Noname manuscript No.
(will be inserted by the editor)

Choosing the most relevant level sets for depicting a sample of densities

Pedro Delicado · Philippe Vieu

Received: date / Accepted: date

Abstract When exploring a sample composed with a set of bivariate density functions, the question of the visualisation of the data has to front with the choice of the relevant level set(s). The approach proposed in this paper consists in defining the optimal level set(s) as being the one(s) allowing for the best reconstitution of the whole density. A fully data-driven procedure is developed in order to estimate the link between the level set(s) and their corresponding density, to construct optimal level set(s) and to choose automatically the number of relevant level set(s). The method is based on recent advances in functional data analysis when both response and predictors are functional. After a wide description of the methodology, finite sample studies are presented (including both real and simulated data) while theoretical studies are reported to a final appendix.

Keywords Additive Functional Model; Bivariate Density Representation; Functional Data Analysis; Function on Function Regression; Optimal Level Set.

1 Introduction

An usual way for visualizing a bivariate density function consists in plotting, for various levels $\alpha_1, \dots, \alpha_J$, the corresponding level sets $C_{\alpha_1}, \dots, C_{\alpha_J}$. However, the question of deciding which values for the parameters α_j have to be selected is a crucial one, since some value may highlight interesting feature of the density that could be hidden by using other values. This appears clearly

P. Delicado
Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain, E-mail: pedro.delicado@upc.edu

P. Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France
E-mail: philippe.vieu@math.univ-toulouse.fr

from the example depicted in Figure 1. Parts a) and b) of Figure 1 show two densities exhibiting rather different structural shapes: one is clearly unimodal while the second one presents two different modes. Parts c) and d) of Figure 1 represent the same two densities by means of 6 level sets (those having probability content α_i are 0.02, .1, .25, .50, .75, .95). In this example, high level sets ($\alpha = .95$ or .75) do not let appear any difference between both densities and middle level sets ($\alpha = 0.5$ or 0.25) show a small difference but still do not detect the main structural difference between both densities since both seem to have a single mode. In counterpart, small level sets ($\alpha = 0.1$ or 0.02) are clearly depicting these main structural changes. This simple example shows the importance of the choice of the level(s) when representing a bivariate density function. Data-driven ways for driving this choice for depicting a single density have been studied in the literature and various methods have been proposed (see [6] for general discussion).

The purpose of this paper is to address this question in the situation when not only a single density function f has to be represented but when a sample $f_1, \ldots, f_N$ of densities has to be analysed. While apparently the problem looks similar there are at least three major differences. First of all note that the question of representing a sample of densities by level sets is even more important than for a single density, since in an obvious way there is no means for making interpretable any joint representation of surfaces like the ones depicted in parts a) and b) of Figure 1. Secondly, plots presenting many different level sets (like the plots c) and d) in Figure 1) become totally uninformative when one has to represent jointly various densities, and so the question of how choosing in an automatic way a few levels sets of interest is even more crucial than for representing a single density. Last but not least point to be highlighted, is the fact that the question of data-driven levels cannot be solved in a simple and naive way just by applying repeatedly existing methods for a single density representation, because in order to detect possible differences and/or similarities between some of the f_i it is necessary to find probability contents α_j which are common to the whole sample. As far as we know, this question of optimal level sets representation for a family of densities has not been widely studied in the literature (see however [6] for a first approach).

The route followed in this paper consists in modelling the links between a level set C_α and its underlying density function f as a regression problem. If one wishes to select only one level, this is a regression problem with functional response (namely, the density f) and a single functional covariate (namely, the set C_α). In the general framework when one wishes to use various level sets, this is also a functional regression model but with multi-functional covariates (namely, $C_{\alpha_1}, \ldots, C_{\alpha_J}$). This regression problem involves variables being of a high degree of complexity, and has therefore to be modelled in a flexible nonlinear way and nonparametric modelling is a natural candidate for that. Moreover as motivated in earlier literature for multivariate data in [23] the presence of more than one covariable may do necessary the use of additive structure for

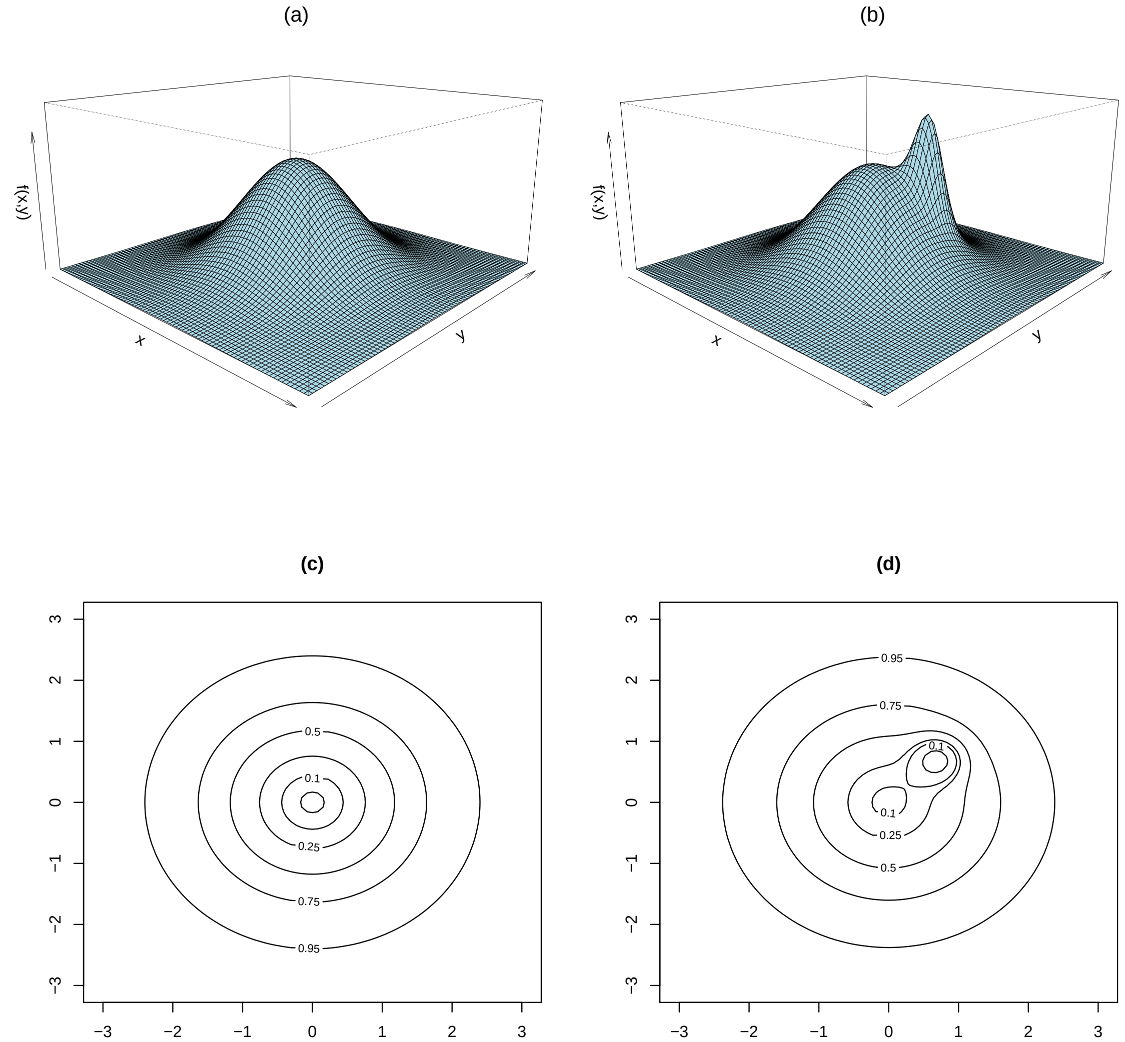

Fig. 1 Two bivariate density functions (plots a and b), and their level sets representations (plots c and d).

dealing with possible sparseness effects. Recent advances on functional data analysis allow to develop as well nonparametric ideas to propose flexible models for one functional covariable (see [11], [12]) as well as additive ideas to deal with multi-functional problems (see e.g. [1], [13] or [14]). These ideas will allow for estimating the regression link between the sets $C_{\alpha_1}, \ldots, C_{\alpha_J}$ and the density f and then in solving the problem of data-driven selection of the levels by choosing the values $(\alpha_1, \dots, \alpha_J)$ for which the regression estimates lead to the best prediction of the density f. Of course, the number of level sets J plays a crucial role since it should be sufficiently high to reflect as much as possible informations on the densities but also sufficiently small for providing interpretable plots. In a natural way, the additive structure proposed in this paper allows to use penalised least squares techniques for constructing a data-driven choice of J.

The paper is organised as follows. In a first attempt, after having stated a precise definition for a single optimal level set, the methodology for estimating it is developed in Section 2. Then, this is extended in Section 3 by means of functional additive ideas for estimating various optimal levels $(\alpha_1, \dots, \alpha_J)$. As pointed in Subsection 3.3, one will also derive data-driven estimation of the dimensional parameter J, and as far as we know this is the first paper in which a solution to this problem is proposed. We decided to put the mathematical study of the asymptotic properties of the estimates into a final Appendix, and to emphasise rather on the computational issues (see Section 4) and on the finite sample behaviour of the methods (see Section 5). To improve the reading, the paper is firstly presented in the simplest situations when the densities $f_1, \dots, f_N$ are fully known, but it is available for set of estimated densities. More precisely, Section 6 will discuss how the methodology directly extends to situation when each density is itself obtained by a preliminary smoothing procedure. In Section 7 a real data example is introduced in the context of electoral data in Spain: the joint density of two relevant variables (participation in an election contest, and proportion of votes cast for a particular option) is represented by level curves in several municipalities, and one applies all the previous methodology to these Spanish electoral data.

2 Data-driven optimal level set

In this section one starts with the simple problem of choosing a single level set C_α for depicting densities data (one will address in Section 3 the more general question of deciding whether more than one levels, and how many, can be of interest and how estimating them). We are in the situation in which one has a sample of bivariate densities available: $f_i, i = 1, \dots, N$. For each value of $\alpha \in]0, 1[$ we denote by $C_{\alpha,i}$ the level set with probability content α associated with each density f_i, and by C_α the level set with probability content α associated with an other generic density f. Precisely, C_α satisfies the probability condition

$$\int_{C_\alpha} f = \alpha$$

and can be written, for some ψ (depending on α) on the form:

$$C_\alpha = \{(x_1, x_2) \in \mathbb{R}^2, f(x_1, x_2) > \psi\}.$$

The idea is to find a functional link between the set C_α and the corresponding density f, and then to use this functional link for choosing the most relevant value of α. This section is structured in the following way. Section 2.1 fixes some general notations while Section 2.2 presents the methodology for selecting in a data-driven way the most relevant value (let say $\hat{\alpha}$). Theoretical study of the procedure is reported to Section A in the Appendix.

