Advances in Data Analysis and Classification (2022) 16:761–796
https://doi.org/10.1007/s11634-021-00457-4

REGULAR ARTICLE

Nonparametric estimation of directional highest density regions

Paula Saavedra-Nieves[1] · **Rosa M. Crujeiras**[1]

Received: 30 September 2020 / Revised: 17 July 2021 / Accepted: 17 July 2021 /
Published online: 10 October 2021

Abstract
Highest density regions (HDRs) are defined as level sets containing sample points of relatively high density. Although Euclidean HDR estimation from a random sample, generated from the underlying density, has been widely considered in the statistical literature, this problem has not been contemplated for directional data yet. In this work, directional HDRs are formally defined and plug-in estimators based on kernel smoothing and associated confidence regions are proposed. We also provide a new suitable bootstrap bandwidth selector for plug-in HDRs estimation based on the minimization of an error criteria that involves the Hausdorff distance between the boundaries of the theoretical and estimated HDRs. An extensive simulation study shows the performance of the resulting estimator for the circle and for the sphere. The methodology is applied to analyze two real data sets in animal orientation and seismology.

Keywords Bootstrap · Confidence regions · Directional data · Hausdorff distance · Highest density regions · Kernel density estimation · Level sets

Mathematics Subject Classification 62G05 · 62G07

1 Introduction

Set estimation is focused on the reconstruction of a set (or the approximation of any of its characteristic features such as its boundary or its volume) from a random sample of points. One of the specific topics in this area is concerned with the estimation of sets directly related to density functions such as level sets. Mathematically, for a given level $t > 0$, the goal is to reconstruct the unknown set

$$G_g(t) = \{x \in \mathbb{R}^d : g(x) \geq t\} \tag{1}$$

✉ Paula Saavedra-Nieves
paula.saavedra@usc.es

[1] Universidade de Santiago de Compostela, Santiago de Compostela, Spain

from a random sample of points of a density function g on $\mathbb{R}^d$. This topic has received considerable attention in the statistical literature, specially since the notion of population clusters was established in Hartigan (1975) as the connected components of the set in (1). This cluster definition relies clearly on the user-specified level t, so for addressing this problem, an algorithm for estimating the smallest level with more than a single connected component was proposed in Steinwart (2015). For a general review on clustering, see (Anderberg 1973; Everitt 1993; Cuevas and Fraiman 1997) and (Rinaldo and Wasserman 2001).

The number of clusters is a basic feature for a statistical population. However, the problem of its estimation is not always taken into account in cluster analysis where it is usually chosen by the practitioner as a first step. Since the number of clusters is equal to the number of connected components of a level set, a very natural estimator for this populational parameter is the number of the connected components of the level set reconstruction. This perspective that solves the problem of selecting this unknown population parameter is considered, for instance, in Cuevas et al. (2000), Cuevas et al. (2001) and Biau et al. (2007).

Level set estimation theory has been mainly established for a density supported on an Euclidean space such as in Eq. (1) with very few contributions in other domains. Cuevas et al. (2006) consider the estimation of level sets for general functions (not necessarily a density) providing some consistency theoretical results and showing a level set on the sphere for illustration. More recently, the reconstruction of density level sets on manifolds is studied in Cholaquidis et al. (2020). Through some simulations, the behavior of the proposed method is analyzed on the torus and on the sphere.

Unfortunately, for most applications, the specific value of the level t in (1) is fully unknown by the practitioner. In addition, areas of the distribution support where g is close to zero (*non-effective support*) are usually of limited interest for applications. If the practitioner establishes the probability content instead of the level t, a new kind of density level sets emerges known as highest density regions (HDRs) (see Box and Tiao 1973 and Hyndman 1996). The estimation of HDRs involves further complexities given that the threshold of this particular type of level sets must be determined from the established probability content. Perhaps due to its practical importance, HDRs plug-in reconstruction from the linear kernel density estimator has been widely studied considering also the problem of selecting an appropriate bandwidth specifically devised for the HDR reconstruction (see, for instance, Baíllo and Cuevas 2006 or Samworth and Wand 2010). However, as far as we know, the notion of HDR has not been introduced for directional data yet. Therefore, the main goals of this work are to (1) generalize HDRs definition to the directional setting, (2) establish a plug-in procedure for HDRs reconstruction from the proposal of a new bootstrap bandwidth for a well-known directional kernel density estimator that can be seen as the first specific selector for directional HDRs, (3) check its performance through an extensive simulation study analysing the effect of considering a smoothing parameter not specifically designed for HDR estimation and (4) apply this methodology to analyze data on animal orientation and on seismology.

One may argue that such an absence of a general and effective proposal for directional HDRs estimation may be due to a lack of practical interest, but this is far from the truth, so let us present two application examples that motivate the developments

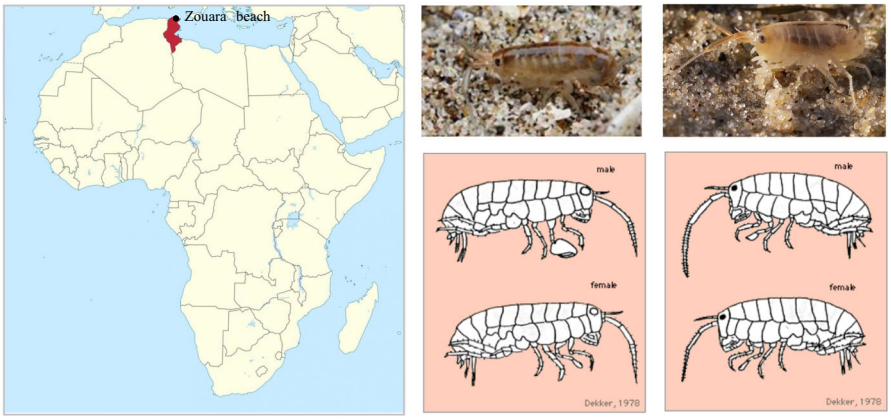


Fig. 1 Geographical location of Zouara beach (right). Talorchestia brito (center) and Talitrus saltator (left)

in this work. The first one concerns a problem from animal orientation studies and the second one is related to earthquakes occurrences. Both datasets are available in the R package `HDiR`.[1]

1.1 Some motivating examples

Animal orientation example. Behavioral plasticity is considered by biologists as a feature of adaptation to changing beach environments. In particular, orientation is an adaptation characteristic that can not be modified by a single factor. Nonetheless, experts found some regularities in the orientation of sandhoppers and other animals from beach environments by changing one factor at a time under other controlled conditions.

For instance, the orientation of two sandhoppers species (Talitrus saltator and Talorchestia brito) is analyzed in Scapini et al. (2002). Both species are shown in Fig. 1. Bottom pictures can be found in Dekker (1978). Comparing the two species through regresion procedures, Scapini et al. (2002) conclude that Talitrus saltator showed more differentiated orientations, depending on the time of day, period of the year and sex, with respect to Talorchestia brito. Moreover, it seems that Talitrus saltator shows a higher flexibility (variation) of orientation than Talorchestia brito under the same environmental conditions, supporting the hypothesis that the former has a higher level of terrestrialization. As an illustration, Fig. 2 (left panel) shows the 36 orientation points (slightly jittered) corresponding to males of the specie Talitrus saltator measurements during the noon in April. It also contains the 77 angles (slightly jittered) when the measures are taken in October (Fig. 2, right panel). Differences in the distribution on the circle of these two samples can be easily observed. Therefore, the month of the year seems to play a significant role in sandhoppers behavior. In particular, two clusters for October measurements can be detected around the angle π but they are not present for the April sample. Similar comments could be done for the

[1] https://CRAN.R-project.org/package=HDiR.

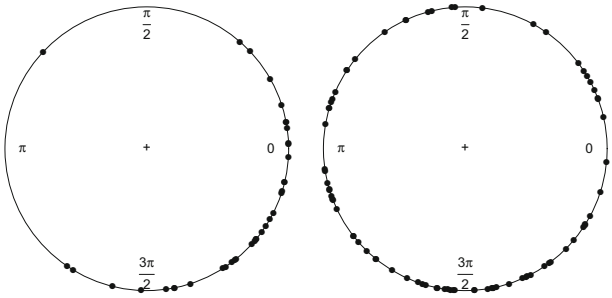


Fig. 2 Orientation data (slightly jittered) corresponding to males of the specie Talitrus saltator registered in the noon in April (left) and October (right)

situation registered around the angles $\pi/2$. HDRs reconstruction (with low probability content) would allow to determine the biggest modes of the distribution, and then, its clusters. Therefore, HDRs can be seen as a useful alternative to analyze sandhoppers orientation.

Earthquakes occurrences. The European-Mediterranean Seismological Centre (EMSC)[2] is a non-governmental and non-profit organisation that has been established in 1975 at the request of the European Seismological Commission. Since the European-Mediterranean region has suffered several destructive earthquakes, there was a need for a scientific organisation to be in charge of the determination, as quickly as possible (within one hour of the earthquake occurrence), of the characteristics of such earthquakes. These predictions are based on the seismological data received from more than 65 national seismological agencies, mostly in the Euro-Med region. Figure 3 (left) shows the geographical coordinates (red points), downloaded from EMSC website, of a total of 272 medium and strong world earthquakes registered between 1^{th} October 2004 and 9^{th} April 2020. The magnitude of all these events is at least 2.5 degrees on the Richter scale. Of course, these planar points correspond to spherical coordinates on Earth. Due to the important damages that earthquakes cause, cluster detection of HDRs could be also useful to identify, from a real dataset, where earthquakes are specially likely. This information is key for decision-making, for example, to update construction codes guaranteeing a better building seismic-resistance. An interactive representation of the sphere can be seen in Appendix D.

1.2 Paper organization

This paper is organized as follows. Section 2 contains some background ideas on directional level set estimation including some discussion on error measurements and some existing consistency results in the directional setting that will be really useful to extent the definition of HDRs for directional data in Sect. 3. There, plug-in estimators and the corresponding confidence regions are also established. Concretely, we consider the plug-in methods based on a well-known directional kernel density estimator, which requires a smoothing parameter (bandwidth) for its practical implementation.

[2] European-Mediterranean Seismological Centre: www.emsc-csem.org.

Fig. 3 Distribution of earthquakes around the world between October 2004 and April 2020 (left). HDR contour obtained from the sample of world earthquakes registered between October 2004 and April 2020 (right)

An appropriate bootstrap bandwidth selector, the first one specifically designed for directional HDRs estimation, is also introduced in this section. Section 4 presents an extensive simulation study illustrating the performance of the resulting plug-in reconstruction for the HDRs (for circular and spherical domains) considering the new bandwidth selector. These results are compared with those obtained with directional smoothing parameters not specifically designed for HDR estimation. In Sect. 5, the proposed methodology is applied to analyze the two real data examples presented in the Introduction. Finally, some conclusions and ideas for further research are presented in Sect. 6. Appendix A and the supplementary material that completes this work include further information on the datasets. Appendix B specifies the parameters taken for the construction of the spherical densities in the simulation study. Appendix C contains some additional results of simulations. Appendix D collects the description of the bandwidth selectors considered in the simulation study. All the methods presented in this paper, along with the real data examples, are accessible in `HDiR` package.

2 Some background on directional level sets

The specific problem of reconstructing density level sets in the directional setting is reviewed in this section: the definition of directional level set is introduced jointly with a plug-in estimator. Based on the real data and simulated examples, some discussion about how to measure the estimation error and some asymptotic results are also included.

2.1 On directional level sets

Consider a random vector X taking values on a $d-$dimensional unit sphere S^{d-1} with density f. Given a level $t > 0$, the directional level set is defined as:

$$G_f(t) = \{x \in S^{d-1} : f(x) \geq t\}. \tag{2}$$

The nature of different level sets is shown in Fig. 4, which represents $G_f(t)$ in grey color for three different circular densities and three different values of the level t. The threshold t is represented through a dotted grey line. Note that, if large values of t are considered (bottom row in Fig. 4), $G_f(t)$ coincides with the greatest modes. However, for small values of t, the level set $G_f(t)$ is virtually equal to the support of the distribution.

It is important to noticed that, following (Hartigan 1975), we may also establish the concept of cluster in directional setting as the connected components of the level set $G_f(t)$. With this view in mind, note that the density represented in the second row of Fig. 4 presents four connected components for all of the considered values for t, determining four population clusters.

