Contribute 4

Optimal level sets for representing a bivariate density function

Pedro Delicado, Philippe Vieu

Abstract We deal with the problem of representing a bivariate density function by level sets. The choice of which levels are used in this representation are commonly arbitrary (most usual choices being those with probability contents .25, .5 and .75). Choosing which level is (or which levels are) of most interest 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 approach we develop is based on minimum distance ideas.

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

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

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, [3] or [7] as well as the accompanying R packages `sm` and `ks`, respectively). Other authors ([12], [13], [14], [15]) draw the density contour levels at equally spaced heights (see also the R package `KernSmoth`, associated with [15]).

Pedro Delicado
Universitat Politècnica de Catalunya, Barcelona, Spain, e-mail: pedro.delciado@upc.edu

Philippe Vieu
Université Paul Sabatier, Toulouse, France, Italy, e-mail: philippe.vieu@math.univ-toulouse.fr

We consider the following problem: given a bivariate density function f, choose the combination of values J and $\alpha_1, \ldots, \alpha_J$ defining the *best* (in some sense) graphical representation of f. In some cases, the value of J can be fixed in advance; for instance, when only one level set is used to represent a density. The exact meaning of *best graphical representation* is specified later. 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. It can also be said that the *visual distance* between f and its graphical representation using the chosen density level sets must be minimised.

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. 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. 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).

This contribution will be centered around the theoretical properties of the optimal level sets defined in Section **4.1** (see Theorem 4**.1**) and on their finite sample behaviour (both on simulated and real data). In addition, one will also shortly discuss some alternative method for constructing level sets as well as some tracks for future researches as presented in Section **4.2**.

4.1 Optimal level sets for a bivariate density

We 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 explore the following approach: to choose 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: $C_{\alpha_1}, \ldots, C_{\alpha_J}$.

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

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

where Δ denotes the symmetric difference between sets, λ is the Lebesgue measure in $\mathbb{R}^2$ and μ_f is the probability measure in $\mathbb{R}^2$ having f as a density function. There exist other distances between sets that could be used as an alternative (Hausdorff's distance, for instance; for more details on distances between sets used in set estimation see, e.g., [5]).

An appealing way to choose values $\alpha_1, \ldots, \alpha_J$ is by solving this minimisation problem:

$$\min_{0 < \alpha_1 < \cdots < \alpha_J < 1} \int_0^1 d(C_u, C_{\alpha_{j(u)}}) du \tag{4.1}$$

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 4.1. *For $d = d_f$, the optimal solution to problem (4.1) 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 (4.1) is α_j^λ, $j = 1, \ldots, J$, such that

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

Observe that α_j^f, the optimal values when using $d = d_f$, do not depend on f. This is no longer true when using $d = d_\lambda$. 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 bivariate density f is 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 [5] and [6]. Of particular interest for us are the works of [2] and [4], which deal with the convergence of the plug-in density level estimating sets $C_{\alpha,n} = \{x \in \mathbb{R}^2 : \hat{f}_n(x) \geq \gamma_\alpha, n\}$, 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, [2] obtain rates of convergence for $\Pr\{Z \in C_{\alpha,n}\} - \alpha$, where $Z \sim f$ is independent of $\hat{f}_n$ (see [11], for similar results under weaker assumptions). [1] proves that $d_\lambda(C_{\alpha,n}, C_\alpha)$ converges almost surely to 0 while [4] finds the convergence rate.

4.2 Further researches

The problem of representing only one density by some of its density level sets admits a second approach. It can be argued 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 minimise in $\alpha_1, \ldots, \alpha_J$ the distance between this piecewise uniform density and the one we wish to represent by $C_{\alpha_1}, \ldots, C_{\alpha_J}$.

The situation in which one has a sample of bivariate densities available ($f_i, i = 1, \ldots, N$) that must be represented is also of interest. A way for attacking the problem could be to look for the link between the densities and their corresponding level sets. Because both of them are functional objects, the recent advances in FDA (see the books [8], [9] and [10]) would be helpful.

Bibliography

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

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

[3] A.W. Bowman and A. Azzalini (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

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

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

[6] A. Cuevas and R. Fraiman (2010) Set estimation. In W. Kendall and I. Molchanov, editors, *New Perspectives in Stochastic Geometry*, chapter 11, pages 374–397. Oxford University Press, Oxford.

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

[8] F. Ferraty and P. Vieu (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Verlag.

[9] L. Horváth and P. Kokoszka (2012) *Inference for functional data with applications*, Springer.

[10] J.O. Ramsay and B. W. Silverman (2005) *Functional Data Analysis*. Second edition Springer, New York.

[11] Q. Ren and M. Mojirsheibani (2008) Nonparametric estimation of level sets under minimal assumptions. *Statistics & Probability Letters*, 78:3029–3033.

[12] D.W. Scott (1992) *Multivariate Density Estimation*. John Wiley & Sons.

[13] B.W. Silverman (1986) *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman & Hall/CRC.

[14] J.S. Simonoff (1996) *Smoothing Methods in Statistics*. Springer Verlag.

[15] M.P. Wand and M.C. Jones (1995) *Kernel Smoothing*, volume 60. Chapman & Hall/CRC.