2.1 Some notation

In a standard way, each level set C_α is supposed to be an element of

$$\mathcal{C} = \{\text{compact subsets of } \mathbb{R}^2\},$$

and a way for measuring the proximity between two elements C_1 and C_2 of $\mathcal{C}$ consists in using the semi-metric

$$d_\lambda(C_1, C_2) = \lambda(C_1 \Delta C_2),$$

λ being the Lebesgue measure on $\mathbb{R}^2$. In an other hand, each density function f is supposed to be an element of

$$\mathcal{D} \subset \{\text{Bivariate densities on } \mathbb{R}^2 \text{ having compact level sets}\}.$$

To measure the similarities between two density functions one can use standard functional distances. For seek of simplicity of presentation, in this paper we only consider L_2-type measures of proximity by considering, for a given known link operator ϕ[1], the following metric on the space $\mathcal{F} \subset \phi(\mathcal{D})$,

$$d(g_1, g_2) = \left(\int_{\mathbb{R}^2} (g_1(x) - g_2(x))^2 dx \right)^{1/2}, \forall g_1, g_2 \in \mathcal{F},$$

that can also be denoted by $||g_1 - g_2||$. Examples of operators ϕ and guidelines for choosing it in practice will be given in Section 4.

2.2 Data-driven optimal level set choice

Let us for the moment fix the value of α. Looking for the link between a level set with probability content α and its corresponding density f can be addressed by means of a regression model

$$g = R_\alpha(C_\alpha) + \epsilon_\alpha, \tag{2.1}$$

where $g = \phi(f)$ for a fixed known operator ϕ, ϵ_α is a random element of $L^2(\mathbb{R}^2)$ with zero mean, and $R_\alpha : \mathcal{C} \to L^2(\mathbb{R}^2)$. The main feature is to be a functional

[1] Note that the choice of this operator belongs to the user. Letting this choice open increases the flexibility of the method. From a mathematical point of view, ϕ should be a one to one correspondence such that $\mathcal{F} = \phi(\mathcal{D}) \subset L^2(\mathbb{R}^2)$. An example of natural choice is $\phi = \log$ with $\mathcal{D}$ such that for any $f \in \mathcal{D}$ one has that $\log(f) \in L^2(\mathbb{R}^2)$.

on functional regression problem, in the sense that both the explanatory variable (i.e., the level set C_α) and the response (i.e. the function g) are complex mathematical objects lying into infinite dimensional spaces. Identifiability of the model is reported to Section A in the Appendix.

Our wish is to estimate the value of α_0 leading to the best representation of the density function, then it is natural to construct estimates of the operators R_α and then to look for the value α for which the corresponding estimate gives the best prediction of the response g. Precisely, each nonlinear regression operator R_α is estimated by means of the following functional kernel regressor. Given the sample of densities $\{f_i, i = 1, \ldots, N\}$ we consider the corresponding pairs $(g_i, C_{\alpha,i})$, and the operator R_α is estimated at any new element $c \in \mathcal{C}$ by

$$\hat{R}_{h,\alpha}(c) = \frac{\sum_{i=1}^N g_i K(\frac{d_\lambda(C_{\alpha,i},c)}{h})}{\sum_{i=1}^N K(\frac{d_\lambda(C_{\alpha,i},c)}{h})}, \qquad (2.2)$$

which is the weighted average of the g_i for which the corresponding level set $C_{\alpha,i}$ is closed from c, closeness having to be understood with respect to the semi-distance d_λ between level sets. The form of the weights is defined by means of an univariate real kernel function K, while the parameter $h = h(n) > 0$ acts as a smoothing parameter. Comments including practical guidelines for choosing the various parameters of these estimates will be given along Section 4.

Technical conditions ensuring good asymptotic properties of these estimates are recalled in the Appendix, but it is worth being pointed at this stage that (as in any nonparametric problem) the quality of the estimate is directly linked with the smoothing parameter (i.e. with the bandwidth h). That means that the choice of α will be strongly impacted by the smoothing parameter h. Therefore, both quantities h and α must be selected simultaneously. If we denote by $\Theta_N = H_N \otimes \mathcal{A}_N$ the set of possible values for the pair (h, α), natural choices for h and α are defined to be those (assumed to exist) leading to the smallest error when estimating the operator R by the kernel one $\hat{R}_{h,\alpha}$, namely:

$$(h_1, \alpha_1) = \arg \min_{(h,\alpha)\in\Theta_N} Err(h, \alpha)$$

where

$$Err(h, \alpha) = \sum_{i=1}^N ||R(C_{\alpha,i}) - \hat{R}_{h,\alpha}(C_{\alpha,i})||^2 W(C_{\alpha,i}),$$

W being some given weight function. Practical guidelines for choosing the various parameters entering in this method (namely W, H_N and $\mathcal{A}_N$) will be given along Section 4.

Unfortunately, the values (h_1, α_1) are uncomputable in practice and the main challenge is to be able to construct data-driven approximations of (h_1, α_1).

As previously described in [15], a way for choosing parameters in nonparametric regression is to use cross-validation ideas (these ideas have been recently extended for choosing h in the functional setting which is our purpose in this paper). Because the level α has to be selected jointly with the bandwidth h, one do that in the following way:

$$(\hat{h}, \hat{\alpha}) = \arg \min_{(h,\alpha)\in\Theta_n} \mathrm{CV}(h,\alpha), \tag{2.3}$$

with

$$\mathrm{CV}(h,\alpha) = \sum_{i=1}^{N} ||g_i - \hat{R}_{h,\alpha}^{-i}(C_{\alpha,i})||^2 W(C_{\alpha,i}).$$

In this formula $\hat{R}_{h,\alpha}^{-i}$ is the leave-one-out version of the estimate $\hat{R}$. The theoretical assesment of this procedure will be given in Theorem 1 (see Section A in the Appendix) which states that the cross-validated probability content $\hat{\alpha}$ gives (asymptotically) the same minimal error as the uncomputable value α_1. As a direct consequence of this result one will get the consistency of the data-driven selected probability content $\hat{\alpha}$ towards the true theoretical optimal one α_0 (see Corollary 2 in Section A of Appendix).

3 Choosing more than one level set

At this stage, the methodology described in Section 2 before provides an automatic way for choosing an optimal level set for representing the density sample $f_1, \ldots, f_N$. In order to improve the representation of these densities, one wishes to use more than one level sets (let say $\alpha_1, \ldots, \alpha_J$) and the aim of this Section 3 is to extend the previous methodology to this setting. The main difficulty comes from the fact that, in regression problems with multi-covariates the nonparametric modelling suffers from sparseness effects and new models have to be developed. We will describe in Section 3.1 how additive models ideas can be useful for this purpose, while Section 3.2 will construct optimal data-driven choices for the probability contents $\alpha_1, \ldots, \alpha_J$. Finally, and this is not the least point, the visualisation constraints put in force the importance of the dimensional parameter J. A good choice of J has to balance both the need for high degree of informations about the densities (that means that J should be large enough) and the wish for insuring interpretable plots (J should be small enough for that). Based on functional adaptation of model selection ideas, Section 3.3 will present an optimal data-driven way for estimating J.

3.1 The methodology

Additive ideas have been developed in multivariate nonparametric analysis in order to balance the trade-off between flexibility of the model and sparseness of the data (see for instance [23]). Here, because the problem is a multi-functional one, additive modelling becomes a natural candidate for modelling in a nonparametric way the fact that more than one level set may lead to better reconstruction of the response g than a single one. The model is defined as follows:

$$g = \mathcal{R}_J(C_{\alpha^1,\dots,\alpha^J}) + \epsilon = \sum_{j=1}^{J} R^j_{\alpha^j}(C_{\alpha^j}) + \epsilon. \tag{3.1}$$

Identifiability will be discussed in Section B of the Appendix. Estimation in this additive model can be performed in some iterative way by using a kernel operatorial estimate at each step like the one defined in (2.2). Specifically, given J, and given some vector of probability contents $(\alpha^1, \dots, \alpha^J)$ as well as some vector of bandwidths $(h^1, \dots, h^J)$, the additive estimates are constructed by putting $\hat{R}^1_{h^1,\alpha^1}$ as defined in (2.2), and then by regressing successive residuals in the following way:

$$\hat{R}^j_{h^j,\alpha^j}(c) = \frac{\sum_{i=1}^N Q_i^{j-1} K(\frac{d_\lambda(C_{\alpha^j,i},c)}{h^j})}{\sum_{i=1}^N K(\frac{d_\lambda(C_{\alpha^j,i},c)}{h^j})}, \tag{3.2}$$

with

$$Q_i^{j-1} = g_i - \sum_{k=1}^{j-1} \hat{R}^k_{h^k,\alpha^k}(C_{\alpha^k,i})$$

In a functional framework, this procedure is proposed in [13], in which various asymptotic properties of the estimates of the additive components are given. In this paper the reader will find asymptotic results concerning each functional additive estimated component $\hat{R}^j_{h^j,\alpha^j}$. Here, our wish is to discuss briefly how the methodology presented before in this paper may lead directly to a data-driven way for estimating the number of terms to be involved into the additive model. Of course, the data-driven choice of this dimensionality parameter cannot be addressed without taking into consideration the question of choosing the various bandwidths h^j and the various probability contents α^j.

3.2 Data-driven choices of the relevant levels

In a first attempt, we select the parameters (h^j, α^j) jointly step by step by means of the cross-validation ideas presented above in Section 2.2. Specifically, the first probability content and the first bandwidth are chosen as defined in (2.3), and subsequently at each step the pair (h^j, α^j) is chosen as:

$$(\hat{h}^j, \hat{\alpha}^j) = \arg \min_{(h,\alpha)\in\Theta_N} \mathrm{CV}^j(h,\alpha), \tag{3.3}$$

with

$$\mathrm{CV}^j(h,\alpha) = \sum_{i=1}^{N} ||Q_i^{j-1} - \hat{R}^{-i}_{\hat{h}^{j-1},\hat{\alpha}^{j-1}}(C_{\alpha,i})||^2 W(C_{\alpha,i}).$$

In this formula, for any k, $\hat{R}^{-i}_{h^k,\alpha^k}$ is the leave-one-out version of the estimate $\hat{R}^k_{h^k,\alpha^k}$. In Section B of the Appendix (see Theorem 3) one states the consistency of this data-driven selected level.