Plug-in estimation is the most natural and common choice for reconstructing density level sets in the Euclidean space. A review of other existing estimation alternatives can be seen in Rodríguez-Casal and Saavedra-Nieves (2019). Plug-in methods are devised to reconstruct the level set in (1) as

$$\hat{G}_g(t) = \{x \in \mathbb{R}^d : g_n(x) \geq t\}$$

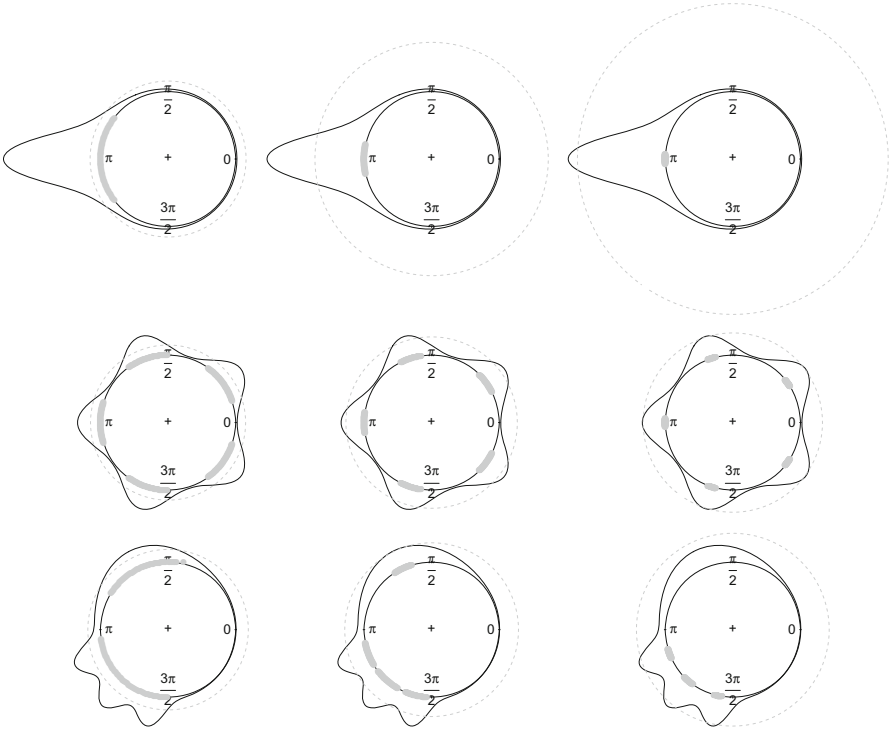


Fig. 4 For thee different circular densities, $G_f(t)$ for $t = t_1$ (first column), $t = t_2$ (second column) and $t = t_3$ (third column) verifying $0 < t_1 < t_2 < t_3$. Equivalently, $L(f_\tau)$ for $\tau = 0.2$ (first column), $\tau = 0.5$ (second column) and $\tau = 0.8$ (third column)

where g_n usually denotes the classical kernel estimator for Euclidean data (see Parzen 1962 and Rosenblatt 1956). This methodology, which has received considerable attention (see, for instance, Tsybakov 1997; Baíllo 2003; Mason and Polonik 2009; Rigollet and Vert 2009; Mammen and Polonik 2013; Polonik 2013 or Chen et al. 2017) can be easily generalized to the directional setting. Given a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\} \in S^{d-1}$ of the unknown directional density f, the corresponding level set $G_f(t)$ in (2) can be reconstructed as

$$\hat{G}_f(t) = \{x \in S^{d-1} : f_n(x) \geq t\} \tag{3}$$

where f_n denotes a nonparametric directional density estimator. Following the classical ideas for real-valued random variables, a kernel estimator on S^{d-1} is provided in Bai et al. (1989) ($d > 2$) who also proved strong pointwise consistency, uniform consistency, and L_1-norm consistency of the estimator (see also Hall et al. 1987 and Klemelä 2000 for further results). Following Bai et al. (1989), from a random sample $\mathcal{X}_n$ on a $d-$dimensional sphere the directional kernel density estimator at a point $x \in S^{d-1}$ is defined as

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{vM}(x; X_i; 1/h^2), \tag{4}$$

where $1/h^2 > 0$ is concentration parameter and K_{vM} denotes the von Mises-Fisher kernel density (see Appendix B for explicit formulae). The consideration of a von Mises kernel in Eq. (4) is not the only option and it is particularly interesting to point out the use of a wrapped-normal kernel in the circular setting. In this case, Huckemann et al. (2016) proved that this kernel guarantees the monotonicity on the number of modes with respect to the smoothing parameter, something that also happens for the gaussian kernel in the linear case. It may be argued that such a kernel shoud be used in our problem. Nevertheless, it is computationally more expensive and our practical experience shows that results in practice are quite similar.

Note that the kernel estimator in (4) can be viewed as a mixture of von Mises-Fisher. Furthermore, the concentration parameter $1/h^2$ plays an analogous role to the bandwidth in the Euclidean case. For small values of $1/h^2$, the density estimator is oversmoothed. The opposite effect is obtained as $1/h^2$ increases: with a large value of $1/h^2$, the estimator is clearly undersmoothing the underlying target density. Hence, the choice of h is a crucial issue. For simplicity, in what follows, we refer to h as bandwidth parameter. Several approaches for selecting h in practice, in circular and even directional settings, have been proposed in the literature (see Appendix D). All the existing proposals aim to minimize some error criterion on the target density, but none of them is specifically designed focusing on the reconstruction of a directional level set.

Figure 5 shows three plug-in estimators $\hat{G}_f(t)$ for models (black colour) and levels t_1, t_2 and t_3 (dotted grey line) considered in Fig. 4. Kernel density estimators (grey color) in (4) have been determined from samples of size 250 considering the proposal in Oliveira et al. (2012) as bandwidth parameter.

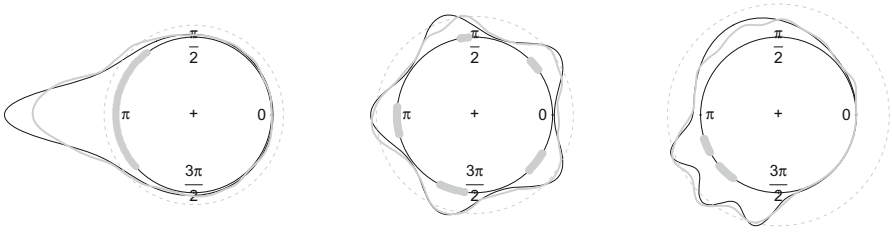


Fig. 5 Plug-in density level sets $\hat{G}_f(t)$ from $\mathcal{X}_{250}$ for three different circular densities with $t = t_1$ (first column), $t = t_2$ (second column) and $t = t_3$ (third column) verifying $0 < t_1 < t_2 < t_3$

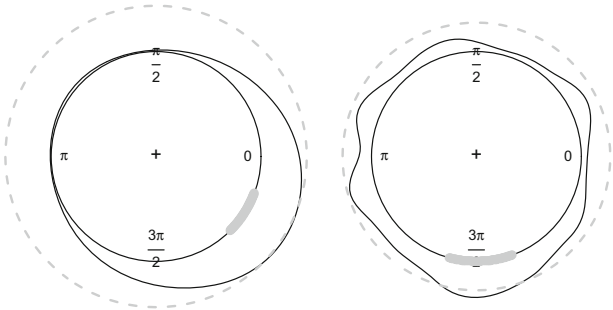


Fig. 6 Plug-in level set estimators obtained from the orientation samples corresponding males of the specie Talitrus saltator registered in the noon in April (left) and October (right)

For instance, for the sandhoppers example, Fig. 6 shows the plug-in estimators obtained for the two samples of sandhoppers represented in Fig. 2. It is possible to detect the largest modes of the two sample distributions corresponding to April and October samples. These results allow us to confirm the differences between the two populations. The largest cluster of April orientations is located around the angle $7\pi/4$. However, the pattern observed for October registries is completely different. Although an only cluster is identified around the angle $3\pi/2$, if the level t decreases slightly two additional groups can be detected around the angles $3\pi/4$ and $5\pi/4$, respectively.

Regarding the earthquakes illustration, Fig. 3 (right) shows the plug-in contour in blue obtained from the selected sample of world earthquakes considered. Chosing a convenient value of the level t, the greatest mode of sample distribution is identified in the Southeast of Europe. Countries such as Italy, Greece or Turkey (located withint this cluster) are the most affected areas in the recent past.

2.2 Error measures and consistency results on directional level sets

The level set $\hat{G}_f(t_3)$ represented in Fig. 5 (third column) presents two connected components. However, Fig. 4 (right plot in the bottom row) shows that the theoretical level set $G_f(t_3)$ has exactly three components. Therefore, the estimation error is considerably large. Distances between sets are the common criteria considered in set estimation to measure the discrepancies between the theoretical region to be estimated

and the corresponding reconstruction. Of course, this is also applicable when the goal is to estimate level sets or HDRs.

The distance in measure d_μ between two Borel sets A and B in $\mathbb{R}^d$ is defined as

$$d_\mu(A, B) = \mu(A \triangle B) \tag{5}$$

where μ denotes the Lebesgue measure and $A \triangle B$, the symmetric difference of A and B calculated as $(A \cap B^c) \cup (A^c \cap B)$ with A^c representing the complementary of A. Consistency results for directional plug-in estimators have been already obtained in the literature for this distance. For the estimator established in (3) defined on S^2, Cuevas et al. (2006) and Cholaquidis et al. (2020) check that $\lim_{n\to\infty} d_\mu(G_f(t), \hat{G}_f(t)) = 0$, a.s., and $\lim_{n\to\infty} d_\mu(G_f(t), \hat{G}'_f(t)) = 0$, a.s., where $\hat{G}'_f(t) = \{x \in S^{d-1} : f'_n(x) \geq t\}$ and f'_n denotes the kernel estimator (for manifolds with boundary) proposed in Berry and Sauer (2017). From the definition in (5), it easy to check that the distance in measure d_μ does not penalize those level set estimators that have an isolated point as a connected component or any other set with null Lebesgue measure. Additionally, the undersmoothing caused by the choice of a small bandwidth value may provoque that the estimator $\hat{G}_f(t)$ presents non-significant connected components with small Lebesgue measure. In this case, d_μ would not be as effective as, for instance, the Hausdorff distance in detecting this situation.

Let us recall that, if A and B are now non-empty compact sets in $\mathbb{R}^d$, the Hausdorff distance between A and B is established as follows

$$d_H(A, B) = \max\left\{\sup_{x\in A} \rho\left(\{x\}, B\right), \sup_{y\in B} \rho\left(\{y\}, A\right)\right\} \tag{6}$$

where $\rho(\{x\}, B) = \inf_{y\in B}\{\rho(x, y)\}$ being $\rho(x, y)$ the distance between two points. Note that the definition of the Hausdorff distance is very general and depending on the selection of the distance ρ, different error criteria emerge. Usually, ρ corresponds to the chordal distance (Euclidean distance in $\mathbb{R}^d$, ρ_1).

Remark 1 Other natural choices such as the geodesic distance (great circle, ρ_2) could be considered in Eq. (6). Hopf-Rinow Theorem states that ρ_1 and ρ_2 induce the same topology on S^{d-1}. Figure 1 in Jeong et al. (2017) illustrates that $\rho_1(x, y) \leq \rho_2(x, y)$ for any pair of points x, y in the unit circle. Following Lemma 3 in Boissonnat et al. (2019), a general upper bound for the $\rho_2(x, y)$ for all $x, y \in S^{d-1}$ depending on $\rho_1(x, y)$. Specifically, it is possible to prove that $\rho_2(x, y) \leq \arcsin(\rho_1(x, y))$ for all $x, y \in S^{d-1}$ when the constant r is equal to $1/2$.

The metric d_H is not completely successful in detecting differences in shape properties. In other words, two sets can be very close in Hausdorff distance and still show quite different shapes. This typically happens where the boundaries ∂A and ∂B are far apart, no matter the proximity of A and B. So a natural way to reinforce the notion of visual proximity between two sets provided by Hausdorff distance is to account also for the proximity of the respective boundaries. This error criterion has been also

considered for establishing consistency results of several directional plug-in reconstructions. Cuevas et al. (2006) prove that $\lim_{n\to\infty} d_H(\partial G_f(t), \partial \hat{G}_f(t)) = 0$, a.s., when the Hausdorff distance is defined from ρ_1. If the Hausdorff distance involves ρ_2, (Cholaquidis et al. 2020) prove that $\lim_{n\to\infty} d_H(\partial G_f(t), \partial \hat{G}'_f(t)) = 0$ and $\lim_{n\to\infty} d_H(G_f(t), \hat{G}'_f(t)) = 0$, a.s. The existing monotone relationship between chordal and geodesic distances guarantees the consistency of the plug-in estimator in Cholaquidis et al. (2020) also when the Hausdorff distance depends on ρ_1 instead of ρ_2. Therefore, if the target is the reconstruction of a set, the Hausdorff metric (defined from the chordal distance) can be seen as a suitable error criteria in the directional setting.

