Second International Workshop on
Functional and Operatorial Statistics.
Santander, June 16-18, 2011

Dimensionality reduction for samples of bivariate density level sets: An application to electoral results

Pedro Delicado

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

pedro.delicado@upc.edu

Abstract

A bivariate density can be represented as a density level set containing a fixed amount of probability (0.75, for instance). Then a functional dataset where the observations are bivariate density functions can be analyzed as if the functional data are density level sets. We compute distances between sets and perform standard Multidimensional Scaling. This methodology is applied to analyze electoral results.

1. Introduction

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 for 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 parties voters usually participate in elections less than right parties voters. In this work we analyze the relationship between position on the left-right political dimension and the willingness to vote. Given that individual data are not available we use aggregated data at level of electoral districts (”mesas electorales” in Spanish: 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 2004 Spanish general elections.

For each electoral district the available information allows us to define these two variables: participation (proportion of potential voters that finally vote) and proportion of votes for right parties. Observe that this last variable is not exactly the same as the proportion of potential voters with right political ideology. Unfortunately we only know what is voting people that vote indeed. Nevertheless, if the size of the electoral district is small compared with the size of the city it is sensible to believe that both quantities should be similar. We assume that, given the electoral district, the political orientation (left-right) is independent from the decision of voting or not.

We consider the 50 cities in Spain with the bigger numbers of electoral districts (157 districts or more). For each of these cities we have a list of observations of the bivariate random variable (*participation*, *proportion of votes for right parties*), an observation for each electoral district. We use then a kernel density estimator to obtain from this list an estimation of the joint distribution of these two variables in each of the 50 cities considered in our study. Therefore we have a functional dataset of length 50 consisting on bivariate densities.

A preliminary dimensionality reduction step is usually very useful to perform the exploratory analysis of functional datasets. Given that the dataset we are considering consists on bivariate densities, it is possible to adapt the dimensionality reduction techniques considered in Delicado (2011) for functional datasets formed by unidimensional densities. Nevertheless we propose here an alternative way.

A bivariate density $f(x, y)$ is frequently represented by some of its *density level sets*, defined as $L(c) = \{(x, y) \in \mathbf{R}^2 : f(x, y) \geq c\}$, for $c > 0$, or just their boundaries, in a contour plot. Bowman and Azzalini (1997) propose to display only the contour level plots that contain specific probabilities (they use 0.25, 0.50 or 0.75, reminiscing a boxplot) as a effective way to characterize the shape of a bivariate density. The role of density level sets is also relevant in the area of set estimation (see Cuevas and Fraiman 2009).

Bowman and Azzalini (1997, Section 1.3) give a very nice illustration of the use of density level sets for exploratory data analysis. They study data on aircraft designs from periods 1914-1935, 1936-1955 and 1956-1984. They obtain the first two principal components and represent their joint density using a single level plot (that corresponding to probability 0.75) for each period. In a single graphic Bowman and Azzalini (1997, Figure 1.8) are able to summarize the way in which aircraft designs have changed over the last century.

We borrow this way to summarize a bivariate density (the density level plot corresponding to probability 0.75). Therefore our functional dataset is finally formed by 50 such density level sets. As an example Figure 1 shows the density level sets corresponding to the 5 largest municipalities in Spain, jointly with the density level set corresponding to the whole country as a reference. The standard correlation coefficient for each case has been annotated. It is clear that there is a considerable variability between different level sets. Moreover the relationship between participation and vote orientation is clearer when considering homogeneous sets of electoral districts (those corresponding to a specific city) than when considering the whole country (top left panel).