3.3 Data-driven choice of the number of relevant levels

A naive way for choosing the number of relevant levels would consist in minimising (now, over J) once again an error criterion, but this would lead to obvious over-estimation of J_0 and would avoid for easy interpretation of the outputs. This phenomenon is typical of any dimensionality estimation problem (see for instance [24] for a general discussion in the standard multivariate analysis) and an usual way to solve the problem is to introduce some dimensional penalty. Precisely in our problem, one may use a penalised version of the cross-validation criterion defined before, and this leads to choosing J in the following way

$$\hat{J} = \arg\min_{J\geq 1} \mathrm{PCV}(J), \tag{3.4}$$

where

$$\mathrm{PCV}(J) = \mathrm{CV}^J(\hat{h}^J, \hat{\alpha}^J)(1 + J\lambda_N).$$

The dimensionality penalty parameter λ_N must tend to 0 to prevent the good asymptotic behaviour of the cross-validation criterion from being affected. So we assume that:

$$\lim_{N\to\infty} \lambda_N = 0. \tag{3.5}$$

Practical guidelines for choosing this penalty in practice are described along Section 4 while theoretical assessment of the procedure is provided along Section B of the Appendix (see Corollary 4).

4 Computational issues

4.1 Choosing the parameters of the nonparametric estimates

As in any nonparametric problem, the behaviour of kernel estimates does not depend so much on the weighting function K and the usual Epanechnkow kernel

$$K(u) = \frac{3}{4}(1-u^2),\ u \in [-1, +1] \tag{4.1}$$

can be used, while the choice of the bandwidth needs much more attention. While cross-validation is known to be an optimal tool for data-driven bandwidth choice in functional data situations (see [19]), the choice of the set H_N on which the selection has to be made (see Section 2.2) is a crucial point, since it should be sufficiently large to capture easily the optimal bandwidth but should not be too large for trivial implementation reasons. The way for balancing this trade-off is to use k-NN (i.e. k nearest neighbours) ideas. As it has been theoretically proved in [17] the estimate (2.2) has the same asymptotic behaviour as the following k-NN one:

$$\hat{R}_{h_k,\alpha}(c) = \frac{\sum_{i=1}^N g_i K(\frac{d_\lambda(C_{\alpha,i},c)}{h_k})}{\sum_{i=1}^N K(\frac{d_\lambda(C_{\alpha,i},c)}{h_k})}, \tag{4.2}$$

where $h_k = \min\{h, \#\{i = 1, \ldots, N, d_\lambda(C_{\alpha,i}, c) \leq h\} = k\}$, in such a way that looking for an optimal value of h can be reduced to the question of choosing an optimal value of k which is a much easier problem in the sense that k is a discrete parameter lying into a finite set K_N. In a concrete way, the criterion CV (see again Section 2.2) has to be minimised over

$$H_N = \{h_k, k \subset K_N\} \text{ for some } K_N \subset \{1, \ldots, N\}.$$

Note that K_N can be chosen as being rather small. For instance, for a sample of size $N = 50$ one could reasonably use

$$H_N = \{4, 7, 10, 13, 16\}. \tag{4.3}$$

The same k-NN procedure can be obviously used for the kernel additive regressors defined in (3.2).

4.2 Chosing other parameters

The other parameters to be chosen are not really statistical parameters but depend more on the results one wishes to have. For instance the size of the set $\mathcal{A}_N$ of possible probability contents does not need to be very large to provide interpretable results. As a matter of example, for a size $N = 50$ a reasonable choice can be

$$\mathcal{A}_N = \{.05, .15, .25, .35, .45, .55, .65, .75\}. \tag{4.4}$$

As it is usual with cross-validation methods, the weight functions $W(.)$ have no strong influence on the results, and we used the simplest choice

$$W(C) = 1 \text{ if } d_\lambda(C, \emptyset) \leq d_\lambda(C_1, \emptyset) \text{ and } W(C) = 0 \text{ if not,}$$

where C_1 is the common support of the simulated density functions. Finally, concerning the operator ϕ, its choice depends really on the goals of the study and on the features of the densities that one wishes to highlight. The simplest choice is $\phi(h) = f$. In other problems, if one is more interested in the variations of the densities than on their exact values, other choices can be to use derivative operators $\phi(f) = d^k f / d^k$. An other natural choice, is to look at logarithms of densities and in this case one can take

$$\phi(f) = \log f. \tag{4.5}$$

4.3 Choosing the penalization

As motivated in a general multivariate analysis in [24], least square penalisation is an usual way for dealing with dimensionality choices. The question when applying a penalised cross-validation criterion such as in (3.3) is the choice of the penalty term λ_N. While this can be a difficult task in very high dimensional problems and specific sophisticated penalty have to be introduced such as the usual SCAD'one[2] defined in [7] or [8] (see also [5] or [22] for recent advances), it is much more easy to deal with in the purpose of this paper since for obvious interpretability reasons the number J of relevant levels has necessarily to be very small. This is why we decided to use the simplest choice, as already used in other dimensionality problems involving functional data (see for instance [9] and references therein):

$$\lambda_N = \gamma/(\log N), \text{ for some } \gamma > 0. \tag{4.6}$$

5 Finite sample size studies

5.1 Some simulated samples of densities

We have simulated a random sample of N bivariate densities, each of them being the density of the mixture of two bivariate normal random variables, truncated at the square $[-3.035, 3.035] \times [-3.035, 3.035]$. The simulation model is rather general in order to cover various different situations and it is defined from the following generic expression for the generated densities (before truncation):

$$f(x,y) = \nu\varphi(x,y;\mu_1,\mu_2,I_2)+(1-\nu)\varphi(x,y;\mu_1+\rho\cos\theta,\mu_2+\rho\sin\theta,\sigma I_2), \tag{5.1}$$

where φ is the density function of a bivariate normal random variable (with obvious notation) and I_2 is the identity matrix of size 2. The various parameters of the model insure a very wide scope of possible shapes for the generated densities, covering for instance as well standard unimodal Gaussian functions like the one depicted in part a) of Figure 1 (in this case, $\nu = 1$ and $\mu_1 = \mu_2 = 0$) as bimodal densities like the one depicted in part b) of Figure 1 (in this case, $\nu = .92$, $\mu_1 = 1$, $\mu_2 = 2$, $\sigma = .25$, $\rho = 1$ and $\theta = \pi/4$).

Let us point several characteristics of densities $f(x, y)$ defined by (5.1). Observe that (excepted when $\nu = 0$ or 1) they are a mixture of two unimodal densities, being bimodal for ρ sufficiently large (or σ small). Figure 2 represents the general form of such densities, highlighting the role of each parameter entering in the model, the level sets being used for the representation of f being at this stage arbitrarily chosen to be those with probability contents α equal to 0.02, 0.1, 0.25, 0.5, 0.75 and 0.95.

Note that two of these densities with common parameter ρ can be transformed into each other by a rotation and a translation. Given two of these densities differing only in parameter θ (a rotation with centre (μ_1, μ_2) transforms one into the other), their level sets with probability content $\alpha \geq 0.75$ are almost equal, while those corresponding to small values of α (say $\alpha \leq 0.1$) present large differences. On the other hand, two densities f and f' with location parameters $\mu = (\mu_1, \mu_2)$ and $\mu' = (\mu'_1, \mu'_2)$, with $\|\mu - \mu'\|$ moderate or large (say $\|\mu - \mu'\| \geq 1$) will present large differences in level sets corresponding to large values of α (say $\alpha \geq 0.5$), but level sets corresponding to small values

[2] Smoothly Clipped Absolute Deviation

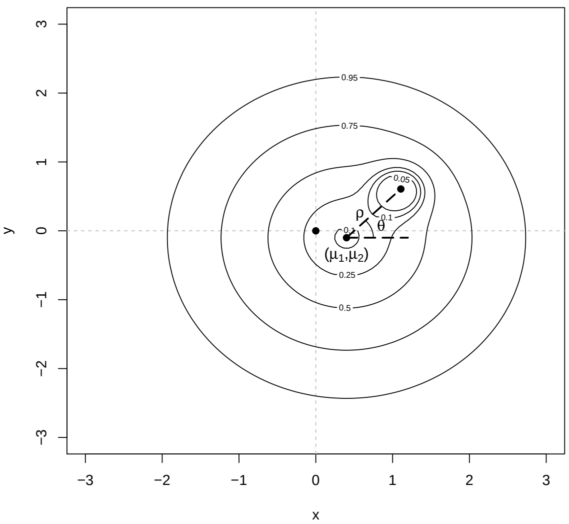

Fig. 2 General shape of a density function of the form of (5.1).

of α may not be so large (for instance, the level sets with probability content $\alpha = 0.05$ of f and f' are almost coincident when $\mu_1 = -1$, $\mu_1' = 1$, $\mu_2 = \mu_2' = 0$, $\rho = \rho' = 1$, $\theta = 0$, $\theta' = \pi$, because both have their highest mode at $(0, 0)$).

5.2 Presentation of the models

In order to cover a large variety of situations we have considered 7 different models, obtained in the following way. We have generated random samples of densities according to (5.1) by taking random values of θ, μ_i, $i = 1, 2$, and ρ (the other parameters being fixed to $\nu = 0.05$ and $\sigma = 0.25$). Specifically, $\theta \sim U(0, 2\pi)$, $\rho \sim U(1 - r, 1 + r)$ and $\mu_i \sim U(1 - m, 1 + m)$, $i = 1, 2$. We have considered 7 different cases (or models) to generate random densities according to (5.1), corresponding to specific choices of r and m as summarised in Table 1.

Table 1 The 7 different models for two bivariate normal densities samples

Model	1	2	3	4	5	6	7
r	0	0	0	0	0.25	0.25	0.25
m	0	0.1	0.25	1	0.1	0.25	1

One may observe that in the simplest case (Model 1) the densities only differ in θ, while in Models 2, 3 and 4 differences in location and orientation are allowed (but ρ is still fixed to be equal to 1 in these three models). Finally, Models 5, 6 and 7 describe more complex situations in which all parameters may vary from one density to each other.

5.3 Presentation of the study

One aims is to show the finite sample behaviour of the methods presented before in this paper. So, for each among these 7 models, a random sample of $N = 50$ density functions has been generated, and this learning sample is used to compute the various different estimates. Moreover to assess the validity of the method we have also generated an independent second test sample $f_i^{\text{ts}}, i = 1, \dots, N_{\text{ts}}$ of size N_{ts}. Because the role of N_{ts} is of small interest for our purpose we just took as size for this testing sample the simplest one $N_{\text{ts}} = N$. Then in each case, five additive models (one for each $J \in \{1, \dots, 5\}$; see equation (3.1)) are fitted for each of the 7 samples, as explained in Section 3. Observe that the additive model corresponding to $J = 1$ coincides with the nonparametric model (2.1) described in Section 2. All along the study the various parameters entering into the statistical procedures have been chosen automatically according to the guidelines described along Section 4 before.

