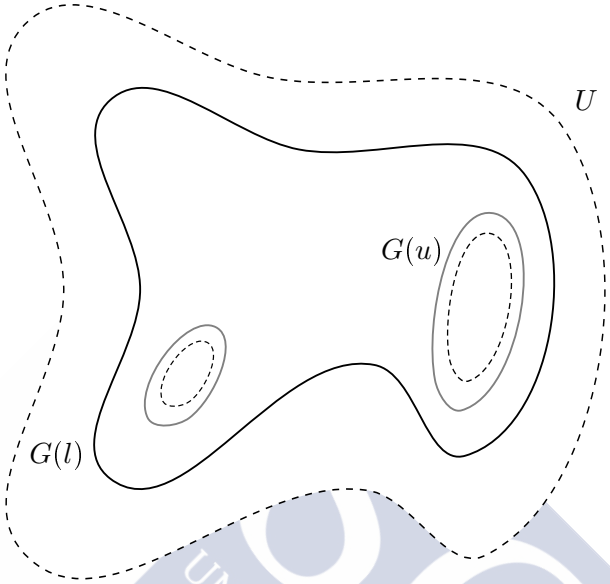


Figure B.1: Elements in Proposition B.0.2. $G(u)$ in gray, $G(l)$ in black and the open set U in dashed line.

Proof. According to the proof of Theorem 3 in Walther (1997), one can find $\upsilon > 0$ such that $G(l) \backslash \operatorname{Int}(G(u)) \oplus B_\upsilon[0] \subset U$, see Figure B.1. In addition, the kernel K satisfies the assumptions in Theorem 3.1 in Stute (1984). Following Walther (1997), one can prove that there exists a compact set C such that $G(l) \backslash \operatorname{Int}(G(u)) \oplus B_{\upsilon/2}[0] \subset C$ and such that if h_n is a sequence of the order $(\log n/n)^{1/(d+2p)}$ then it is verified that

$$\sup_C |f_n - K * f| = O\left(n^{-p/(d+2p)}\right), \text{ almost surely,} \tag{B.1}$$

$$\sup_C |K * f - f| = O\left(h_n^p\right), \tag{B.2}$$

where we write $K * f$ for $\int h_n^{-d} K((\cdot - x)/h_n) f(x)\, dx$. Equations (B.1) and (B.2) correspond to equations (13) and (14) in Walther (1997). By the triangle inequality, we can guarantee that, almost surely,

$$\sup_C |f_n - f| = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{p/(d+2p)}, n^{-p/(d+2p)}\right\}\right)$$

$$= O\left(\left(\frac{\log n}{n}\right)^{p/(d+2p)}\right). \qquad \square$$

Proposition B.0.3 corresponds to equation (15) in Walther (1997). The behavior of the kernel density estimator is studied in the complement of the compact set C.

Proposition B.0.3. *Let C the compact set in Proposition B.0.2. Under assumptions (A) and (K) established in Chapter 4, there exists $w > 0$ verifying that*

$$\mathbb{P}\left(\inf_{G(u)\cap C^c} f_n(x) > u + \frac{w}{2},\ \textit{eventually}\right) = 1$$

and

$$\mathbb{P}\left(\sup_{G(l)^c\cap C^c} f_n(x) < l - \frac{w}{2},\ \textit{eventually}\right) = 1.$$

Proposition B.0.4 corresponds to Lemma 2 (b) in Walther (1997). It establishes some interesting relationships between level sets with close enough thresholds.

Proposition B.0.4. *Under assumption (A) established in Chapter 4, there exists a constant $c > 0$ such that if t_1 and t_2 are such that $l \leqslant t_1 < t_2 \leqslant u$ and $t_2 - t_1 \leqslant (m/2)c$ then*

$$G(t_1) \ominus B_{\frac{2}{m}(t_2-t_1)}[0] \subset G(t_2) \subset G(t_1) \ominus B_{\frac{1}{2m}(t_2-t_1)}[0]$$

and

$$G(t_2) \oplus B_{\frac{1}{2m}(t_2-t_1)}[0] \subset G(t_1) \subset G(t_2) \oplus B_{\frac{2}{m}(t_2-t_1)}[0].$$

Finally, Proposition B.0.5 is presented. It corresponds to Lemma 3 in Walther (1997). This result is the only one used in Chapter 3 too.

Proposition B.0.5. *Let $K \subset \mathbb{R}^d$ be a compact set, $r > 0$ and let $\mathcal{X}_n$ be a i.i.d. sample generated from a distribution with density function f. Let $\mathcal{G}_K(r)$ be the family of sets defined in Definition 4.5.4.*

1. *If $f \geqslant b > 0$ on $A \in \mathcal{G}_K(r)$ and $0 < \epsilon < \min\{\overline{r}/2, r\}$ then*

$$\mathbb{P}\left(A \oplus B_{\overline{r}-2\epsilon}[0] \not\subset (A \cap \mathcal{X}_n) \oplus B_{\overline{r}}[0]\right)$$
$$\leqslant D\left(\epsilon, A \oplus B_{\overline{r}}[0]\right) \exp\left(-nab \min\{\overline{r} - \epsilon, r\}^{(d-1)/2} \epsilon^{(d+1)/2}\right).$$

where

$$D\left(\epsilon, A \oplus B_{\overline{r}}[0]\right) = \max\{card\ V : V \subset A \oplus B_{\overline{r}}[0],\ |x-y| > \epsilon \textit{ for different } x, y \in N\}$$

and a is a dimensional constant.

