UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 10-35
Statistics and Econometrics Series 19
September 2010

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

SENSITIVITY AND ROBUSTNESS IN MDS CONFIGURATIONS FOR MIXED-TYPE DATA: A STUDY OF THE ECONOMIC CRISIS IMPACT ON SOCIALLY VULNERABLE SPANISH PEOPLE

Aurea Grané and Rosario Romera

Abstract:

Multidimensional scaling (MDS) techniques are initially proposed to produce pictorial representations of distance, dissimilarity or proximity data. Sensitivity and robustness assessment of multivariate methods is essential if inferences are to be drawn from the analysis. To our knowledge, the literature related to MDS for mixed-type data, including variables measured at continuous level besides categorical ones, is quite scarce. The main motivation of this work was to analyze the stability and robustness of MDS configurations as an extension of a previous study on a real data set, coming from a panel-type analysis designed to assess the economic crisis impact on Spanish people who were in situations of high risk of being socially excluded. The main contributions of the paper on the treatment of MDS configurations for mixed-type data are: (i) to propose a joint metric based on distance matrices computed for continuous, multi-scale categorical and/or binary variables, (ii) to introduce a systematic analysis on the sensitivity of MDS configurations and (iii) to present a systematic search for robustness and identification of outliers through a new procedure based on geometric variability notions.

Keywords: Gower Distance, MDS configurations, Mixed-type Data, Outliers Identification, Related Metric Scaling, Survey Data.

AMS subject classification: 62-07, 62-09, 62F35, 62F40, 62P25.

Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. E-mail: A. Grané, aurea.grane@uc3m.es (Corresponding author), R. Romera, rosario.romera@uc3m.es
This work has been partially supported by Spanish grants MTM2009-13985-C02-01 (Spanish Ministry of Science and Innovation), SEJ2007-64500 (Spanish Ministry of Education) and S2007/HUM-0413 (Comunidad de Madrid).

Sensitivity and robustness in MDS configurations for mixed-type data: A study of the economic crisis impact on socially vulnerable Spanish people

Aurea Grané Rosario Romera

Statistics Department. Universidad Carlos III de Madrid

Abstract

Multidimensional scaling (MDS) techniques are initially proposed to produce pictorial representations of distance, dissimilarity or proximity data. Sensitivity and robustness assessment of multivariate methods is essential if inferences are to be drawn from the analysis. To our knowledge, the literature related to MDS for mixed-type data, including variables measured at continuous level besides categorical ones, is quite scarce. The main motivation of this work was to analyze the stability and robustness of MDS configurations as an extension of a previous study on a real data set coming from a panel-type analysis designed to assess the economic crisis impact on Spanish people who were in situations of high risk of being socially excluded. The main contributions of the paper on the treatment of MDS configurations for mixed-type data are: (i) to propose a joint metric based on distance matrices computed for continuous, multi-scale categorical and/or binary variables, (ii) to introduce a systematic analysis on the sensitivity of MDS configurations and (iii) to present a systematic search for robustness and identification of outliers through a new procedure based on geometric variability notions.

AMS subject classification: 62-07, 62-09, 62F35, 62F40, 62P25

Keywords: Gower Distance, MDS configurations, Mixed-type Data, Outliers Identification, Related Metric Scaling, Survey Data.

1 Introduction

The techniques collectively known as multidimensional scaling (MDS) have the purpose of constructing a set of points in a Euclidean space whose interdistances are either equal (metric or classical MDS) or approximately equal (nonmetric MDS) to

Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Spain.
E–mail: A. Grané, aurea.grane@uc3m.es; R. Romera, rosario.romera@uc3m.es.
This work has been partially supported by Spanish grants MTM2009-13985-C02-01 (Spanish Ministry of Science and Innovation), SEJ2007-64500 (Spanish Ministry of Education) and S2007/HUM-0413 (Comunidad de Madrid).
Date: July 23, 2010.

those in a given matrix of dissimilarities, in such a way that the interpoint distances approximate the interobject dissimilarities as closely as possible. MDS original purpose was as a descriptive tool, to visualize such distance data with low-dimensional pictorial representation. Firstly, these techniques are an essential and powerful tool for the representation of stimulus-attribute relationships everywhere in the behavioural sciences (see Piccarella and Lior 2007 for a recent application to sequence analysis to describe life course trajectories); Secondly, they pose a number of challenges to current statistical theory (for an interesting survey on this issue we refer the reader to Ramsay 1982). Various possible measures of approximation between interpoint distances and interobject dissimilarities can be used, each resulting in a different MDS configuration. General context references are Borg and Groenen (2005), Cox and Cox (2000) and Krzanowski and Marriott (1994) as well as Gower and Hand (1996).