3. Multidimensional Scaling for density level datasets

The functional data we are analyzing are sets (density level sets). When looking for a dimensionality reduction technique for this kind of data it is much more natural to turn

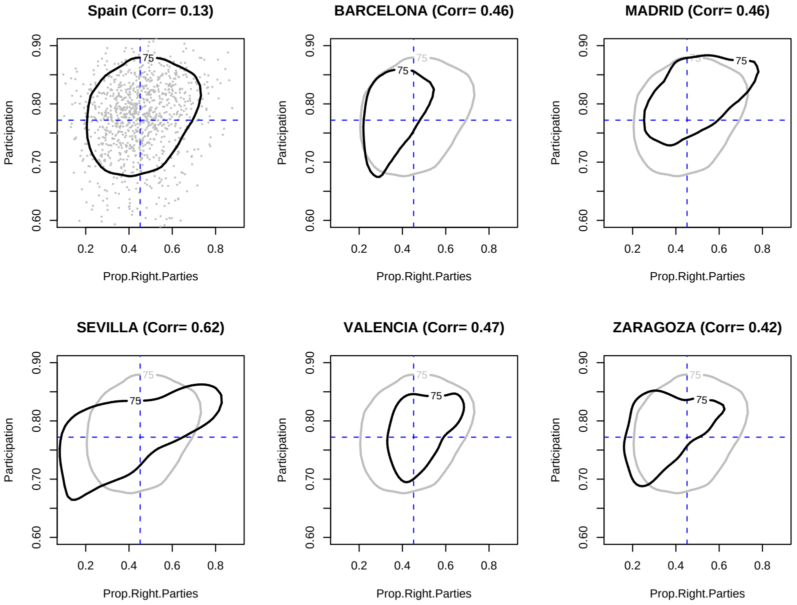

Figure 1: Example of 6 density level sets.

to Multidimensional Scaling (MDS) than to some kind of Principal Component Analysis (PCA). The main reason is that there exist several well known definitions of distance between sets but there is not a clear Hilbert space structure on the set of sets allowing to define PCA for datasets of sets.

Let us revise the essentials of Multidimensional Scaling (for more details see Borg and Groenen 2005, for instance). This dimensionality reduction technique requires an inter-individual distance matrix (or a dissimilarity matrix) as the only information about data. Assume that there are n individuals and that Δ is the $n \times n$ matrix with element (i, j) equal to $d_{ij} \geq 0$, the dissimilarity between individuals i and j. Assume that for $q \leq n$ there exists a $n \times q$ data matrix X such that the Euclidean distance between the i-th and j-th rows of X is d_{ij}. We say that X is a *Euclidean configuration* of Δ. Such a configuration does not always exist. When it does, Δ is said to be *Euclidean*. In this case X can be chosen by having orthogonal columns known as *principal coordinates*.

Define the $n \times n$ matrix D with element (i, j) equal to d_{ij}^2. It can be proved that Δ is Euclidean if and only if $Q = -(1/2)PDP$ is positive definite, where $P = I - (1/n)\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is a centering matrix ($\mathbf{1}$ is the $n \times 1$ vector of ones). In this case, let $Q = V\Lambda V^{\mathrm{T}}$ be the spectral decomposition of Q (V is an $n \times n$ orthonormal matrix, and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$). Let $\tilde{X}_q = V_q \Lambda_q^{1/2}$, where V_q is formed by the first q columns of V and $\Lambda_q = diag(\lambda_1, \ldots, \lambda_q)$. Then $Q \approx \tilde{X}_q \tilde{X}_q^{\mathrm{T}}$ and $\tilde{X}_q$ is a q-dimensional approximate Euclidean configuration of Δ. Dimensionality reduction usually aims at visualizing data, which requires a plane representation, and implies $q = 2$.