3 HDRs in the directional setting

As noted in the Introduction, the level t is usually unknown in (1) and, for practical purposes, the practitioner chooses the probability content of the set instead of the level t. These particular class of level sets widely considered for Euclidean data are the so-called HDRs (see Box and Tiao 1973; Hyndman 1996 or Samworth and Wand 2010). However, as far as we know, HDRs were not defined in the directional context yet. Motivating the need for a proper extension of this notion and the proposal of adequate estimation tools can be easily justified. Figure 7 (top) shows four different 50% circular regions (regions containing 50% of the probability, empirically approximated) for the kernel density estimator f_n represented in grey. Although all of them have probability content equal to 50%, they exhibit completely different shapes. Therefore, it is obvious that there exists an infinite number of ways to choose a region with given coverage probability and in a general scenario, it may not be clear which region must be chosen. The same happens for real-valued random variables, and (Hyndman 1996) suggests that HDRs are the best subset to summarize a probability distribution.

The usual purpose in summarizing a probability distribution by a region of the sample space is to delineate a comparatively small set which contains most of the probability, although the density may be nonzero over infinite regions of the sample space. Therefore, as in the Euclidean case, it is necessary to decide what properties the region has to verify. The following conditions are natural:

(C1) The region should occupy the smallest possible volume in the sample space.
(C2) Every point inside the region should have probability density at least as large as every point outside the region.

Following (Box and Tiao 1973), conditions (C1) and (C2) are equivalent and lead to regions called HDRs. Definition 1 formalizes this concept in the directional context taking into account the second criterion.

Definition 1 Let f be a directional density function on S^{d-1} of a random vector X. Given $\tau \in (0, 1)$, the $100(1-\tau)\%$ HDR is the subset

$$L(f_\tau) = \{x \in S^{d-1} : f(x) \geq f_\tau\} \tag{7}$$

where f_τ can be seen as the largest constant such that

$$\mathbb{P}(X \in L(f_\tau)) \geq 1 - \tau \tag{8}$$

with respect to the distribution induced by f.

According to Polonik (1997) and García et al. (2003) in the Euclidean context, $L(f_\tau)$ is the minimum volume level set with probability content at least $(1 - \tau)$. Figure 4 shows the HDR $L(f_\tau)$ in grey for three different circular densities and three different values of τ. The threshold f_τ is represented through a dotted grey line. Note that, if large values of τ are considered, $L(f_\tau)$ is equal to the greatest modes and, therefore, the most differentiated clusters can be easily identified. However, for small values of τ, $L(f_\tau)$ is almost equal to the support of the distribution.

3.1 Plug-in estimation of directional HDRs

The first step to reconstruct the HDR in Definition 1 for a given $\tau \in (0, 1)$ is to estimate the threshold f_τ. As in the Euclidean case, numerical integration methods could be also used in the directional setting in order to approximate its value. However, when the dimension increases, the computational cost becomes a major issue due to the complexity of the numerical integration algorithms considered on high dimensional spaces. An alternative approach for estimating f_τ with a feasible computational cost is described next.

As before, let X be a random vector with directional density f and let $Y = f(X)$ be the random vector obtained by transforming X by its own density function. Since $\mathbb{P}(f(X) \geq f_\tau) = 1-\tau$, f_τ is exactly the $\tau-$quantile of Y, following (Hyndman 1996), f_τ can be estimated as a sample quantile from a set of independent and identically distributed random vectors with the same distribution as Y.

In particular, if $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ denotes a set of independent observations in S^{d-1} from a density f, $\{f(X_1), \ldots, f(X_n)\}$ is a set of independent observations from the distribution of Y. Let $f_{(j)}$ be the $j-$th largest value of $\{f(X_i)\}_{i=1}^n$ so that $f_{(j)}$ is the (j/n) sample quantile of Y. We shall use $f_{(j)}$ as an estimate of f_τ. Specifically, we choose $\hat{f}_\tau = f_{(j)}$ where $j = \lfloor \tau n \rfloor$. Cadre et al. (2009) study the convergence of $\hat{f}_\tau$ to f_τ in the linear setting.

Obviously, if f is a known function, the observations can be pseudorandomly generated and the estimation of f_τ could be made arbitrarily accurate by increasing n. In practice, f is often unknown and we have as only information a random sample of points $\mathcal{X}_n$ from an unknown density f. From this sample, we propose first to determine the kernel estimator f_n in (4). If n is large enough, then calculate the set $\{f_n(X_1), \ldots, f_n(X_n)\}$ in order to estimate f empirically. If n is moderate, it may be preferable to generate observations $\mathcal{X}_n = \{X_1, \ldots, X_N\}$ of large size N from f_n. For small values of n it may not be possible to get a reasonable density estimate. Besides, with few observations and no prior knowledge of the underlying density, there seems little point in attempting to summarize the sample space (see Wand and Jones 1995 for some discussion on the number of observations needed for a reasonable linear density estimate). Note that the problem here is not with the density quantile algorithm (that

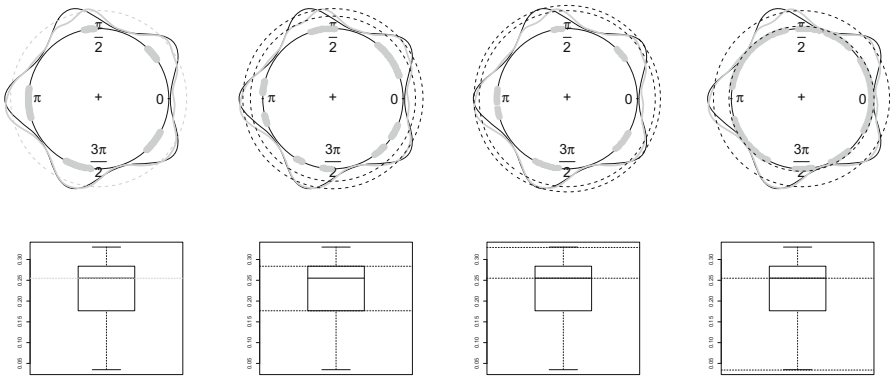


Fig. 7 50% circular regions obtained from the circular kernel estimator f_n (Grey color) obtained from a sample $\mathcal{X}_{250}$. Boxplots of $\{f_n(X_1), \ldots, f_n(X_{250})\}$ and quantiles (dotted lines) that determine the 50% regions (bottom)

give results to an arbitrary degree of accuracy given a density), but with estimating the density from insufficient data.

Once the threshold f_τ is estimated, plug-in methods reconstruct the $100(1-\tau)\%$ HDR namely $L(f_\tau)$ in (7) as

$$\hat{L}(\hat{f}_\tau) = \{x \in S^{d-1} : f_n(x) \geq \hat{f}_\tau\}. \tag{9}$$

Figure 7 shows the circular kernel estimator f_n (grey color) calculated from a sample $\mathcal{X}_{250}$ generated from the second model (black color) in Fig. 4 and different empirically approximated 50% circular regions (grey color, top). The boxplot of the transformed values denoted by $\{f_n(X_1), \ldots, f_n(X_{250})\}$ is also shown (bottom). The dotted lines represent the quantiles that determine the corresponding 50% (probability coverage) circular region. Note that only the estimated HDR (left), $\hat{L}(\hat{f}_\tau)$, is able to show the existence of the five existing modes.

Apart from the consistency of $\hat{f}_\tau$, Cadre et al. (2009) establish the exact convergerce rate (considering the distance in measure d_μ as error criteria) for Euclidean HDRs. The extension of these results to the directional setting does not seem straightforward. However, if $\hat{f}_\tau-$consistency remains true, we could prove that $\hat{L}(\hat{f}_\tau)$ is also a d_H-consistent estimator of $L(f_\tau)$ in S^2 under the assumptions of Corollary 1 and condition (T) in Cuevas et al. (2006). To complete the proof is only necessary to apply a triangle inequality on $d_H(L(f_\tau), \hat{L}(\hat{f}_\tau))$.

3.1.1 Confidence regions for estimated HDRs

The density quantile algorithm detailed above for approximating the threshold f_τ involves an empirical approximation. Then, it is convenient to compute some uncertainty limits on the estimated regions.

For the simplest case of X being a circular random variable (following Hyndman 1996), standard asymptotic results for a sample in Cox and Hinkley (1979) allow

to prove that $\hat{f}_\tau$ is asymptotically normally distributed with mean f_τ and variance $\tau(1-\tau)/(n[F(f_\tau)]^2)$ where

$$F(y) = y \sum_{i=1}^{n(y)} |f^{'}(z_i)|^{-1}$$

and $\{z_i\}$ denote those points in the sample space of X such that $f(z_i) = y$, $i = 1, 2, \ldots, n(y)$.

Alternatively, a bootstrap algorithm can be easily designed to compute confidence regions for estimated HDRs. The procedure is detailed in Algorithm 1.

Algorithm 1: Bootstrap procedure to estimate HDRs confidence regions.

Inputs: $\mathcal{X}_n$, τ, the number of bootstrap resamples B and the confidence level α.

1. Calculate the directional kernel density estimador f_n from $\mathcal{X}_n$.

2. Initialize the bootstrap procedure:

foreach $b \in \{1, ..., B\}$ **do**

- Generate a bootstrap sample of size n from f_n, $\mathcal{X}^*_{b,n}$.
- Calculate the directional kernel density estimador f^*_n from $\mathcal{X}^*_{b,n}$.
- Determine $f^*_n(\mathcal{X}^*_{b,n})$.
- Calculate the threshold $\hat{f}^*_{\tau,b}$ as the τ-quantile of $f^*_n(\mathcal{X}^*_{b,n})$.

end

3. Determine the thresholds $\hat{f}_{\tau_1}$ and $\hat{f}_{\tau_2}$ as the $(\alpha/2)-$ and $(1-\alpha/2)-$quantiles of the vector $\hat{f}^*_\tau$.

4. Compute $\hat{L}(\hat{f}_{\tau_1})$ and $\hat{L}(\hat{f}_{\tau_2})$ from f_n.

Output: Confidence bands, $\hat{L}(\hat{f}_{\tau_1})$ and $\hat{L}(\hat{f}_{\tau_2})$.

As an illustration, Fig. 8 shows the estimated confidence regions using the asymptotic approach (first row, in dark red color) and the bootstrap procedure (second row, in purple color) for three different values of τ when $\alpha = 0.05$ and $B = 250$. Cross validation bandwidths introduced in Hall et al. (1987) were used as smoothing parameters for circular density estimation in both approaches.

3.2 A suitable bootstrap bandwidth selector

The plug-in reconstruction of the directional HDRs in (9) involves the calculation of the kernel density estimator in (4) that is known to be heavily dependent on the selection of h. The existing methods for selecting an optimal value for h aim for minimizing some error criterion on the target density f, but they are not specifically

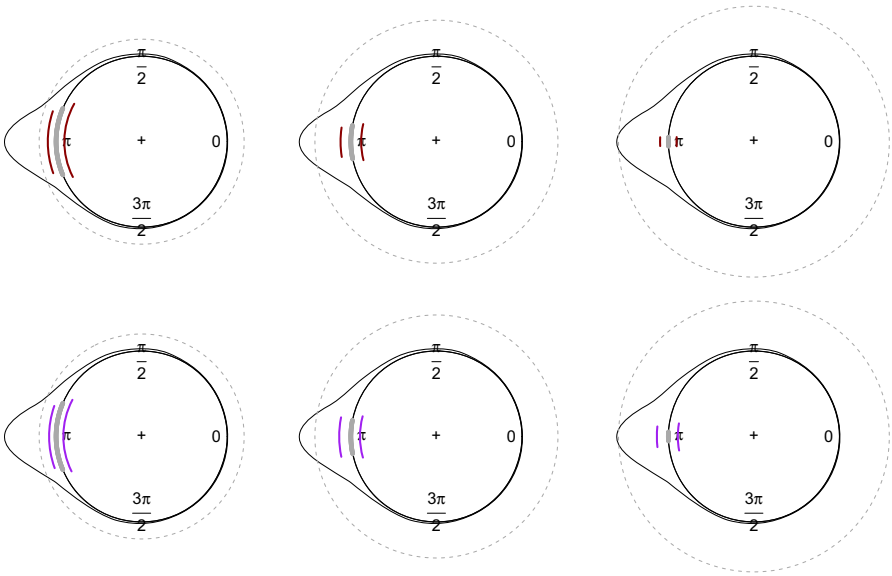


Fig. 8 95 % Confidence regions considering the asymptotic approach (first row, in dark red color) and the bootstrap procedure with $B = 250$ (second row, in purple color) from $\mathcal{X}_{500}$ of a circular density with τ_1 (first column), τ_2 (second column) and τ_3 (third column) verifying $0 < \tau_1 < \tau_2 < \tau_3 < 1$

designed for the estimation of HDRs. The goal of this section is to propose the first selector of h specifically designed for HDRs reconstruction.