In the MDS framework it is quite natural to handle proximity measures treating both qualitative and quantitative variables. This feature is of primary interest in data collected through questionnaires, since data are often of mixed type obtained as measures of variables at different levels, e.g. continuous, nominal, ordinal or ratio-scale. The joint treatment of quantitative and qualitative variables can be achieved if dissimilarities are suitable defined (see Ramsay 1980). The well-known Gower's general similarity coefficient (see Gower 1971) considers mixtures of numerical continuous, categorical and binary variables, but does not consider ordinal variables. For several recent versions of Gower's coefficient covering ordinal variables we refer the reader to Cox and Cox (2000) and references therein. Nevertheless, the additive treatment of the variables of Gower's based similarity coefficients results in a lack of consideration of the association between variables. In this paper we contribute in this line proposing a joint metric constructed via *related metric scaling* from three different distance matrices computed on continuous, multi-scale categorical and binary variables, respectively.

As Krzanowski (2006) pointed out the results of any data analysis require additional information about the stability of the solution and the prime interest in MDS is in assessment of stability of the points in the MDS configuration. When a distributional model is not available, as occurs in most cases, then either bootstrapping or crossvalidation are methods commonly employed to search for stability. Bootstrap technique views the observed data as a sample from some unknown underlying distribution, and provides a sampling framework for analyzing the stability of the sample statistics by using repeated sampling with replacement from the observed data. Crossvalidation technique computes the variability of the calculated statistics by leaving out of the computations different portions of the observed data. The most common case is the leave-one-out crossvalidation, i.e. when each individual is omitted in turn from the data. Thus, in MSD we need to compare configurations obtained either from different bootstrap samples or from different omissions of the crossvalidation procedure. A problem arise when any individuals are omitted from the dissimilarity matrix, thus the corresponding points will be missing in the MDS configuration and to compare the ensemble of configurations to quantify the variability of points can be a difficult task. DeLeeuw and Meulman (1986) have implemented a viable solution based on leave-one-out crossvalidation for tackling sensitivity issues in MDS, and there is not much literature on this matter to date. Solaro (2010) presents an interesting appli-

cation to customer satisfaction which is based on the latter approach for variables measured at different levels, e.g. nominal, ordinal or ratio-scale.
Despite a sensitivity study can also reveal possible influences of "abnormal" observations in a multivariate data analysis, there is a body of specific tools developed to search for multivariate outliers in data analysis. Outliers are observations that appear to break the pattern or grouping shown by the majority of the observations. Most conventional multivariate methods are sensitive to outliers due to the fact that they are based on least squares or similar criteria where even one outlier can deteriorate the model. Therefore, it is important to, firstly, identify outliers and, secondly, decide whether the outliers should be accomodated or rejected in the modeling process. On the other hand, the aim of robust methods is to reduce or remove the effect of outlying data points and let the remainder to built the desired results. Robust methods provide a powerful methodology extending a conventional "manual" analysis and elimination of outliers by using "conventional" outliers diagnosis. Most of the robust methods are developed for continuous variables and they make extensively use of robust estimates of location and covariance, i.e., M-estimators, Stahel-Donoho estimator, multivariate trimming, minimum volume-estimator or S-estimators. For an overview of *multivariate outliers detection and robustness* we refer the interested reader to Hubert, Rousseeuw, and Vanden Branden (2005) and references therein. For outliers detection in high-dimensional multivariate data see also Peña and Prieto (2007). The latter approach has been successfully applied for outliers detection in a regression context (González, Peña, and Romera 2009). Nevertheless, there is a lack of suitable robust techniques dealing with mixed-type data that can be applied in the context of MDS, and this is in fact the primary interest of this paper. We propose a procedure to search for robustness of the different MDS configurations considered for mixed-type data. In addition, we propose a test to identify multivariate outliers in the context of mixed-type data, depending on a distance-based quantity, closely related to the concept of geometric variability. The main contributions of the paper on the treatment of MDS configurations for mixed-type data are: (i) to propose a joint metric based on distance matrices computed for continuous, multi-scale categorical and/or binary variables, (ii) to introduce a systematic analysis on the sensitivity of MDS configurations and (iii) to present a systematic search for robustness and identification of outliers through a new procedure based on geometric variability notions.
A primary motivation for this study was to find common characteristics of the participants of several social programs carried out throughout Spain attending to some continuous and categorical variables, that contain information of the incidence of the 2008 economic crisis on people that are under risk of being socially excluded. Since the information is of different type, we are interested in computing distance matrices for mixed-type data (using, Gower's similarity coefficient) or either combining different distance matrices, avoiding redundant information (related metric scaling). Through leave-one-out crossvalidation procedures we analyze their stability and robustness in four contaminations of the original data set. We conclude that, in case of having mixed-type data with more than one continuous variable, MDS configurations obtained via related metric scaling are preferable to those computed from Gower's metric. The four contaminated data sets are also used to study the effectiveness of the proposed test to identify multivariate outliers in the context of mixed-type data.