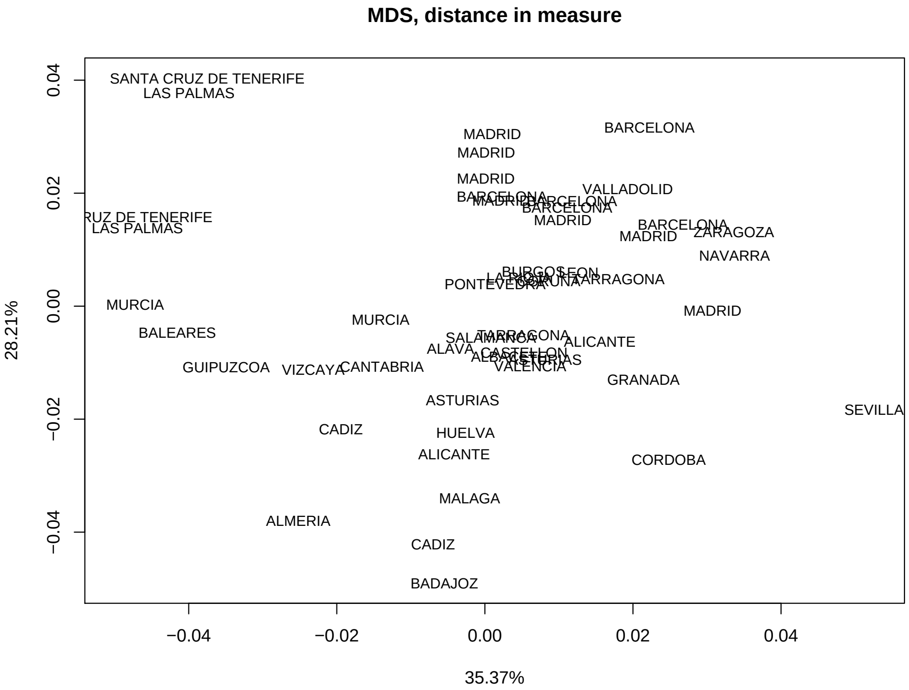

Figure 2: Plane of the first two principle coordinates obtained from the MDS.

Two distances between sets used frequently (see Cuevas 2009, for instance) are:
Distance in measure: Given $U, V \subseteq \mathbf{R}^2$, $d_\mu(U,V) = \mu(U \,\Delta\, V)$, where $U \,\Delta\, V = (U \cup V) - (U \cap V)$ is the symmetric difference of U and V, and μ is the Lebesgue measure in $\mathbf{R}^2$.
Hausdorff metric: Given $U, V \subseteq \mathbf{R}^2$, $d_H(U,V) = \inf\{\varepsilon > 0 : U \subseteq B(V,\varepsilon), V \subseteq B(U,\varepsilon)\}$, where for $A \subseteq \mathbf{R}^2$, $B(A,\varepsilon) = \cup_{x \in A} B(x,\varepsilon)$, and $B(x,\varepsilon)$ is the closed ball with center x and radius ε in $\mathbf{R}^2$.

Distance in measure is easily computed from the output of standard two dimensional density estimation routines (we use the R library sm, accompanying the book of Bowman and Azzalini, 1997). The computation of Hausdorff metric is not so direct. This is the reason why in this work we use distance in measure between density level sets. Once the distance matrix is calculated the MDS procedure follows in a standard way.

2. Analyzing electoral behavior

Figure 2 represents the plane of the first two principle coordinates obtained from the

MDS analysis of the distance in measure matrix between the 50 density level sets in our study. The labels used in this graphic indicate the province where the 50 big cities are placed (observe that some of them belong to the same province). The percentage of variability explained by these two principal coordinates is around 60%, so it could be interesting to explore additional dimensions. There is not any nonlinearity pattern neither clustering structure.

In order to have a better interpretation of these first two principle coordinates additional graphics are helpful. Jones and Rice (1992) propose the following way to represent functional principal coordinates (or principal components). They suggest picking just three functional data in the dataset: the data corresponding to the median principal coordinate score, and those corresponding to quantiles α and $(1-\alpha)$ of these score values (α close to zero guarantees that these functional data are representative of extreme values of principal component scores). Alternatively, functional data corresponding to the minimum and maximum scores could go with the median score functional data. This is exactly what we represent in Figure 3, using blue color for the minimum, black color for the median and red color for the maximum.

The first principal coordinate goes from negative relationship between participation and proportion of votes to right parties (a city in the province of Santa Cruz de Tenerife) to almost independence (a city in the province of Barcelona) to a positive relationship (Sevilla). The interpretation of the second principal coordinate is not so clear. We observe that the area of the density level sets decreases when moving from the minimum scores (Badajoz) to the maximum (a city in the province of Santa Cruz de Tenerife, different from that cited when talking about the first principal coordinate), but a deeper analysis should be done in order to establish a clearer interpretation.

References

[1] Borg, I., Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. Springer-Verlag, New York (2005).