A bootstrap bandwidth selector focused on the problem of reconstructing HDRs is introduced in what follows. The idea is to use an error criterion that quantifies the differences between the theoretical region and its plug-in reconstruction. In the real-valued setting, Samworth and Wand (2010) propose one of the first bandwidth selectors for HDRs estimation studying an relatively uncommon distance (depending on both μ and g) between these sets. In this work, we consider the classical Hausdorff distance (introduced in Sect. 2.2) between the boundaries of the HDR and the corresponding estimator.

In the directional case, the closed expression of $d_H(\partial L(f_\tau), \partial \hat{L}(\hat{f}_\tau))$ is not known. However, it could be estimated through a bootstrap procedure. Therefore, a new bandwidth selector can be established as

$$h_1 = \arg\min_{h>0} \mathbb{E}_B \left[d_H(\partial L^*(\hat{f}_\tau^*), \partial \hat{L}(\hat{f}_\tau)) \right] \tag{10}$$

where $\mathbb{E}_B$ denotes the bootstrap expectation with respect to random samples $\mathcal{X}_n^* = \{X_1^*, \ldots, X_n^*\}$ generated from the directional kernel f_n that, of course, is dependent on a pilot bandwidth and also on the choice of the distance ρ in Eq. (6).

Figure 12 shows the theoretical HDR for model S3 (see Sect. 4.2) when $\tau = 0.5$ (first and second columns). Moreover, the plug-in estimator $\hat{L}(\hat{f}_\tau)$ obtained from a sample of size $n = 1000$ and considering the bandwidth proposed in García-Portugués (2013) when $\tau = 0.5$ is also represented (third column). Note that, for this sample

size, only the largest mode is detected. In this particular case, the Hausdorff error is smaller if the HDR is reconstructed from a cross-validation bandwidth designed for density estimation (fourth and fifth columns). A relevant issue appears when h_1 is estimated from imprecise HDR estimators. Remember that the minimization procedure considered for determining h_1 involves the boundary of the set $\hat{L}(\hat{f}_\tau)$. If this set is poorly approximated the resulting bandwidth surely will not provide competitive results. Therefore, largest sample sizes will be considered in this section for avoiding this problem.

Another point that is worth to mention is that diverse bandwidths selectors emerge from the consideration of the different choices of ρ in the definition of the Hausdorff distance. In fact, other bandwidths could be defined if, for example, d_H in (10) is replaced by a completely different error criteria such as the distance in measure d_μ that, unlike Hausdorff distance, does not take in account the connected components of a set only composed by a isolated point. Therefore, we could propose as many bandwidths as existing distances between sets attending to the specific properties and characteristics of each distance.

4 Simulation study

The performance of the proposed bandwidth selector is explored in this section. As it has been mentioned, there exist other bandwidth selectors for directional kernel density estimation (see Appendix D), although not specifically designed for HDR reconstruction. We will also check the impact of considering some of these selectors in the HDR plug-in estimation. Specifically, the selector h_1 established in (10) was implemented considering the chordal distance ρ_1 that, as we show in Sect. 2.2, guarantees good asymptotic properties of directional level sets. The code for computing it is available in the R library `HDiR`. All the other bandwidths are implemented in the R packages `NPCirc`[3] and `Directional`[4]. Sects. 4.1 and 4.2 contain the results obtained in circular and spherical settings, respectively. Some additional results of simulations are also contained in Appendix C.

4.1 Estimation of circular HDRs

A collection of 9 circular densities (models C1 to C9) have been considered in this simulation study. These models are mixture of different circular distributions and they correspond to densities 5, 6, 7, 8, 10, 11, 16, 19 and 20 fully described in Oliveira et al. (2014). Figure 9 shows these densities and the thresholds f_τ for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$ through dotted circles.

A total of 250 random samples of sizes $n = 500$ and $n = 1000$ were generated for each of these models. From each sample, circular HDRs are reconstructed for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$. Results for $\tau = 0.2$ and $\tau = 0.8$ are exposed in Appendix C.1. The behavior of plug-in methods that emerge from the consideration of different

[3] https://CRAN.R-project.org/package=NPCirc.

[4] https://CRAN.R-project.org/package=Directional.

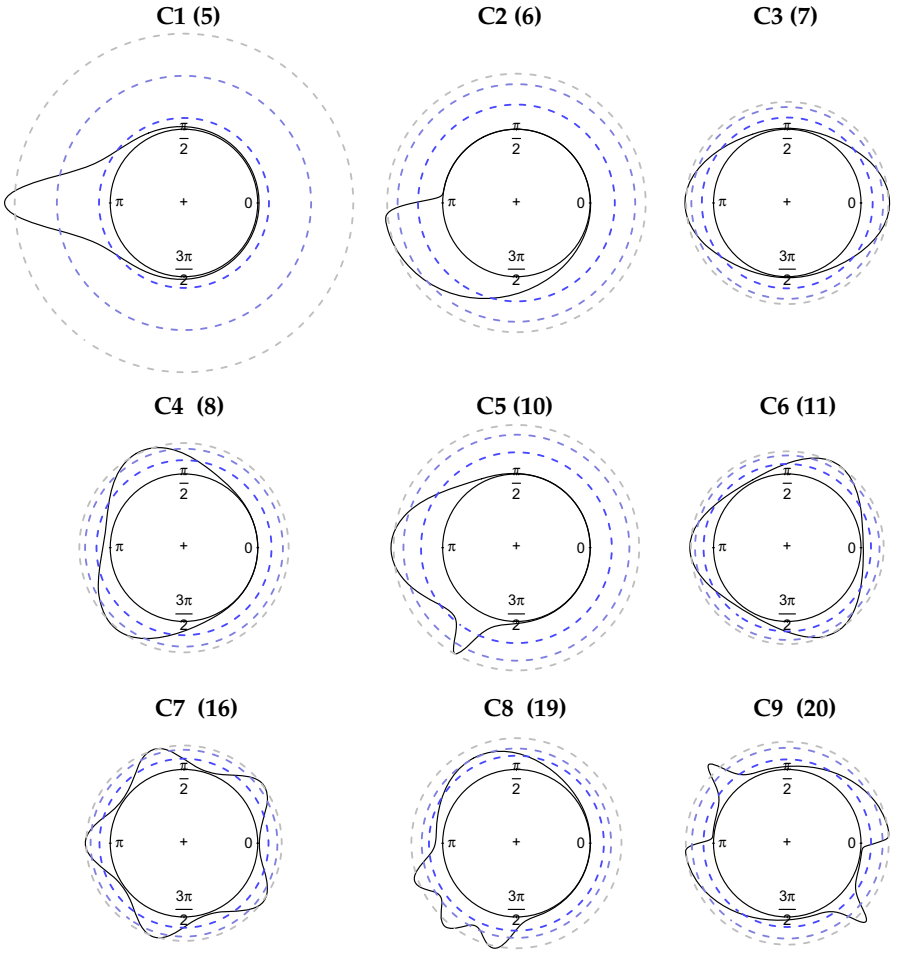


Fig. 9 Circular density models for simulations. Dotted circles represent the threshold f_τ when $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$, respectively

bandwidth parameters will be checked. Apart from h_1, we will consider the circular rule-of-thumb by Taylor (2008) (h_2); its improved version by Oliveira et al. (2012) (namely h_3); cross-validation methods (likelihood h_4 and least squares h_5) introduced by Hall et al. (1987) and the bootstrap bandwidth (h_6) presented by Di Marzio et al. (2011). Note that for computing h_1, a pilot bandwidth is required. In this study, h_3 has been taken as a pilot, and $B = 200$ resamples are considered for obtaining h_1.

For each method and each sample, the estimation error is measured by computing the Hausdorff distance (d_H) between the boundaries of estimated HDR and the frontier of theoretical set. As a reference, note that the maximum value of this criteria in S^1 is 2 (the length of the diameter of the circle).

Tables 1 and 2 show the means and the standard deviations of the 250 estimation errors obtained when $\tau = 0.5$ from samples of sizes $n = 500$ and $n = 1000$, respec-

Table 1 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.5, n = 500$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.027	0.015	0.120	0.092	**0.113**	0.052	0.107	0.055	0.425	0.290
h_2	**0.026**	0.014	0.103	0.042	1.303	0.372	0.104	0.047	0.655	0.103
h_3	**0.026**	0.015	0.100	0.088	0.120	0.054	0.112	0.056	**0.417**	0.307
h_4	**0.026**	0.015	**0.090**	0.054	0.123	0.060	0.112	0.056	0.588	0.219
h_5	**0.026**	0.015	0.091	0.054	0.122	0.059	0.108	0.054	0.633	0.154
h_6	**0.026**	0.015	0.104	0.051	**0.113**	0.049	**0.103**	0.048	0.659	0.054

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.141	0.063	0.206	0.216	**0.504**	0.409	0.243	0.211
h_2	1.427	0.085	1.313	0.060	0.777	0.385	1.327	0.200
h_3	0.141	0.058	**0.192**	0.221	0.554	0.407	**0.213**	0.215
h_4	0.142	0.062	0.221	0.270	0.665	0.407	0.403	0.363
h_5	0.143	0.063	0.221	0.270	0.667	0.404	0.403	0.363
h_6	**0.136**	0.056	0.207	0.252	0.658	0.380	1.273	0.286

Table 2 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.5, n = 1000$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.019	0.011	0.071	0.050	**0.080**	0.033	0.073	0.031	**0.410**	0.307
h_2	**0.018**	0.010	0.086	0.030	1.262	0.340	**0.070**	0.028	0.661	0.080
h_3	0.019	0.010	**0.062**	0.041	0.083	0.033	0.077	0.034	0.414	0.307
h_4	**0.018**	0.010	0.078	0.031	0.084	0.038	0.076	0.033	0.628	0.159
h_5	**0.018**	0.010	0.079	0.031	0.085	0.038	0.075	0.034	0.631	0.153
h_6	0.019	0.010	0.073	0.035	0.081	0.032	0.074	0.029	0.646	0.115

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	**0.095**	0.042	0.105	0.065	**0.488**	0.421	0.133	0.105
h_2	1.425	0.082	1.306	0.047	0.673	0.357	1.301	0.191
h_3	0.101	0.043	0.095	0.063	0.569	0.390	**0.117**	0.098
h_4	0.099	0.040	0.104	0.116	0.650	0.377	0.297	0.319
h_5	0.099	0.041	0.104	0.116	0.650	0.377	0.297	0.319
h_6	0.097	0.038	**0.092**	0.036	0.624	0.357	0.168	0.210

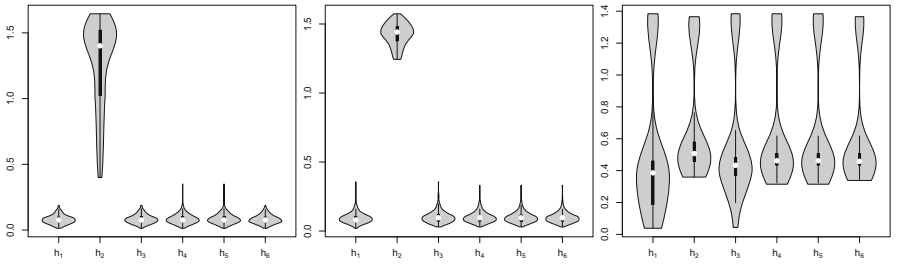


Fig. 10 Violin plots of Hausdorff errors for models C3, C6 and C8 when $\tau = 0.5$ and $n = 1000$. Note that due to the behaviour of h_2, the scales of these figures are different

tively. Bold numbers correspond to the lowest mean errors obtained for each density. Taking into account the variety of models considered, exhibiting different features, it is not surprising that all of the bandwidth selectors are the best ones for some model, showing h_1 a competitive behavior in all cases. In fact, it is the best one for models C3, C5, C6 and C8 (with $n = 1000$).

Figure 10 shows the violin plots of Hausdorff errors obtained for some of the simulation models when $\tau = 0.5$ ($n = 1000$). It shows that h_2 is the selector that presents a worst behavior for models C3 and C6. Furthermore, its variance is again specially large for model C3.

Finally, it is worth to mention that results in Appendix C.1 shows that the competitive behavior of h_1 improves considerably when high values of τ are selected. This is not a minor question when the goal is to estimate the biggest modes of a distribution.