2. Further, if $f \geqslant b > 0$ *on* K, $0 < \epsilon < \min\{\overline{r}/3, 1\}$ *and* $r \geqslant \overline{r} - 2\epsilon$ *then*

$$\mathbb{P}\left(A \oplus B_{\overline{r}-3\epsilon}[0] \not\subset (A \cap \mathcal{X}_n) \oplus B_{\overline{r}}[0] \text{ for some } A \in \mathcal{G}_K(r)\right)$$
$$\leqslant D\left(\epsilon, K \oplus B_{\overline{r}}[0]\right) D\left(\frac{\epsilon}{10\overline{r}}, S^{d-1}\right) \exp\left(-nab(\overline{r} - 2\epsilon)^{(d-1)/2}(\epsilon/2)^{(d+1)/2}\right)$$

where S^{d-1} *denotes the unit sphere.*

Summary in Galician
Resumo en galego

A *estimación de conxuntos* abre un capítulo relativamente recente da estatística matemática onde a xeometría xoga un papel moi relevante. Esta teoría ten como finalidade estimar un conxunto no espazo Euclidiano a partir dunha mostra aleatoria de puntos cuxa distribución está intimamente relacionada con el. A resolución deste tipo de problemas ten aplicacións interesantes na análise clúster (ver Hartigan, 1975), en control de calidade (ver Devroye e Wise, 1980 ou Baíllo et al., 2000) ou na análise de imaxes para reconstruir, por exemplo, o hábitat dunha planta ou dunha especie animal (ver De Haan e Resnick, 1994). Para unha revisión en profundidade, ver Cuevas e Fraiman (2010). Neste traballo centrarémonos no problema de estimación do soporte e de conxuntos de nivel. Existen distintas alternativas na literatura dependendo das condicións de forma asumidas sobre o conxunto a reconstruir. Se non dispoñemos de ningunha información a priori, será preciso considerar estimadores flexibles que nos permitan abordar eficientemente a maior cantidade de situacións posibles. Noutro caso, se restrinximos a familia de conxuntos a estimar, poderemos traballar con estimadores máis sofisticados, que se adapten mellor as restricións xeométricas establecidas. A maioría destes estimadores dependen fortemente da elección de parámetros de suavizado ao igual que sucede no contexto da estimación funcional non paramétrica. O obxectivo principal desta tese consiste en estimalos de xeito automático e consistente para, logo, analizar o comportamento dos estimadores resultantes dos conxuntos a reconstruir.

Antes de revisar os métodos de estimación para o soporte e para os conxuntos de nivel que existen na literatura, imos establecer criterios que nos permitan evaluar a calidade dos mesmos. As distancias entre conxuntos miden a proximidade e similitude do estimador ao conxunto teórico a reconstruir. Existen varias posibilidades para definir a distancia entre conxuntos tales coma a distancia en medida, en medida ponderada ou a distancia Hausdorff. Se A e C son dous conxuntos de Borel acotados, defínese a distancia en medida entre A e C como

$$d_\mu(A, C) = \mu(A \triangle C),$$

onde μ denota a medida de Lebesgue e $\triangle$, a diferencia simétrica, isto é,

$$A \triangle C = (A \backslash C) \cup (C \backslash A).$$

Polo tanto, $d_\mu(A, C)$ é unha medida útil para cuantificar a similitude entre os conxuntos A e C en termos de contidos. De forma máis xeral, se f é unha función de densidade en $\mathbb{R}^d$ e A e C, dous conxuntos de Borel (non necesariamente acotados) entón é posible definir a distancia en medida ponderada

$$d_{\mu_f}(A, C) = \int_{A \triangle C} f(t)\ dt.$$

Intuitivamente, $d_{\mu_f}(A, C)$ representa a probabilidade de que unha observación da variable aleatoria con función de densidade f pertenza só a un dos dous conxuntos A e C. En xeral, a distancia en medida ponderada concede máis peso nas rexións onde os datos tenden a ser máis densos. A distancia Hausdorff está definida sobre o espazo de subconxuntos non baleiros e compactos en $\mathbb{R}^d$. Sexan $A, C \subset \mathbb{R}^d$

$$d_H(A, C) = \max \left\{ \sup_{a \in A} d(a, C), \sup_{c \in C} d(c, A) \right\},$$

onde $d(a, C) = \inf\{\|a - c\| : c \in C\}$. Ou equivalentemente,

$$d_H(A, C) = \inf\{\varepsilon > 0 : A \subset C \oplus B_\varepsilon(0),\ C \subset A \oplus B_\varepsilon(0)\},$$

onde $B_\varepsilon(0)$ é a bola aberta de centro 0 e radio ε e $\oplus$ denota a suma de Minkowski con $C \oplus B_\varepsilon(0) = \{c + b : c \in C,\ b \in B_\varepsilon(0)\}$. Neste caso, $d_H(A, C)$ cuantifica a proximidade física entre os conxuntos A e C. Pode probarse que a distancia Hausdorff é unha métrica, ver Sección 2.4 en Edgar (1990) ou Sección 1.4 en Matheron (1975) para máis detalles.

A **estimación do soporte** é quizáis o problema máis sinxelo da estimación de conxuntos. Formalmente, ocúpase de estimar o soporte $S \subset \mathbb{R}^d$ dunha distribución de probabilidade absolutamente continua $\mathbb{P}_X$ da variable aleatoria X a partir dunha mostra aleatoria simple $\mathcal{X}_n = \{X_1, ..., X_n\}$ de X. Se non se asume ningunha restrición de forma sobre S, Chevalier (1976) e Devroye e Wise (1980) propuxeron un estimador moi simple para o soporte

$$\bigcup_{i=1}^{n} B_{\epsilon_n}[X_i],$$

onde $B_{\epsilon_n}[X_i]$ denota a bola pechada de centro X_i e radio ϵ_n. Ver tamén Korostelëv e Tsybakov (1993), Cuevas e Rodríguez-Casal (2004) e Biau et al. (2008) onde o comportamento deste estimador foi analizado. O problema de seleccionar o parámetro de

suavizado ϵ_n para incorporar información a priori sobre S no estimador foi considerado en Baíllo et al. (2000) e en Baíllo and Cuevas (2001). Asumiron que S era conexo e estrelado, respectivamente, incorporando estas informacións a priori no estimador de Devroye e Wise (1980).