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

[3] Cuevas, A. and Fraiman, R. Set estimation. In: *New Perspectives in Stochastic Geometry*, W. Kendall and I. Molchanov, eds. Oxford University Press, Oxford (2009).

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

[5] Delicado, P. Dimensionality reduction when data are density functions, *Computational Statistics and Data Analysis*, **55** (1), 401-420 (2011).

[6] Jones, M.C., Rice, J.A. Displaying the important features of large collections of similar curves. *The American Statistician*, **46** (2), 140-145 (1992).

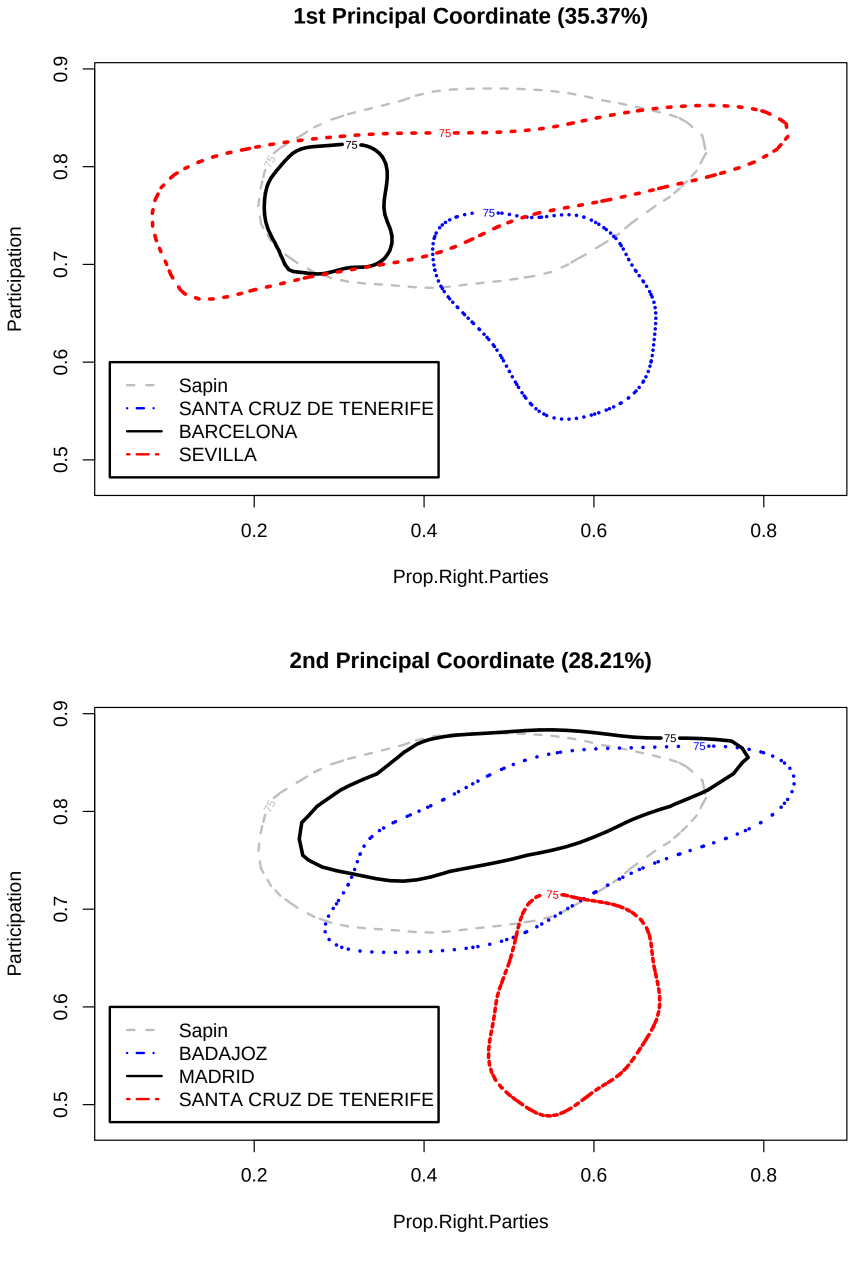

Figure 3: Helping to the interpretation of the first two principle coordinates.