5.4 The additive methodology in action

As motivated in Section 4.1 the various kernels entering into our procedure have been chosen according to (4.1) and the operator ϕ is the one defined in (4.5). For each model, we compute the estimates based on the learning sample. This includes the following calculations:

- $\hat{\alpha}^j$, $\hat{k}^j$, according to the guidelines described in (4.3) and (4.4). Observe that for any $J \in \{1, \dots, 5\}$ the value of j goes from 1 to J, and that $\hat{\alpha}^j$ and $\hat{k}^j$ are the same for all $J \geq j$.
- $\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$: The cross-validation (leave-one-out) estimate of the mean squared prediction error. The quantity $\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)$ has been defined in equation (3.3).
- $R^2_{\text{CV},j} = 1 - \text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/\text{TSS}_{\text{CV},j}$, where

$$\text{TSS}_{\text{CV},j} = \sum_{i=1}^{N} \|Q_i^{j-1} - (1/(N-1)) \sum_{l \neq i} Q_l^{j-i}\|^2$$

  is the total sum of squares estimated by leave-one-out (Q_i^{j-1} is defined in equation (3.2)).

Then, to see how the various fitted additive models behave, we have applied these estimates to the testing sample, and we have computed the following quantities:

- $\mathrm{MSPE}_{\mathrm{ts},J} = (1/N_{\mathrm{ts}}) \sum_{i=1}^{N_{\mathrm{ts}}} \|g_i^{\mathrm{ts}} - \hat{g}_i^{\mathrm{ts},J}\|^2$, the mean squared prediction error in the testing sample.
- $R^2_{\mathrm{ts},J} = 1 - \mathrm{MSPE}_{\mathrm{ts},J}/\mathrm{MSS}_{\mathrm{ts}}$, where

$$\mathrm{MSS}_{\mathrm{ts}} = (1/N_{\mathrm{ts}}) \sum_{i=1}^{N_{\mathrm{ts}}} \|g_i^{\mathrm{ts}} - (1/N) \sum_{l=1}^{N} g_l\|^2$$

is the mean sum of squares estimated in the test sample.

Observe that, for $J = 1$, the statistics $\mathrm{CV}^1(\hat{k}^1, \hat{\alpha}^1)/N$ is able to estimate the mean squared prediction error of the nonparametric functional regression model (2.1) for new independent densities, but this does not remain true for the statistics $\mathrm{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$, $j = 2, \ldots, J$, when $J \geq 2$, because $\mathrm{CV}^j(\hat{k}^j, \hat{\alpha}^j)$ corresponds to intermediate kernel regressions where the responses are no longer the densities. This is an additional reason for having introduced a second testing sample, and so $\mathrm{MSPE}_{\mathrm{ts},J}$ is an estimate of the mean squared prediction error for additive models for any J and can therefore be used as a tool for comparing the behavior of the various fitted models.

The results obtained for these various statistics are presented in Table 2.

5.5 Comments on the results

A summary of the main conclusions of the results in Table 2 can be drawn as follows. First of all, we can observe that the optimal α values increase with the variability in μ_i, $i = 1, 2$, as was expected: larger level sets indicate large differences in location more clearly. Secondly, the effect of randomness in parameter ρ is limited: cases 5, 6 and 7 are similar to cases 2, 3 and 4, respectively. Thirdly, in case 1, almost all the variability between densities is explained by using only one level set (that with probability content $\alpha = 0.25$), while in cases 2, 3, 5 and 6, $\hat{J} = 2$ components are required in the additive model: the first optimal probability content $\hat{\alpha}^1$ is large (0.65 or 0.75) and explains the differences in location between densities; then the level sets corresponding to $\hat{\alpha}^2$ are smaller ($0.05 \leq \hat{\alpha}^2 \leq 0.25$) and detect the orientation of the highest mode with respect to the lowest one. For cases 4 and 7 (maximum variability in μ_i, $i = 1, 2$), it may be $\hat{J} = 3$, with $\hat{\alpha}^1 = \hat{\alpha}^2 = 0.75$ (using a different number of neighbours: $\hat{k}^1 \neq \hat{k}^2$) and $\hat{\alpha}^3$ equal to 0.15 or 0.05.

5.6 Controlling the dimensionality

Finally, the question of graphical representation of the densities sample is strongly linked with the choice of the dimensional parameter J, too large value of J leading to unexploitable plots while too small ones would not reflect enough information on the densities. As motivated in Section 3.3, estimating J can be performed by means of penalised least square criterion. Precisely, we

Table 2 Mixture of two bivariate normal densities. The left hand side of the table shows the parameters r and m used to generate the 7 different types of mixtures densities. A summary of the results from the fitted additive models are shown on the right-hand side of the table.

Case	r	m		$J=1$	$J=2$	$J=3$	$J=4$	$J=5$
1	0	0	$\hat{\alpha}^j$, $\hat{k}^j$	.25, 4	.05, 10	.35, 16	.35, 16	.35,16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	.0096	.0023	.0016	.0015	.0015
			$R^2_{\text{cv},j}$	.9482	-.0105	-.2324	-.2646	-.2790
			$\text{MSPE}_{\text{ts},J}$	.0173	.0130	.0125	.0122	.0120
			$R^2_{\text{ts},J}$	.9127	.9344	.9371	.9384	.9393
2	0	.1	$\hat{\alpha}^j$, $\hat{k}^j$	.65, 7	.05, 13	.75, 16	.25, 16	.75, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	.2054	.0463	.0311	.0288	.0257
			$R^2_{\text{cv},j}$	.8005	.3128	-.0134	-.2127	-.20934
			$\text{MSPE}_{\text{ts},J}$	.2193	.1516	.1397	.1309	.1273
			$R^2_{\text{ts},J}$	.8145	.8717	.8818	.8893	.8923
3	0	.25	$\hat{\alpha}^j$, $\hat{k}^j$	.75, 7	.05, 10	.65, 16	.55, 16	.75, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	.4276	.1407	.0808	.0734	.0659
			$R^2_{\text{cv},j}$	.9267	.3069	.0019	-.2002	-.2362
			$\text{MSPE}_{\text{ts},J}$	.3662	.2705	.2520	.2468	.2364
			$R^2_{\text{ts},J}$	.9291	.9476	.9512	.9522	.9542
4	0	1	$\hat{\alpha}^j$, $\hat{k}^j$	.75, 7	.75, 4	.15, 16	.75, 16	.25, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	3.4968	1.2958	.2996	.2485	.2382
			$R^2_{\text{cv},j}$	.9560	.1424	-.1165	-.2000	-.2371
			$\text{MSPE}_{\text{ts},J}$	3.5649	1.5322	1.4718	1.4159	1.3912
			$R^2_{\text{ts},J}$	.9567	.9814	.9821	.9828	.9831
5	.25	.1	$\hat{\alpha}^j$, $\hat{k}^j$	.65, 7	.05, 7	.75, 16	.75, 16	.25, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	.2210	.0439	.0204	.0190	.0181
			$R^2_{\text{cv},j}$	.7319	.3974	-.0216	-.2320	-.2599
			$\text{MSPE}_{\text{ts},J}$	.2820	.1869	.1697	.1645	.1579
			$R^2_{\text{ts},J}$	.6945	.7976	.8162	.8218	.8290
6	.25	.25	$\hat{\alpha}^j$, $\hat{k}^j$	.75, 7	.25, 10	.75, 16	.05, 16	.75, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	.4700	.1432	.0780	.0681	.0452
			$R^2_{\text{cv},j}$	.8849	.3253	-.0932	-.1581	-.2092
			$\text{MSPE}_{\text{ts},J}$	.5335	.3751	.3338	.3024	.2870
			$R^2_{\text{ts},J}$	.9132	.9390	.9457	.9508	.9533
7	.25	1	$\hat{\alpha}^j$, $\hat{k}^j$	.75, 7	.75, 4	.05, 16	.65, 16	.65, 16
			$\text{CV}^j(\hat{k}^j, \hat{\alpha}^j)/N$	2.9903	1.1448	.2745	.1286	.1255
			$R^2_{\text{cv},j}$	.9646	.0780	-.1033	-.2170	-.2539
			$\text{MSPE}_{\text{ts},J}$	3.9657	1.8748	1.8355	1.8175	1.7951
			$R^2_{\text{ts},J}$	.9528	.9777	.9782	.9784	.9786

have used the penalised cross-validation technique by following the practical guidelines described in Section 4.3 for choosing the penalty (i.e. λ_N was selected as in (4.6) with $\gamma = 1$. [3]

[3] Other choices of γ have been tested and the method turns not to be too much sensitive on this parameter. For instance, in the situation depicted in Figure 3, any value of $\gamma \leq 0.6$ leads to the same minimum $J = 3$.

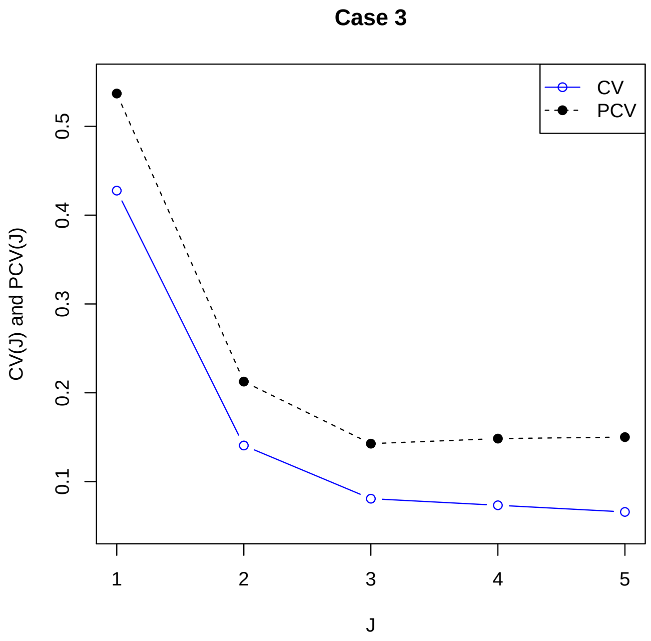

Fig. 3 CV and its penalised version as functions of J

To see more on this question of choosing J, let us look more in detail on the specific Model 3 (other models have been analysed in the same way and behave similarly). For this model, Figure 3 gives more insight on choosing J. This figure plots both the standard criterion and its penalised version.