Reconstrucións máis sofisticados poden usarse se existe información adicional sobre o conxunto. Rényi e Sulanke (1963) e Rényi e Sulanke (1964) estudiaron o caso no que $S \subset \mathbb{R}^2$ é convexo. Propuxeron como estimador a envoltura convexa dos puntos mostrais, $\mathrm{conv}(\mathcal{X}_n)$. Korostelëv e Tsybakov (1993) ou Dümbgen e Walther (1996) analizaron o comportamento deste estimador.

Sen embargo, a convexidade pode ser un condición de forma demasiado restrictiva na práctica. Se S non é convexo, $\mathrm{conv}(\mathcal{X}_n)$ podería non ser a mellor opción. Por iso, é preciso considerar unha propiedade xeométrica máis flexible, a $r-$convexidade onde $r > 0$. En lugar de asumir, coma no caso convexo, a existencia dun hiperplano separador para cada punto exterior, asumimos que existe unha bola aberta separadora de radio r. Polo tanto, se un conxunto é $r-$convexo entón tamén é $r'-$convexo para calquera $0 < r' \leqslant r$. Se supoñemos que o soporte S é $r-$convexo, un estimador natural para S sería a envoltura $r-$convexa da mostra

$$C_r(\mathcal{X}_n) = \bigcap_{\{B_r(x):B_r(x)\cap\mathcal{X}_n=\varnothing\}} (B_r(x))^c.$$

Pode probarse facilmente que $C_{r'}(\mathcal{X}_n) \subset C_r(\mathcal{X}_n)$ se, de novo, $0 < r' \leqslant r$. Por outra banda, a $r-$convexidade de S implica que unha bola de radio r roda libremente en $\overline{S^c}$, ver Cuevas et al. (2012). É dicir, para cada punto $s \in \partial S$ existe $x \in \mathbb{R}^d$ tal que $s \in B_r[x] \subset \overline{S^c}$. A propiedade de rodamento libre garantiza certa suavidade na fronteira. Para analizar en detalle as relacións existentes entre a $r-$convexidade e a propiedade de rodamento libre, ver Walther (1997). Rodríguez-Casal (2007) probou que, se unha bola de radio r rodaba libremente en S e en $\overline{S^c}$, $d_\mu(S, C_r(\mathcal{X}_n)) = O((\log n/n)^{2/(d+1)})$, case seguro. As mesmas tasas de converxencia foron obtidas para $d_H(S, C_r(\mathcal{X}_n))$ e $d_H(\partial S, \partial C_r(\mathcal{X}_n))$. Anque a familia de $r-$convexos é máis ampla que a de convexos, as tasas obtidas son da mesma orde que as da envoltura convexa para estimar soportes convexos, ver Dümbgen e Walther (1996). Sen embargo, este estimador presenta unha forte limitación. Na práctica, S é descoñecido e, como consecuencia, r tamén. Mandal e Murthy (1997) propuxeron un método para estimar r a partir de $\mathcal{X}_n$ só válido no caso bidimensional. Nótese que se r está demasiado próximo a cero entón $C_r(\mathcal{X}_n)$ coincide practicamente con $\mathcal{X}_n$. Sen embargo, se r toma un valor demasiado grande, $C_r(\mathcal{X}_n)$ podería ser case igual a $\mathrm{conv}(\mathcal{X}_n)$. De feito, sería posible atopar un *spacing* (bola pechada dentro de $C_r(\mathcal{X}_n)$) con área máis ou menos grande que non interseca a $\mathcal{X}_n$.

Por outra banda, a **estimación de conxuntos de nivel** ocúpase de reconstruir conxuntos $G(t) = \{x \in \mathbb{R}^d : f(x) \geqslant t\}$ onde f denota a función de densidade de X e

$t > 0$. Na maioría das aplicacións, os usuarios precisan garantir que o conxunto de nivel ten un contido en probabilidade fixado máis grande ou igual ca $1 - \tau$ onde $\tau \in (0, 1)$. Neste caso, o valor de t é descoñecido e é desexable estimar:

$$L(\tau) = \{x \in \mathbb{R}^d : f(x) \geqslant f_\tau\}$$

onde

$$f_\tau = \sup \left\{ y \in (0, \infty) : \int_{-\infty}^{\infty} f(t) \mathbb{I}_{\{f(t) \geqslant y\}} \geqslant 1 - \tau \right\}.$$

Dous pasos son precisos para reconstruir $L(\tau)$ de forma automática a partir de $\mathcal{X}_n$ xerada, en este caso, por f. Primeiro, f_τ debe ser estimado para satisfacer o contido en probabilidade fixado. Logo, unha das tres metodoloxías diferentes que existen para reconstruir un conxunto de nivel debe ser seleccionada.

Se f_n denota un estimador nonparamétrico, usualmente, o estimador tipo núcleo entón o estimador de f_τ podería calcularse de tres xeitos diferentes. Unha opción pasa por considerar métodos de integración numérica para resolver a ecuación

$$\int_{\{f_n \geqslant t\}} f_n(x) dx = 1 - \tau$$

en t. Este algoritmo podería ser ineficiente dende un punto de vista computacional. Para resultados de consistencia, ver Cadre (2006). Sen embargo, o método proposto por Hyndman (1996) ten un coste computacional menor e pode resultar verdadeiramente útil para dimensión xeral. Estima f_τ calculando o cuantil τ da distribución empírica de $f_n(X_1), ..., f_n(X_n)$. Ver Cadre et al. (2009), para resultados de consistencia. A última alternativa consiste en imitar o procedemento empírico proposto en Walther (1997).

A continuación, detallamos brevemente as tres metodoloxías de estimación de conxuntos de nivel. Elexir un algoritmo ou outro depende, ao igual que na estimación do soporte, das restricións de forma asumidas.

A estimación *plug-in* é a elección máis natural para estimar $L(\tau)$ cando non existe información sobre a xeometría do conxunto. Consiste en reemplazar f por f_n. É ben sabido que f_n depende fortemente da elección da ventá, ver Wand e Jones (1995). Polo tanto, o problema práctico da metodoloxía plug-in é a selección da mesma. Este problema foi considerado por vez primeira no contexto da estimación de conxuntos de nivel por Baíllo e Cuevas (2006). Ver tamén os selectores descritos en Samworth e Wand (2010) ou Singh et al. (2009).