The paper is organized as follows: In Section 2 we review the principal characteristics of the employed methods. In Section 3 we describe leave-one-out crossvalidation procedures to study the sensitivity and robustness of these configurations and in Section 4 we apply these methodologies to a real data set. In Section 5 we study the test to identify multivariate outliers in the context of mixed-type data and we conclude in Section 6.

2 Multidimensional scaling for mixed-type data

In this section we review the main properties of the employed methods. Given n p-dimensional vectors $\{\mathbf{z}_i, 1 \leq i \leq n\}$ containing the information of n individuals, we compute a distance matrix $\mathbf{D}$, with entries $d(\mathbf{z}_i, \mathbf{z}_j)$, for $1 \leq i, j \leq n$. Since this information can be either of qualitative or quantitative nature, or both, it is crucial the adequacy of the dissimilarity function used in the computation of $\mathbf{D}$. In this work we will discuss two alternative ways of obtaining $\mathbf{D}$ when the data is of mixed-type. Suppose that we are interested in obtaining a principal coordinate representation for the set of n individuals, provided that $\mathbf{D}$ satisfies the Euclidean property. Let $\mathbf{D}^{(2)}$ be the matrix of squared distances and consider the double-centered inner product matrix

$$\mathbf{G} = -\frac{1}{2}\mathbf{H}\,\mathbf{D}^{(2)}\,\mathbf{H},$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\,\mathbf{1}'$ is the $n \times n$ centering matrix, $\mathbf{I}$ is the identity matrix of order n and $\mathbf{1}$ is the $n \times 1$ vector of ones. The Euclidean requirement is equivalent to the positive semi-definiteness of $\mathbf{G}$, hence to the existence of an $\mathbf{X}$ such that $\mathbf{G} = \mathbf{X}\,\mathbf{X}'$, called in this context a *centered Euclidean configuration* of $\mathbf{D}$, meaning that $\mathbf{1}'\,\mathbf{X} = 0$ and that the squared Euclidean interdistances $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ between the rows $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $\mathbf{X}$ coincide with the corresponding entries in $\mathbf{D}$. The verification of this equivalence involves some simple algebra for which we refer to any standard textbook on Multidimensional Scaling, such as Borg and Groenen (2005).
The principal coordinate representation (or classical metric scaling) is obtained when matrix $\mathbf{X}$ is computed through the spectral decomposition of $\mathbf{G}$, that is:

$$\mathbf{X} = \mathbf{T}\,\mathbf{\Lambda}^{1/2},$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\mathbf{G}$, ordered in decreasing order, and $\mathbf{T}$ is the matrix whose columns are the corresponding eigenvectors. The number of nonzero eigenvalues will be at most $n-1$, and in most practical cases these eigenvalues will be positive (Krzanowski 2006). If some of them are negative, then $\mathbf{D}$ is non-Euclidean and some of the axes in the representation are imaginary. In this case, a possible solution is to consider the transformation $\tilde{\mathbf{D}}^{(2)} = \mathbf{D}^{(2)} + c\,(\mathbf{1}\,\mathbf{1}' - \mathbf{I})$, where $c \geq 2|\lambda|$ and λ is the negative eigenvalue of maximum module, which assures an Euclidean configuration for $\tilde{\mathbf{D}}^{(2)}$.
In the following we describe two techniques for computing either $\mathbf{D}$ or $\mathbf{G}$ when data is of mixed-type.