The decreasing form of the criterion CV is decreasing, as expected, and if the dimension was selected by this criterion one would have taken higher values of J leading to a very small gain in terms of prediction (see Table 2) and to hardly representable results. By means of the penalised approach the criterion PCV exhibits a clear global minimum leading to a selected dimension $\hat{J} = 2$. This value allows for a good representation of the densities (as attested by the low prediction error appearing in Table 2), as well as an easy visualization of the outputs (this will be commented later in Section 5.7). The next Figure 4 shows that the Mean Square Prediction Error in the second testing sample as a function of J ($\text{MSPE}_{\text{ts},J}$) and its penalised version ($\text{PMSPE}_{\text{ts},J}$) exhibit roughly the same shape as $\text{CV}(J)$ and $\text{PCV}(J)$, respectively (up to a vertical shift). The curves are decaying rather fast until the optimal value. Then the penalized versions grow up (PCV growing slower $MSPE$) while the standard versions do not grow but are rather stable after the optimal value. These facts highlight the good behaviour of cross-validation for approximating the true unknown error and show also how penalised cross-validation is efficient for choosing J. In particular the penalization avoids for selecting a too high and uninformative dimension.

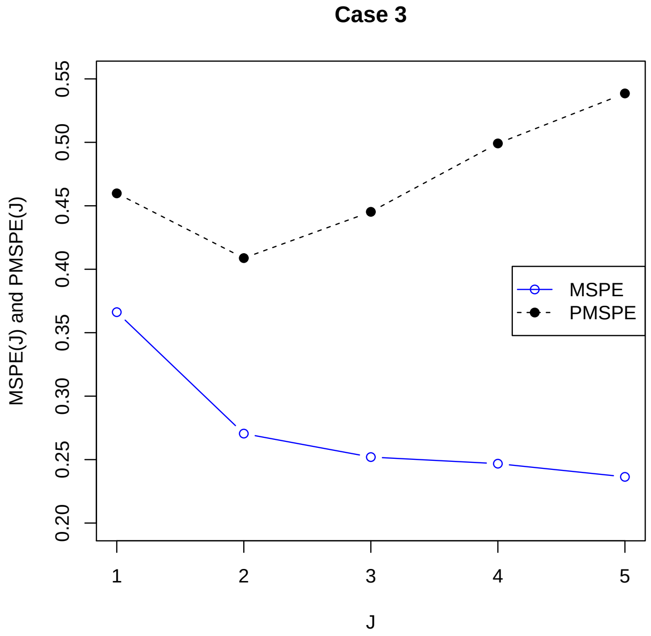

Fig. 4 MSPE and its penalised version as function of J.

5.7 Visualisation of the results

Again for Model 3, we present in Figure 5 the representation of the sample of 50 densities. As discussed before, the dimension was selected to be $\hat{J} = 2$ and the corresponding relevant levels were 0.75 and 0.05.

Clearly the pair of selected levels $(\hat{\alpha}_1, \hat{\alpha}_2) = (0.75, 0.05)$ provides a good representation of the sample. The high probability level set $C_{0.75}$ looks similar for all densities generated by Model 3 and so reflects the main common feature in the sample (remind that in this model one has $\rho = 1$ for all densities). In counterpart, the small probability level set $C_{0.05}$ detects the main difference between the densities in this sample, that is the different location of the main peak (remind that in this Model 3, the angle parameter θ changes from a density to each other).

6 What about samples of estimated densities?

In some practical situations it may be the case that the densities f_i are not known but are rather estimated by means of some usual preliminary bivariate nonparametric smoothers. The aim of this short Section 6 is to show how the general methodology, presented before along Sections 2 and 3 for known den-

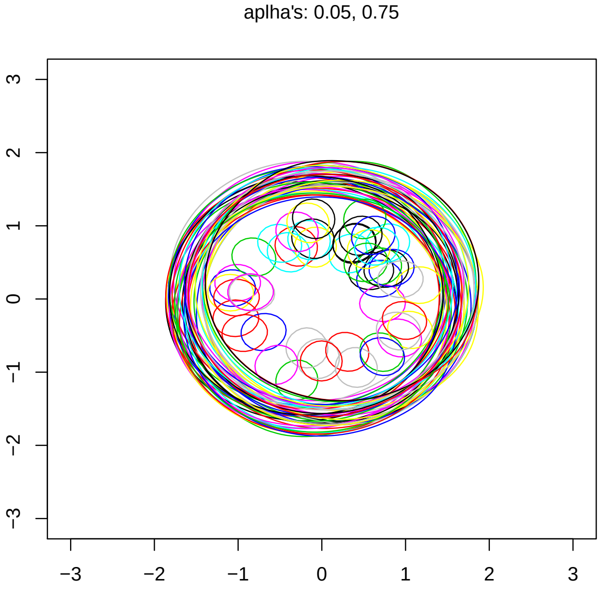

Fig. 5 Representation of a sample of $N = 50$ densities according to Model 3 using $J = 2$ data-driven level sets.

sities, behave similarly for estimated densities.

Assume that each f_i is estimated by means of a finite number of observed pairs $X_{i,j} \in \mathbb{R}^2, j = 1, \ldots, n_i$. In practice one often has unbalanced data in the sense that the densities f_i are built from samples not necessarily having the same size. For a simple presentation of the following, we put forward the following hypothesis:

$$\forall i = 1, \ldots, N, \; n_i \sim K_i n.$$

Similarly, for each α, the level set $C_{\alpha,i}$ of the density f_i is not observed but rather estimated by means of the plug-in estimator derived from each estimated density. We will denote by $f_{i,n}$ the estimate of each unknown density function f_i and by $C_{\alpha,i,n}$ the i-th estimated set with probability content α.

Basically, applying the functional nonparametric methods (both the single method in Section 2.2 and the additive one presented in Section 3) to the new data $(f_{i,n}, C_{\alpha,i,n})$ rather than to the theoretical ones $(f_i, C_{\alpha,i})$ will have no influence on the theoretical properties of the estimated parameters. The reason for this is rather simple: since the smoothing procedures leading to the construction of $(f_{i,n}, C_{\alpha,i,n})$ are bi-dimensional nonparametric problems, they induce an estimation error which is of much smaller order than the method itself (since this is subsequently an infinite dimensional nonparametric problem).

Technical assumptions quantifying these ideas are presented and commented along Section C in the Appendix, along which we will give some general result (see Theorem 5) stating the asymptotic properties of the procedure.

7 An application to real electoral data

The most important way of political participation for people in democratic countries is certainly to vote in electoral calls. Nevertheless the participation in elections is usually far from 100%: many people decide not going to vote for several reasons. A relevant question is if there exists some relationship between the political ideology of a given voter and its decision of going or not to vote in a particular election. In Spain it is given as a fact that potential left-wing parties voters usually participate in elections less than right-wing parties voters. We analyze the relationship between position on the left-right wing political dimension and the willingness to vote. Given that individual data are not available we use aggregated data at level of polling stations (lists of around 1000 people that vote at the same ballot box because they live in the same small area). Specifically we use electoral results from 2011 Spanish general elections.

For each polling station the available information allows us to define these two variables: participation (proportion of potential voters that finally vote) and proportion of votes for right-wing parties. Observe that this last variable is not exactly the same as the proportion of potential voters with right-wing political ideology. Unfortunately we only know what do vote the people who actually vote. Nevertheless, if the size of the polling station is small compared with the size of the city it is sensible to believe that both quantities should be similar. This is because in big cities, conditioning on the polling station of a voter is almost equivalent to conditioning on many economical, educational and sociological variables that simultaneously determine both decisions: *to vote or not to vote?* and *what to vote?*.

We formally need the following assumption:

> The political orientation (left-right wing) of a potential voter and his/her decision of voting or not are conditionally independent, given the polling station where he/she votes.

This assumption allows us to use observed aggregated data at the level of polling stations to study the joint distribution of *political orientation* and *participation.*

We consider the 100 cities in Spain with the largest number of polling stations (82 or more). For each of these cities we have a list of observations of the bivariate random variable (*proportion of votes for right-wing parties*, *participation*), an observation for each polling station. We use then a bivariate kernel density estimator to obtain from this list an estimation of the joint distribution of these two variables at each of the 100 cities considered in our study. Therefore we have a functional dataset of length 100 consisting on

Table 3 Results for the electoral data example. Maximums of $R^2_{\mathrm{ts},J}$ in J are in bold face type.

Data set		$J=1$	$J=2$	$J=3$	$J=4$	$J=5$
Raw data	$\hat{\alpha}^j$	.95	.65	.15	.95	.05
	$\hat{k}^j$	7	4	7	7	10
	$R^2_{\mathrm{ts},J}$	**.5972**	.5910	.5784	.5684	.5638
Centered data	$\hat{\alpha}^j$	.95	.65	.15	.85	.75
	$\hat{k}^j$	10	4	4	4	4
	$R^2_{\mathrm{ts},J}$	.6953	**.7333**	.7226	.7225	.7188

bivariate densities. As a matter of illustration, Figure 6 shows the density level sets corresponding to the 6 municipalities with the largest numbers of polling stations. In top plots probability content is $\alpha = 0.5$ which is the recommended level set by the quick rule given in [6] when using only one level set, while in bottom plots one has $\alpha_1 = 0.25$ and $\alpha_2 = 0.75$ as recommended by the quick rule given in [6] when using two level sets.

For applying our procedure, we divide the set of 100 density functions in two subsets: a *learning sample* with the $N = 50$ cities with odd numbers in the list of municipalities sorted by number of polling stations, and a *test sample* with the other. Five additive models (one for each $J \in \{1, \dots, 5\}$; see equation 3.1) are fitted with the learning sample, as explained in Section 3. The statistical procedures are similar to those used in Section 5. Then, for each model, we compute the estimates based on the learning sample to determine $\hat{\alpha}^j$, $\hat{k}^j$, $j = 1, \dots, J$, $J = 1, \dots, 5$. At the end, to see how the various fitted additive models behave, we have applied these estimates to the testing sample, and we have computed $R^2_{\mathrm{ts},J}$, $J = 1, \dots, 5$. All of these quantities have been defined in Section 5, and the results obtained for these various statistics are summarized in Table 3.

The first three rows of the table correspond to the raw data. It is apparent that the best additive model includes only one level set with probability content $\alpha = 0.95$. This is a clear indication that the main differences between densities are in their support, and they may reflects differences in location, dispersion and/or shape. The upper panel of Figure 7 shows the density level set with probability content $\alpha = 0.95$ for the 6 cities with the largest numbers of polling stations. These graphics show that these six densities are different in location (compare Madrid and Barcelona, for instance), dispersion (Valencia is more concentrated than Madrid, for instance) and shape (Sevilla is quite different from the rest, but also Barcelona, Málaga and Valencia are quite different from Madrid and Zaragoza). These differences where less clear in the upper panel of Figure 6 (here the main findings are the differences in location and the different shape of Sevilla), where $\alpha = 0.5$ is used, according to the quick rule proposed in [6]. Other procedures proposed in that paper (namely, those based on ranks of distances) lead to choose a probability content of at least 0.9, in agreement with what we obtain here with the additive model (3.1).