Os métodos de *exceso de masa* asumen que o investigador ten información a priori sobre a forma do conxunto de nivel $G(t)$. Esta metodoloxía foi proposta por Hartigan (1987) e Müller e Sawitzki (1987). Ver tamén Polonik (1995). A idea base destes algoritmos é moi sinxela: $G(t)$ maximiza o funcional

$$H_t(B) = \mathbb{P}(B) - t\mu(B).$$

H_t pode ser estimado empiricamente. Por tanto, se $G(t)$ pertence a unha familia de conxuntos dada entón o estimador é o máximo da versión empírica do funcional previo sobre a familia de conxuntos considerada.

A última metodoloxía é un *híbrido* das dúas anteriores. Ao igual que os métodos de exceso de masa, asume restricións de forma sobre a clase de conxuntos considerada e, como os métodos plug-in, precisa estimar f de forma nonparamétrica. Walther (1997) propuxo o método de suavizado granulométrico para reconstruir $L(\tau)$ adaptando o estimador do soporte de Devroye e Wise (1980) ao contexto da estimación de conxuntos de nivel asumindo $r-$convexidade como restrición de forma. Na práctica, o estimador é unha unión de bolas pechadas de radio r con centros nos puntos de $\mathcal{X}_{n,+}(\hat{f}_\tau)$ que distan a lo menos r dos puntos de $\mathcal{X}_{n,-}(\hat{f}_\tau)$, onde $\mathcal{X}_{n,+}(\hat{f}_\tau) = \{X \in \mathcal{X}_n : f_n(X) \geqslant \hat{f}_\tau\}$, $\mathcal{X}_{n,-}(\hat{f}_\tau) = \mathcal{X}_n \backslash \mathcal{X}_{n,+}(\hat{f}_\tau)$ e $\hat{f}_\tau$ é un estimador de f_τ. Por suposto, a principal desvantaxe deste método é o descoñecemento do parámetro r.

No **Capítulo 1** introducimos os problemas de estimación do soporte e de conxuntos de nivel e revisamos con detalle as dúas ferramentas matemáticas básicas na estimación de conxuntos, distancias e propiedades xeométricas. A continuación, resumimos brevemente os principais resultados obtidos neste traballo de investigación. Finalmente, mostramos os conxuntos de datos empregados. Primeiro, preséntanse dous conxuntos de datos reais. Decribiremos un procedemento para xerar mostras uniformes sobre as rexións de auga contidas en dúas imaxes do Mar de Aral tomadas en 2000 e 2011 polo satélite Terra da NASA. A continuación, preséntase un conxunto de datos derivado do estudo realizado en Henderson et al. (2002). Contén 1221 pares de coordenadas de residencia para 233 casos de leucemia e 988 controis rexistrados entre 1982 e 1998 en Lancashire e Greater Manchester. En canto os modelos de simulación, descríbese un conxunto de 18 densidades un-dimesionais (ver Marron e Wand, 1992 e Berlinet e Devroye, 1994) e tres modelos de soportes $r-$convexos contidos en $\mathbb{R}^2$.

No **Capítulo 2** realizamos unha revisión bibliográfica sobre a estimación do soporte e de conxuntos de nivel. Ademais, propóñense dous métodos híbridos novos para estimar conxuntos convexos e $r-$convexos, respectivamente. Ao igual que o método de suavizado granulométrico, ámbolos dous adaptan estimadores do soporte ó contexto da estimación de conxuntos de nivel. O primeiro, estima o conxunto de nivel calculando $\text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau))$. O segundo, $C_r(\mathcal{X}_{n,+}(\hat{f}_\tau))$. Nótese que ningún dos dous algoritmos ten en conta a información almacenada en $\mathcal{X}_{n,-}(\hat{f}_\tau)$ sobre o complementario do conxunto de nivel. Nun extenso estudo de simulación amosamos o compartamento práctico das tres metodoloxías existentes para conxuntos de nivel das 18 densidades un-dimensionais consideradas xa que algúns dos algoritmos non foron extendidos a dimesión superior. Na comparativa, incluironse os selectores clásicos de ventá de Sheather e Jones e validación cruzada para ser comparados cos métodos plug-in específicos para estimar conxuntos de

nivel. Os resultados obtidos amosan que os métodos plug-in clásicos, onde a ventá é seleccionada para estimar f, son máis competitivos que os selectores específicos de Baíllo e Cuevas (2006), Samworth e Wand (2010) ou Singh et al. (2009). Ademais, se non se asumen condicións de forma sobre o conxunto de nivel, son as mellores alternativas para reconstruilo. Noutro caso, os métodos de Müller e Sawitzki ou os híbridos poden ser considerados. A primeira metodoloxía non proporcionou resultados demasiado competitivos. Sen embargo, unha das súas principais ventaxas é que non precisa suavizar os datos para reconstruir o conxunto de nivel. En canto os métodos híbridos, os resultados obtidos son bastante prometedores á hora de estimar conxuntos $r-$convexos a pesar de que r, ao ser descoñecido, foi fixado de antemán. Seleccionalo de forma automática podería mellorar de novo os resultados obtidos.