2.1 Gower's general similarity coefficient

A very popular metric for mixtures of numerical continuous, multi-state categorical and binary variables is the one based on Gower's general similarity coefficient (Gower 1971), which for two p-dimensional vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ is equal to

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |z_{ih} - z_{jh}|/R_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \tag{1}$$

where $p = p_1 + p_2 + p_3$, p_1 is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, α is the number of matches for the p_3 multi-state categorical variables, and R_h is the range of the h-th continuous variable. The entries of matrix $\mathbf{D}^{(2)}$ are computed as

$$d^2(\mathbf{z}_i, \mathbf{z}_j) = 1 - s_{ij}. \tag{2}$$

Gower (1971) proved that (2) satisfies the Euclidean requirement.

2.2 Related metric scaling

Like all distance functions satisfying additivity with respect to variables, the distance based on Gower's general similarity coefficient implicitly ignores any association (e.g. correlation) between variables (Gower 1992, Krzanowski 1994). On the other hand, *related metric scaling* (Cuadras and Fortiana 1998) is a multivariate technique that allows to obtain a unique representation of a set of individuals from several distance matrices computed on the same set of individuals. The method is based on the construction of a joint metric that satisfies several axioms related to the property of identifying and discarding redundant or repeated information.

Given a set of $m \geq 2$ matrices of squared distances measured on the same group of n individuals, $\{\mathbf{D}_\alpha^{(2)}\}_{\alpha=1,\ldots,m}$, the first requirement in the construction of the joint metric is that all matrices $\mathbf{D}_\alpha$ have the same *geometric variability*. This concept was introduced by Cuadras and Fortiana (1995) as a variant of Rao's diversity coefficient (Rao 1982a, Rao 1982b) and, given a distance matrix $\mathbf{D}_\alpha$, its sample version is:

$$V_{D_\alpha} = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_\alpha^2(\mathbf{z}_i, \mathbf{z}_j). \tag{3}$$

They proved that if $\mathbf{X}_\alpha$ is an Euclidean configuration of $\mathbf{D}_\alpha$, i.e., a principal coordinate representation, then the total variability of $\mathbf{X}_\alpha$, that is the trace of the covariance matrix of $\mathbf{X}_\alpha$ coincides with the geometric variability of $\mathbf{D}_\alpha$. Note that the condition of equal geometric variability can always be assumed to hold, since multiplying a distance matrix by an appropriate constant amounts to a change of measurement unit.

For each distance matrix $\{\mathbf{D}_\alpha^{(2)}\}_{\alpha=1,\ldots,m}$, we consider its doubly-centered inner product matrix

$$\mathbf{G}_\alpha = -\frac{1}{2} \mathbf{H}\, \mathbf{D}_\alpha^{(2)}\, \mathbf{H}, \quad \text{for } \alpha = 1, \ldots, m,$$

and obtain the joint metric as that whose doubly-centered inner product matrix is:

$$\mathbf{G} = \sum_{\alpha=1}^{m} \mathbf{G}_\alpha - \frac{1}{m} \sum_{\alpha \neq \beta} \mathbf{G}_\alpha^{1/2}\, \mathbf{G}_\beta^{1/2}. \tag{4}$$

Note that the principal coordinates are computed directly from matrix (4). In case it is necessary, we can recover matrix $\mathbf{D}^{(2)}$ with the following formula:

$$\mathbf{D}^{(2)} = \mathbf{g}\,\mathbf{1}' + \mathbf{1}\,\mathbf{g}' - 2\,\mathbf{G}, \tag{5}$$

where $\mathbf{g} = diag(\mathbf{G})$.
Based on these ideas, in Section 4, we construct the joint metric from $m = 3$ different distance matrices: the first one for continuous variables, the second one for multi-state categorical variables and the last one for binary variables.

3 Sensitivity and robustness of MDS configurations

Krzanowski (2006) proposed a leave-one-out crossvalidation procedure, based on the method by DeLeeuw and Meulman (1986), in order to study the sensitivity of MDS configurations, and illustrated it with some biometric examples. In this Section we review this methodology with the aim to compare the stability and robustness of the MDS configurations proposed in Sections 2.1 and 2.2. A real data set application can be found in Section 4.

