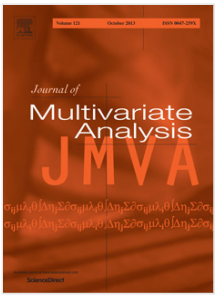
Accepted Manuscript



Optimal level sets for bivariate density representation

Pedro Delicado, Philippe Vieu

PII: S0047-259X(15)00093-7
DOI: http://dx.doi.org/10.1016/j.jmva.2015.04.005
Reference: YJMVA 3920

To appear in: *Journal of Multivariate Analysis*

Received date: 13 October 2014

Please cite this article as: P. Delicado, P. Vieu, Optimal level sets for bivariate density representation, *Journal of Multivariate Analysis* (2015), http://dx.doi.org/10.1016/j.jmva.2015.04.005

Optimal level sets for bivariate density representation

Pedro Delicado[a,*], Philippe Vieu[b]

[a] *Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain.*
[b] *Institut de Mathématiques, Université Paul Sabatier, Toulouse, France.*

Abstract

In bivariate density representation there is an extensive literature on level set estimation when the level is fixed, but this is not so much the case when choosing which level is (or which levels are) of most interest. This is an important practical question which depends on the kind of problem one has to deal with as well as the kind of feature one wishes to highlight in the density, the answer to which requires both the definition of what the optimal level is and the construction of a method for finding it. We consider two scenarios for this problem. The first one corresponds to situations in which one has just a single density function to be represented. However, as a result of the technical progress in data collecting, problems are emerging in which one has to deal with a sample of densities. In these situations, the need arises to develop joint representation for all these densities, and this is the second scenario considered in this paper. For each case, we provide consistency results for the estimated levels and present wide Monte Carlo simulated experiments illustrating the interest and feasibility of the proposed method.

Keywords: Bivariate density representation, functional data analysis, minimum distance estimation, multidimensional scaling, nonparametric density estimation.

1. Introduction

Let f be a bivariate probability density function. For $\alpha \in]0,1[$ we define the density level set with probability content α as

$$C_\alpha = \{x \in \mathbf{R}^2 : f(x) \geq \gamma_\alpha\},$$

where γ_α is such that

$$\int_{C_\alpha} f(x)dx = \alpha.$$

*Corresponding author
Email addresses: pedro.delicado@upc.edu (Pedro Delicado), philippe.vieu@math.univ-toulouse.fr (Philippe Vieu)

When needed, we will write C_α^f to make explicit the dependence of C_α on f. A standard way to represent the bivariate density f graphically is by drawing in the same graph density level sets corresponding to several values $\alpha_1, \ldots, \alpha_J$, or just their boundaries (see, for instance, Bowman and Azzalini 1997 or Duong 2007 as well as the accompanying R packages `sm` and `ks`, respectively). Other authors (Silverman 1986, Scott 1992, Wand and Jones 1995, Simonoff 1996) draw the density contour levels at equally spaced heights (see also the R package `KernSmoth`, associated with Wand and Jones 1995).

In this paper we consider the following problem: given a bivariate density function f (respectively, N bivariate density functions $f_1, \ldots, f_N$) and fixed an integer $J \geq 1$, choose the combination of values $\alpha_1, \ldots, \alpha_J$ defining the *best* (in some sense) graphical representation of f (resp., $f_1, \ldots, f_N$). The exact meaning of *best graphical representation* is specified in Sections 2 and 3. For the moment, an informal way to express this concept is to say that the chosen density level sets must reflect *as well as possible* the shape of f (resp., $f_1, \ldots, f_N$). It can also be said that the *visual distance* between f (or $f_1, \ldots, f_N$) and its (their) graphical representation using the chosen density level sets must be minimized.

Representing bivariate densities by one level set (in this case $J = 1$) allows us to draw more than one bivariate density function in the same graph. This kind of graphs is helpful in different situations, such as:

- Several samples of the same bivariate random variable X are taken at different times (or in different regions, or more in general, in different conditions). A nonparametric estimation of the density of X is derived from each sample. A graph that enables possible changes in the distribution of X across different scenarios to be visualized consists in representing the estimated densities in the same graph, each by a density level set. A very nice example can be found in Bowman and Azzalini (1997). They study data on aircraft designs from the periods 1914-1935, 1936-1955 and 1956-1984, originally explored in Bowman and Foster (1993). They obtain the first two principal components (identified as "size" and "speed adjusted by size", respectively) and represent their joint density by using only a level plot (corresponding to probability 0.75) for each period. The authors are able to summarize the way in which aircraft designs have changed over the last century in a single graph (reproduced here in Figure 1, top panel).

- Assume that a functional principal component analysis (FPCA) is performed from the set of bivariate densities $f_1, \ldots, f_N$. In FPCA, for one-dimensional functions it is standard for representing the principal functions graphically superposing three functions in the same plot: the mean function and the mean function plus (and minus) the principal function (multiplied by a constant). See, for instance, Ramsay and Silverman (2005, Section 8.3.1). In order to plot a similar graph when addressing with bivariate density functions, we need a way to represent three such functions in the same graph. The use of a level set for representing each function is a simple and effective choice. A related example (using multi-

dimensional scaling instead of FPCA) can be found in Delicado (2011a). The levels sets used there have a probability content of 0.75.

In other situations, it could be interesting to have more than one level set (in this case $J > 1$) for depicting some feature of the density, as in the following example in which $J = 3$:

- When the number of bivariate density functions to be represented is large, and when each density is recorded at a different time and the elapsed time between two consecutive densities is short, a convenient way to represent them is by an animated graph, in which each image corresponds to the graph of each bivariate density. In this case it is appropriate to represent each density by just a few density level sets (for instance, 3). Therefore the animated graph shows how the level sets evolve over time. Let us consider the aircraft example once more. The animated graph provided as supplementary material (see Appendix B) represents a set of 52 bivariate densities (we use the `animate` LaTeX package from Holoček and Sojka 2004). For $i = 1, \ldots, 52$, the i-th density is estimated from data corresponding to aircrafts produced between year $\min\{i+1913, 1956\}$ and year $\max\{i + 1932, 1935\}$, so that the periods 1914-1935, 1936-1955 and 1956-1984 are particular cases ($i = 1, 23, 52$). The density level sets corresponding to probabilities 0.25, 0.5 and 0.75 are drawn for each density. This dynamic graph is an attractive way to visualize the development of aircraft design and complements the static view (Figure 1, top panel).

However, for $J = 1$ as well as for more than one level set, an important open question is to determine which level(s) should be used. Nowadays, it is standard to represent a bivariate density function (either known or nonparametrically estimated from a random sample) by plotting $J = 3$ of its density level sets, usually those corresponding to $\alpha = 1/4$, $1/2$ and $3/4$ (by analogy with the univariate boxplots), as in the aircraft example discussed above. Bowman and Azzalini (1997) call these plots 'sliceplots', and refer the reader to Bowman and Foster (1993) for further details. A relevant question is to determine whether the choice of these three values of α is sensible (under some criterion) or if there exists an alternative better choice. This of course depends on what is meant by a *good* level set.

The paper is organized in two main sections (Sections 2 and 3), concluding remarks (Section 4) and an Appendix (with two parts). Section 2 deals with the case of having only one bivariate density function to be represented ($N = 1$). In this case $J > 1$ is normally used (the choice of J will be given by computational or even aesthetic or perceptual considerations: too many level sets in the same graph makes it difficult to appreciate). We develop two approaches for quantifying the quality of a level set (and therefore for constructing the optimal representation). The first one (subsection 2.1) uses distances between level sets, while the second one (subsection 2.2) is based on distances between density functions. In each case we discuss the methodologies, show their performance on simulated data (in particular, when $J = 3$, one sees that the usual choices

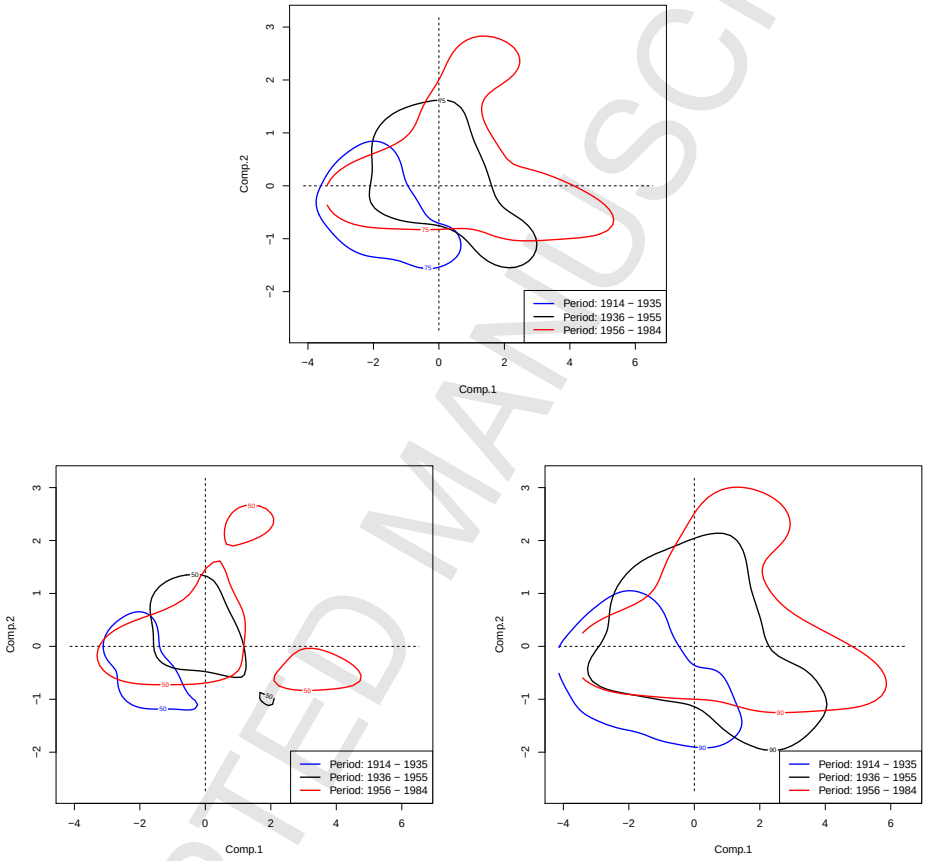


Figure 1: Aircraft data. The estimated bivariate densities of the first two principal components are represented by density level sets.
Top panel: Level sets with content 0.75 for three periods.
Bottom left panel: Level sets with content 0.5 (optimal value according to (2.1) for $J = 1$).
Bottom right panel: Level sets with content 0.9 (optimal value according to (2.6) for $J = 1$ using Hellinger distance).

$\alpha = 1/4$, $1/2$ and $3/4$ are not always the most relevant) and give some asymptotic results, paying attention to the most current situation when the density to be represented is not the real density but rather an estimated one. Subsection 2.3 points out that the proposals introduced in Section 2 are not well suited to the case of having several density functions to be represented.