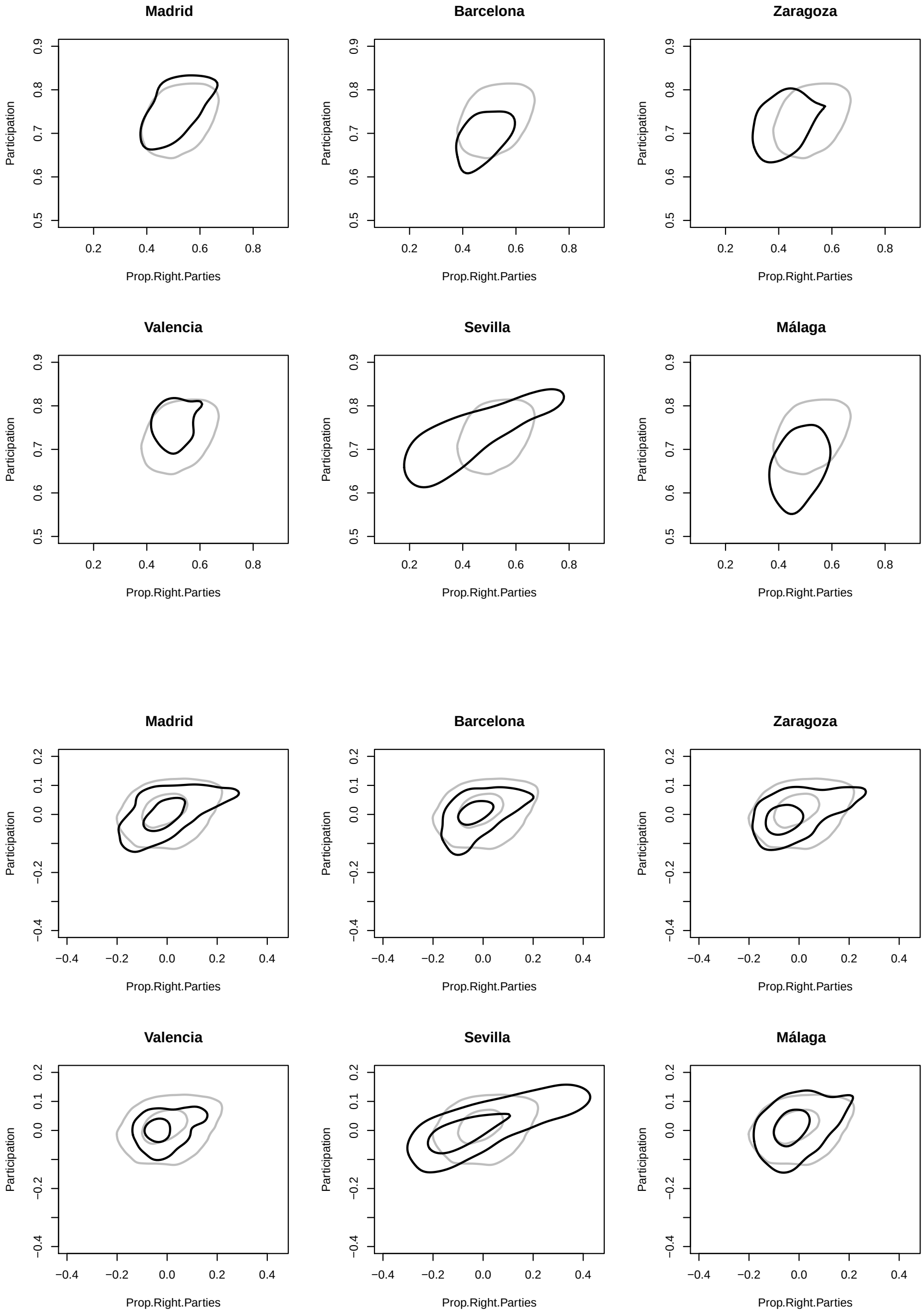

Fig. 6 Density level sets corresponding to the 6 municipalities with the largest number of polling stations. In grey, the level sets corresponding to the whole country. Probability contents have been chosen according to the quick rule proposed in [6]. *Upper panel:* Raw data, the probability content is $\alpha = 0.50$. *Lower panel:* Centred data, probability contents are $\alpha_1 = 0.25$ and $\alpha_2 = 0.75$.

In order to remove the "differences in location" effect from the analysis, we consider the estimated density functions from the centred data at municipality level, so that all the densities are centred at $(0, 0)$. The last three rows in Table 3 show the corresponding results. Now two level sets are required in the additive model, having probability contents $\alpha_1 = 0.95$ and $\alpha_2 = 0.65$. Again the presence of the largest allowed probability content indicates big difference in density supports. Following the recommendation of the second fitted additive model, the lower panel of Figure 7 shows the density level sets with probability contents $\alpha_1 = 0.95$ and $\alpha_2 = 0.65$ corresponding to the centred data of the 6 municipalities with the largest numbers of polling stations. Now the differences among cities in dispersion (Valencia is the most concentrated, and Sevilla the most dispersed) and shape (all the cities appear to be different from the other) are emphasised, and they are clearer than when following the quick rule proposed in [6] for 2 levels sets ($\alpha_1 = .25$ and $\alpha_2 = 0.75$; see the lower panel of 6). When using the procedures based on ranks of distances proposed in [6], the probability contents that are obtained are $\alpha_1 = 0.67$ and $\alpha_2 = 0.95$, very close to those obtained here with the additive model (3.1). The advantage of our proposals in this paper over those in [6] is that now we have a rule for choosing the number J of required level sets. In the present example this number is $J = 2$, according to Table 3.

8 Conclusions

This paper has used recent advances on nonparametric statistics for functional variables and on dimension estimation, for constructing in a data-driven way a reasonably small number of level sets for representing a family of bivariate density functions. From an applied point of view, its is shown through a wide scope of simulated models that the method combines both easiness of representation of the outputs and visualisation of the main features of the densities. This appealing finite sample behaviour goes together with theoretical optimality results. Finally it is worth being noted that this is, as far as we know, the first paper using nonparametric functional data methodology for dealing with objects being much more complicated than standard one dimensional curves. In most of applications in functional regression the explanatory variable is a one dimensional curve (see for instance [20], [21], [12], [16]) and the response is scalar. The nonparametric methodology has been extended and used also to responses being also a one dimensional curve (see eg [11] or [10]), but in the situation depicted in this paper both the explanatory variables (which is a compact two dimensional set) and the response one (which is a bivariate density) are not simple one-dimensional curves.

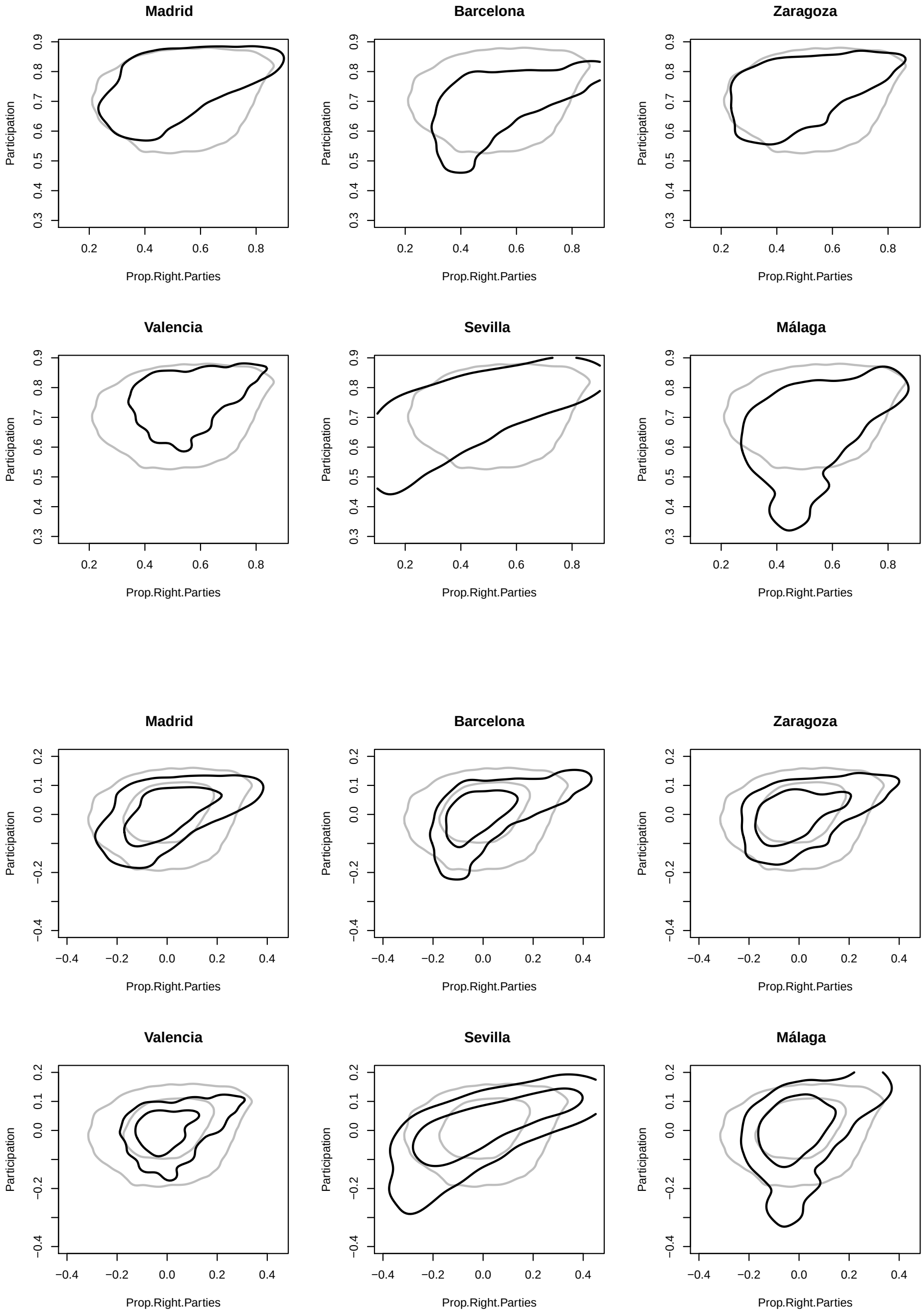

Fig. 7 Example of density level sets corresponding to the 6 municipalities with the largest number of polling stations. In grey, the level sets corresponding to the whole country. Probability contents have been chosen according to the additive model (3.1); see also Table 3. *Upper panel:* Raw data, the probability content is $\alpha = 0.95$. *Lower panel:* Centred data, probability contents are $\alpha_1 = 0.95$ and $\alpha_2 = 0.65$.