3.1 Sensitivity analysis

We start by reviewing the method proposed by Krzanowski (2006). Let $\mathbf{D}^{(2)}$ be a matrix of squared distances computed on n individuals, $\mathbf{G}$ the corresponding doubly-centered inner product matrix, $\mathbf{X} = \mathbf{T}\,\mathbf{\Lambda}^{1/2}$ the principal coordinates, where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{G}$, ordered in decreasing order and $\mathbf{T}$ is the matrix of the corresponding eigenvectors. Let us call *initial configuration* to the n-point Euclidean configuration given by $\mathbf{X}$.
Suppose that we are interested in evaluating the influence of the i-th individual on the other $n-1$ individuals, in the sense that how the exclusion of the i-th individual from the original data set can affect the Euclidean coordinates of the $n-1$ points. For this purpose, we start by deleting the i-th individual from the original data set, then we compute anew the matrix of squared distances between the other $n-1$ individuals and finally obtain the principal coordinates. We denote with the subindex (i) the previous matrices without the i-th individual. Once the principal coordinates $\mathbf{X}_{(i)}$ are computed, we project the i-th point on the $(n-1)$-Euclidean configuration using Gower's interpolation formula (Gower 1968):

$$\mathbf{x} = \frac{1}{2}\,\mathbf{\Lambda}_{(i)}^{-1}\,\mathbf{X}_{(i)}'\,(\mathbf{g}_{(i)} - \mathbf{d})',$$

where $\mathbf{g}_{(i)} = diag\,\mathbf{G}_{(i)}$ and $\mathbf{d}$ is the $(n-1)\times 1$ vector containing the squared distances between the i-th individual and the other $n-1$ individuals. Finally, in order to compare this "augmented" configuration, $\mathbf{X}_{(i)}^{\star}$, with the initial one, it is necessary to assure that both configurations are correctly aligned.
The crossvalidation assessment of configuration stability follows directly by repeating the above process and leaving out each individual (each row of $\mathbf{X}$) in turn, so that we can have n "augmented" configurations, $\{\mathbf{X}_{(i)}^{\star}, i = 1, \ldots, n\}$, that are compared with $\mathbf{X}$ by superimposing them, provided that they are correctly aligned.

When n is moderately large, the $n(n+1)$ points may overload the diagram. Hence, in this case Krzanowski (2006) proposes to surround each point with the smallest hypersphere that contains a given percentage (for example, 90% or 95%) of the cross-validatory replicate points. The radii of these hyperspehers are computed as the appropriate quantiles of the Euclidean distances between the original coordinates of the points and the coordinates of their replicates. Hence, small hyperpheres indicate a very stable point, whereas large ones a very unstable one.

3.2 Robustness analysis

A usual way to proceed in studying the robustness of a method is to compare the results of its application in two different scenarios: the first one involves the original (raw or simulated) data set, whereas in the second one, a percentage of outliers is artificially introduced in the original data set. In this work we are interested in comparing the robustness of the MDS configurations obtained from mixed-type data using the metrics proposed in Sections 2.1 and 2.2. For this purpose, we consider different contaminations of the original data set (see Section 4 below) and, for each contamination and each MDS configuration, we analyze the 95%-stability regions through leave-one-out crossvalidation procedures.
Most of the recent robust techniques are developed for high-dimensional multivariate data, which is not the case for example, for data collected through questionnaires. We focus primary on this type of data and we proceed as follows.
In the application of Section 4, the joint metric is computed from three different distance matrices: For multi-state categorical variables we consider Sokal-Michener's similarity coefficient and for binary variables Jaccard's similarity coefficient. For the continuous variables we construct a robust estimation of Mahalanobis' distance that consists of estimating the entries in the covariance matrix in the following way: the variance of the j-th continuous variable is estimated from a 5%-trimmed sample, as suggested by Tuckey (1960). Gnanadesikan (1997) proposes a simple idea for estimating the covariance between two variables, Z_j and Z_k, based on the identity

$$cov(Z_j, Z_k) = \frac{1}{4}\left(var(Z_j + Z_k) + var(Z_j - Z_k)\right).$$

Then, one robust estimator for the covariance between Z_j and Z_k may be obtained from

$$s^*_{jk} = \frac{1}{4}(\hat{\sigma}^{*2}_{+} - \hat{\sigma}^{*2}_{-}),$$

where $\hat{\sigma}^{*2}_{+}$ and $\hat{\sigma}^{*2}_{-}$ are robust estimators of the variances of $Z_j + Z_k$ and $Z_j - Z_k$, respectively, and may be obtained by the method mentioned above.
Nevertheless, our method can be extended by replacing the robust distance used in the application, i.e. the Mahalanobis' robust distance, by any other robust distance applied to the continuous variables.