No **Capítulo 3** propoñemos un novo método automático para estimar soportes $r-$convexos asumindo que a mostra se distribúe uniformemente en S. Dacordo cos comentarios previos, o parámetro de suavizado r pode resultar bastante influinte nas estimacións. Un método descriptivo gráfico é introducido como primeira aproximación ó problema de selección do parámetro de suavizado r. Logo, presentamos un algoritmo estocástico para seleccionar o seu valor óptimo, r_0, cando S non é convexo, onde

$$r_0 = \sup\{\gamma > 0 : C_\gamma(S) = S\}.$$

Se S fora convexo, $r_0 = \infty$. Ademais, se r_0 é máximo do conxunto $\{\gamma > 0 : C_\gamma(S) = S\}$ entón é posible garantir que S é tamén r_0-convexo. Neste caso, está claro que $C_r(\mathcal{X}_n)$, con $r < r_0$, non sería un estimador admisible xa que infraestimaría S respeto de $C_{r_0}(\mathcal{X}_n)$. Isto débese a que $C_r(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$ e entón, $d_\mu(C_{r_0}(\mathcal{X}_n), S) \leqslant d_\mu(C_r(\mathcal{X}_n), S)$ (as mesmas conclusións mantéñense para a distancia de Hausdorff). Polo tanto, o único parámetro admisible sería $r = r_0$. Por outra banda, se $r > r_0$ entón $C_r(\mathcal{X}_n)$ podería sobreestimar S considerablemente

Para estimar r_0, empleamos o test de uniformidade proposto por Berrendero et al. (2012). Dado $r^{'} > 0$, o contraste rexeita que $\mathcal{X}_n$ é uniforme sobre $C_{r^{'}}(\mathcal{X}_n)$ se $C_{r^{'}}(\mathcal{X}_n)$ contén unha bola ou *spacing* de área suficientemente grande que non interseca a $\mathcal{X}_n$. Neste traballo, asumiremos unha aproximación oposta. Supoñemos que $\mathcal{X}_n$ se distribúe uniformemente en S. Entón, se existe un *spacing* demasiado grande en $C_{r^{'}}(\mathcal{X}_n)$ deduciremos que estamos sobreestimando S. Un valor inferior a $r^{'}$ debería ser considerado. En definitiva, propoñemos estimar r_0 como

$$\hat{r}_0 = \sup\{\gamma > 0 : \text{ A hipótese nula de uniformidade é aceptada sobre } C_\gamma(\mathcal{X}_n)\}.$$

Probamos que, con probabilidade tendendo a un, $\hat{r}_0 \geqslant r_0$. Ademais, $\hat{r}_0$ converxe a r_0, en probabilidade baixo (R^r_λ) como condición de forma:

(R_λ^r) Unha bola pechada de radio $\lambda > 0$ roda libremente en S e unha bola pechada de radio $r > 0$ roda libremente en $\overline{S^c}$.

Baixo (R_λ^r), probamos que o parámetro r_0 é un máximo. Unha vez que a consistencia de $\hat{r}_0$ como estimador do parámetro r_0 foi establecida, cómpre analizar a calidade de $C_{\hat{r}_0}(\mathcal{X}_n)$ como estimador do soporte S. Baixo certas condicións de regularidade, se o límite da distancia de Hausdorff entre S e $C_{\overline{r}}(S)$ é cero cando $\overline{r} \geqslant r_0$ entón $C_{\hat{r}_0}(\mathcal{X}_n)$ é un estimador consistente en Hausdorff do soporte S en probabilidade. Os mesmos resultados mantéñense para a distancia en medida. Sen embargo, se a hipótese anterior non se cumple, dado que $\hat{r}_0 \geqslant r_0$, foi preciso considerar $C_{r_n(\mathcal{X}_n)}$ como estimador do soporte onde $r_n = \nu\hat{r}_0$ con $\nu \in (0,1)$. Para tal elección, mantéñense as tasas obtidas por Rodríguez-Casal (2007) con r coñecido. Nun estudo de simulación compáranse a nosa proposta e o método de Mandal e Murthy (1997) considerando tres modelos distintos, diferentes valores para os tamaños mostrais e para o nivel de significación do contraste empregado. En xeral, o noso algoritmo presenta un mellor comportamento global para estimar r_0. O método de Mandal e Murthy (1997) infraestima r_0, principalmente para valores altos do tamaño de mostra. Por outra banda, considerar valores do nivel de significación moi próximos a cero reduce o número de datos atípicos de forma considerable. Ademais, redúcese tamén o risco de rexeitar a hipótese nula de uniformidade cando é certa e, se o modelo de simulación considerado non é demasiado complexo, os resultados obtidos son lixeiramente mellores. Finalmente, mostramos un exemplo con datos reais onde propoñemos un procedemento útil para analizar se o Mar de Aral perdeu auga nos últimos anos empregando as dúas imaxes do satélite Terra da NASA tomadas en 2000 e 2011. Tal e como cabía esperar, os resultados obtidos permiten concluir que o Mar de Aral perdeu auga. De feito, a superficie de auga en 2000 é sobre o triple da superficie de auga en 2011.

No **Capítulo 4** presentamos un método novo e automático para estimar conxuntos de nivel $G(t)$ $r-$convexos. O método da envoltura $r-$convexa definido no Capítulo 1 presenta bos resultados a pesar de que o parámetro r é descoñecido. O valor óptimo de r depende, neste caso, do nivel t considerado,

$$r_0(t) = \sup\{\gamma > 0 : C_\gamma(G(t)) = G(t)\}.$$

De novo, se $G(t)$ é convexo $r_0(t) = \infty$. Modificaremos lixeiramente o método da envoltura $r-$convexa presentado no Capítulo 1 para obter un estimador automático de $r_0(t)$ e de $G(t)$. Empregando a información sobre $G(t)^c$ propoñemos estimar $r_0(t)$ coma,

$$\hat{r}_0(t) = \sup\{\gamma > 0 : C_\gamma(\mathcal{X}_n^+(t)) \cap \mathcal{X}_n^-(t) = \varnothing\},$$

onde

$$\mathcal{X}_n^+(t) = \{X \in \mathcal{X}_n : f_n(X) \geqslant t + D_n\} \text{ e } \mathcal{X}_n^-(t) = \{X \in \mathcal{X}_n : f_n(X) < t - D_n\}$$

e $D_n = M(\log n/n)^{p/(d+2p)}$ para unha constante $M > 0$ suficientemente grande. De novo probamos que, con probabilidade un, $\hat{r}_0(t) \geqslant r_0(t)$ e $\hat{r}_0(t)$ converxe a $r_0(t)$ e

$$d_H(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{p/(d+2p)}, \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right\}\right), \text{ case seguro.}$$

As tasas obtidas coinciden coas do método de suavizado granulométrico cando o parámetro r é coñecido. Se r é descoñecido, o método de Walther (1997) proporciona as mesmas tasas cun termo de penalización. Isto non sucede para o método proposto xa que o paramétro de suavizado é estimado de forma automática a partir de $\mathcal{X}_n$.