4.2 Estimation of spherical HDRs

For the spherical scenario, 9 density models have been considered. These models, namely S1 to S9, are mixtures of von Mises-Fisher densities on the sphere, allowing to represent complex structures showing multimodality and/or asymetry. Parameters of mixtures are fully established in Table 8 in Appendix B for reproducibility. Moreover, these densities are also implemented in the R package `HDiR`. Figure 11 shows them and the corresponding HDRs for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$.

For sample sizes $n = 500$, $n = 1500$ and $n = 2500$, 200 random samples were generated from models S1 to S9. From each sample, HDRs are reconstructed for $\tau = 0.2, \tau = 0.5$ and $\tau = 0.8$. As before, results for $\tau = 0.2$ and $\tau = 0.8$ are contained in Appendix C.2. The performance of different plug-in methods that emerge from the consideration of different bandwidth parameters discussed in this work is checked. Apart from h_1, cross-validation bandwidth selectors for data on a sphere S^{d-1} (h_5) and the plug-in bandwidth selector introduced by García-Portugués (2013) (h_7) are taken into account. In this case, a total of $B = 50$ resamples are established for estimating the proposed bootstrap bandwidth h_1, taking h_5 as a pilot bandwidth.

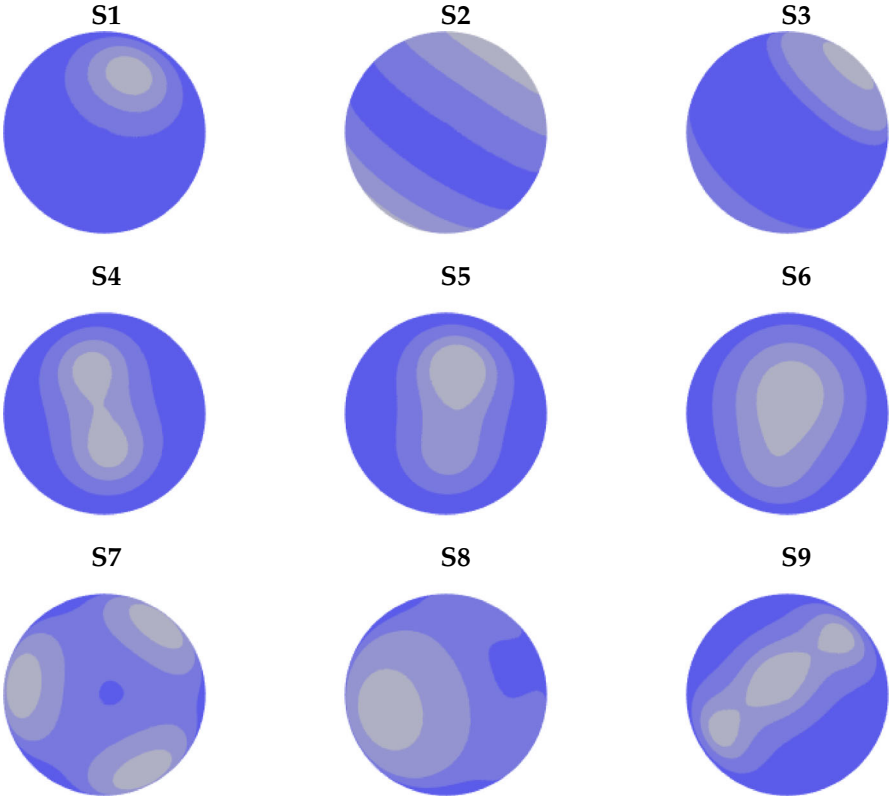


Fig. 11 Finite mixtures of von Mises-Fisher spherical models for simulations. HDRs are represented for $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$

For each method and each sample, the estimation error is again measured calculating the Hausdorff distance between the boundaries of estimated HDR and the frontier of theoretical set. As reference, note that the maximum value of both criteria S^2 is also 2. In this case, this upper bound coincides exactly with the length of the diameter of the sphere.

Tables 3, 4 and 5 contains the results for $\tau = 0.5$ when $n = 500$, $n = 1500$ and $n = 2500$, respectively. Bold numbers correspond to the lowest mean errors obtained for each density. The proposed selector h_1 is the best or second best in all cases. In fact, h_1 and h_5 usually behave similarly and h_7 is the worst selector for S3.

Figure 13 contains the violin plots of Hausdorff errors for some of the considered models when $\tau = 0.5$ and $n = 1500$. Remark that h_1 and h_5 usually present similar results, see densities S5 and S9. However, h_1 is clearly more competitive for models S1 and S8.

To conclude, simulations in Appendix C.2 allows to confirm the good performance of the selector h_1 when small or big values of τ are considered for spherical data.

Table 3 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.5, n = 500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	
h_1	**0.044**	0.018	0.843	0.249	0.924	0.567	**0.113**	0.033	0.131	0.048
h_5	0.069	0.020	**0.796**	0.245	**0.888**	0.552	0.118	0.035	0.130	0.047
h_7	0.082	0.022	0.880	0.181	1.497	0.514	0.115	0.031	**0.127**	0.045

	S6		S7		S8		S9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.136	0.038	0.307	0.074	**0.140**	0.070	0.172	0.061
h_5	**0.135**	0.043	**0.292**	0.066	0.220	0.212	0.174	0.056
h_7	0.145	0.045	0.313	0.074	0.147	0.091	**0.149**	0.049

Table 4 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.5, n = 1500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.032**	0.008	0.568	0.155	0.779	0.590	**0.077**	0.024	**0.092**	0.030
h_5	0.048	0.013	**0.536**	0.129	**0.591**	0.392	0.080	0.021	**0.092**	0.030
h_7	0.057	0.014	0.648	0.144	1.473	0.505	0.079	0.020	**0.092**	0.029

	S6		S7		S8		S9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.093	0.027	0.223	0.052	0.095	0.046	0.103	0.032
h_5	**0.086**	0.023	**0.218**	0.052	0.125	0.117	0.111	0.032
h_7	0.093	0.023	0.223	0.055	**0.093**	0.027	**0.098**	0.024

Table 5 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.5, n = 2500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.026**	0.005	**0.437**	0.131	0.760	0.595	**0.064**	0.019	**0.074**	0.022
h_5	0.042	0.009	0.458	0.123	**0.458**	0.243	0.066	0.017	0.076	0.024
h_7	0.050	0.012	0.523	0.124	1.495	0.508	0.066	0.017	0.076	0.024

	S6		S7		S8		S9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.088	0.027	0.181	0.052	**0.076**	0.048	0.085	0.028
h_5	**0.076**	0.023	**0.178**	0.048	0.113	0.136	0.088	0.025
h_7	0.082	0.023	0.180	0.049	0.081	0.053	**0.081**	0.020

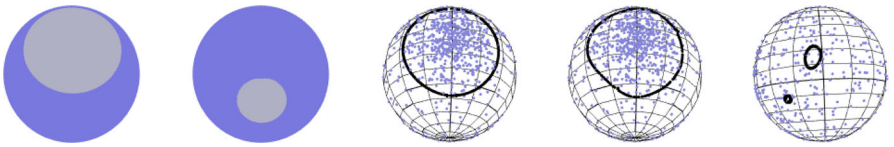


Fig. 12 Theoretical HDR for model S3 when $\tau = 0.5$ (first and second columns). Sample of size $n = 1000$ of model 3 (blue color) and corresponding plug-in estimators (black color) when $\tau = 0.5$ considering h_7 (third column) and h_5 (fourth and fifth columns) as smoothing parameters. Note that the last two columns show two views of the sphere

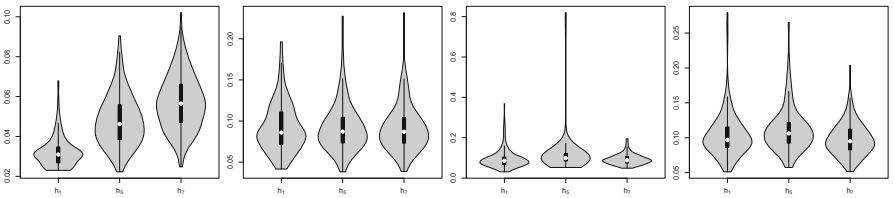


Fig. 13 Violin plots of Hausdorff errors for models S1, S5, S8 and S9 when $\tau = 0.5$ and $n = 1500$. Note that the scales of these figures are different

5 Real data analysis

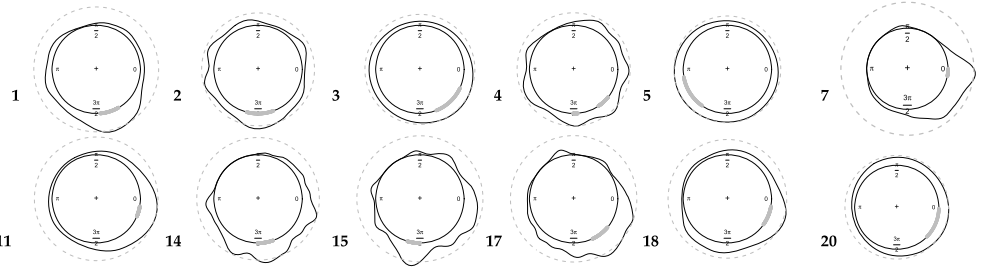
The proposed methodology is now applied to the two real datasets presented in the Introduction, exemplifying the aplicability of the method for circular and spherical data.

5.1 Behavioral plasticity of sandhoppers

Adaptation to changing beach environments for the real example on sandhoppers introduced in Sect. 1.1 is analyzed from HDRs estimation perspective. HDRs are estimated for $\tau = 0.8$ disaggregating the sandhoppers data taking into account the categories of variables specie, sex, time of day and month of year. As consequence, a total of 24 set estimators are determined, numbered E1 to E24. Variables combinations yielding this group classification are presented in Table 7 in Appendix A.

Note that the estimated HDRs correspond to the largest modes of the orientation distributions. Hausdorff distances between the boundaries of these 24 sets are able to establish the degree of dissimilarity of HDRs. In general, large distances between the boundaries of two sets indicate the existence of modes in different directions. If the categories of all variables with the exception of one are fixed, it is possible to check if the different values of the non-fixing variable has some influence in sandhoppers orientation through the comparison of the estimated HDRs. The upper triangular matrix in Table 6 contains the Hausdorff distances (defined from ρ_1) between boundaries. The largest distances are represented in blue color. Grey color is used in order to depict the next largest values. Furthermore, Table 6 (top) contains some of the estimated HDRs that present the largest distances.

Table 6 Upper triangular matrix contains the Hausdorff distances between boundaries of sets from 1 to 24. HDRs representations for $\tau = 0.8$ for some sets between 1 and 24 (top)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1		0.35	0.60	0.36	1.35	1.53	1.21	0.68	0.51	0.42	1.02	0.60	0.30	0.19	0.52	0.52	0.35	0.84	0.51	0.97	0.30	0.67	1.34	0.75
2			0.76	0.53	1.07	1.63	1.47	1.00	0.84	0.76	1.31	0.92	0.60	0.25	0.35	0.85	0.69	1.14	0.84	1.12	0.64	0.83	1.47	0.91
3				0.38	1.55	1.07	0.97	0.40	0.33	0.29	0.76	0.31	0.30	0.77	1.07	0.24	0.26	0.56	0.23	0.40	0.38	0.14	0.84	0.36
4					1.37	1.27	1.28	0.76	0.60	0.51	1.09	0.68	0.34	0.55	0.86	0.61	0.44	0.91	0.60	0.76	0.39	0.31	1.14	0.40
5						1.93	1.97	1.77	1.68	1.63	1.91	1.73	1.53	1.28	1.05	1.69	1.59	1.84	1.68	1.77	1.56	1.58	1.93	1.63
6							0.85	1.00	1.33	1.31	0.84	1.15	1.32	1.64	1.81	1.15	1.29	0.84	1.18	0.71	1.37	1.01	0.26	0.93
7								0.60	0.76	0.84	0.23	0.68	1.00	1.29	1.49	0.75	0.91	0.44	0.76	0.60	0.95	1.02	0.61	1.06
8									0.41	0.38	0.38	0.18	0.44	0.85	1.15	0.18	0.35	0.18	0.21	0.31	0.46	0.46	0.76	0.51
9										0.09	0.58	0.24	0.26	0.61	0.86	0.24	0.16	0.58	0.20	0.71	0.21	0.40	1.13	0.49
10											0.63	0.20	0.18	0.52	0.82	0.20	0.08	0.55	0.16	0.68	0.13	0.36	1.09	0.45
11												0.46	0.80	1.11	1.32	0.53	0.70	0.21	0.55	0.38	0.75	0.82	0.60	0.86
12													0.35	0.69	1.00	0.08	0.25	0.35	0.09	0.49	0.30	0.38	0.92	0.42
13														0.49	0.81	0.28	0.10	0.60	0.26	0.69	0.08	0.38	1.11	0.46
14															0.34	0.69	0.52	1.01	0.66	1.13	0.41	0.84	1.48	0.92
15																1.00	0.84	1.29	0.97	1.39	0.74	1.14	1.68	1.21
16																	0.18	0.35	0.04	0.49	0.29	0.30	0.92	0.35
17																		0.52	0.16	0.66	0.11	0.34	1.07	0.42
18																			0.39	0.16	0.63	0.62	0.60	0.67
19																				0.52	0.25	0.29	0.95	0.34
20																					0.76	0.46	0.46	0.51
21																						0.45	1.17	0.53
22																							0.86	0.25
23																								0.91
24																								

In particular, Hausdorff distance between regions 5 and 11 is equal to 1.91 (close to 2, the maximum value of Hausdorff distance). According to Table 7, the variable configuration 5 corresponds to the largest orientation modes for females of the specie Talitrus saltator when the orientation is measure in noon during October. Region 11 refers to same measurements taken in April. Therefore, the month can be seen as variable that has influence on the orientation for sandhoppers.