4 An application to real data

In this Section we apply the techniques described in Section 2 and Section 3 to a real data set obtained from a collaborative project with a NGO. A social questionnaire

is administered to the people attended by this institution, who are in situations of greater fragility. It measures their situation in regards to 5 fields: personal, familial, economic, social, and environmental/housing. Depending on the results, the participants interviewed are situated in ranges of moderate, high, very high or extreme risk. The average of these results is a comprehensive indicator of vulnerability. In November 2008, this institution decided to set up a panel-type study in order to observe the rate of the economic crisis on people in situations of high risk, who participated in social programs throughout the country. The first wave of interviews was finished in May 2009 and the second one in October 2009. The findings of the second wave of interviews shed light on its incidence among the most vulnerable segments, which possess the least number of resources to develop capacities to confront the crisis. A comparative analysis of the results of the first wave, which was conducted during the month of May 2009, makes it possible for us to establish comparisons and to determine the direction of the evolution of this process. The main findings of this study are in a Technical Report (in Spanish and English), available at http://www.sobrevulnerables.es/sobrevulnerables/boletines.do?method=inicio .

4.1 Description of the data set

In this work we focuss only on some of the responses of $n = 438$ individuals that participated all along the panel-type study. The participants are distributed in several social programs:

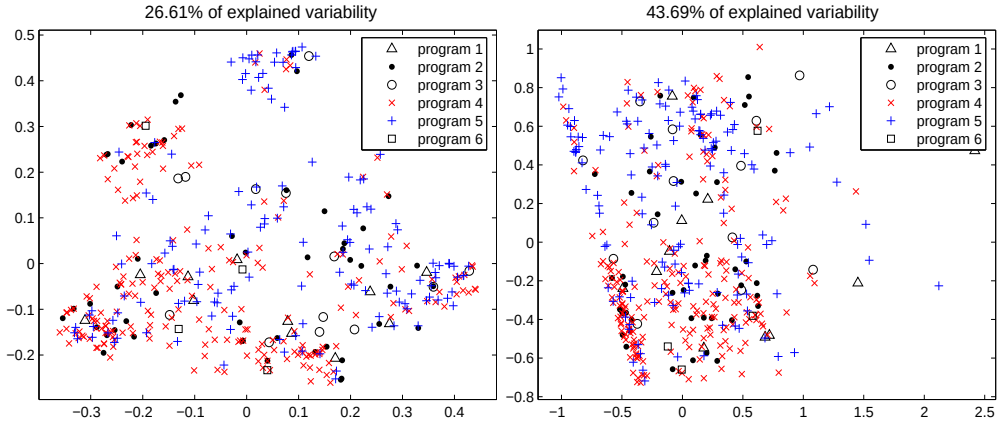
Social programs under study:	
1. Drug Dependency Assistance	4. Immigrants and Refugees
2. Fight Against Poverty and Social Exclusion	5. Aged People
3. Assistance for Disabled People	6. Inmates and Former Inmates

For the sake of simplicity and without loss of generality, we consider the information related to the following eight mixed-type variables:

type	description	values/categories (% frequency distribution)
continuous	monthly income (in euros)	from 0 to 3000 euros
continuous	age (in years)	from 20 to 94
binary	housing problems	no (22.1), yes (77.9)
binary	sex	male (23.1), female (76.9)
categorical	employment status	employed (27.4), autonomous (1.4), unemployed (29.5), retired (27.9), has never worked (13.8)
categorical	how often they receive substantial economic aid	never (24.9), rarely (18.7), sometimes (21.7), often (25.6), always (9.1)
categorical	expectations for their life in general for the next 12 months	better (25.1), worse (16.7), same (37.0), do not know (21.2)
categorical	economic expectations for the next 12 months	better (21.5), worse (17.6), same (39.7), do not know (21.2)

Let $\mathbf{Z}$ be the 438×8 data matrix (hereafter, original data set), containing the responses to the variables described above. The information concerning to social programs is not used to derive the distances between individuals, but to label them.

Figure 1: MDS configuration using Gower's and joint metrics.



(a) Gower's metric (b) Joint metric

4.2 MDS configurations for mixed-type data: Gower's vs. related metric scaling

We start by analyzing the results of the application of the two techniques described in Section 2 to matrix $\mathbf{Z}$. Hereafter, we call Gower's metric to the distance matrix derived using formula (2) and joint metric to the distance matrix obtained from formula (5). Once MDS configurations are obtained, we use the additional information about the social program to label the individuals.