Na práctica e como primeiro paso, sería natural determinar os conxuntos $\mathcal{X}_n^+(t)$ e $\mathcal{X}_n^-(t)$. Sen embargo, ámbolos dous dependen da sucesión D_n. Un procedemento bootstrap foi proposto neste traballo para seleccionala de forma automática a partir da mostra orixinal $\mathcal{X}_n$. Para ilustrar a metodoloxía presentada, empregaremos as mostras de casos de leucemia e controis correspondentes coas áreas de Lancashire e Greater Manchester. Resulta moi interesante coñecer se a incidencia xeográfica desta enfermidade mostra algunha tendencia ó clustering no espazo xeográfico considerado. Por exemplo, é conveniente analizar se os casos de leucemia adoitan ocorrer preto doutros casos. Este problema cobrou moita importancia nos últimos anos, á luz das preocupacións levantadas sobre posibles conexións entre incidencia de certas enfermidades e potenciais fontes de contaminación do medio ambiente. Estudiamos os conxuntos de nivel para as mostras dos casos e dos controis. Os resultados obtidos permiten observar un exceso na intensidade dos casos respeto da poboación asociada, moi probablemente, a factores industriais.

Finalmente, as fórmulas das densidades un-dimensionais empregadas como modelos de proba no estudo de simulación do Capítulo 2 son mostradas no Apéndice A. No Apéndice B, recóllense unha serie de resultados teóricos que aparecen en Walther (1997). Son verdadeiramente útiles para simplificar as probas obtidas nos Capítulos 3 e 4.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliography

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rodgers, W.H. and Tuckey, J.W. (1972). Robust Estimates of Location: Survey and Advances. Princeton University Press.

[2] Arias-Castro, E. and Rodríguez-Casal, A. (2014). On the estimation of the perimeter of a domain with smooth boundary. Technical Report.

[3] Baíllo, A. (2003). Total error in a plug-in estimator of level sets. Statistics & Probability Letters, 65, 411–417.

[4] Baíllo, A., Cuesta-Albertos, J.A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. Statistics & Probability Letters, 53, 27–35.

[5] Baíllo, A. and Cuevas, A. (2001). On the estimation of a star-shaped set. Advances in Applied Probability, 33, 717–726.

[6] Baíllo, A. and Cuevas, A. (2006). Parametric versus nonparametric tolerance regions in detection problems. Computational Statistics, 21, 523–536.

[7] Baíllo, A., Cuevas, A. and Justel, A. (2000). Set estimation and nonparametric detection. Canadian Journal of Statistics, 28, 765–782.

[8] Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates. Publications de l'Institut de Statistique de l'Université de Paris, 38, 3–59.

[9] Berrendero, J. R., Cuevas, A. and Pateiro-López, B. (2012). A multivariate uniformity test for the case of unknown support. Statistics and Computing, 22, 259–271.

[10] Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. Journal of the Royal Statistical Society, 154, 143–155.

[11] Biau, G., Cadre, B. and Pelletier, B. (2008). Exact rates in density support estimation. Journal of Multivariate Analysis, 99, 2185–2207.

[12] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 353–360.

[13] Bräker, H. and Hsing, T. (1998). On the area and perimeter of a random convex hull in a bounded convex set. Probability Theory and Related Fields, 111, 517–550.

[14] Cadre, B. (2006). Kernel estimation of density level sets. Journal of Multivariate Analysis, 97, 999–1023.

[15] Cadre, B., Pelletier, B. and Pudlo, P. (2009). Clustering by estimation of density level sets at a fixed probability. Available in arXiv 00397437.

[16] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. Computational Statistics & Data Analysis, 17, 153–176.

[17] Cuevas, A. (1990). On pattern analysis in the non-convex case. Kybernetes, 19, 26–33.

[18] Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. Canadian Journal of Statistics, 28, 367–382.

[19] Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. Annals of Statistics, 25, 2300–2312.

[20] Cuevas, A. and Fraiman, R. (2010). Set estimation. In Kendall, W. S. and Molchanov, I. S., editors, New Perspectives in Stochastic Geometry, 374–397. Oxford University Press.

[21] Cuevas, A., Fraiman, R. and Pateiro-López, B. (2012). On statistical properties of sets fulfilling rolling-type conditions. Advances in Applied Probability, 44, 311–329.

[22] Cuevas, A., Fraiman, R. and Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. Annals of Statistics, 35, 1031-1051.

[23] Cuevas, A., González-Manteiga, W. and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. Australian & New Zealand Journal of Statistics, 48, 7–19.

[24] Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. Advances in Applied Probability, 36, 340–354.

[25] Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society, 52, 73–104.

[26] Chernoff, H. (1964). Estimation of the mode. Annals of the Institute of Statistical Mathematics, 16, 31–41.

[27] Chevalier, J. (1976). Estimation du support et du contour du support d'une loi de probabilité. Annales de l'Institut Henri Poincaré, 12, 339–364.

[28] DasGupta, A., Ghosh, J.K. and Zen, M.M. (1995). A general method for constructing confidence sets in arbitrary dimensions: with applications. Annals of Statistics, 23, 1408–1432.

[29] De Haan, L. and Resnick, S. (1994). Estimating the home range. Journal of Applied Probability, 31, 700–720.

[30] Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. SIAM Journal on Applied Mathematics, 38, 480–488.

[31] Diggle, P.J. (2013). Statistical Analysis of Spatial and Spatio-Temporal Point Patterns (third edition). Boca Raton:Chapman and Hall/CRC Press.

[32] Diggle, P. J. Index of Point Pattern Datasets. Retrieved February 12, 2014 from http://www.lancaster.ac.uk/staff/diggle/pointpatternbook/datasets/.