Hausdorff distance between regions 5 and 6 is equal to 1.93. According to Table 7, set 6 also corresponds to the HDR for females of the specie Talitrus saltator but, in this case, when the orientation is registered in morning during October. Then, the moment of the day also seems a factor with influence on the sandhoppers behavior.

Several cells in Table 6 are represented in pink color. All of them corresponds to considerable large values of distances (larger than 1.00) and they are used to analyze briefly the influence of each of the variables in the dataset. Under the same values of the rest of variables Talitrus saltator and Talorchestia brito present different behaviors. For instance, distances between sets 5 and 17 or 3 and 15 correspond to this situation. Sets 5 and 17 can be compared using their representations in Table 6 (top).

The importance of the sex variable for the specie Talitrus saltator can be also seen considering the Hausdorff distances of the sets 2 and 5, 3 and 6 or 18 and 15. According to images in Table 6, these sets present their largest modes in completely different directions. Note that the role of the variable month is clearly remarkable. The relatively high values of the distances between sets 1 and 7 and 6 and 12 or 14 and 20 for the species Talitrus saltator and Talorchestia brito also corresponds to the existence of modes in different directions. Finally, the importance of the moment of the day for the Talitrus saltator can be studied through the distances between sets 4 and 5 or 4 and 6. Remark that set 4 has two connected components while set 5 only presents one.

Finally, the analysis of Hausdorff distances for the two species of sandhoppers shows that the median of the Talitrus saltator in Hausdorff distance is 0.76, clearly bigger than the median of Talorchestia brito that is equal to 0.52. Therefore, Talitrus saltator presents more differentiated orientations, depending on the time of day, period of year and sex, with respect to Talorchestia brito. Therefore, conclusions in Scapini et al. (2002) are corroborated from this perspective.

5.2 Earthquakes distribution on Earth

According to the theory of plate tectonics, Earth is an active planet. Its surface is composed of about 15 individual plates that move and interact, constantly changing and reshaping Earth's outer layer. These movements are usually the main cause of volcanoes and earthquakes. Seismologists have related these natural phenomena to the boundaries of tectonic plates because they tend to occur there, see Selley et al. (2004). In fact, the concentration of earthquake epicenters traces the filamentary network of fault lines and, consequently, they could be analyzed alternatively from the perspective of nonparametric filamentary structure estimation (see, for instance, Genovese et al. 2012). Moreover, tectonic hazards can provoque important damages (destroy buildings, infrastructures or even cause deaths). Therefore, it is important to detect which

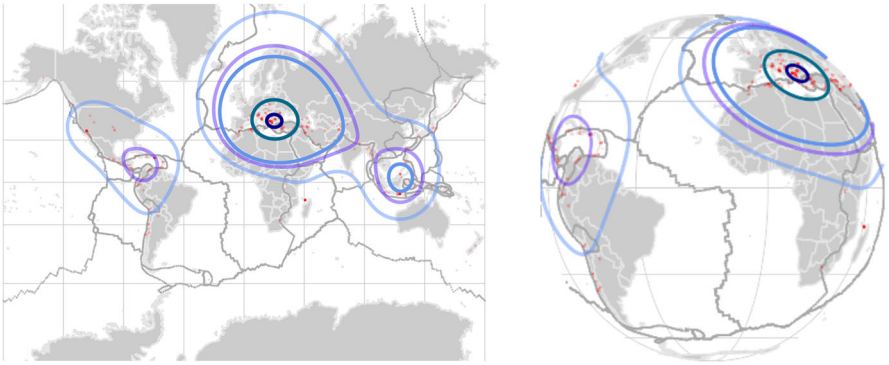


Fig. 14 Contours of HDRs for $\tau_1 = 0.1$, $\tau_2 = 0.3$, $\tau_3 = 0.5$, $\tau_4 = 0.7$ and $\tau_5 = 0.9$ obtained from the sample of world earthquakes registered between October 2004 and April 2020

areas are specially risky. As an illustration, the recent world earthquakes distribution is analyzed next through HDRs estimation.

Figure 14 shows the margins of the tectonic plates (grey color) and the geographical coordinates (red points) of a total of 272 medium and strong earthquakes registered between 1th October 2004 and 9th April 2020 already introduced in Sect. 1.1. Note that most of events are exactly located on the plates boundaries.

Our main goal is to detect which areas are really problematic nowadays. In Sect. 1.1, we show that the largest mode is located on the Southeast Europe considering a value of $\tau = 0.8$. However, a more general view on earthquakes distribution could be obtained if more HDRs are reconstructed for a range of values of τ. Specifically, they were estimated choosing $\tau_1 = 0.1$, $\tau_2 = 0.3$, $\tau_3 = 0.5$, $\tau_4 = 0.7$ and $\tau_5 = 0.9$. The bandwidth parameter used is the proposed in García-Portugués (2013). The corresponding contours are also represented in Fig. 14 using blue colors. An interactive representation of these HDRs can be seen in Appendix D.

The two smallest contours (dark blue colors) corresponds to density regions with probability at least $1 - \tau_5 = 0.1$ and $1 - \tau_4 = 0.3$, respectively. Therefore, they match with the greatest modes of earthquakes world distribution and they identify the more risky parts of the world. They are located on Europe. Concretely, on the boundaries intersection for the Eurasian and African Plates. Note that the second of these regions even includes the frontier of the Arabian Plate. Contours for $\tau_2 = 0.3$ and $\tau_3 = 0.5$ are related to Indo-Australian Plate and margins of Philippine Sea and Pacific Plates appears when $\tau_1 = 0.1$.

As for America, the most problematic area is detected in Central America. Concretely, it is mainly located on the frontiers of Cocos, Nazca and Caribbean Plates. According to the contours shown, this region belongs to the zone of the world where the 70% $(1 - \tau_2\%)$ of earthquakes are registered. If $\tau_1 = 0.1$ is considered then Pacific, North and South American plates appears as risky areas.

6 Conclusions and discussion

The main goals of this work are to extend the definition of HDRs for directional data and propose a plug-in estimator based on a new bootstrap bandwidth selector that is focused on HDRs reconstruction. The route designed to reach this goal can be summarized as follows: (1) Extending the definition of HDRs for directional data, (2) proposing general HDRs plug-in estimators and two different procedures for estimating confidence regions, (3) introducing the first specific selector of the bandwidth parameter for directional HDRs reconstruction, (4) studying the practical behavior of the plug-in estimators (using the new selector and other classical directional bandwidths not specifically designed for HDR reconstruction) and (5) applying the plug-in reconstruction of HDRs to the real data on sandhoppers orientation and earthquakes.

Some further research on the proposed estimator and some natural extensions of this work are discussed. The performance of the procedures for estimating the HDRs confidence regions should be compared, for instance, through simulations. Additionally, consistency results on the proposed HDR estimator and the bootstrap bandwidth selector could be explored following the scheme in Cadre et al. (2009). Regarding the procedure for bandwidth selection, there are two natural extensions. Firstly, as it has been mentioned along the paper, other distances may be used. Secondly, the consideration of the kernel density estimates proposed in Di Marzio et al. (2011) (torus) and García-Portugués et al. (2013) (cylinder) enables the adaptation of our proposal to these settings.

Note also that in the Introduction, we refer to the notion of cluster as the number of connected components of the probability density. With this view in mind, an estimator of the number of directional clusters can be given by the number of connected components of the HDRs plug-in estimator. In addition, (two or more) directional densities could be also compared using the ideas explored in this work: we may compare the discrepancy between directional HDRs estimations, for instance, measuring distances between boundaries. The simple geometric structure of estimators could be used to compute the procedure and calibrate the test using re-sampling schemes.

Finally, earthquakes on Earth could be analyzed following alternative approaches. Note that contour lines in Fig. 14 do not clearly follow the geometry of tectonic plates. A possible cause of this behavior is that earthquakes occur very close to the boundary of the density support (that is, the frontiers of the tectonic plates) and this issue may produce a bias in the estimator. For manifolds with known boundaries, Theorem 3.1 in Berry and Sauer (2017) provides a consistent estimate of the density both in the interior and the frontier, reducing the bias for density evaluations closed to the boundaries. Since the concentration of earthquakes epicenters traces the filamentary network of fault lines, following (Genovese et al. 2012), the performance of nonparametric filament estimators should be also checked for further insight in this problem.

Acknowledgements R.M. Crujeiras and P. Saavedra-Nieves acknowledge the financial support of Ministerio de Economía y Competitividad and Ministerio de Ciencia e Innovación of the Spanish government under grants MTM2016-76969P, MTM2017-089422-P, PID2020-118101GBI00 and PID2020-116587GB-I00 and ERDF. Authors also thank Elena Vázquez Abal for her help, Prof. Felicita Scapini for providing the sandhoppers data (collected under the support of the European Project ERB ICI8-CT98-0270), the computational resources of the CESGA Supercomputing Center and the referees for the constructive comments which have improved the paper.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

A Further details on the datasets

A.1 Levels to the estimated HDRs disaggregating the sandhoppers variables

The orientation of two sandhoppers species (Talitrus saltator and Talorchestia brito) is analyzed in Scapini et al. (2002). The experiment was carried out on the exposed non-tidal sand of Zouara beach located in the Tunisian northwestern coast. Apart from the specie and the orientation angles, this dataset contains information about other variables such as sex (male, female), month (April, October) and moment of the day (morning, afternoon and noon) when the experiment was done. We refer to Scapini et al. (2002) and Marchetti and Scapini (2003) for further details on the dataset and the experimental design.

Table 7 contains the associated levels to the 24 estimated HDRs when variables are disaggregated.

Table 7 Associated levels to the 24 estimated HDRs

Variables		Males			Females		
levels		Afternoon	Noon	Morning	Afternoon	Noon	Morning
Talitrus saltator	October	E1	E2	E3	E4	E5	E6
	April	E7	E8	E9	E10	E11	E12
Talorchestia brito	October	E13	E14	E15	E16	E17	E18
	April	E19	E20	E21	E22	E23	E24

A.2 Interactive representation of HDRs for eathquakes on Earth

An interactive representation of HDRs contours for earthquakes around the world is contained in the supplementary material.

B Simulated spherical models

The circular models considered in the Simulation Study correspond to models 5, 6, 7, 8, 10, 11, 16, 19 and 20 in Oliveira et al. (2012). The spherical models have been constructed as mixtures of von Mises-Fisher distributions. Specifically, the von Mises-Fisher density is given by

$$K_{vM}(x;\mu;\kappa) = C_d(\kappa)\exp\{\kappa x^T\mu\}, \quad \text{with} \quad C_d(\kappa) = \frac{\kappa^{\frac{d-1}{2}}}{(2\pi)^{\frac{d+1}{2}}\mathcal{I}_{\frac{d-1}{2}}(\kappa)}$$

where $\mu \in S^{d-1}$ is the directional mean, $\kappa > 0$ the concentration parameter around the mean, T stands for the transpose operator and $\mathcal{I}_p$ is the modified Bessel function of order p, given by

$$\mathcal{I}_p(z) = \frac{(\frac{z}{2})^p}{\pi^{1/2}\Gamma(p+1/2)}\int_{-1}^{1}(1-t^2)^{p-1/2}e^{zt}dt$$

where $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$, with $p > -1$.

The spherical models S1 to S9 are obtained as mixtures, given by:

$$f_S = \sum_{m=1}^{M}\pi_m K_{vM}(x;\mu_m;\kappa_m), \quad \pi_m \geq 0, \ \sum_{m=1}^{M}\pi_m = 1.$$

The combination of mean and concentration parameters, as well as the weights π_m considered are specified in Table 8.