A Asymptotic theory for the procedure in Section 2

Identifiability of the model and main theoretical result

The complexity of the data makes necessary to attack the problem in a much flexible way, and this is the reason why we adopted a nonparametric strategy. Said with other words, in the model (2.1) R_α is a nonlinear operator from $\mathcal{C}$ into $\mathcal{F}$ and ϵ_α is a functional error $\epsilon_\alpha = \epsilon_\alpha(x)$ satisfying

$$\forall x \in \mathbb{R}^2, E\epsilon_\alpha(x) = 0 \text{ and } E\epsilon_\alpha(x)^2 < \infty.$$

This regression model will serve for giving a precise definition of what is the best level. Basically, the wish is to represent the density function f (or more generally its transformation $g = \phi(f)$) in an optimal way and this can be obtained by using as level α_0 the one leading to the smallest variance in the error term of the regression model. Precisely, one will define an optimal probability content α_0 in the following way:

$$\forall \alpha \neq \alpha_0, \forall x \in \mathbb{R}^2, \mathrm{Var}(\epsilon_{\alpha_0}(x)) < \mathrm{Var}(\epsilon_\alpha(x)). \tag{A.1}$$

that, it is implicitly supposed that such a value α_0 exists and is unique.

The next Theorem 1 states the asymptotic optimality of the procedure.

Theorem 1 *Under the conditions (A.4)-(A.12) one has*

$$\left| \frac{Err(\hat{h}, \hat{\alpha})}{Err(h_1, \alpha_1)} \right| \to 1, \ \ a.s. \tag{A.2}$$

As a direct consequence of this result, one has the following corollary which states the main theoretical result of this section that is the consistency of the data-driven selected level towards the true theoretical one.

Corollary 2 *Under the conditions of Theorem 1 and if in addition (A.13) and (A.14) hold, then one has*

$$\hat{\alpha} \to \alpha_0, \textit{in probability.} \tag{A.3}$$

Commented assumptions for Theorem 1 and Corollary 2

The asymptotic properties of the estimated operator $\hat{R}_{h,\alpha}$ can be derived directly from the recent advances obtained in nonparametric regression when both explanatory and response variables are functional (see for instance [11] for asymptotic normality and [10] for uniform rates of convergence). These consistency results are not our purpose here since we wish rather to discuss how this method may allow us to select the probability content α. The technical assumptions necessary for ensuring consistency properties of the estimated operator $\hat{R}_{h,\alpha}$ are recalled below (see [10]).

- *Conditions on the explanatory variable.* There is a function F such that for t small enough, one has

$$\begin{aligned} 0 < c_1 F(t) < & \inf_{\{0 \leq \alpha \leq 1, c \in \mathcal{C}\}} P(d_\lambda(C_\alpha, c)) \leq t) \\ \leq & \sup_{\{0 \leq \alpha \leq 1, c \in \mathcal{C}\}} P(d_\lambda(C_\alpha, c)) \leq t) \\ < & \ c_2 F(t) \ < \ \infty, \end{aligned} \tag{A.4}$$

and

$$\forall s \in [0, 1], \lim_{t \to 0} \frac{F(st)}{F(t)} \text{ exists.} \tag{A.5}$$

- *Conditions on the response variable.* The response variable g has conditional moments $\sigma^k_{x,\alpha}(c) = E(|g(x)|^k | C_\alpha = c)$ satisfying:

$$\exists B < \infty, \forall x \in \mathbb{R}^2, \forall \alpha \in [0,1], \forall k > 0, \sigma^k_{x,\alpha}(.) \leq Bk! < \infty, \tag{A.6}$$

$$\sigma^2_{x,\alpha}(.) \text{ is continuous and bounded from below, uniformly in } x \text{ and } \alpha. \tag{A.7}$$

- *Conditions on the regression operator.* The regression operators R_α are submitted to a smooth nonparametric model that consists in the assumption that there exist $A < \infty$ and $0 < \beta \leq 1$ such that:

$$\forall \alpha \in [0,1], \forall (c,c') \in \mathcal{C}^2, ||R_\alpha(c) - R_\alpha(c')|| \leq A d_\lambda(c,c')^\beta. \tag{A.8}$$

- *Conditions on the parameters of the estimate.* The bandwidth h is assumed to be such that for some $0 < \delta < 1$ and some $D < \infty$:

$$\lim_{N\to\infty} h = 0 \text{ and } \lim_{N\to\infty} F(h) \sim DN^{-\delta}, \tag{A.9}$$

while the kernel K should be such that for t small enough and for some $0 < c_3 < c_4 < \infty$:

$$\begin{aligned} 0 < c_3 F(t) < & \inf_{\{0\leq\alpha\leq 1, c\in\mathcal{C}\}} E\left(K\left(\frac{d_\lambda(C_\alpha,c)}{t}\right)\right) \\ \leq & \sup_{\{0\leq\alpha\leq 1, c\in\mathcal{C}\}} E\left(K\left(\frac{d_\lambda(C_\alpha,c)}{t}\right)\right) \\ < & \, c_4 F(t) < \infty. \end{aligned} \tag{A.10}$$

These conditions have been widely used and their high degree of generality was already pointed out in the functional nonparametric literature discussed before. The main reason for not discussing them here is that they are only needed in order to ensure the good asymptotic behaviour of the estimators $\hat{R}_{h,\alpha}$. In other words, in what remains of the paper, the set of conditions (A.4)-(A.10) could be changed into any other set of conditions ensuring both asymptotic normality and almost sure convergence of the estimates $\hat{R}_{h,\alpha}$.

These were conditions, for fixed α, to control the asymptotic behaviour of the estimate. Now, additional conditions are necessary for ensuring the asymptotic optimality of the data-driven procedure for selecting α. These are as follows: first of all, in order to preserve generality, we let the cardinality of the set of possible levels to grow up to infinity with the sample size N through the following condition:

$$\exists \tau > 0, card(\Theta_N) = O(N^\tau). \tag{A.11}$$

The weight function W satisfies, for some constants c_5 and c_6, the following usual conditions

$$0 < W(.) \leq c_5 < \infty, \text{ and } W(C) = 0 \text{ if } d_\lambda(C, \varnothing) > c_6. \tag{A.12}$$

All these previous assumptions are needed to obtain Theorem 1, while in order to obtain Corollary 2 one needs the following additional assumptions:

$$\alpha_0 \in \mathcal{A}_N, \text{ for } N \text{ large enough.} \tag{A.13}$$

and

$$\forall \epsilon > 0, \exists \eta > 0, \forall \alpha \in \mathcal{A}_n, |\alpha - \alpha_0| > \eta \Rightarrow E(R_\alpha(C) - R(C))^2 > \epsilon. \tag{A.14}$$

Condition A.13 just means that the set on which α is selected is sufficiently rich, while Condition A.14 insures that two different levels contain sufficiently different information on the underlying density.

B Asymptotic theory for the procedure of Section 3

Identifiability of the model

To ensure the identifiability of the model (3.1, we introduce additive versions of the usual uniqueness condition (A.1). We assume that there is some integer $J_0 \geq 1$ and some vector $(\alpha_0^1, \dots, \alpha_0^{J_0})$ such that if either $J \neq J_0$ or $(\alpha^1, \dots, \alpha^{J_0}) \neq (\alpha_0^1, \dots, \alpha_0^{J_0})$ then

$$\mathrm{Var}(g - \sum_{j=1}^{J_0} R^j_{\alpha_0^j}(C_{\alpha_0^j})) < \mathrm{Var}(g - \sum_{j=1}^{J} R^j_{\alpha^j}(C_{\alpha^j})) < \infty. \tag{B.1}$$

Of course, because of the additive structure, as long as J_0 satisfies (B.1) then any integer greater than J_0 does so, too. To avoid this problem, it is assumed that J_0 is the smallest integer such that (B.1) holds.

In the next result, one sets the consistency of the data-driven selected level defined in (3.3) towards the true theoretical optimal ones. Its proof consists in iterative using of the results stated before when one single level set had to be selected (see Corollary 2).

Asymptotic behaviour of the procedure

Theorem 3 *Consider the model defined by (B.1) and assume that the conditions (A.4)-(A.7), (A.10)-(A.12) hold. If in addition, (A.8) holds for each operator $R^j_{\alpha^j}$, if (A.9) holds for each bandwidth h^j and if (A.13) and (A.14) hold for any α_0^j, then one has*

$$\forall j,\ \hat{\alpha}^j \to \alpha_0^j, \textit{in probability} . \tag{B.2}$$

Corollary 4 *Under the conditions of Theorem 3, and if in addition (3.5) holds, then one has*

$$P\left[\hat{J} \geq J_0\right] = 1,\ \textit{for } N \textit{ large enough.} \tag{B.3}$$

C Asymptotic theory for the procedure in Section 6

Commented assumptions for Theorem 5

We may quantify these ideas by means of the following general assumptions:

$$\forall i = 1, \dots, N, ||f_{i,n} - f_i||^2 = o_p\left(\frac{1}{NF(h)}\right) \tag{C.1}$$

and

$$\forall i = 1, \dots, N, \forall 0 \leq \alpha \leq 1, d_\lambda(C_{\alpha,i,n}, C_{\alpha,i}) = o_p\left(\frac{1}{NF(h)}\right). \tag{C.2}$$

Of course, there is a doubly asymptotic problem to address (asymptotics on n and N). That means that, to give sense to both conditions before, one has to assume that:

$$n \equiv n_N \to \infty, \text{ as } N \to \infty. \tag{C.3}$$

Theorem 5 states precisely that the parameters estimated when using the sampled densities $f_{i,n}$ have asymptotic properties similar to those of the parameters that could have been theoretically estimated when using the true unknown densities f_i. However, before that, it is worth pointing out the high degree of generality of the conditions (C.1) and (C.2). The functional nonparametric literature (see, for instance, the book by [12]) provides evidence that in most cases one has $F(h) = O_p(h)$ (other cases being more of mathematical relevance than really linked to practical purposes), while the standard quadratic errors in functional literature (see, for instance, [11]) are known to be of the form

$$h^{2\beta} + \frac{1}{NF(h)}, \tag{C.4}$$

and so can be optimised (when $F(h) \sim Ch)$) and becomes (because $\beta \leq 1$)

$$N^{-\frac{2}{3}}. \tag{C.5}$$

On the other hand, the standard literature on bivariate nonparametric estimation shows that most density smoothers (kernel, splines, wavelets, etc.) may reach, if each density has for instance k continuous derivatives, the usual optimal rate of convergence

$$n^{-\frac{2k}{2k+1}}. \tag{C.6}$$

Looking at (C.5) and (C.6), it is easy to see that (C.1) is satisfied as soon as each density f_i has more than 1 derivative. In the same way, the literature on level sets estimation has established that under mild conditions $\lim_n d_\lambda(C_{\alpha,i,n}, C_{\alpha,i}) = 0$, almost surely or in probability, and furthermore rates of convergence have already been derived ([2], [3], [4], [18]). For instance, for bivariate density functions, [2] shows that the error of estimation has the rate

$$n^{-\eta}, \quad \text{for any } 0 < \eta < \kappa/(4 + 2\kappa), \tag{C.7}$$

when h is of exact order $(\log n/n)^{(1+\kappa)/(4+2\kappa)}$, where $\kappa > 0$ is a parameter controlling the steepness of the target density function f: assumption (F1) in [2] establishes that there exists an interval $[a, b] \subseteq (\inf_x f(x), \sup_x f(x))$ and a positive constant K such that for all c in $[a, b]$

$$P(|f(X) - c| < \epsilon) \leq K\epsilon^\kappa$$

when X is a random variable with density f. Intuition tells us that the steeper f is (large κ), the faster the rates will be. Assuming that $\kappa > 4$ and looking at (C.5) and (C.7) one also sees that (C.2) is satisfied.