Consequently, in Section 3 we consider the representation of several densities ($N > 1$). We deal first with the case of representing each density function with only one level set (subsection 3.1) and then we consider the case of several level sets (subsection 3.2). The main idea is that the distances between the bivariate density functions $f_1, \ldots, f_N$ and the distances between their graphical representation using level sets must be as similar as possible. Subsection 3.3 shows that our proposals are still valid when the real density functions are unknown. Subsection 3.4 gathers together the simulated examples corresponding to Section 3.

Finally, Section 4 summarizes the main conclusions drawn from the paper. To facilitate the reading of the paper, the technical issues (including the mathematical assumptions and the proofs) are all addressed together in Appendix A. Appendix B describes the on-line supplementary material.

2. Optimal level sets for a single density

We now consider the problem of representing only one density by some of its density level sets. We assume that J has been fixed in advance and we wish to make the best choice of $\alpha_1, \ldots, \alpha_J$. There is no single way for specifying what *best* might mean. We consider two possibilities: the first chooses the J density level sets that best represent the whole family of level sets $\{C_\alpha : \alpha \in]0, 1[\}$, in the sense that each non-plotted C_α is close to the nearest level among those that are plotted: $C_{\alpha_1}, \ldots, C_{\alpha_J}$. This is developed in Section 2.1.

In the second approach, we argue that each collection of level sets $C_{\alpha_1}, \ldots, C_{\alpha_J}$ naturally defines a piecewise uniform bivariate density function. Our proposal is to minimize in $\alpha_1, \ldots, \alpha_J$ the distance between this piecewise uniform density and the one we wish to represent by $C_{\alpha_1}, \ldots, C_{\alpha_J}$. Section 2.2 deals with this idea. Some artificial data examples are used to illustrate both approaches in practice.

In a different context Marron and Tsybakov (1995) proposed a visual distance between a univariate density functions and its nonparametric estimations, that otherwise is difficult to be extended to the bivariate case.

2.1. Optimality based on distances between density level sets

We consider the following distances between sets $A, B \subseteq \mathbf{R}^2$:

$$d_\lambda(A, B) = \int_{A \Delta B} dx = \lambda(A \Delta B), \quad d_f(A, B) = \int_{A \Delta B} f(x) dx = \mu_f(A \Delta B),$$

where Δ denotes the symmetric difference between sets, λ is the Lebesgue measure in $\mathbf{R}^2$ and μ_f is the probability measure in $\mathbf{R}^2$ having f as a density

function. There exist other distances between sets that could be used as an alternative (Hausdorff's distance, for instance, or its L^p version defined in Baddeley 1992; for more details on these and other distances between sets see, e.g., Cuevas 2009 or Cuevas and Fraiman 2010 and references therein).

A natural way to choose values $\alpha_1, \ldots, \alpha_J$ is by solving this minimization problem:

$$\min_{0<\alpha_1<\cdots<\alpha_J<1} \int_0^1 d(C_u, C_{\alpha_{j(u)}})du \tag{2.1}$$

where d is either d_λ or d_f, and $j(u)$ is such that

$$d(C_u, C_{\alpha_{j(u)}}) = \min_{j=1\ldots J} d(C_u, C_{\alpha_j}),$$

that is, $C_{\alpha_{j(u)}}$ is the closest set to C_u among the sets $C_{\alpha_1}, \ldots, C_{\alpha_j}$.

Theorem 1. *For $d = d_f$, the optimal solution to problem (2.1) is*

$$\alpha_j^f = \frac{2j-1}{2J},\ j = 1, \ldots, J.$$

Assume now that the support of f, say C_1, is compact. For $d = d_\lambda$ the optimal solution to problem (2.1) is α_j^λ, $j = 1, \ldots, J$, such that

$$\frac{\lambda(C_{\alpha_j^\lambda})}{\lambda(C_1)} = \frac{2j-1}{2J},\ j = 1, \ldots, J.$$

The proof of this theorem (in the Appendix) shows that the choice of $d = d_f$ results in some kind of probability transform (for $v > u$, $d_f(C_u, C_v) = \mu_f(C_v) - \mu_f(C_u) = v - u$) that ties our proposal with the k-median problem for the uniform distribution over $[0, 1]$ (see Lemma 4 in the Appendix). The main implication of this fact is that α_j^f, the optimal values when using $d = d_f$, do not depend on f, what is no longer true when using $d = d_\lambda$, making the definition of optimal α_j's based on $d = d_f$ much more appealing than the other alternative.

For the first values of J the optimal α_j^f are the following:

J	$\alpha_j^f,\ j = 1, \ldots, J$
1	1/2
2	1/4, 3/4
3	1/6, 1/2, 5/6

We see that when $J = 3$ level sets are plotted, the optimal values (in this sense) for α_j are not those that are commonly used (0.25, 0.5 and 0.75). The lower left panel of Figure 1 represents three bivariate densities (one corresponding to each of the three periods defined by Bowman and Azzalini 1997) using the optimal value of α for $J = 1$. The most notable difference with respect to the top panel is that the rapid development between the first and second period is now more apparent (the corresponding level sets are almost disjoint), while the development between the second and third period took place in three directions:

specialization in larger aircraft, specialization in faster aircraft, and recovering smaller and slower aircraft, such as those manufactured at the beginning of the century.

The bivariate density f it not commonly known in practice. We normally observe n independent data coming from f and we define an estimator $\hat{f}_n$ of f based on these data ($\hat{f}_n$ is usually a nonparametric estimator of the kernel type). Then the level sets finally plotted are not those belonging to f but those belonging to $\hat{f}_n$ (which are known as plug-in density level estimators). Short reviews on level set estimation can be found in Cuevas (2009) and Cuevas and Fraiman (2010). Of particular interest for us are the works of Baíllo et al. (2001) and Cadre (2006), which deal with the convergence of the plug-in density level estimating sets $C_{\alpha,n} = \{x \in \mathbf{R}^2 : \hat{f}_n(x) \geq \gamma_{\alpha,n}\}$, with $\int_{C_{\alpha,n}} \hat{f}_n(x)dx = \alpha$, to the density level set C_α of f, where $\hat{f}_n$ is a kernel density estimator of f based on n independent copies of the random variable X with density f. Specifically, Baíllo et al. (2001) obtain rates of convergence for $P\{Z \in C_{\alpha,n}\} - \alpha$, where $Z \sim f$ is independent of $\hat{f}_n$ (see Ren and Mojirsheibani 2008, for similar results under weaker assumptions). Baíllo (2003) proves that $d_\lambda(C_{\alpha,n}, C_\alpha)$ converges almost surely to 0 while Cadre (2006) finds the convergence rate. All these results give theoretical support to the use of estimated level sets in Theorem 1.

Different approaches to density level sets estimation, not based on the plug-in principle, have been explored by Polonik (1995), Tsybakov (1997) and Willett and Nowak (2007). These authors fix the value of the density functions at the boundary of the level set (instead of fixing the desired probability content) and they estimate directly the level set, without be interested in the estimation of the whole density. In these papers, as well as in Baíllo et al. (2001), Baíllo (2003) or Cadre (2006), the level set estimation is the main objective, whereas our main goal is to decide which level sets should be represented.

2.2. Optimality based on distances between bivariate densities

In the previous section we introduced an optimality criterion (see equation (2.1)) based exclusively on distances between density level sets. In some sense equation (2.1) is a kind of location problem in the set of all possible level sets. Nevertheless this optimality criterion (2.1) disregards an important fact: when we plot J density level sets in a graph we are seeking to represent a bivariate probability density function f (or the induced probability measure μ_f). Therefore, it is desirable that the graph is as close as possible (in some sense) to the target density f. A natural way to measure closeness between a graph of density level sets and a density function is to regard such a graph as itself defining a bivariate density. Then distance measures between bivariate densities (or bivariate distributions) can be used.

Let us assume that the support of f is a known compact set C_1. To simplify the exposition, we consider for the moment that we are looking for only one level set ($J = 1$) with probability content $\alpha \in [0, 1]$. We wish to associate a bivariate density to the level set C_α and we are choosing it from among the family $\mathcal{G}_{f,\alpha}$ of

Table 1: Truncated standard bivariate normal. Values of α^*, the solution to Problem (2.2), for different choices of distances between densities.

D:	L_1 norm	L_2 norm	Hellinger	L_2 norm of logs	Kullback-Leibler	Symmetric Kullback-Leibler
α^*:	0.75	0.66	0.83	0.95	0.83	0.84

bivariate density functions having C_α as the level set with probability content α:

$$\mathcal{G}_{f,\alpha} = \{g \text{ density with support } C_1 : \mu_g(C_\alpha) = \alpha,$$
$$g(x) \geq g(y) \text{ for all } x \in C_\alpha,\ y \in C_1 \setminus C_\alpha\}.$$

In this family, the maximum entropy distribution is that having the piecewise uniform distribution at C_α and $C_1 \setminus C_\alpha$ as density function (see, for instance, Bernardo and Smith 1994, pages 208-209):

$$g_{f,\alpha}(x) = \frac{\alpha}{\lambda(C_\alpha)} I_{C_\alpha}(x) + \frac{1-\alpha}{\lambda(C_1) - \lambda(C_\alpha)} I_{C_1 \setminus C_\alpha}(x).$$

Observe that $g_{f,0} = g_{f,1}$ are both equal to the density of the uniform distribution over C_1. Given that $g_{f,\alpha}(x)$ is, in some sense, the least informative density in $\mathcal{G}_{f,\alpha}$, this is the density function that we associate to level set C_α.

Let D be a distance function between bivariate density functions. In order to choose an optimal value of α we propose solving the following minimization problem:

$$\min_{0 \leq \alpha \leq 1} D(f, g_{f,\alpha}). \tag{2.2}$$

Observe that our goal is to represent the density f by the level set that defines the piecewise uniform distribution that is as close as possible to f. There are many ways to define a distance between two bivariate density functions (see Delicado 2011b for a commented bibliography on this topic).

Let us give a numerical example of the resolution of problem (2.2). Consider the probability density function of a truncated standard bivariate normal random variable, truncated at the square $[-3.035, 3.035] \times [-3.035, 3.035]$, having probability 0.99 under the standard bivariate normal distribution. Problem (2.2) has been solved for several choices of distance D in order to obtain the optimal value α^*. The results are shown in Table 1. Observe that optimal αs are always over 0.66, the value obtained using L_2 norm. For L_1 distance the optimal value α^* is 0.75, the value used by Bowman and Azzalini (1997) to produce Figure 1 (top panel). The other distances produce larger values of α^*. This indicates that these distances are especially sensitive to discrepancies in low density areas, an unsurprising fact given that three of them involve the logs of density values.