Table 8 Parameters for the mixtures of von Mises-Fisher spherical distributions S1 to S9 used in the Simulation Study

Model	μ	κ	Weights
S1	(0, 0, 1)	10	1
S2	(0, 0, 1); (0, 0, −1)	1; 1	1/2; 1/2
S3	(0, 0, 1); (0, 0, −1)	10; 1	1/2; 1/2
S4	(0, 0, 1); (0, $1/\sqrt{2}$, $1/\sqrt{2}$)	10; 10	1/2; 1/2
S5	(0, 0, 1); (0, $1/\sqrt{2}$, $1/\sqrt{2}$)	10; 10	2/5; 3/5
S6	(0, 0, 1); (0, $1/\sqrt{2}$, $1/\sqrt{2}$)	10; 5	1/5; 4/5
S7	(0, 0, 1); (0, 1, 0); (1, 0, 0)	5; 5; 5	1/3; 1/3; 1/3
S8	(0, 0, 1); (0, 1, 0); (1, 0, 0)	5; 5; 5	2/3; 1/6; 1/6
S9	(0, 0, 1); (0, $1/\sqrt{2}$, $1/\sqrt{2}$); (0, 1, 0)	10; 10; 10	1/3; 1/3; 1/3

C Additional simulation results

C.1 Circular HDRs estimation

Tables 9 and 10 show the results for $\tau = 0.2$ (for $n = 500$ and $n = 1000$, respectively). In this case, h_1 (being a competitive selector in all the scenarios) is the best one for models C3 and C5 (with $n = 1000$). Note that h_2 presents a poor behavior for models C3, C6, C7, C8 and C9, and h_6 performance is also unsatisfactory for models C5 and C9 ($n = 500$), although it improves with sample size.

Tables 11 and 12 contain the results obtained for $\tau = 0.8$ when $n = 500$ and $n = 1000$, respectively. According to Table 12, h_1 is the best selector for five models (C2, C6, C7, C8 and C9). It is clear that the new selector improves its results when large values of τ are considered and, therefore, largest modes are identified.

Figure 15 shows the violin plots of Hausdorff errors obtained for some of the simulation models when $\tau = 0.2$ ($n = 1000$). If $\tau = 0.2$, h_2 is the selector that presents a worst behavior for models C3 and C6. Furthermore, its variance is always specially large.

Table 9 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.2$, $n = 500$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.086	0.064	0.070	0.071	0.089	0.037	0.084	0.031	0.086	0.053
h_2	**0.067**	0.039	**0.049**	0.032	1.443	0.309	0.099	0.038	0.206	0.153
h_3	0.094	0.060	0.065	0.059	0.090	0.036	0.085	0.031	**0.081**	0.053
h_4	0.075	0.042	0.051	0.036	0.090	0.036	0.085	0.031	0.097	0.072
h_5	0.075	0.041	0.051	0.036	0.091	0.036	0.084	0.031	0.190	0.150
h_6	0.093	0.058	**0.049**	0.032	**0.087**	0.033	**0.081**	0.031	0.358	0.110

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.098	0.038	**0.081**	0.026	0.154	0.098	0.145	0.065
h_2	1.786	0.110	1.812	0.064	0.251	0.094	1.758	0.115
h_3	0.100	0.038	0.084	0.027	0.153	0.093	0.142	0.063
h_4	0.099	0.038	**0.081**	0.025	**0.128**	0.067	**0.131**	0.049
h_5	0.100	0.038	**0.081**	0.025	**0.128**	0.067	**0.131**	0.049
h_6	**0.097**	0.037	0.082	0.026	0.156	0.091	1.686	0.372

Table 10 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.2$, $n = 1000$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.058	0.040	0.040	0.034	**0.061**	0.025	0.058	0.024	**0.060**	0.028
h_2	**0.049**	0.027	0.036	0.021	1.428	0.334	0.066	0.024	0.097	0.066
h_3	0.059	0.039	0.037	0.028	0.063	0.024	0.058	0.025	0.061	0.024
h_4	0.053	0.028	0.036	0.021	0.063	0.024	0.057	0.024	0.077	0.037
h_5	0.053	0.028	0.036	0.021	0.063	0.024	**0.056**	0.023	0.083	0.052
h_6	0.059	0.039	**0.035**	0.020	**0.061**	0.024	0.057	0.024	0.174	0.148

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	0.072	0.028	0.057	0.016	0.101	0.060	0.115	0.042
h_2	1.798	0.106	1.820	0.057	0.179	0.047	1.759	0.118
h_3	0.073	0.028	0.057	0.016	0.097	0.060	**0.110**	0.044
h_4	0.072	0.028	**0.056**	0.016	**0.091**	0.043	0.114	0.041
h_5	0.072	0.028	**0.056**	0.016	**0.091**	0.043	0.114	0.041
h_6	**0.071**	0.028	0.057	0.016	0.107	0.038	0.111	0.042

Table 11 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.8$, $n = 500$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.022	0.014	0.151	0.108	**0.691**	0.812	**0.610**	0.679	0.079	0.075
h_2	0.020	0.013	0.184	0.063	1.849	0.282	0.905	0.756	0.058	0.036
h_3	0.022	0.014	**0.141**	0.106	0.705	0.835	0.635	0.701	0.088	0.073
h_4	**0.019**	0.013	0.155	0.081	0.685	0.823	0.634	0.700	0.070	0.045
h_5	**0.019**	0.013	0.157	0.081	0.726	0.844	0.650	0.709	0.061	0.041
h_6	0.022	0.014	0.184	0.060	0.784	0.873	0.673	0.721	**0.051**	0.031

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	**1.189**	0.664	**1.076**	0.333	0.353	0.243	**0.837**	0.436
h_2	1.747	0.059	1.820	0.068	0.356	0.060	1.809	0.188
h_3	1.246	0.644	1.143	0.236	**0.342**	0.247	0.983	0.363
h_4	1.269	0.632	1.171	0.262	0.404	0.195	1.101	0.317
h_5	1.246	0.644	1.171	0.262	0.407	0.186	1.101	0.317
h_6	1.310	0.606	1.162	0.251	0.396	0.096	1.752	0.251

Table 12 Means (M) and standard deviations (SD) of 250 errors in Hausdorff distance for $\tau = 0.8$, $n = 1000$ and $B = 200$

	C1		C2		C3		C4		C5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	0.015	0.012	**0.127**	0.075	0.472	0.678	0.441	0.594	0.056	0.038
h_2	**0.013**	0.011	0.157	0.042	1.878	0.231	0.561	0.707	0.045	0.029
h_3	0.016	0.012	0.120	0.070	0.461	0.692	**0.410**	0.588	0.064	0.041
h_4	**0.013**	0.011	0.146	0.046	0.467	0.699	0.412	0.593	0.050	0.032
h_5	**0.013**	0.011	0.146	0.046	**0.447**	0.680	0.433	0.614	0.050	0.032
h_6	0.016	0.012	0.136	0.050	0.460	0.702	0.431	0.613	**0.043**	0.027

	C6		C7		C8		C9	
	M	SD	M	SD	M	SD	M	SD
h_1	**0.983**	0.721	**1.048**	0.30	**0.252**	0.220	**0.708**	0.416
h_2	1.746	0.058	1.825	0.063	0.341	0.051	1.810	0.188
h_3	1.011	0.724	1.086	0.242	0.257	0.219	0.897	0.334
h_4	1.004	0.727	1.112	0.241	0.360	0.176	1.038	0.199
h_5	0.998	0.728	1.112	0.241	0.360	0.176	1.038	0.199
h_6	1.025	0.724	1.086	0.242	0.396	0.077	0.957	0.289

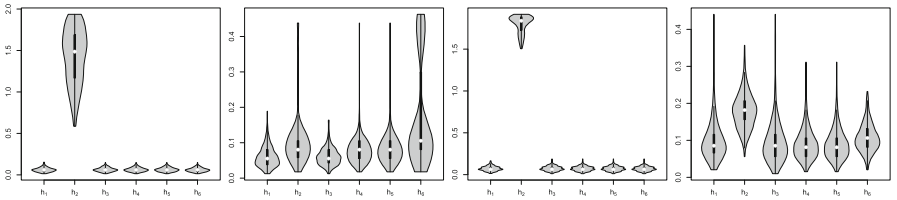


Fig. 15 Violin plots of Hausdorff errors for models C3, C5, C6 and C8 for $\tau = 0.2$ and $n = 1000$. Note that due to the behaviour of h_2, the scales of these figures are different

C.2 Spherical HDRs estimation

Tables 13 and 14 show the means and the standard deviations of the 200 estimation errors obtained when $\tau = 0.2$ from samples of sizes $n = 1500$ and $n = 2500$, respectively. Bold numbers correspond to the lowest mean errors obtained for each density. Except for model S8, h_1 is the best or shows a competitive performance.

Tables 15 and 16 show the means and the standard deviations of the 200 estimation errors obtained when $\tau = 0.8$ from samples of size $n = 1500$ and $n = 2500$, respectively. Although results for S7 are not good when h_1 is considered, this selector is again the best or competitive with h_5. As for h_7, results are remarkably poor in S2 and S6.

Figure 16 contains the violin plots of Hausdorff errors for models S3, S4, S6 and S9 when $\tau = 0.2$ and $n = 2500$. Note that he performance of selector h_1 is considerably good.

Table 13 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.2$, $n = 1500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.033**	0.010	**0.757**	0.261	0.445	0.168	**0.065**	0.014	**0.074**	0.018
h_5	0.052	0.012	0.764	0.238	0.598	0.145	0.072	0.018	0.078	0.020
h_7	0.063	0.013	1.224	0.243	**0.371**	0.072	0.073	0.017	0.078	0.019
	S6		S7		S8		S9			
	M	SD	M	SD	M	SD	M	SD		
h_1	**0.075**	0.019	0.748	0.241	0.302	0.118	0.081	0.017		
h_5	0.080	0.022	**0.651**	0.335	0.275	0.082	0.088	0.018		
h_7	0.089	0.023	0.921	0.199	**0.263**	0.087	**0.079**	0.016		

Table 14 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.2$, $n = 2500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.029**	0.009	0.650	0.212	0.340	0.089	**0.059**	0.013	**0.061**	0.014
h_5	0.044	0.010	**0.648**	0.199	0.484	0.097	0.063	0.014	0.067	0.015
h_7	0.053	0.011	1.137	0.245	**0.306**	0.053	0.062	0.014	0.067	0.016
	S6		S7		S8		S9			
	M	SD	M	SD	M	SD	M	SD		
h_1	**0.064**	0.016	0.667	0.272	0.247	0.099	0.070	0.015		
h_5	0.071	0.018	**0.556**	0.336	0.230	0.073	0.075	0.016		
h_7	0.079	0.017	0.889	0.229	**0.218**	0.073	**0.069**	0.014		

Table 15 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.8$, $n = 1500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.030**	0.012	0.629	0.359	**0.040**	0.014	0.138	0.042	**0.132**	0.078
h_5	0.054	0.018	**0.537**	0.295	0.053	0.018	**0.134**	0.041	0.154	0.101
h_7	0.068	0.023	0.915	0.533	**0.040**	0.012	**0.134**	0.041	0.155	0.102
	S6		S7		S8		S9			
	M	SD	M	SD	M	SD	M	SD		
h_1	**0.089**	0.023	0.462	0.231	**0.063**	0.021	0.243	**0.109**		
h_5	0.110	0.033	0.212	0.143	0.083	0.025	**0.233**	0.123		
h_7	0.128	0.039	**0.195**	0.160	0.073	0.022	**0.233**	0.124		

Table 16 Means (M) and standard deviations (SD) of 200 errors in Hausdorff distance for $\tau = 0.8$, $n = 2500$ and $B = 50$

	S1		S2		S3		S4		S5	
	M	SD	M	SD	M	SD	M	SD	M	SD
h_1	**0.023**	0.008	0.403	0.171	0.031	0.008	0.122	0.033	**0.115**	0.073
h_5	0.043	0.013	**0.399**	0.128	0.047	0.011	**0.117**	0.033	0.133	0.097
h_7	0.054	0.017	0.670	0.475	**0.030**	0.009	**0.117**	0.032	0.135	0.098

	S6		S7		S8		S9	
	M	SD	M	SD	M	SD	M	SD
h_1	**0.081**	0.023	0.234	0.200	**0.051**	0.016	0.201	0.086
h_5	0.096	0.032	0.166	0.056	0.066	0.018	**0.186**	0.090
h_7	0.112	0.037	**0.140**	0.047	0.059	0.017	0.194	0.110

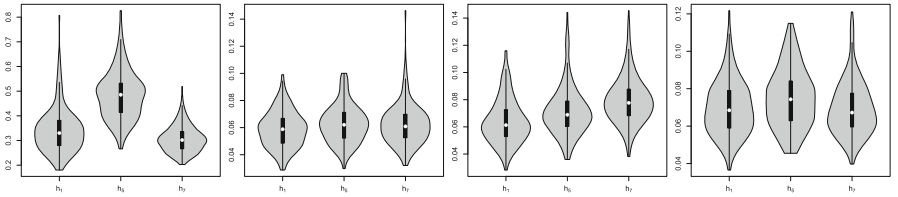


Fig. 16 Violin plots of Hausdorff errors for models S3 and S4, S6 and S9 when $\tau = 0.2$ and $n = 2500$. Note that the scales of these figures are slightly different

D Some details on the directional bandwidth selectors

We briefly revise in this section some bandwidth selection methods designed for kernel density estimation. Although these methods do not focus on HDRs, but on the reconstruction of the whole density curve, it may be argued that they could also be used for constructing the proposed plug-in estimator. The performance of our proposal is compared in all the simulated scenarios with different bandwidth selectors for circular and spherical data.