[33] Diggle, P. J., Gatrell, A. C. and Lovett, A. A. (1990). Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology. Spatial epidemiology. London: Pion.

[34] Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximations of convex sets. Advances in Applied Probability, 28, 384–393.

[35] Duong, T. (2007). Ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. Journal of Statistical Software, 21, 1–16.

[36] Eddy, W.F. and Hartigan, J.A. (1977). Uniform convergence of the empirical distribution function over convex sets. Annals of Statistics, 5, 370–374.

[37] Edelsbrunner, H. (To appear). Alpha shapes — a survey. In Tessellations in the Sciences.

[38] Edelsbrunner, H., Kirkpatrick, D. G. and Seidel, R. (1983). On the shape of a set of points in the plane. Information Theory, IEEE Transactions on, 29, 551–559.

[39] Edgar, G. A. (1990). Measure, Topology, and Fractal Geometry. Undergraduate Texts in Mathematics. Springer-Verlag, New York.

[40] Gardner, A. B., Krieger, A. M., Vachtsevanos, G. and Litt, B. (2006). Oneclass novelty detection for seizure analysis from intracranial EEG. Journal of Machine Learning Research, 7, 1025–1044.

[41] Gayraud, G. and Rousseau, J. (2005). Rates of convergence for a bayesian level set estimation. Scandinavian Journal of Statistics, 32, 639–660.

[42] Geffroy, J. (1964). Sur un problème d'estimation géométrique. Publications de l'Institut de statistique de l'Université de Paris, 13, 191–210.

[43] Goldenshluger, A. and Zeevi, A. (2004). The hough transform estimator. Annals of Statistics, 32, 1908–1932.

[44] Grübel, R. (1988). The length of the short. Annals of Statistics, 16, 619–628.

[45] Hartigan, J.A. (1975). Clustering Algorithms. John Wiley & Sons, New York.

[46] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. Journal of the American Statistical Association, 82, 267–270.

[47] Henderson, R., Shimakura, S. and Gorst, D. (2002). Modeling spatial variation in leukemia survival data. Journal of the American Statistical Association, 97, 965–972.

[48] Huo, X. and Lu, J. C. (2004). A network flow approach in finding maximum likelihood estimate of high concentration regions. Computational Statistics & Data Analysis, 46, 33–56.

[49] Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50, 120–126.

[50] Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. Computational Statistics & Data Analysis, 50, 760–774.

[51] Janson, S. (1987). Maximal spacings in several dimensions. Annals of Probability, 15, 274–280.

[52] Jiménez, R. and Yukich, J.E. (2011). Nonparametric estimation of surface integrals. Annals of Statistics, 39, 232–260.

[53] Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. Journal of the Royal Statistical Society, 47, 559–573.

[54] Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. Journal of the American Statistical Association, 97, 692–701.

[55] Klemelä, J. (2004). Visualization of multivariate density estimates with level set trees. Journal of Computational and Graphical Statistics, 13, 599–620.

[56] Klemelä, J. (2006). Visualization of multivariate density estimates with shape trees. Journal of Computational and Graphical Statistics, 15, 372–397.

[57] Korostelëv, A. and Tsybakov, A. (1993). Minimax Theory of Image Reconstruction. Lecture Notes in Statistics, 82. Springer-Verlag, New York.

[58] Lientz, B.P. (1970). Results on nonparametric modal intervals. SIAM Journal on Applied Mathematics, 19, 356–366.

[59] Mammen, E. and Polonik, W. (2013). Confidence regions for level sets. Journal of Multivariate Analysis, 122, 202–214.

[60] Mandal, D. P. and Murthy, C. A. (1997). Selection of alpha for alpha-hull in $\mathbb{R}^2$. Pattern Recognition, 30, 1759–1767.

[61] Markou, M. and Singh, S. (2003). Novelty detection: a reviewpart 1: statistical approaches. Signal Processing, 83, 2481–2497.

[62] Marron, J. and Wand, M. (1992). Exact mean integrated squared error. Annals of Statistics, 20, 712–736.

[63] Mason, D. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. Annals of Applied Probability, 19, 1108–1142.

[64] Matheron, G. (1975). Random Sets and Integral Geometry. Wiley, New York.

[65] Molchanov, I.S. (1998). A limit theorem for solutions of inequalities. Scandinavian Journal of Statistics, 25, 235–242.

[66] Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint No. 398, SFB123, University Heidelberg.

[67] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. Journal of the American Statistical Association, 86, 738–746.

[68] NASA, Earth Observatory. Shrinking Aral Sea. Retrieved February 12, 2013 from http://earthobservatory.nasa.gov/Features/WorldOfChange/aral_sea.php.

[69] Nolan, D. (1991). The excess-mass ellipsoid. Journal of Multivariate Analysis, 39, 348–371.

[70] Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association, 85, 66–72.

[71] Pateiro-López, B. (2008). Set Estimation Under Convexity Type Restrictions. Universidade de Santiago de Compostela.

[72] Pateiro-López, B. and Rodríguez-Casal, A. (2010). Generalizing the convex hull of a sample: the R package alphahull. Journal of Statistical Software, 34, 1–28.

[73] Pateiro-López, B. and Rodríguez-Casal, A. (2013). Recovering the shape of a point cloud in the plane. TEST, 22, 19–45.

[74] Perkal, J. (1956). Sur les ensembles $\varepsilon-$convexes. Colloquium Mathematicae, 4, 1–10.

[75] Polonik (1995). Measuring mass concentration and estimating density contour clusters-an excess mass approach. Annals of Statistics, 23, 855–881.

[76] Reitzner, M. (2003). Random polytopes and the efron-stein jackknife inequality. Annals of Probability, 31, 2136–2166.

[77] Renka, R. J. (1996). Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation package. ACM Transactions on Mathematical Software, 22, 1–8.

[78] Rényi, A. and Sulanke, R. (1963). Über die konvexe Hülle von n zufällig gewählten Punkten. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 2, 75–84.

[79] Rényi, A. and Sulanke, R. (1964). Über die konvexe Hülle von n zufällig gewählten Punkten. II. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 3, 138–147.