The problem (2.2) is a population problem (it assumes that density f is known). Let us consider its sampling version, obtained by replacing f by an estimation. Let $\{f_n\}_n$ be a sequence of random density functions approximating f (the most common case being that $f_n = \hat{f}_n$ is a nonparametric estimation of f derived from a size n sample from a random variable with density f). Let α^* be the solution to (2.2) and let $\hat{\alpha}_n$ be the solution to the following minimization problem:

$$\min_{0<\alpha<1} D(f_n, g_{f_n,\alpha}). \tag{2.3}$$

The next theorem establishes the convergence of $\hat{\alpha}_n$ to α^*. Technical assumptions as well as the proof are reported in the Appendix.

Theorem 2. *Let f be a bivariate density function with compact support C_1 and let f_n be a sequence of random bivariate density functions with support C_1. Let D be a distance between density functions for which assumptions Ass.1, Ass.2, Ass.3 and Ass.4 are verified. Let α^* be the solution to problem (2.2) and let $\{\hat{\alpha}_n\}_n$ be a sequence of solutions for problem (2.3). Then*

$$\lim_{n\to\infty} \hat{\alpha}_n = \alpha^* \text{ almost surely.}$$

Compactness assumption for the support of f in the previous theorem is required to define the piecewise uniform density functions $g_{f,\alpha}$. For the case of non-compact support we recommend to work with a version of f truncated to the level set $C^f_{1-\varepsilon}$, for ε small enough, namely $f_{1-\varepsilon}$, that does have compact support so Theorem 2 applies to it. The relationship between the level sets of f and $f_{1-\varepsilon}$ is that $C^f_\alpha = C^{f_{1-\varepsilon}}_{\alpha/(1-\varepsilon)}$, for $\alpha < 1-\varepsilon$, because when $X \sim f$

$$\Pr(X \in C^f_\alpha | X \in C^f_{1-\varepsilon}) = \frac{\alpha}{1-\varepsilon}.$$

Theorem 2 can be extended to the case of choosing $J \geq 1$ level sets. Let $\Theta = \{\theta = (\alpha_1, \ldots, \alpha_J) \in \mathbf{R}^J : 0 \leq \alpha_1 \leq \cdots \leq \alpha_J \leq 1\}$. For $\theta \in \Theta$ define

$$g_{f,\theta} = \sum_{j=0}^{J} \frac{\alpha_{j+1} - \alpha_j}{\lambda(C_{\alpha_{j+1}}) - \lambda(C_{\alpha_j})} I_{C_{\alpha_{j+1}} \setminus C_{\alpha_j}}(x),$$

where $\alpha_0 = 0$, $\alpha_{J+1} = 1$ and $C_{\alpha_0} = \emptyset$. Observe that the functions $g_{f,\theta}$ is the density function of a piecewise uniform distribution. Now the analogue to problem (2.2) is

$$\min_{\theta\in\Theta} D(f, g_{f,\theta}) \tag{2.4}$$

with optimum θ^*, and the version of (2.3) is

$$\min_{\theta\in\Theta} D(f_n, g_{f_n,\theta}) \tag{2.5}$$

with solution $\hat{\theta}_n$. In this context it can be proved that $\lim_n \hat{\theta}_n = \theta^*$ (almost surely) following the same arguments used in the proof of Theorem 2.

Table 2: Truncated standard bivariate normal. Values of α_j^*, $j = 1, \ldots, J$, for $J = 1, \ldots, 4$, the solutions to Problem (2.4), for different choices of distances between densities.

J	L_1 norm $\alpha_j^*, j = 1, \ldots, J$	Symmetric Kullback-Leibler $\alpha_j^*, j = 1, \ldots, J$	L_2 norm of logs $\alpha_j^*, j = 1, \ldots, J$
1	0.75	0.84	0.95
2	0.55, 0.88	0.67, 0.94	0.77, 0.95
3	0.43, 0.73, 0.93	0.51, 0.80, 0.95	0.63, 0.86, 0.95
4	0.32, 0.61, 0.82, 0.95	0.25, 0.61, 0.83, 0.95	0.53, 0.78, 0.89, 0.95

Let us now give an example of solution to problem (2.4). Consider again the density function of the truncated standard bivariate normal. We solved the problem (2.4) for the same distances used before when solving problem (2.2) and J going from 1 to 4. In general we can say that for J=1,2,3, the optimal α values are ordered as follows, according to the distance used,

$$L_2 < L_1 < \text{Hellinger} \approx \text{K-L} \approx \text{Sym. K-L} < L_2\text{logs}.$$

For $J = 4$, all the distances lead to similar optimal α values, except L_2 norm between logs, which gives larger values. In general, the optimal α values are in $[0.3, 0.95]$. The lowest value found for an optimal α is 0.20 (using Hellinger distance and $J = 4$). Table 2 shows a representative part of the results obtained.

Observe that the probability contents of the optimal level sets (the optimal solution to problems (2.2) or (2.4)) depend on the density f that we are representing. Therefore, if two or more densities must be represented in the same graph it would not to be a good idea to choose the optimal αs by solving problems (2.2) or (2.4) separately for each density. It would be much more sensible to look for a common α value (or a common $\theta = (\alpha_1, \ldots, \alpha_J)$) and use it to represent all the densities available. The next subsection shows how the proposals put forward in this section for one density can be adapted when we have two or more densities. Nevertheless, this adaptation has some limitations that are overcome in Section 3.

2.3. Difficulties with managing several densities

Consider now the case of N density functions, $f^1, \ldots, f^N$, each to be represented by J level sets, having common probability contents $\alpha_1 \leq \cdots \leq \alpha_J$. For $i = 1, \ldots, N$, let $\{f_n^i\}_n$ be a sequence of density functions approaching f^i. Then we consider the minimization problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} D(f_n^i, g_{f_n^i, \theta}), \tag{2.6}$$

Table 3: Aircraft data. Kernel density estimation for three periods. Values of α_n^N (with $N = 3$), solutions to Problem (2.6), for different choices of distances between densities.

D:	L_1 norm	L_2 norm	Hellinger	L_2 norm of logs	Kullback-Leibler	Symmetric Kullback-Leibler
$\hat{\alpha}_n^N$:	0.81	0.72	0.90	0.95	0.88	0.95

with optimum solution $\hat{\theta}_n^N$. An appropriate modification of Theorem 2 would guarantee that $\hat{\theta}_n^N$ converges (almost surely) to the solution θ_N^* of

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} D(f^i, g_{f^i,\theta}). \tag{2.7}$$

As an application of problem (2.6), consider again the example of the Aircraft data classified into three periods. We wish to represent $N = 3$ densities (one for each period) with one level set ($J = 1$). Table 3 shows the optimal α_n^N for different choices of distances between densities. Observe that the results are similar to those obtained for the bivariate truncated normal (see Table 1). The lower right panel of Figure 1 shows the level sets with probability content 0.9 (optimal value according to Hellinger distance) for the three densities.

This approach for when N densities must be represented simultaneously has the following characteristic: the common $\theta = (\alpha_1, \ldots, \alpha_J)$ that we are seeking tries only to provide a good individual representation of the densities involved: it does not attempt to highlight the differences between them. An extreme case arises when N different bivariate density functions share the same density level set with probability content α_1 but they differ from each other in the level set corresponding to probability content α_2 (for an example of this situation, see Case 1 of simulated density functions in Section 3.4). Assume additionally that the value of α_1 is the optimal one for each individual representation (that is, α_1 solves the N individual optimization problems (2.2)). Then α_1 is also the solution of the optimization problem (2.7) for $J = 1$. Nevertheless the representation of the N density functions by their α_1 level set only produces the superposition of N identical level sets, leading to the false conclusion that the N density functions are similar. A better global representation is obtained using, for instance, the N level sets with probability contents α_2, because these sets are different as the density functions are (see Figure 3 for such an example).

We address this problem in the following Section 3, where we propose an alternative approach.

3. Optimal representation of several densities by level sets

Let us consider N bivariate density functions, $f^1, \ldots, f^N$; it could be the case that they form a random sample corresponding to observations of a common

stochastic process. Given D, a distance function between bivariate density functions, we define the distance matrix

$$\mathbf{D} = \left(d_{ij} = D(f^i, f^j)\right)_{i,j=1,\ldots,N}$$

reflecting the inter-distances between any pair of density functions f^i and f^j in the previous list.

First we deal with the problem of representing the density functions using only a level set ($J = 1$). Secondly we look into the case of using several level sets for representing each density ($J > 1$).

3.1. Representation with a single level set

Let $\alpha \in [0, 1]$ and let C^i_α be the level set of f_i with probability content α, for $i = 1, \ldots, n$. For a pair of density functions f_i and f_j we define

$$\delta^{\alpha,d}_{ij} = d(C^i_\alpha, C^j_\alpha)$$

where d is a distance function between sets (see Section 2.1). For the case of d being the distance in probability, that depends on the specific density function in use, we take

$$d(C^i_\alpha, C^j_\alpha) = d_{f_i}(C^i_\alpha, C^j_\alpha) + d_{f_j}(C^i_\alpha, C^j_\alpha). \tag{3.1}$$

Alternatively, we can use the density functions corresponding to piecewise uniform distributions, defined in Section 2.2, to define

$$\delta^{\alpha,D}_{ij} = D(g_{f^i,\alpha}, g_{f^j,\alpha}),$$

where D is a distance between bivariate density functions.

Let $\boldsymbol{\delta}^\alpha$ be the $N \times N$ distance matrix whose elements are δ^α_{ij} computed as $\delta^{\alpha,d}_{ij}$ or as $\delta^{\alpha,D}_{ij}$. Our objective is to represent the N density functions $f^1, \ldots, f^N$ by their level set having common probability content α, choosing α in an optimal way according to a specified criterion.

A first natural criterion for choosing α comes from borrowing ideas from Multidimensional Scaling (MDS; see, for instance, Borg and Groenen 2005). Given a $N \times N$ distance matrix, the objective of MDS is to find a low dimensional configuration X (that is, a $N \times p$ matrix, with p small) such that the Euclidean distance between rows i and j in X is *as similar as possible* to the element (i, j) in the starting distance matrix. Then the i-th object (associated with the i-th row and column of the distance matrix) is represented in $\mathbf{R}^p$ with coordinates given by the i-th row of X. We propose accordingly to choose α in such a way that the distance matrix $\boldsymbol{\delta}^\alpha$ between the level set representations is *as similar as possible* to the distance matrix $\mathbf{D}$ between bivariate density functions. Essentially, our proposal is in the same line as the Generalized MDS introduced by Bronstein et al. (2006), where the objective is to find configurations in two different metric spaces with similar distance matrices.

One way to specify the meaning of *similar* in MDS is through the *normalized Stress*, a measure of the relative error made when the distance matrix $\mathbf{D}$ is approximated by $\boldsymbol{\delta}^\alpha$, defined as

$$\sigma_n(\boldsymbol{\delta}^\alpha, \mathbf{D}) = \frac{\sum_{i<j}(\delta_{ij}^\alpha - D_{ij})^2}{\sum_{i<j} D_{ij}^2}.$$

Allowing for scale changes, a more convenient measure is $\min_{b>0} \sigma_n(b\boldsymbol{\delta}^\alpha, \mathbf{D})$. It can be proved (Borg and Groenen 2005) that

$$\min_{b\in\mathbf{R}} \sigma_n(b\boldsymbol{\delta}^\alpha, \mathbf{D}) = 1 - c(\boldsymbol{\delta}^\alpha, \mathbf{D})^2$$

where

$$c(\boldsymbol{\delta}^\alpha, \mathbf{D}) = \frac{\sum_{i<j} \delta_{ij}^\alpha D_{ij}}{\left(\sum_{i<j}(\delta_{ij}^\alpha)^2 \sum_{i<j} D_{ij}^2\right)^{1/2}}, \tag{3.2}$$

is known as *Tucker's coefficient of congruence* between the elements of both distance matrices (see, for instance, Borg and Groenen 2005, page 248). This implies that $c(\boldsymbol{\delta}^\alpha, \mathbf{D})^2$, the square of the Tucker's coefficient of congruence, is the coefficient of determination R^2 of the least squares linear regression passing by the origin of the elements of $\mathbf{D}$ against the elements of $\boldsymbol{\delta}^\alpha$ (Borg and Groenen 2005).

Therefore, a first approach to choose an optimal α, according to the normalized Strees (with possibly a change of scale) is to solve one of the two following equivalent optimization problems:

$$\min_{\alpha\in[0,1]} \min_{b\in\mathbf{R}} \sigma_n(b\boldsymbol{\delta}^\alpha, \mathbf{D}) \Longleftrightarrow \max_{\alpha\in[0,1]} c(\boldsymbol{\delta}^\alpha, \mathbf{D})^2. \tag{3.3}$$

Nevertheless this approach suffers from a practical disadvantage: The range of values $c(\boldsymbol{\delta}^\alpha, \mathbf{D})$ that are observed in practice is narrower than the theoretical one ($[0, 1]$, for non-negative quantities). As an example, consider m pairs of data (u_i, v_i) coming from a random variable (U, V), U and V being independent and uniformly distributed on $[0, 1]$. For large values of m the Tucker's coefficient of congruence for these data will be close to $E(UV)/E(U^2) = 3/4$. Therefore the relevant range of the coefficient of congruence is $[0.75, 1]$ when the data have uniform marginals.

Alternatives to the Tucker's coefficient of congruence with better practical performance are the Pearson's correlation coefficient or the Spearman's rank correlation coefficient (the last one being able to detect positive relations between pairs of variables, even if they are non-linear). Given the $m = N(N-1)/2$ elements δ_{ij}^α, $i < j$, we define r_{ij}^α as the rank of δ_{ij}^α among the m elements. Analogously, we define R_{ij} as the rank of D_{ij} among the m elements over the diagonal of matrix $\mathbf{D}$. The Spearman's rank correlation coefficient of δ_{ij}^α and D_{ij}, with $i > j$, is defined as the Pearson's correlation coefficient between r_{ij}^α and R_{ij}, with $i < j$, and it can be computed as (Gibbons and Chakraborti 2003, page 423)

$$S(\boldsymbol{\delta}^\alpha, \mathbf{D}) = \frac{12\sum_{i<j} r_{ij}^\alpha R_{ij}}{m(m^2-1)} - \frac{3(m+1)}{m-1}.$$

It is easy to see that the Tucker's coefficient of congruence between r_{ij}^{α} and R_{ij}, with $i < j$, is also a monotone increasing function of $\sum_{i<j} r_{ij}^{\alpha} R_{ij}$. Therefore to look for the value of α maximizing the Spearman's rank correlation coefficient is equivalent to look for the value of α maximizing the Tucker's coefficient of congruence between the ranks of distances (nevertheless, a Spearman's rank correlation coefficient equal to 0 corresponds to a Tucker's coefficient equal to $(3/4)(1 + 1/(2m + 1))$, close to 0.75 for large m).

Taking into account the previous considerations, our proposal is to look for the value of α that solves the following optimization problem:

$$\max_{\alpha \in [0,1]} S(\boldsymbol{\delta}^{\alpha}, \mathbf{D})^2. \tag{3.4}$$

In practice, this is equivalent to maximize $S(\boldsymbol{\delta}^{\alpha}, \mathbf{D})$ (this is the case, for instance, when all the densities have the same support; then for $\alpha = 1$ all the level sets are equal and $D_{ij} = 0$ for all i, j; then $S(\boldsymbol{\delta}^{1}, \mathbf{D}) = 0$). The advantage of squaring in (3.4) is that this way the objective function can be computed as the coefficient of determination R^2 of the simple linear regression of R_{ij}, $i > j$, against r_{ij}^{α}, $i > j$.

3.2. Representation with more than one level sets

Now we consider the case of using several level sets for representing each density f_i, $i = 1, \ldots, N$. Let $0 \leq \alpha_1 \leq \cdots \leq \alpha_J \leq 1$ be J probability contents, for a given J. We call $\theta = (\alpha_1, \ldots, \alpha_J)$, as in Section 2.2. For a given bivariate density function f we define

$$\mathbf{C}_{\theta}^{f} = (C_{\alpha_1}^{f}, \ldots, C_{\alpha_J}^{f}),$$

the family of level sets of f with probability contents given by the elements of θ. For $i = 1, \ldots, N$, we denote $\mathbf{C}_{\theta}^{f^i}$ by $\mathbf{C}_{\theta}^{i}$.

Given two families of level sets, $\mathbf{C}_{\theta}^{i}$ and $\mathbf{C}_{\theta}^{j}$, there are several ways to define a distance between them. For instance, for a given distance function d between sets (see Section 2.1), we can define

$$\delta_{ij}^{\theta,d} = \delta^{d}(\mathbf{C}_{\theta}^{i}, \mathbf{C}_{\theta}^{j}) = \sum_{\alpha_h \in \theta} \delta_{ij}^{\alpha_h},$$

with $\delta_{ij}^{\alpha_h} = d(C_{\alpha_h}^{f^i}, C_{\alpha_h}^{f^j})$.

Another alternative is to use the density functions corresponding to piecewise uniform distributions, defined in Section 2.2, to define

$$\delta_{ij}^{\theta,D} = \delta^{D}(\mathbf{C}_{\theta}^{i}, \mathbf{C}_{\theta}^{j}) = D(g_{f^i,\theta}, g_{f^i,\theta}),$$

where $g_{f,\theta}$ has been defined after Theorem 2 and D is a distance between bivariate density functions. Let be $\boldsymbol{\delta}^{\theta}$ the $N \times N$ distance matrix whose elements are δ_{ij}^{θ} computed as $\delta_{ij}^{\theta,d}$ or as $\delta_{ij}^{\theta,D}$. Following the previous reasoning that led us to equation (3.4), we propose to solve the following problem:

$$\max_{\theta \in \Theta} S(\boldsymbol{\delta}^{\theta}, \mathbf{D})^2. \tag{3.5}$$

Nevertheless we set out the following alternative approach to choose the optimal θ. Recall that $S(\boldsymbol{\delta}^\alpha, \mathbf{D})^2$ is the coefficient of determination R^2 of a simple linear regression involving ranks of distances. Therefore, instead of pooling the distances $\delta_{ij}^{\alpha_h}$, $\alpha_h \in \theta$, to define $\delta_{ij}^{\theta,d}$, we propose to fit a multiple linear regression involving ranks of distances, and to take the coefficient of determination as the objective function to be maximized. To be specific, consider the multiple linear regression where the response is the set of ranks R_{ij}, $i < j$, defined before, with J explanatory variables, the ranks $r_{ij}^{\alpha_h}$ of the J distances $\delta_{ij}^{\alpha_h}$, $\alpha_h \in \theta$. Let $R_S^2(\theta)$ the corresponding coefficient of determination. Our second proposal is to solve the following optimization problem:

$$\max_{\theta \in \Theta} R_S^2(\theta). \tag{3.6}$$

3.3. Case of estimated densities

Let us add a few lines about the common case of not knowing exactly the densities f_i, $i = 1, \ldots, N$. Instead we assume that for $i = 1, \ldots, N$, there is a sequence $\{f_n^i\}_n$ of density functions approaching f_i as n goes to infinity. As in Section 2 the most frequent situation is that $f_n^i = \hat{f}_{i,n}$ is a nonparametric estimation of f_i derived from a size $m_{i,n}$ sample from a random variable with density f_i, with $\liminf_n \min_i m_{i,n}/n > 0$ and $\limsup_n \max_i m_{i,n}/n < \infty$.

Consider the following version of problem (3.4), where $\boldsymbol{\delta}^\alpha$ and $\mathbf{D}$, defined from f_i, $i = 1, \ldots, N$, has been substituted by their counterparts, say $\boldsymbol{\delta}_n^\alpha$ and $\mathbf{D}_n$, defined from f_n^i, $i = 1, \ldots, N$:

$$\max_{\alpha \in [0,1]} S(\boldsymbol{\delta}_n^\alpha, \mathbf{D}_n)^2. \tag{3.7}$$

The following Theorem tells us that solving problem (3.7) is asymptotically equivalent to solving problem (3.4). The main idea is that the number of densities N is fixed, even if n goes to infinity. The proof is reported in the Appendix. Extensions of this result for covering problems (3.5) and (3.6) are straightforward.

Theorem 3. *Let D and d be distances between density functions and level sets, respectively, for which assumptions Ass.1 and Ass.5 are verified for any density function f_i and sequence $\{f_n^i\}_n$, $i = 1, \ldots, N$. Assume that there are no ties in distances: for $i < j$ and $k < l$, with $(i, j) \neq (k, l)$, we have that $D(f_i, f_j) \neq D(f_k, f_l)$ and that*

$$\inf_{0<\alpha<1} |d(C_\alpha^{f_i}, C_\alpha^{f_j}) - d(C_\alpha^{f_k}, C_\alpha^{f_l})| > 0.$$

Let α^* be the solution to problem (3.4) and let $\hat{\alpha}_n$ be the solution to problem (3.7). Then

$$\lim_{n \to \infty} \hat{\alpha}_n = \alpha^* \text{ almost surely.}$$

3.4. Some Monte Carlo experiments

The examples we present consist of sets of N bivariate density functions f_i, $i = 1, \ldots, N = 50$, such that f_i is a mixture of three bivariate normal densities, truncated at the square $[-3.035, 3.035] \times [-3.035, 3.035]$. The generic expression for these densities (before truncation) is

$$f(x, y) = \sum_{j=0}^{2} \eta_j \phi_2(x, y; \mu_{1,j}, \mu_{2,j}, \sigma_j^2 I_2), \tag{3.8}$$

where I_2 is the identity matrix of size 2 and we denote by $\phi_2(x, y; \mu_1, \mu_2, \Sigma)$ the density function of a bivariate normal centered at (μ_1, μ_2) with variance matrix Σ, evaluated at $(x, y) \in \mathbf{R}^2$. The mean vectors are defined as $(\mu_{1,0}, \mu_{2,0}) = (0, 0)$ and for $j = 1, 2$,

$$(\mu_{1,j}, \mu_{2,j}) = \rho_j(\cos\theta_j, \sin\theta_j).$$

The way we generate random densities according to (3.8) is by taking independent random values of θ_j and ρ_j, for $j = 1, 2$. Specifically, $\theta_j \sim U(0, 2\pi)$, $\rho_j \sim U(r_j - .1, r_j + .1)$ (r_j is a fixed value in $\{0, 1, 2\}$). We have considered 3 different cases (or models) to generate random densities according to (3.8), corresponding to specific choices of r_j, σ_j and η_j, for $j = 1, 2$ ($\eta_0 = 1 - \eta_1 - \eta_2$ and we use always $\sigma_0 = 1$). The left hand side column of Table 4 shows the parameters used to generate the different cases of mixture densities. The left column of panels in Figure 2 shows an example of f_i for each considered case. The level sets used for the representation of f_i are those with probability contents α equal to 0.05, 0.1 (this level set has two non-connected subsets for Cases 1 and 2), 0.25, 0.5, 0.75 and 0.95.

Let us remark several characteristics of densities $f(x, y)$ defined by (3.8). We start examining those corresponding to Case 1 (upper left panel in Figure 2 is an example). Densities corresponding to this case are very close to a mixture of two bivariate normal densities, those corresponding to $j = 0$ and $j = 2$ in (3.8), while the component corresponding to $j = 1$ acts as a slight random perturbation. They are bimodal indeed. Given two of these densities, their difference depends mainly on how different their parameters θ_2 are. Their level sets with probability content $\alpha \geq 0.75$ are almost equal, while those corresponding to small values of α (say $\alpha \leq 0.25$) present large differences.

Densities corresponding to Case 2 (middle left panel in Figure 2) are clearly the mixture of 3 bivariate normal densities, with centers at distances 0, 2 and 1 from the origin, respectively. They are bimodal because the mixture component in (3.8) corresponding to $j = 1$ does not produce a third mode. The differences in parameters θ_1 and θ_2 are the responsible of the distances between densities corresponding to Case 2. Two generic densities following this model have all their level sets different. Differences between levels sets with small probability contents (say $\alpha \leq 0.5$) reflect differences in parameter θ_2, while dissimilarity between levels sets with larger probability contents respond to differences in parameter θ_1.

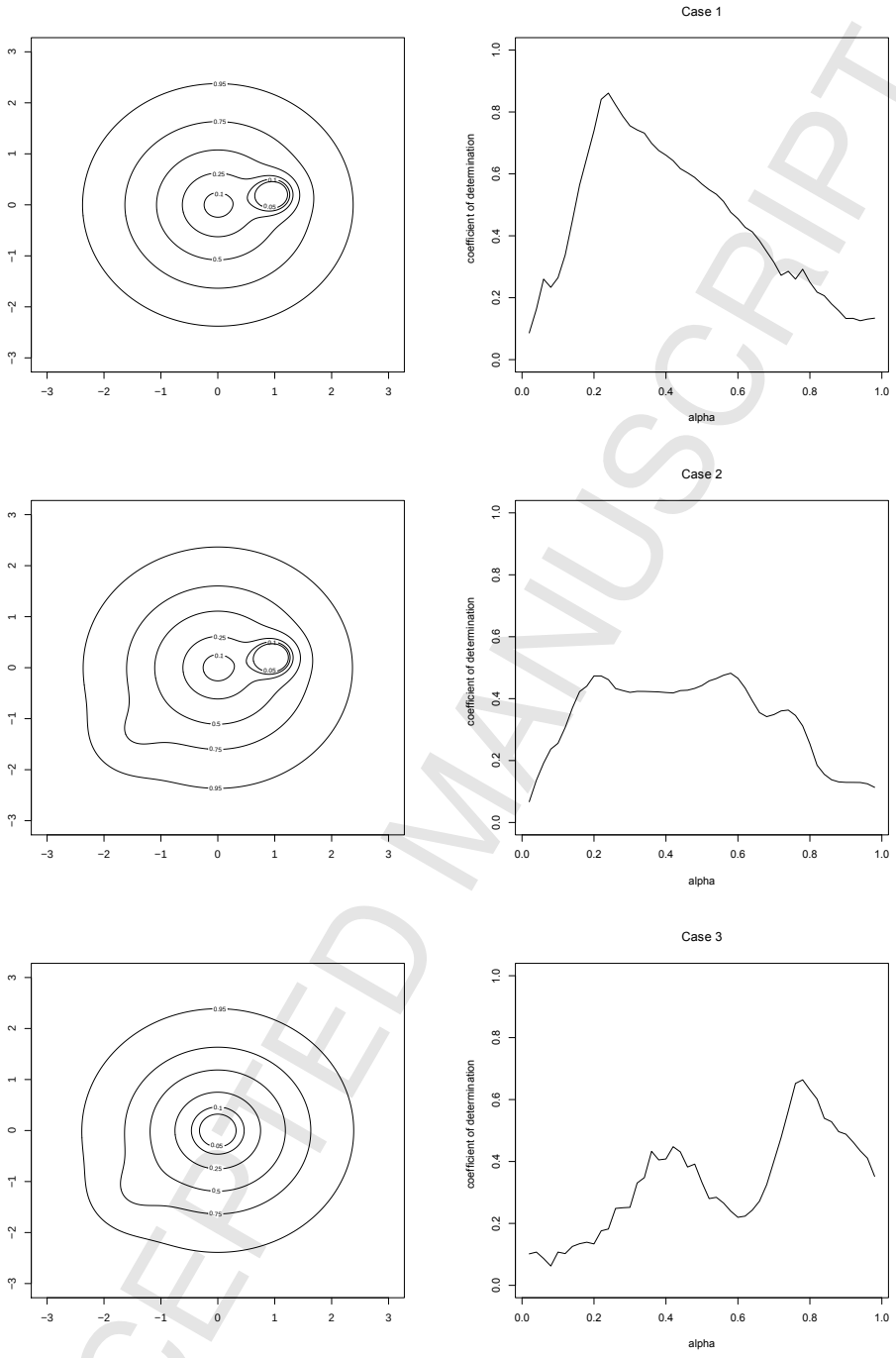


Figure 2: Mixture of three bivariate normal densities according to model (3.8). First, second and third rows correspond respectively to Cases 1, 2 and 3 described in Table 4. *Left column:* For each case, level sets corresponding to one of the $N = 50$ simulated densities. *Right column:* For each case, graphic of $R^2_S(\alpha)$ as a function of α.

Regarding Case 3 (lower left panel in Figure 2), it is similar to Case 2 when removing the main mode, one unit length far from origin. Densities following Case 3 are unimodal and they differ mainly in level sets with large probability content (say $\alpha \geq 0.7$). Parameter θ_1 is the main responsible of differences between densities.

For each of the three sets of density functions following Cases 1, 2 and 3, respectively, using L_1 distance as the distance D between densities and distance in probability (3.1) as the distance d between level sets, we are interested in the coefficients of determination $R^2_S(\theta)$, where $\theta = (\alpha_1, \ldots, \alpha_J)$, and $J \in \{1, 2, 3, 4\}$ is the number of level sets we are looking for. For $J = 1$, the graphics at the right column of Figure 2 show the values of $R^2_S(\alpha)$ for α going from 0.02 to 0.98, with increments of 0.02.

For Case 1 (upper right panel in Figure 2) the maximum of $R^2_S(\alpha)$ is achieved at $\alpha^* = 0.24$, with a value of $R^2_S(\alpha^*) = 0.8609$. For Case 2 (middle right panel in Figure 2) the function $R^2_S(\alpha)$ has its maximum at $\alpha^* = 0.58$ and it is equal to 0.4822 (much lower than in Case 1). There is an additional local maximum at $\alpha = 0.2$, with almost the same value as the global maximum. A third local maximum is at $\alpha = 0.74$. Regarding Case 3 (lower right panel in Figure 2) the maximum (0.6638) is at $\alpha^* = 0.78$. There is another local maximum around $\alpha = 0.4$.

We have solved the optimization problem (3.6) for J, the dimension of θ, in $\{1, 2, 3, 4\}$. We have used a quasi-Newton method which allows box constraints for the optimization variables. Specifically we have used the implementation of the algorithm proposed by Byrd et al. (1995) provided by the function `optim` of R (R Core Team 2013) when the method `‘‘L-BFGS-B’’` is selected.

The results of the optimization procedure are shown at the right hand side of Table 4. For each case and each J, the optimal value $\theta^* = (\alpha^*_1, \ldots, \alpha^*_J)$ and the corresponding value of the objective function $R^2_S(\theta^*)$ are printed. We can see that for Case 1 the optimum values do not vary very much when J goes from 1 to 4. Then we conclude that for this case it is enough to use only one level set to have a good representation of the density functions. On the contrary, Cases 2 and 3 seems to require two level sets for doing the task, and Case 2 may need even a third level set. In general, the values of α^*_j are close to the local maximums of the functions $R^2_S(\alpha)$ (Figure 2, right column) but there are exceptions to this rule.

We conclude this section by comparing the results presented above with those obtained when using the proposals made in Section 2. When we consider a single density function according to the model (3.8), and on applying the techniques described in Section 2 for the choice of the optimal level sets to represent this function, the results we obtain are very similar to those obtained for the truncated standard bivariate normal (see Tables 1 and 2), the mixture density being (at least) 90% equal to the density of the standard bivariate normal. In particular, the methods proposed in Section 2 do not give rise to the representation of any level set with probability content $\alpha < 0.3$. Similar conclusions are drawn when solving problem (2.5) jointly for the N densities in

Table 4: Mixture of three bivariate normal densities according to model (3.8). The left hand side of the table shows the parameters r_j, σ_j and η_j, for $j = 1, 2$ ($\eta_0 = 1 - \eta_1 - \eta_2$) used to generate 3 different types of mixtures densities. The optimum $\theta^* = (\alpha_1^*, \ldots, \alpha_J^*)$ when solving the optimization problem (3.6) and the value of the objective function for it, are shown on the right-hand side of the table.

Case	Parameters		J	α_1^*	α_2^*	α_3^*	α_4^*	$R_S^2(\theta)$
1	$r_1 = 0$	$r_2 = 1$	1	0.244				0.8753
	$\sigma_1 = 1$	$\sigma_2 = 0.25$	2	0.209	0.340			0.8960
	$\eta_1 = 0.05$	$\eta_2 = 0.05$	3	0.197	0.290	0.431		0.9017
			4	0.212	0.404	0.595	0.787	0.9254
2	$r_1 = 2$	$r_2 = 1$	1	0.571				0.4840
	$\sigma_1 = 0.4$	$\sigma_2 = 0.25$	2	0.228	0.763			0.7572
	$\eta_1 = 0.025$	$\eta_2 = 0.025$	3	0.199	0.622	0.800		0.8326
			4	0.191	0.407	0.614	0.845	0.8446
3	$r_1 = 2$	$r_2 = 0$	1	0.780				0.6638
	$\sigma_1 = 0.4$	$\sigma_2 = 1$	2	0.409	0.816			0.8312
	$\eta_1 = 0.025$	$\eta_2 = 0.05$	3	0.230	0.653	0.869		0.8366
			4	0.184	0.831	0.837	0.860	0.8600

the samples.

As an example, we consider Case 1 of densities according to (3.8) and replicate the analysis we have reported in Table 4 for $J = 1$. In this case, we use the value for α that were found to be optimal when solving problem (2.7) for these densities and L_1 distance, which coincides (up to the second decimal digit) with that reported for the truncated bivariate normal (Table 2): 0.75. We compute then $R_S^2(\alpha)$ for $\alpha = 0.75$ and we obtain a value of the coefficient of determination in the regression involving ranks of distances equal to 0.0021. Comparing this very low value with the value 0.8753 in Table 4 (last column, first row) confirms that solving problem (3.6) provides values of α_j leading to grater coefficient of determination than when using proposals introduced in Section 2, which were not designed for this purpose (see subsection 2.3).

In addition to the expected high coefficients of determination, the proposals put forward in this Section 3 also possess a remarkable descriptive power when representing a large number N of densities in a single graph, as may be observed in panels (a) and (b) of Figure 3. They represent $N = 50$ densities according to model (3.8), Case 1. One level set is used for each density ($J = 1$). Panel (a) uses the value $\alpha = .75$ obtained when solving problem (2.7), whereas in panel (b) α is 0.244, the solution of problem (3.6). Panel (b) clearly shows that these densities have a common mode around the origin and an additional mode at distance approximately 1 from the origin, very different from density

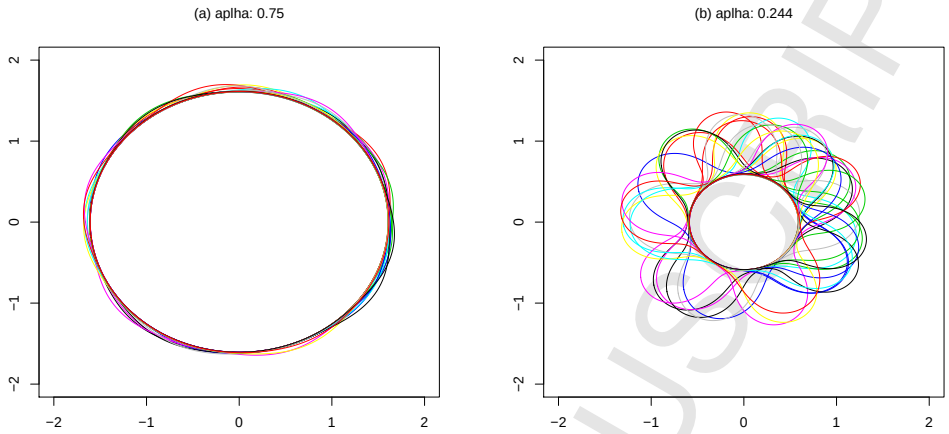


Figure 3: (a) Representation of $N = 50$ densities according to (3.8), Case 1, using the value $\alpha = .75$ obtained when solving problem (2.5). (b) The same as (a) when using $\alpha = 0.244$, the solution of problem (3.6). (In the interest of a better visualization both graphics use the frame $[-2, 2] \times [-2, 2]$ instead of $[-3, 3] \times [-3, 3]$, as in Figure 2.)

to density. Little can be learned from panel (a) about the differences between these densities.

An analogous simulation study has been done solving problem (3.5) instead of problem (3.6). We have used $\delta^{\theta}_{i,j} = \delta^{\theta,D}_{i,j}$ and D equal to the L_1 distance between densities. The results (not reported here) are qualitatively similar to those presented in this Section. In particular the graphics of the squared rank correlations $S(\boldsymbol{\delta}^{\alpha}, \mathbf{D})^2$ as functions of α for Cases 1, 2 and 3 are similar to those displayed in the right column of Figure 2.

4. Conclusions

Level sets are known to be nice graphical tools for visualizing density functions. Usually, levels are fixed arbitrarily (most usual choices being those with probability contents 0.25, 0.5 and 0.75), but changing the levels may lead to different conclusions from the same data. To the best of our knowledge, this paper is the first to respond to the natural questions: How can levels be chosen? Can such a choice be made in some data-driven way to take into account both the specificity of the available data and the kind of statistical problem one has to deal with?

We study separately two scenarios. In both cases we show that our proposals provide good theoretical properties and ease of implementation as well as a satisfactory practical finite sample performance. In particular, emphasis is given to the fact that our selected levels can detect information that standard level choices (like those with probability contents 0.25, 0.5 and 0.75) may hide.

In the first scenario we deal with the case when there is only one density function to be represented. Our proposals here are based on the minimum distance between sets or between density functions. Our main findings are the following. We have presented a quick method to choose the J optimal probability contents that does not depend on the specific density to be represented (for instance, for $J = 3$ they are 1/6, 1/2 and 5/6); in general, higher values for probability contents α are obtained when using methods that depend on the density to be represented, and they also depend on the specific distance in use (L_1 distance leading to not so high values of α).

In the second scenario we consider the case of many densities to be represented. Our approach here is related with generalized MDS. Several possibilities have been analyzed and, finally, our proposal consists on maximizing the coefficient of determination of a linear regression involving the ranks of distances between the densities to be represented and the ranks of the distances between their level sets. As practical advice, we recommend to solve the problem (3.6) using L_1 distance between density functions and distance in probability between level sets. Moreover using only one level set for each density function usually gives good results when representing several densities in the same graphic.

Appendix A. Theoretical issues

Theoretical issues in Section 2

We need the following Lemma to prove Theorem 1:

Lemma 4. *Let $\beta_0 = 0$, $\beta_J = 1$. Consider the problem*

$$\min_{0<\alpha_1<\beta_1\cdots<\beta_{J-1}<\alpha_J<1} \sum_{j=1}^{J} \int_{\beta_{j-1}}^{\beta_j} |u - \alpha_j| du.$$

The optimal solution is

$$\hat{\beta}_j = \frac{j}{J}, j = 1, \ldots, J-1, \hat{\alpha}_j = \frac{\hat{\beta}_{j-1} + \hat{\beta}_j}{2} = \frac{2j-1}{2J}, j = 1, \ldots, J.$$

Proof of the Lemma 4: For $\alpha_1, \ldots, \alpha_J$ fixed, the optimal values of β_j are

$$\beta_j = \frac{\alpha_j + \alpha_{j+1}}{2}, j = 1, \ldots, J-1.$$

For $\beta_1, \ldots, \beta_{J-1}$ fixed, the optimal values of α_j are

$$\alpha_j = \frac{\beta_{j-1} + \beta_j}{2}, j = 1, \ldots, J.$$

Then, using both equations we have for $j = 1, \ldots, J-1$ that

$$2\beta_j = \frac{\beta_{j-1} + \beta_j}{2} + \frac{\beta_j + \beta_{j+1}}{2},$$

then

$$\beta_j = \frac{\beta_{j-1} + \beta_{j+1}}{2}, j = 1, \ldots, J-1,$$

and then

$$\beta_j - \beta_{j-1} = c, j = 1, \ldots, J,$$

for some constant c. This, jointly with the boundary conditions $\beta_0 = 0$ and $\beta_J = 1$, leads to the solution: $\hat{\beta}_j = j/J, j = 1, \ldots, J-1$. The values for optimal α are easily derived and the proof finishes.

Proof of Theorem 1. We start by assuming that $d = d_f$. Taking into account that $C_u \subseteq C_v$ for all $0 < u < v < 1$ and that $d_f(A, B) = \mu_f(B \setminus A) = \mu_f(B) - \mu_f(A)$ when $A \subseteq B$, we have that for all $0 < u < v < 1$

$$d_f(C_u, C_v) = \mu_f(C_v) - \mu_f(C_u) = v - u = |u - v|.$$

Then the minimization problem (2.1) is equivalent to the following:

$$\min_{0<\alpha_1<\cdots<\alpha_J<1} \int_0^1 |u - \alpha_{j(u)}| du = \min_{0<\alpha_1<\cdots<\alpha_J<1} \left\{ \int_0^{(\alpha_1+\alpha_2)/2} |u - \alpha_1| du \right.$$

$$\left. + \sum_{j=1}^{J-1} \left(\int_{(\alpha_j+\alpha_{j-1})/2}^{(\alpha_j+\alpha_{j+1})/2} |u - \alpha_j| du \right) + \int_{(\alpha_{J-1}+\alpha_J)/2}^1 |u - \alpha_J| du \right\}.$$

This problem is equivalent to the k-median problem for the uniform distribution over $[0, 1]$ (in this case with $k = J$). The solution to this problem is easily found (see Lemma 4) to be

$$\alpha_j^* = \frac{2j-1}{2J}, \; j = 1, \ldots, J.$$

Let now $d = d_\lambda$. Observe that problem (2.1) in this case is

$$\min_{0<\alpha_1<\cdots<\alpha_J<1} \int_0^1 d_\lambda(C_u, C_{\alpha_{j(u)}}) du.$$

This problem only depends on f because the definition of the collection of level sets C_u, $u \in]0, 1]$ is based on f. Let us assume that there exists a density $g \neq f$, but sharing the collection of level sets with f: for each $u \in]0, 1]$ there exists a unique $v \in]0, 1]$ such that $C_u = C_v^g$, where C_v^g is the level set of g with probability content v. The relation between u and v is one-to-one. In this case, the solution to problem (2.1) when using Lebesgue measure is the same for both f and g.

Let us define a bivariate distribution sharing with f the collection of level sets. For each $\beta \in]0, 1]$ define $C_\beta^* = C_{\alpha(\beta)}$, where $\alpha(\beta)$ is the unique value such that

$$\lambda(C_{\alpha(\beta)}) = \beta \lambda(C_1).$$

Observe that $C_1^* = C_1$. Then the collection

$$\{C_\beta^* : \beta \in]0, 1]\}$$

identifies a probability measure μ_λ in $\mathbf{R}^2$ having these sets as density level sets and verifying

$$\mu_\lambda(C^*_\beta) = \beta = \frac{\lambda(C^*_\beta)}{\lambda(C^*_1)}, \quad \text{for all } \beta \in]0,1].$$

Therefore, for this measure μ_λ the solution to problem (2.1) is the same as if we use either distances between level sets: the Lebesgue measure or the measure in probability μ_λ. Applying then the first part of the theorem we obtain that the optimal values for β_j, $j = 1, \ldots, J$, are

$$\beta_j^{\mu_\lambda} = \beta_j^{\lambda} = \frac{2j-1}{2J}, \; j = 1, \ldots, J.$$

However, we have seen before that the solution to problem (2.1) when using the Lebesgue measure is the same as that for all distributions that share the same collection of density level sets. Therefore, the optimal solution α_j^λ, $j = 1, \ldots, J$, is

$$\alpha_j^\lambda = \alpha(\beta_j^{\mu_\lambda}) = \alpha\left(\frac{2j-1}{2J}\right), \; j = 1, \ldots, J,$$

and the proof concludes.

Commented assumptions for Theorem 2. Before dealing with the proof of Theorem 2, establishing the convergence of $\hat{\alpha}_n$ to α^* as n goes to infinity, let us introduce the following notation. Given a bivariate density ϕ with compact support C_1, we define the statistical parametric model

$$\mathcal{F}_\phi = \left\{ g_{\phi,\alpha}(x) = \frac{\alpha}{\lambda(C^\phi_\alpha)} I_{C^\phi_\alpha}(x) + \frac{1-\alpha}{\lambda(C_1) - \lambda(C^\phi_\alpha)} I_{C_1 \setminus C^\phi_\alpha}(x) : \alpha \in]0,1[\right\}$$
$$\cup \left\{ g_{\phi,0}(x) = g_{\phi,1}(x) = \frac{1}{\lambda(C_1)} I_{C_1}(x) \right\}.$$

Then we can rewrite α^* and $\hat{\alpha}_n$ as

$$\begin{aligned} \alpha^* &= \arg \min_{\alpha \in [0,1]} \{D(f, g_{f,\alpha}) : g_{f,\alpha} \in \mathcal{F}_f\}, \\ \hat{\alpha}_n &= \arg \min_{\alpha \in [0,1]} \{D(f_n, g_{f_n,\alpha}) : g_{f_n,\alpha} \in \mathcal{F}_{f_n}\}. \end{aligned}$$

Observe that α^* determines the closest distribution in the parametric model $\mathcal{F}_f$ to the true distribution f. On the other hand $\hat{\alpha}_n$ is not a minimum distance estimator because the parametric model used for each n, $\mathcal{F}_{f_n}$, changes with n. Let us define a real minimum distance estimator related with α^* and $\mathcal{F}_{f_n}$:

$$\tilde{\alpha}_n = \arg \min_{\alpha \in [0,1]} \{D(f_n, g_{f,\alpha}) : g_{f,\alpha} \in \mathcal{F}_f\}.$$

In the definition of $\tilde{\alpha}_n$, the parametric model is fixed for every n. Cao et al. (1995) study minimum distance parametric estimation when $f_n = \hat{f}_n$ are kernel

estimators of f, based on previous results from Parr and Schucany (1982). These two works use a slightly more general definition of the sequence of minimum distance estimator $\tilde{\alpha}_n$ as any sequence verifying

$$D(f_n, g_{f,\tilde{\alpha}_n}) \leq \inf_{\alpha\in[0,1]} D(f_n, g_{f,\alpha}) + \varepsilon_n, \tag{A.1}$$

where $\{\varepsilon_n\}$ is a sequence of real numbers tending to zero as n goes to infinity.

Let us state three assumptions (essentially the same assumed in Cao et al. 1995) that imply the convergence of minimum distance estimator $\tilde{\alpha}_n$ to the target parameter value α^*:

Ass.1 $D(f_n, f) \to 0$ almost surely as $n \to \infty$.

Ass.2 $D(\alpha) = D(f, g_{f,\alpha})$ has a unique minimum α^*.

Ass.3 For all α_0 in $[0,1]$ and $\{\alpha_r\} \subset [0,1]$ we have that

$$\lim_{r\to\infty} D(f, g_{f,\alpha_r}) = D(f, g_{f,\alpha_0}) \text{ implies } \lim_{r\to\infty} \alpha_r = \alpha_0.$$

Under assumptions Ass.1, Ass.2 and Ass.3, Theorem 1 of Cao et al. (1995) and Theorem 1 of Parr and Schucany (1982) apply, and it follows that

$$\lim_{n\to\infty} \tilde{\alpha}_n = \alpha^*$$

where $\{\tilde{\alpha}_n\}$ is any sequence verifying (A.1).

In order to prove that $\{\hat{\alpha}_n\}$ is converging to α^* as n goes to ∞, in addition to the previous assumptions concerning minimum distance estimation, we need assumptions about the behavior of the level sets of f_n as plug-in estimators of those of f, when distance d between sets and distance D between densities are considered:

Ass.4 Whenever Assumption Ass.1 is fulfilled, the following is also verified:

$$\lim_{n\to\infty} \sup_{\alpha\in[0,1]} D(g_{f_n,\alpha}, g_{f,\alpha}) = 0 \text{ almost surely.}$$

Ass.5 Whenever Assumption Ass.1 is fulfilled, the following is also verified:

$$\lim_{n\to\infty} \sup_{\alpha\in[0,1]} d(C_\alpha^{f_n}, C_\alpha^f) = 0 \text{ almost surely.}$$

The following proposition (with immediate proof) establishes a relationship between the last two assumptions.

Proposition 5. *If there exists a constant M such that for all densities f_1 and f_2*

$$D(g_{f_1,\alpha}, g_{f_2,\alpha}) \leq M d(C_\alpha^{f_1}, C_\alpha^{f_1}).$$

then Ass.5 is a sufficient condition for Ass.4.

Let us now give some remarks about these technical assumptions.

- In practice the most usual choice for approximating sequence $\{f_n\}_n$ is given by the kernel density estimator of f, based on n i.i.d. random variables $X_1, \ldots, X_n$ with common density f, defined as

$$\hat{f}_n(x;h) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),\ x \in \mathbf{R}^2,$$

  where the kernel K is a bivariate probability density, and $h = h(n) > 0$ is known as the *bandwidth* or *smoothing parameter* (for more details see, for instance, Silverman 1986, Scott 1992, Wand and Jones 1995, Simonoff 1996 or Bowman and Azzalini 1997).

  Several conditions on $h = h(n)$ are known to guarantee that $D(\hat{f}_n, f)$ converges almost surely to zero with n for certain choices of distance D. Then the assumption Ass.1 is satisfied. For instance, for the density estimation in $\mathbf{R}^d$ Devroye and Györfi (1985, Theorem 3.1) establish that almost surely consistency in L_1 distance is equivalent to $\lim_n h = 0$ and $\lim_n nh^d = \infty$, and Bertrand-Retali (1978) proves that almost surely consistency in L_∞ is equivalent to $\lim_n h = 0$ and $\lim_n (n/\log n)h^d = \infty$ for any uniformly continuous density f.

- Let the distance between densities D be the L_1 norm. Assume Ass.1 and consider $\alpha \in]0,1[$. Then,

$$\begin{aligned} D(g_{f_n,\alpha}, g_{f,\alpha}) = D\Bigg(&\frac{\alpha}{\lambda(C_\alpha^{f_n})} I_{C_\alpha^{f_n}}(x) + \frac{1-\alpha}{\lambda(C_1) - \lambda(C_\alpha^{f_n})} I_{C_1 \setminus C_\alpha^{f_n}}(x), \\ &\frac{\alpha}{\lambda(C_\alpha^{f})} I_{C_\alpha^{f}}(x) + \frac{1-\alpha}{\lambda(C_1) - \lambda(C_\alpha^{f})} I_{C_1 \setminus C_\alpha^{f}}(x)\Bigg) \\ &= \alpha \left| \frac{1}{\lambda(C_\alpha^{f_n})} - \frac{1}{\lambda(C_\alpha^{f})} \right| \lambda(C_\alpha^{f_n} \cap C_\alpha^{f}) \\ &+ \left| \frac{\alpha}{\lambda(C_\alpha^{f_n})} - \frac{1-\alpha}{\lambda(C_1) - \lambda(C_\alpha^{f})} \right| \lambda(C_\alpha^{f_n} \setminus \cap C_\alpha^{f}) \\ &+ \left| \frac{1-\alpha}{\lambda(C_1) - \lambda(C_\alpha^{f_n})} - \frac{\alpha}{\lambda(C_\alpha^{f})} \right| \lambda(C_\alpha^{f} \setminus \cap C_\alpha^{f_n}) \\ +(1-\alpha) &\left| \frac{1}{\lambda(C_1) - \lambda(C_\alpha^{f_n})} - \frac{1}{\lambda(C_1) - \lambda(C_\alpha^{f})} \right| \lambda(C_1 \setminus (C_\alpha^{f_n} \cup C_\alpha^{f})) = \\ &S_1 + S_2 + S_3 + S_4. \end{aligned}$$

Consider $\alpha \in]0,1[$ and assume that

$$\lim_{n\to\infty} \lambda(C_\alpha^{f_n} \Delta C_\alpha^f) = 0 \text{ almost surely.} \tag{A.2}$$

This implies that

$$\lambda(C_\alpha^{f_n}) \to \lambda(C_\alpha^{f_n}), \lambda(C_1 \setminus (C_\alpha^{f_n} \cup C_\alpha^f)) \to \lambda(C_1 \setminus \cup C_\alpha^f),$$

$$\lambda((C_\alpha^{f_n} \setminus C_\alpha^f)) \to 0, \lambda((C_\alpha^f \setminus C_\alpha^{f_n})) \to 0,$$

as $n \to \infty$. Then $S_i \to 0$, $i = 1, 2, 3, 4$. We conclude that

$$\lim_{n\to\infty} D(g_{f_n,\alpha}, g_{f,\alpha}) = 0 \text{ almost surely for all } \alpha \in]0,1[.$$

Consider now, as in the previous remark, that $f_n = \hat{f}_n$ are kernel density estimators of f based on n independent observations of $X \sim f$. In this context Baíllo (2003) proves that (A.2) occurs for $\alpha \in]0,1[$ under certain conditions on f (uniform continuity, no existence of flat parts, having level sets smooth enough, existence of some moment), on the kernel function (Lipschitz, compact support and having no flat parts) and on the bandwidth $h = h(n)$ (implying $\lim_n h = 0$ and $\lim_n (n/\log n)h^d = \infty$). Under these assumptions Baíllo (2003) also obtains convergence rates for $\lambda_f(C_\alpha^{\hat{f}_n} \Delta C_\alpha^f)$. See also Polonik (1995), Tsybakov (1997), Cadre (2006), Willett and Nowak (2007) and Mason and Polonik (2009) for alternative approaches.

Additional conditions must be assumed in order to have a uniform version of A.2, which would imply that for kernel density estimators assumptions Ass.4 and Ass.1 are verified (see again the previous remark). Nevertheless, in practice the supremum over $\alpha \in [0,1]$ can be replaced by the maximum over a finite fine grid $\{\alpha_1, \ldots, \alpha_K\} \subset]0,1[$. Then the work of Baíllo (2003) guarantees that, under the conditions cited above,

$$\lim_{n\to\infty} \max_{k=1\ldots K} \lambda(C_{\alpha_k}^{\hat{f}_n} \Delta C_{\alpha_k}^f) = 0 \text{ almost surely.}$$

Proof of Theorem 2. We will show that the sequence $\{\hat{\alpha}_n\}_n$ verifies the definition of the minimum distance estimator of α^* given in equation (A.1). Then Theorem 1 by Cao et al. (1995) and Theorem 1 by Parr and Schucany (1982) ends the proof.

Let $\{\tilde{\alpha}_n\}$ be a minimum distance estimator sequence of α^* verifying (A.1). Observe that

$$D(f_n, g_{f,\hat{\alpha}_n}) = D(f_n, g_{f,\tilde{\alpha}_n}) + (D(f_n, g_{f,\hat{\alpha}_n}) - D(f_n, g_{f,\tilde{\alpha}_n}))$$

$$\leq \inf_{\alpha\in[0,1]} D(f_n, g_{f,\alpha}) + \varepsilon_n + a_n = \inf_{\alpha\in[0,1]} D(f_n, g_{f,\alpha}) + \varepsilon_n^*,$$

where $a_n = D(f_n, g_{f,\hat{\alpha}_n}) - D(f_n, g_{f,\tilde{\alpha}_n})$ and $\varepsilon_n^* = \varepsilon_n + a_n$. So it is sufficient to prove that a_n goes to zero almost surely as n goes to infinity. The definitions of $\tilde{\alpha}_n$ and $\hat{\alpha}_n$ imply that

$$0 \leq D(f_n, g_{f,\hat{\alpha}_n}) - D(f_n, g_{f,\tilde{\alpha}_n}) + \varepsilon_n =$$

$$D(f_n, g_{f_n,\tilde{\alpha}_n}) - D(f_n, g_{f,\tilde{\alpha}_n}) + D(f_n, g_{f,\hat{\alpha}_n}) - D(f_n, g_{f_n,\tilde{\alpha}_n}) + \varepsilon_n \leq$$

$$D(f_n, g_{f_n,\tilde{\alpha}_n}) - D(f_n, g_{f,\tilde{\alpha}_n}) + D(f_n, g_{f,\hat{\alpha}_n}) - D(f_n, g_{f_n,\hat{\alpha}_n}) + \varepsilon_n \leq$$

$$2 \sup_{\alpha\in[0,1]} |D(f_n, g_{f_n,\alpha}) - D(f_n, g_{f,\alpha})| + \varepsilon_n \leq 2 \sup_{\alpha\in[0,1]} |D(g_{f_n,\alpha}, g_{f,\alpha})| + \varepsilon_n.$$

The last inequality follows by the reverse triangle inequality. Then, for all n,

$$-\varepsilon_n \leq a_n \leq 2 \sup_{\alpha\in[0,1]} |D(g_{f_n,\alpha}, g_{f,\alpha})|.$$

The right term goes to zero almost surely with n by assumptions Ass.1 and Ass.4. This and the fact that $\lim_n \varepsilon_n = 0$ imply that $\lim_n a_n = 0$ almost surely and the proof concludes.

Proof of Theorem 3. Observe that

$$D(f_i, f_j) \leq D(f_i, f_n^i) + D(f_n^i, f_n^j) + D(f_j, f_n^j),$$

$$D(f_n^i, f_n^j) \leq D(f_i, f_n^i) + D(f_i, f_j) + D(f_j, f_n^j).$$

For all $i < j$, by Assumption Ass.1, with probability 1 the random sequences $\{f_n^i\}_n$ and $\{f_n^j\}_n$ will verify that

$$|D(f_n^i, f_n^j) - D(f_i, f_j)| \leq D(f_i, f_n^i) + D(f_j, f_n^j) \to_n 0.$$

Let $\epsilon = 0.5 * \min\{|D(f_i, f_j) - D(f_k, f_l)| : i < j, k < l, (i,j) \neq (k,l)\}$. Then, with probability 1, there exists n_ϵ such that for all $n \geq n_\epsilon$, and all $i < j$

$$|D(f_n^i, f_n^j) - D(f_i, f_j)| < \epsilon.$$

Then, with probability 1, the ranks R_{ij} obtained when ordering the distances $D(f_i, f_j)$ coincide with those obtained when ordering the distances $D(f_n^i, f_n^j)$ for $n \geq n_\epsilon$. An analogous argument uses Assumption Ass.5 to prove that, with probability 1, for all $\alpha \in (0,1)$ the ranks r_{ij}^α obtained when ordering the distances $d(C_\alpha^{f_i}, C_\alpha^{f_j})$ coincide with those obtained when ordering the distances $d(C_\alpha^{f_n^i}, C_\alpha^{f_n^j})$ when $n \geq n_\nu$, where

$$\nu = 0.5 * \min\{\inf_{0<\alpha<1} |d(C_\alpha^{f_i}, C_\alpha^{f_j}) - d(C_\alpha^{f_k}, C_\alpha^{f_l})| : i < j, k < l, (i,j) \neq (k,l)\}.$$

As a consequence, with probability 1, for $n \geq \max\{n_\epsilon, n_\nu\}$ we have that $\hat{\alpha}_n = \alpha^*$ and the proof concludes.

Appendix B. Supplementary material

For the Aircraft data example a dynamic graph is provided as supplementary material. It represents 52 different periods that go smoothly from the first to the last periods in the left panel. Each density is represented by 3 level sets (with probability contents 0.25, 0.5 and 0.75). The control buttons below the animation allow interaction.

Acknowledgements

We appreciate very much the suggestions and comments of two anonymous referees. Pedro Delicado is very grateful to Antonio Cuevas and Amparo Baíllo for helpful conversations on plug-in estimation of density level sets, and to Alicia Nieto for teaching him to animate graphs. He also acknowledges the financial support of the Spanish Ministry of Education and Science and FEDER (MTM2010-14887, MTM2013-43992-R). We thank Jeff Palmer for reviewing the English of the manuscript. Part of this work was presented to the 3rd IWFOS Conference in Stresa, June 2014 (see Bongiorno et al., 2014).

References

Baddeley, A., 1992. Errors in binary images and an l^p version of the hausdorff metric. Nieuw Archief voor Wiskunde 10 (4), 157–183.

Baíllo, A., 2003. Total error in a plug-in estimator of level sets. Statistics & Probability Letters 65 (4), 411–417.

Baíllo, A., Cuesta-Albertos, J., Cuevas, A., 2001. Convergence rates in nonparametric estimation of level sets. Statistics & Probability Letters 53 (1), 27–35.

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, New York.

Bertrand-Retali, M., 1978. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. Rev. Roumaine Math. Pures Appl. 23, 361–385.

Bongiorno, E., Salinelli, E., Goia, A., Vieu, P., 2014. Contributions in infinite-dimensional statistics and related topics. Società Editrice Esculapio.

Borg, I., Groenen, P., 2005. Modern multidimensional scaling: Theory and applications, 2nd Edition. Springer.

Bowman, A., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, Oxford.

Bowman, A., Foster, P., 1993. Density based exploration of bivariate data. Statistics and Computing 3, 171–177.

Bronstein, A., Bronstein, M., Kimmel, R., 2006. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. Proceedings of the National Academy of Sciences of the United States of America 103 (5), 1168–1172.

Byrd, R. H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Scientific Computing 16, 11901208.

Cadre, B., 2006. Kernel estimation of density level sets. Journal of Multivariate Analysis 97 (4), 999–1023.

Cao, R., Cuevas, A., Fraiman, R., 1995. Minimum distance density-based estimation. Computational Statistics & Data Analysis 20 (6), 611–631.

Cuevas, A., 2009. Set estimation: Another bridge between statistics and geometry. Boletín de Estadística e Investigación Operativa 25 (2), 71–85.

Cuevas, A., Fraiman, R., 2010. Set estimation. In: Kendall, W., Molchanov, I. (Eds.), New Perspectives in Stochastic Geometry. Oxford University Press, Oxford, Ch. 11, pp. 374–397.

Delicado, P., 2011a. Dimensionality reduction for samples of bivariate density level sets: An application to electoral results. In: Ferraty, F. (Ed.), Recent advances in Functional Data Analysis and Related Topics. Springer, pp. 71–76.

Delicado, P., 2011b. Dimensionality reduction when data are density functions. Computational Statistics and Data Analysis 55 (1), 401–420.

Devroye, L., Györfi, L., 1985. Nonparametric Density Estimation: The L_1-View. Wiley, New York.

Duong, T., 2007. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. Journal of Statistical Software 21, 1–16.

Gibbons, J., Chakraborti, S., 2003. Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded. Taylor & Francis.

Holoček, J., Sojka, P., 2004. Animations in pdfTEX-generated PDF. Lecture Notes in Computer Science 3130, 179–191.

Marron, J., Tsybakov, A., 1995. Visual error criteria for qualitative smoothing. Journal of the american Statistical association 90 (430), 499–507.

Mason, D., Polonik, W., 2009. Asymptotic normality of plug-in level set estimates. The Annals of Applied Probability 19 (3), 1108–1142.

Parr, W., Schucany, W., 1982. Minimum distance estimation and components of goodness-of-fit statistics. Journal of the Royal Statistical Society. Series B, 178–189.

Polonik, W., 1995. Measuring mass concentrations and estimating density contour clusters – An excess mass approach. The Annals of Statistics 23 (3), pp. 855–881.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`

Ramsay, J., Silverman, B. W., 2005. Functional Data Analysis, 2nd Edition. Springer, New York.

Ren, Q., Mojirsheibani, M., 2008. Nonparametric estimation of level sets under minimal assumptions. Statistics & Probability Letters 78, 3029–3033.

Scott, D., 1992. Multivariate Density Estimation. Vol. 139. John Wiley & Sons.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Vol. 26. Chapman & Hall/CRC.

Simonoff, J., 1996. Smoothing Methods in Statistics. Springer Verlag.

Tsybakov, A., 1997. On nonparametric estimation of density level sets. The Annals of Statistics 25 (3), 948–969.

Wand, M., Jones, M., 1995. Kernel Smoothing. Vol. 60. Chapman & Hall/CRC.

Willett, R., Nowak, R., 2007. Minimax optimal level-set estimation. Image Processing, IEEE Transactions on 16 (12), 2965–2979.