As in the Euclidean setting, most used techniques for selecting h are based on the minimization of some error criteria that quantify the accuracy of the kernel density estimator. One of the most simple errors to be considered is the mean integrated squared error that can be written as follows:

$$MISE(h) = \mathbb{E}\left[\int_{S^{d-1}} (f_n(x) - f(x))^2 \omega_d(dx)\right], \tag{11}$$

where ω_d denotes the Lebesgue in S^{d-1}. Then, a possibility is to search for the bandwidth that minimizes (11). However, the asymptotic version of $MISE$, $AMISE$, is more commonly used in the literature. A rule of thumb proposed in Taylor (2008) adapts the idea in Silverman (1986) in kernel linear density estimation to the circular setting. The resulting plug-in selector assumes that the data follow a von Mises

distribution to determine the $AMISE$. The bandwidth is chosen by first obtaining an estimation $\hat{\kappa}$ of the concentration parameter κ in the reference density (for example, by maximum likelihood) through the formula

$$h_2 = \left[\frac{4\pi^{1/2}\mathcal{I}_0(\hat{\kappa})^2}{3\hat{\kappa}^2\mathcal{I}_2(2\hat{\kappa})n}\right]^{1/5}.$$

Remark that the parametrization in Taylor (2008) has been adapted to the context of the estimator (4) by denoting by h the inverse of the squared concentration parameter employed in his paper. The poor performance of this rule is sometimes due to the non robust estimation by maximum likelihood of the concentration parameter. An alternative and robustified estimation procedure is considered in Oliveira et al. (2013).

A new selector also devoted to the circular case is established in Oliveira et al. (2012). It improves the performance of the Taylor's proposal allowing for more flexibility in the reference density, considering a mixture of von Mises. This selector is mainly based on two elements. First, the $AMISE$ expansion derived in Di Marzio et al. (2009) for the circular kernel density estimator by the use of Fourier expansions of the circular kernels. This expression has the following form when the kernel is a circular von Mises (the estimator is equivalent to consider $L(r) = e^{-r}$ and h as the inverse of the squared concentration parameter in (4):

$$AMISE(h) = \frac{1}{16}\left[1 - \frac{\mathcal{I}_2(h^{-1/2})}{\mathcal{I}_0(h^{-1/2})}\right]^2 \int_0^{2\pi} f''(\theta)^2 d\theta + \frac{\mathcal{I}_0(2h^{-1/2})}{2n\pi\mathcal{I}_0(h^{-1/2})^2}. \quad (12)$$

The second element is the Expectation–Maximization (EM) algorithm in Banerjee et al. (2005) for fitting mixtures of directional von Mises. The selector, that is denoted by h_3, proceeds as follows: first, apply the EM algorithm to fit mixtures with different number of components; then, choose the fitted mixture with the lowest AIC. Finally, compute the curvature term in (12) using the fitted mixture and seek for the h that minimizes this expression. This value of h is denoted by h_3.

Of course, plug-in rules are not the only alternative to smoothing parameter selection. Some other data-driven directional procedures were already proposed in Hall et al. (1987) using cross-validation ideas. Specifically, Least Squares Cross-Validation (LSCV) and Likelihood Cross-Validation (LCV) bandwidth are introduced, arising as the minimizers of the cross-validated estimates of the squared error loss and the Kullback–Leibler loss, respectively. The selectors have the following expressions:

$$h_4 = \arg\max_{h>0} 2n^{-1}\sum_{i=1}^{n} f_n^{-i}(X_i) - \int_{S^{d-1}} f_n(x)^2\omega_q(dx)$$

and

$$h_5 = \arg\max_{h>0} \sum_{i=1}^{n} \log f_n^{-i}(X_i),$$

where f_n^{-i} represents the kernel estimator computed without the $i-$th observation.

A bootstrap bandwidth selection procedure for data lying on a $d-$dimensional torus is proposed in Di Marzio et al. (2011). If a von Mises kernel is used, then the bootstrap MISE has a closed expression. Then, h_6 is selected as the value that minimizes

$$\int_{S^1} \mathbb{E}_B \left[{f_n}^*(X) - f_n(X) \right]^2 \omega_d(dx)$$

where $\mathbb{E}_B$ denotes the bootstrap expectation with respect to random samples $\{X_1^*, \ldots, X_n^*\}$ generated from $f_n(X)$. A common problem for small samples is that a local minimum may be chosen, as pointed out by Oliveira et al. (2012).

Apart from existing cross-validation procedures in the directional setting, García-Portugués (2013) derives a plug-in directional analogue to the rule of thumb in Silverman (1986) using the properties of the von Mises density. Moreover, it is the optimal $AMISE$ bandwidth for normal reference density and normal kernel. Concretely, if the von Mises kernel is considered and κ is estimated by maximum likelihood,

$$h_7 = \begin{cases} \left[\frac{4\pi^{1/2} \mathcal{I}_0(\hat{k})^2}{\hat{k}[\mathcal{I}_1(2\hat{k}) + 3\hat{k}\mathcal{I}_2(2\hat{k})]n} \right]^{1/5} & \text{in} \quad S^1 \\ \left[\frac{8 \sinh^2(\hat{k})}{\hat{k}[(1+4\hat{k}^2)\sinh(2\hat{k}) - 2\hat{k}\cosh 2\hat{k}]n} \right]^{1/6} & \text{in } S^2. \end{cases}$$

References

Anderberg MR (1973) Cluster analysis for applications. Academic Press

Bai ZD, Rao CR, Zhao LC (1989) Kernel estimators of density function of directional data. Multivar Stat Probab 24–39

Baíllo A (2003) Total error in a plug-in estimator of level sets. Stat Probab Lett 65:411–417

Baíllo A, Cuevas A (2006) Parametric versus nonparametric tolerance regions in detection problems. Comput Stat 21:523–536

Banerjee A, Dhillon IS, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res 6:1345–1382

Berry T, Sauer T (2017) Density estimation on manifolds with boundary. Comput Stat Data Anal 107:1–17

Biau G, Cadre B, Pelletier B (2007) A graph-based estimator of the number of clusters. ESAIM Probab Stat 11:272–280

Boissonnat JD, Lieutier A, Wintraecken M (2019) The reach, metric distortion, geodesic convexity and the variation of tangent spaces. J Appl Comput Topol 3(1):29–58

Box GEP, Tiao GC (1973) Bayesian Inference in Statistical Analysis, Reading. Addison-Wesley

Cadre B, Pelletier B, Pudlo P (2009) Clustering by estimation of density level sets at a fixed probability

Chen YC, Genovese CR, Wasserman L (2017) Density level sets: Asymptotics, inference, and visualization. J Am Stat Assoc 112:1684–1696

Cholaquidis A, Fraiman R, Moreno L (2020) Level set and density estimation on manifolds. arXiv preprint arXiv:2003.05814

Cox DR, Hinkley D. V. (1979) Theoretical statistics. CRC Press

Cuevas A, Febrero M, Fraiman R (2000) Estimating the number of clusters. Can J Stat 28:367–382

Cuevas A, Febrero M, Fraiman R (2001) Cluster analysis: a further approach based on density estimation. Comput Stat Data Anal 36(4):441–459

Cuevas A, Fraiman R (1997) A plug-in approach to support estimation. Ann Stat 25:2300–2312

Cuevas A, González-Manteiga W, Rodríguez-Casal A (2006) Plug-in estimation of general level sets. Aust N Z J Stat 48(1):7–19

Dekker W, Strandvlooien (1978) Tabellenserie van de Strandwerkgemeenschap 24

Di Marzio M, Panzera A, Taylor CC (2009) Local polynomial regression for circular predictors. Stat Probab Lett 79:2066–2075
Di Marzio M, Panzera A, Taylor CC (2011) Kernel density estimation on the torus. J Stat Plan Inference 141:2156–2173
García JN, Kutalik Z, Cho KH, Wolkenhauer O (2003) Level sets and minimum volume sets of probability density functions. Int J Approx Reason 34:25–47
García-Portugués E (2013) Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. Electron J Stat 7:1655–1685
García-Portugués E, Crujeiras RM, González-Manteiga W (2013) Kernel density estimation for directional-linear data. J Multivar Anal 121:152–175
Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L (2012) The geometry of nonparametric filament estimation. J Am Stat Assoc 107(498):788–799
Everitt BS (1993) Cluster Analysis. Arnold-Halsted
Hall P (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. J Multivar Anal 14(1):1–16
Hall P, Watson GS, Cabrera J (1987) Kernel density estimation with spherical data. Biometrika 74(4):751–762
Hartigan J (1975) Clustering algorithms, Wiley Series in Probability and Mathematical Statistics. Wiley
Huckemann S, Kim KR, Munk A, Rehfeldt F, Sommerfeld M, Weickert J, Wollnik C (2016) The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. Bernoulli 22(4):2113–2142
Hyndman RJ (1996) Computing and graphing highest density regions. Am Stat 50:120–126
Jeong J, Jun M, Genton MG (2017) Spherical process models for global spatial statistics. Stat Sci 32(4):501–513
Klemelä J (2000) Estimation of densities and derivatives of densities with directional data. J Multivar Anal 73(1):18–40
Mammen E, Polonik W (2013) Confidence regions for level sets. J Multivar Anal 122:202–214
Marchetti GM, Scapini F (2003) Use of multiple regression models in the study of sandhopper orientation under natural conditions. Estuar Coast Shelf S 58:207–215
Mason DM, Polonik W (2009) Asymptotic normality of plug-in level set estimates. Ann Appl Probab 19:1108–1142
Oliveira M, Crujeiras RM, Rodríguez-Casal A (2012) A plug-in rule for bandwidth selection in circular density estimation. Comput Stat Data Anal 56:3898–3908
Oliveira M, Crujeiras RM, Rodríguez-Casal A (2013) Nonparametric circular methods for exploring environmental data. Environ Ecol Stat 20:1–17
Oliveira M, Crujeiras RM, Rodríguez-Casal A (2014) NPCirc: an R package for nonparametric circular methods. J Stat Softw 61(9):1–26
Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076
Polonik W (1997) Minimum volume sets and generalized quantile processes. Stoch Proc Appl 69:1–24
Polonik W (2013) Confidence regions for level sets. J Multivar Anal 122:202–214
Rigollet P, Vert R (2009) Optimal rates for plug-in estimators of density level sets. Bernoulli 15:1154–1178
Rinaldo A, Wasserman L (2001) Generalized density clustering, Ann. Stat., 38, 2678-2722 (2010) subsets that differ between samples. Cytometry 45:56–64
Rodríguez-Casal A, Saavedra-Nieves P (2019) Minimax Hausdorff estimation of density level sets. arXiv preprint arXiv:1905.02897
Rosenblatt M (1956) Remarks on some nonparametric estimate of a density function. Ann Math Stat 27:832–837
Samworth R, Wand M (2010) Asymptotics and optimal bandwidth selection for highest density region estimation. Ann Statist 38:1767–1792
Scapini F, Aloia A, Bouslama MF, Chelazzi L, Colombini I, ElGtari M, Fallaci M, Marchetti GM (2002) Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, Talitrus saltator and Talorchestia brito, from an exposed Mediterranean beach. Behav Ecol Sociobiol 51(5):403–414
Selley RC, Cocks R, Plimer I (2004) Encyclopedia of geology, Academic Press
Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall
Steinwart I (2015) Fully adaptive density-based clustering. Ann Stat 43:2132–2167

Taylor CC (2008) Automatic bandwidth selection for circular density estimation. Comput Stat Data Anal 52:3493–3500
Tsybakov AB (1997) On nonparametric estimation of density level sets. Ann Stat 25:948–969
Wand MP, Jones MC (1995) Kernel Smoothing. Chapman and Hall