[80] Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability, 15, 1154–1178.

[81] Ripley, B. D. and Rasson, J. P. (1977). Finding the edge of a poisson forest. Journal of Applied Probability, 14, 483–491.

[82] Robertson, T.J. and Cryer, J.D. (1974). An iterative procedure for estimating the mode. Journal of the American Statistical Association, 69, 1012–1016.

[83] Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. Annales de l'I.H.P.- Probabilités & Statistiques, 43, 763–774.

[84] Rodríguez-Casal, A. and Saavedra-Nieves, P. (2014). A fully data-driven method for estimating the shape of a point cloud. ArXiv preprint arXiv:1404.7397.

[85] Roederer, M. and Hardy, R. R. (2001). Frequency difference gating: A multivariate method for identifying subsets that differ between samples. Cytometry, 45, 56–64.

[86] Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics, 9, 65–78.

[87] Saavedra-Nieves, P., González-Manteiga, W. and Rodríguez-Casal, A. (2014). Level set estimation. In Akritas, M.G. et al., editors, Topics in Nonparametric Statistics, Springer Proceedings in Mathematics & Statistics.

[88] Saavedra-Nieves, P., González-Manteiga, W. and Rodríguez-Casal, A. (Under second review). A comparative simulation study of data-driven methods for estimating density level sets. Journal of Statistical Computation and Simulation.

[89] Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. Journal of the American Statistical Association, 74, 329–339.

[90] Samworth, R.J. and Wand , M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. Annals of Statistics, 38, 1767–1792.

[91] Schneider, R. (1988). Random approximation of convex sets. Journal of Microscopy, 151, 211–227.

[92] Schneider, R. (1993). Convex Bodies: The Brunn-Minkowski Theory. Cambridge University Press.

[93] Serra, J. (1984). Image Analysis and Mathematical Morphology. Academic Press, London.

[94] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society: Series B, 53, 683–690.

[95] Simonoff, J. S. (1996). Smoothing Methods in Statistics. Springer.

[96] Singh, A., Scott, C. and Nowak, R. (2009). Adaptive hausdorff estimation of density level sets. Annals of Statistics, 37, 2760–2782.

[97] Stuetzle, W. and Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. Journal of Computational and Graphical Statistics, 19, 397–418.

[98] Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. Annals of Probability, 12, 361–379.

[99] Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. Annals of Statistics, 25, 948–969.

[100] Venter, J.H. (1967). On estimation of the mode. Ibid, 38, 1445–1446.

[101] Wand, M. (2005). Statistical methods for flow cytometric data. Presentation. Available in http:// www.uow.edu.au/ mwand/talks.html.

[102] Wand, M. and Jones, M. (1995). Kernel Smoothing. Chapman & Hall.

[103] Walther, G. (1997). Granulometric smoothing. Annals of Statistics, 25, 2273–2299.

[104] Walther, G. (1999). On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. Mathematical Methods in the Applied Sciences, 22, 301–316.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Notation

$\mathcal{X}_n$	Sample points $\{X_1, ..., X_n\}$, 11
$\mathbb{P}_X$	Distribution probability function of X, 11
$\mathbb{P}_n$	Empirical distribution probability function of X, 58
f	Density function of X, 12
f_n	Kernel density estimator of f, 14
$\mathbb{R}^d$	$d-$dimensional Euclidean space, 11
S	Support of X, 11
$G(t)$	Density level set with threshold t, 12
$L(\tau)$	Density level set with probability content $1-\tau$, 12
$\mathbb{I}_A$	Indicator function of A, 12
f_τ	Threshold of $L(\tau)$, 12
H_t	Excess mass functional, 14
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$, 15
μ	Lebesgue measure, 14
$A \triangle C$	Symmetric difference between A and C, 16
$d(a, C)$	Distance from the point a to the set C, 17
$d_\mu(A, C)$	Distance in measure between A and C, 16
$d_{\mu_f}(A, C)$	μ_f-distance between A and C, 16
$d_H(A, C)$	Hausdorff distance between A and C, 17
$B_r(x)$	Open ball centered at x and radius r, 15
$B_r[x]$	Closed ball centered at x and radius r, 15
$\oplus$	Minkowski addition, 17
$\ominus$	Minkowski subtraction, 17
A^c	Complement of A, 15
$\text{Int}(A)$	Interior of A, 15
$\overline{A}$	Closure of A, 15
∂A	Boundary of A, 15
$\text{conv}(A)$	Convex hull of the set A, 18

H_n	Convex hull of $\mathcal{X}_n$, 36
$C_r(A)$	$r-$convex hull of the set A, 19
(R^r_λ)	Double rolling condition with λ inside the set and with r outside the set, 175
$\eta(a)$	Normal vector at a, where a is a point in the boundary of the set, 22, 98
$\Delta_n(A)$	Maximal spacing in the set A, 106
$V_n(A)$	Volume of the maximal spacing in the set A, 106
r_0	Value of parameter r to be estimated for the support, 94
$r_0(t)$	Value of parameter r to be estimated for $G(t)$, 128
$\mathcal{X}_{n,+}(t)$	$\{X \in \mathcal{X}_n : f_n(X) \geqslant t\}$, 62
$\mathcal{X}_{n,-}(t)$	$\mathcal{X}_n \backslash \mathcal{X}_{n,+}(t)$, 62
D_n	Sequence of order $(\log n/n)^{p/(d+2p)}$, 131
$\mathcal{X}_n^+(t)$	$\{X \in \mathcal{X}_n : f_n(X) \geqslant t + D_n\}$, 131
$\mathcal{X}_n^-(t)$	$\{X \in \mathcal{X}_n : f_n(X) < t - D_n\}$, 131
G_n^+	Density level set $G(t + 2D_n)$, 146
$\mathcal{X}_n^{G^+}$	Set $\mathcal{X}_n \cap G_n^+$, 147
$\mathcal{G}_A^r$	Family of sets contained in A verifying (R_r^r), 147

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA