Journal of Multivariate Analysis 140 (2015) 1–18

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# 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

ARTICLE INFO

*Article history:*
Received 13 October 2014
Available online 17 April 2015

*AMS subject classifications:*
62-09
62G07

*Keywords:*
Bivariate density representation
Functional data analysis
Minimum distance estimation
Multidimensional scaling
Nonparametric density estimation

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.

## 1. Introduction

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

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

where $\gamma_\alpha$ is such that

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

When needed, we will write $C_\alpha^f$ to make explicit the dependence of $C_\alpha$ 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, [8] or [19] as well as the accompanying R packages `sm` and `ks`, respectively). Other authors [30,29,33,31] draw the density contour levels at equally spaced heights (see also the R package `KernSmoth`, associated with Wand and Jones [33]).

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)

[*] Corresponding author.
*E-mail addresses:* pedro.delicado@upc.edu (P. Delicado), philippe.vieu@math.univ-toulouse.fr (P. Vieu).

http://dx.doi.org/10.1016/j.jmva.2015.04.005
0047-259X/© 2015 Elsevier Inc. All rights reserved.

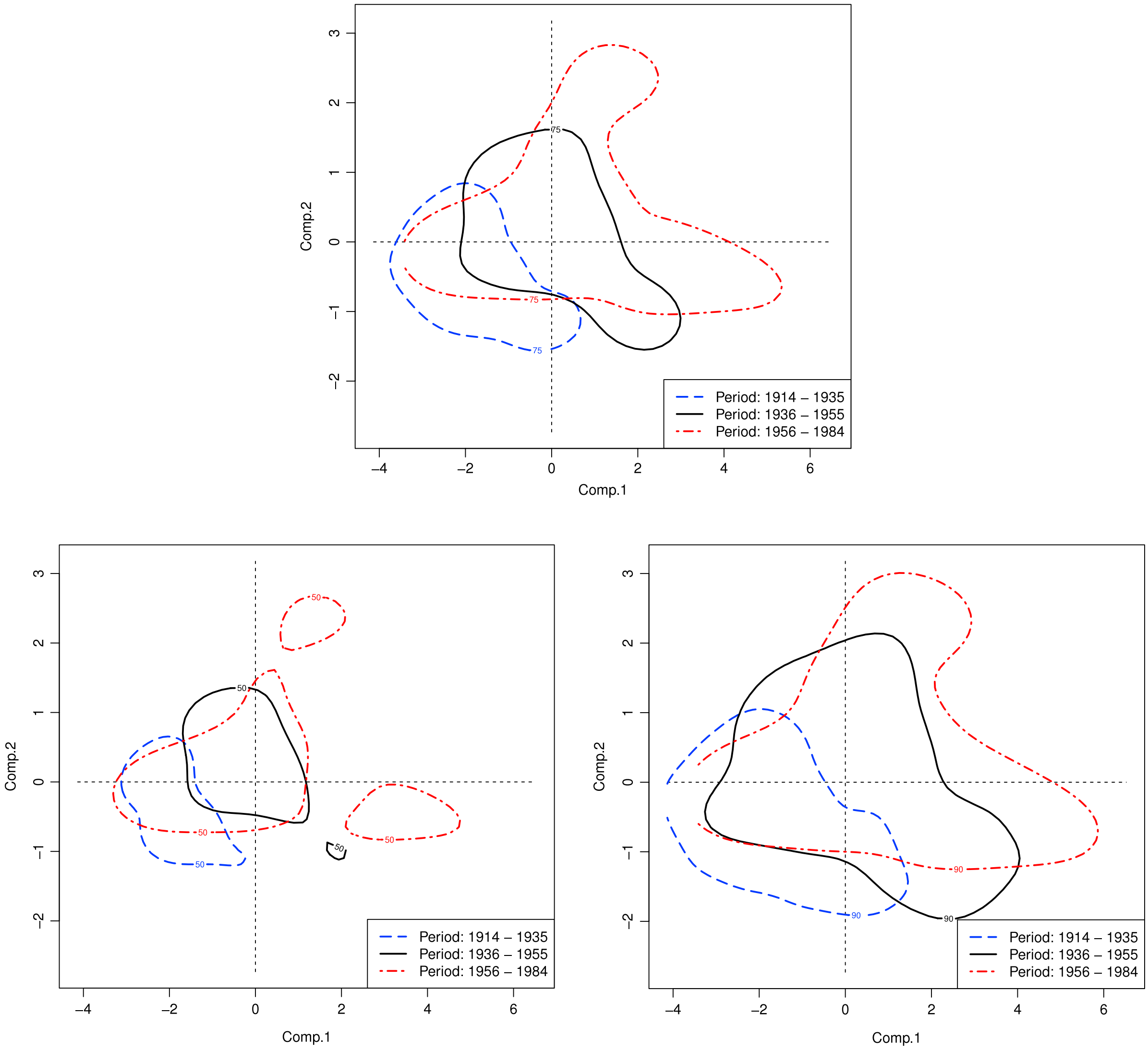

**Fig. 1.** Aircraft data. The estimated bivariate densities of the first two principal components are represented by density level sets. *Top panel:* Level sets with content 0.75 for three periods. *Bottom left panel:* Level sets with content 0.5 (optimal value according to (2.1) for $J = 1$). *Bottom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

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 [8]. They study data on aircraft designs from the periods 1914–1935, 1936–1955 and 1956–1984, originally explored in [9]. 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 Fig. 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