Some asymptotics

In the next theorem one will see basically that, pending to the conditions discussed just before, all the procedures studied before in this paper have the same asymptotic properties in the situation of sample of estimated densities as those they have for samples of known densities.

Theorem 5 *Let us consider the model defined by (2.1) and (A.1) and denote by $\hat{\alpha}_n$ the probability content obtained by minimising (2.3) when changing $(f_i, C_{\alpha,i})$ into $(f_{i,n}, C_{\alpha,i,n})$. Under the conditions of Corollary 2, and if in addition (C.1), (C.2) and (C.3) hold, then we have:*

$$\hat{\alpha}_n \to \alpha_0, \textit{in probability as } N \to \infty. \tag{C.8}$$

Let us consider the model defined by (B.1), and denote by $\hat{\alpha}_n^j$ and $\hat{J}_n$ the parameters obtained by minimising (3.3) and (3.4) when changing $(f_i, C_{\alpha,i})$ into $(f_{i,n}, C_{\alpha,i,n})$. Under the conditions of Corollary 4, and if in addition (C.1), (C.2) and (C.3) hold, then we have:

$$i)\ \forall j,\ \hat{\alpha}_n^j \to \alpha_0^j, \textit{in probability as } N \to \infty, \tag{C.9}$$

and

$$ii)\ P\left[\hat{J}_n \geq J_0\right] = 1,\ \textit{for } N \textit{ large enough.} \tag{C.10}$$

D Proofs

Proof of Theorem 1

We present the proof in a very synthetic way, simply emphasising the specificities linked with the two main contributions in this paper, namely the functional nature of the response and the double simultaneous choice of h and α. For a fixed $x \in \mathbb{R}^2$, results in [19] ensure that the pointwise cross-validation criterion

$$\mathrm{CV}(h,\alpha)_x = \sum_{i=1}^{N} (g_i(x) - \hat{R}^{-i}_{h,\alpha}(C_{\alpha,i})(x))^2 W(C_{\alpha,i}),$$

has the same asymptotic behaviour (up to a constant term independent on h and α) as the pointwise theoretical error

$$Err(h,\alpha)_x = \sum_{i=1}^{N} (R(C_{\alpha,i})(x) - \hat{R}_{h,\alpha}(C_{\alpha,i})(x))^2 W(C_{\alpha,i}),$$

in the sense that, uniformly on h and α, one has

$$\begin{aligned} Err(h,\alpha)_x = \mathrm{CV}(h,\alpha)_x &+ \frac{1}{N}\sum_{i=1}^{N}(g_i(x) - R(C_{\alpha_0,i})(x))^2 W(C_{\alpha,i}) \\ &+ o(Err(h,\alpha)_x), \text{ a.s..} \end{aligned} \tag{D.1}$$

Note now that conditions (A.6) and (A.7) ensure that (D.1) is uniform on x, in such a way that we finally arrive, uniformly on h and α, at

$$Err(h,\alpha) = \mathrm{CV}(h,\alpha) + \frac{1}{N}\sum_{i=1}^{N} E_i^2 + o(Err(h,\alpha)), \text{ a.s.,} \tag{D.2}$$

with

$$E_i^2 = \int (g_i(x) - R(C_{\alpha_0,i})(x))^2 dx W(C_{\alpha,i}).$$

This is enough to obtain

$$\sup_{(h,\alpha,h',\alpha')\in\Theta_n^2} \left| \frac{(Err(h,\alpha) - Err(h',\alpha')) - (\mathrm{CV}(h,\alpha) - \mathrm{CV}(h',\alpha'))}{Err(h,\alpha)} \right| \to 0, \text{a.s..} \tag{D.3}$$

It now suffices to take $(h,\alpha) = (h_1,\alpha_1)$ and $(h',\alpha') = (\hat{h},\hat{\alpha})$ to arrive at the professed result (A.2).

Proof of Corollary 2

The conditions (A.13) and (A.14) allow us to see that one has:

$$\forall \epsilon > 0, \exists \eta > 0, |\alpha - \alpha_0| > \epsilon \Rightarrow Err(\hat{h},\alpha) > \eta.$$

By applying this with $\alpha = \hat{\alpha}$, we obtain

$$\forall \epsilon > 0, \exists \eta > 0, P\left[|\hat{\alpha} - \alpha_0| > \epsilon\right] \le P[Err(\hat{h},\hat{\alpha}) > \eta].$$

On the other hand, by using the equivalence stated in Theorem 1, one obtains

$$\forall \eta > 0, \exists \eta' > 0, P[Err(\hat{h},\hat{\alpha}) > \eta] \le P[Err(h_1,\alpha_1) > \eta'].$$

Finally, standard results in functional regression (see [11]) lead to

$$\forall \eta' > 0, P[Err(h_1,\alpha_1) > \eta'] \to 0.$$

The three last results are enough to prove Corollary 2.

Proof of Theorem 3

This proof is omitted, since it consists only in performing successive iterations of the proof of Corollary 2 stated before. Precisely one can state equivalence results for each direction $j = 1, \dots, J$ between the data-driven criterion CV^j and its theoretical counterpart Err^j defined as

$$Err^j(h, \alpha) = \sum_{i=1}^{N} ||Q_i^j - \hat{R}_{\hat{h}^{j-1}, \hat{\alpha}^{j-1}}(C_{\alpha,i})||^2 W(C_{\alpha,i}).$$

Then, by following the same lines as along Corollary 2 one could state that (B.2) holds for the parameters that would have been selected by minimizing the theoretical errors Err^j and, because of the above mentioned equivalences, that it holds also for the cross-validated parameters as claimed in Theorem 3.

Proof of Corollary 4

Because of (3.5) we can simply do the proof when $\lambda_N = 0$ without loss of generality. When iterating the proof of Theorem 1 the results (D.2) become:

$$\text{PCV}(J) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(g_i - \sum_{j=1}^{J} R^j_{\alpha_0^j}(C_{\alpha_0^j, i}) W(C_{\alpha_0^j, i}) \right) + o(1) \right), \text{ a.s..} \tag{D.4}$$

Finally, the condition (B.1) allows us to obtain

$$\forall J < J_0, \text{PCV}(J) > \text{PCV}(J_0), \text{ a.s.,}$$

and therefore one arrives at

$$\begin{aligned} P\left[\hat{J} < J_0\right] &= \sum_{J=1}^{J_0-1} P[\hat{J} = J] \\ &\leq \sum_{J=1}^{J_0-1} P[\text{PCV}(J) \leq \text{PCV}(J_0)] = 0. \end{aligned}$$

The proof of Theorem 3 is now complete.

Proof of Theorem 5

This proof follows line by line the proofs of Theorems 1 and 3. At each step one can change $(f_i, C_{\alpha,i})$ into $(f_{i,n}, C_{\alpha,i,n})$. All the additional errors terms appearing with this change can be seen to be negligible because of (C.1), (C.2) and (C.4).

Acknowledgements. Pedro Delicado is very grateful to Antonio Cuevas and Amparo Baíllo for helpful conversations on plug-in estimation of density level sets. He also acknowledges the financial support of the Spanish Ministry of Education and Science and FEDER (MTM2013-43992-R). Both authors are grateful to the Institut de Mathmatiques de Toulouse for having offered a Visiting Professor position to the first author, since a large part of this research was initiated during the time of this visit.

References

1. Aneiros-Pérez, G., Cao, R. and Vilar-Fernàndez, J. (2011). Functional methods for time series prediction: a nonparametric approach. *J. Forecast.*, **30**(4), 377-392.
2. Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statistics & Probability Letters*, **65**(4), 411-417.
3. Baíllo, A., Cuesta-Albertos, J., Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statistics & Probability Letters*, **53**(1), 27-35.
4. Cadre, B. (2006). Kernel estimation of density level sets. *J. of Multiv. Analysis*, **99**(4), 999-1023.
5. Choi, S. and Park, J. (2014). Nonparametric additive model with grouped lasso and maximizing area under the ROC curve. *Comput. Statist. Data Anal.*, **77**, 313-325.
6. Delicado, P. and Vieu, P. (2014). Optimal level sets for bivariate density representation. *J. of Multiv. Analysis*, **140**, 1-18.
7. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.
8. Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. of Statist.*, **32**, 928-961.
9. Ferraty, F., Hall, P. and Vieu, P. (2010). Most-predictive design points for functional data predictors. *Biometrika*, **97**(4), 807-824.
10. Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011). Kernel regression with functional response. *Electron. J. Stat.*, **5**, 159-171.
11. Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.*, **109**, 10-28.
12. Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer, New York.
13. Ferraty, F. and Vieu, P. (2009). Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.*, **53**(4), 1400-1413.
14. Goia, A. (2012). A functional linear model for time series prediction with exogenous variables. *Statist. Probab. Lett.*, **82**(5), 1005-1011.
15. Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statist.*, **13**(4), 1465-1481.
16. Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications.* Springer, New York.
17. Kudraszow, N. and Vieu, P. (2013). Uniform consistency of kNN regressors for functional variables. *Statist. Probab. Lett.*, **83**(8), 1863-1870.
18. Mason, D. and Polonik, W. (2009). Asymptotic normality of plug-in level sets estimates. *Annals of Applied Probability*, **19**(3), 1108-1142.
19. Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data: automatic smoothing parameter selection. *J. Statist. Plann. Inference*, **137**(9), 2784-2801.
20. Ramsay J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis.* Springer, New York.
21. Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis, 2nd ed.* Springer, New York.
22. Roberts, S. and Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Comput. Statist. Data Anal.*, **70**, 198-211.
23. Stone, C.J. (1985). Additive regression and their nonparametric models. *The Annals of Statist.*, **13**(2), 689-705.
24. Vieu, P. (1995). Order choice in nonlinear autoregressive models. *Statistics*, **26**(4), 307-328.