In this application, the joint metric has been computed from three different distance matrices. Let $\mathbf{D}_1^{(2)}$, $\mathbf{D}_2^{(2)}$ and $\mathbf{D}_3^{(2)}$ be the three matrices of squared distances measured on the same set of n individuals. In the case of continuous variables, we compute $\mathbf{D}_1^{(2)}$ matrix using the robust version of Mahalanobis' distance described in Section 3.2. For multi-state categorical variables we consider Sokal-Michener's similarity coefficient and for binary variables Jaccard's similarity coefficient, then for $i = 2, 3$ we compute $\mathbf{D}_i^{(2)} = 2\,(\mathbf{1}\,\mathbf{1}' - \mathcal{S}_i)$, where $\mathcal{S}_i$ are the corresponding similarity matrices containing Sokal-Michener's and Jaccard's pairwise similarities.

Panels (a) and (b) of Figure 1 contain the representations of the individuals in the first two principal coordinates computed using Gower's and joint metrics, respectively. We can observe an increase of the percentage of explained variability when using the joint metric (from 26.61% with Gower's metric to 43.69% with the joint metric).

Table 1 contains the correlation coefficients between the original variables and the first three principal axes. We consider Pearson's correlation coefficient for continuous variables, whereas Spearman's correlation coefficient is computed for categorical variables. From Table 1 we see that, when using Gower's metric, the variables with greater influence on the first axis are (in decreasing order) age and employment status, whereas employment status, economic expectations, expectations for their life in general and age are those who have greater influence on the second axis. In the third axis, only monthly income and housing problems have greater influence. In the case of the joint metric, the variables with greater influence on the first two axes, although

Table 1: Correlation between the original variables and the first three principal axes.

	Gower's metric			Joint metric		
original variable	1st C.	2nd C.	3rd C.	1st C.	2nd C.	3rd C.
monthly income	0.3582	0.0111	0.4553	0.9197	0.3859	-0.0616
age	0.6132	0.4179	0.2892	-0.0930	0.9913	-0.0408
housing problems	-0.3839	-0.1435	-0.4054	-0.2495	-0.3082	-0.5315
sex	-0.0549	-0.1463	0.0308	0.1568	-0.0548	0.6971
employment status	0.4830	0.4918	0.0525	-0.2185	0.6287	0.0002
subs. economic aid	-0.0260	0.1604	0.1310	0.0205	-0.0005	0.0548
exp. for their life	0.3827	0.4852	-0.3490	-0.0798	0.1782	0.1254
economic exp.	0.3981	0.4904	-0.2874	-0.0888	0.1938	0.1484

with different order, are monthly income, employment status, age and housing problems, whereas sex and housing problems are those with greater influence on the third axis.

In Figure 2 we depict the same MDS configuration as in Figure 1, but now using those variables more correlated with the principal axes (age, employment status, economic expectations and expectations for their life in general for Gower's metric, and monthly income, age, employment status and housing problems for the joint metric) to label the individuals. We can observe that the joint metric makes better use of the information contained in the continuous variables than Gower's metric (compare, for example, panels (a1) and (b1) of Figure 2). Hence, in the following we interpret only the results concerning to the MDS representation using the joint metric. From panels (b1)–(b4) we can distinguish four different profiles:

P1 People under 50 years, unemployed, with no monthly income,

P2 People under 50 years, either employed or unemployed and mainly under 1000 euro monthly income,

P3 People over 50 years, mainly retired or that had never worked, with no monthly income and housing problems,

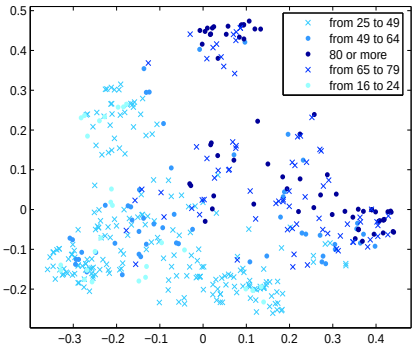
P4 People over 50 years, mainly retired or that had never worked, under 1000 euro monthly income and with housing problems.

From panel (b) of Figure 1 we can conclude that individuals from Immigrants and Refugees program will probably have *P1* or *P2* profiles, while individuals from Aged People program are more likely to have *P3* or *P4* profiles. On the other hand, participants from Drug Dependency Assistance and Inmates and Former Inmates programs present *P2* profile, while participants from Assistance for Disabled People program have *P4* profile. It is very interesting to notice that individuals from Fight Against Poverty and Social Exclusion program spread almost all over profiles.

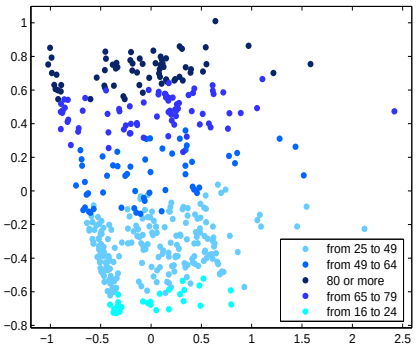
4.3 Sensitivity and robustness analysis

In order to compare the sensitivity of both MDS configurations, we perform the leave-one-out crosvalidation procedure described in Section 3.1. In Figure 3 we depict

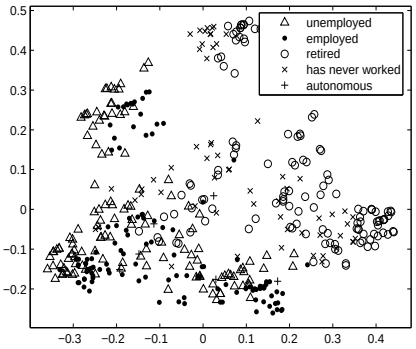
Figure 2: MDS representations using Gower's metric (26.61% of explained variability) and joint metric (43.69% of explained variability). Identification of individuals.



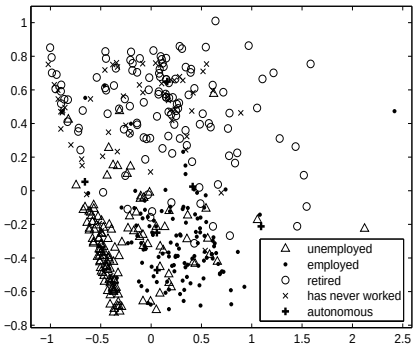
(a1) Age. Gower's metric.



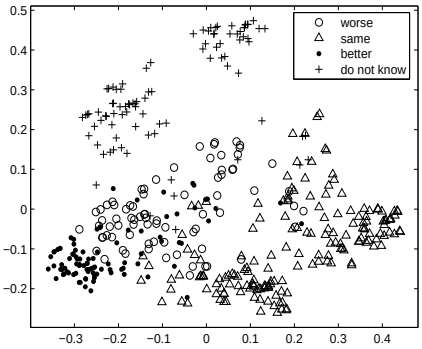
(b1) Age. Joint metric



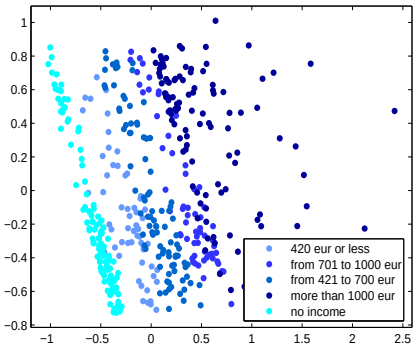
(a2) Employment status. Gower's metric.



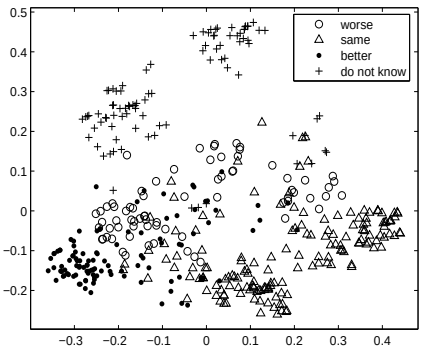
(b2) Employment status. Joint metric.



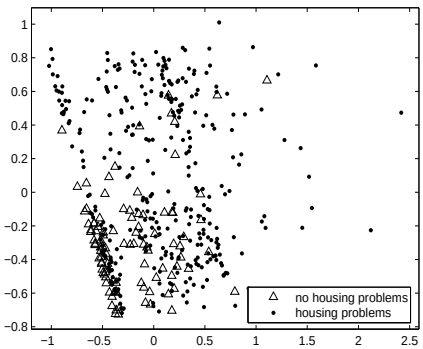
(a3) Economic expectations. Gower's metric.



(b3) Monthly income. Joint metric.

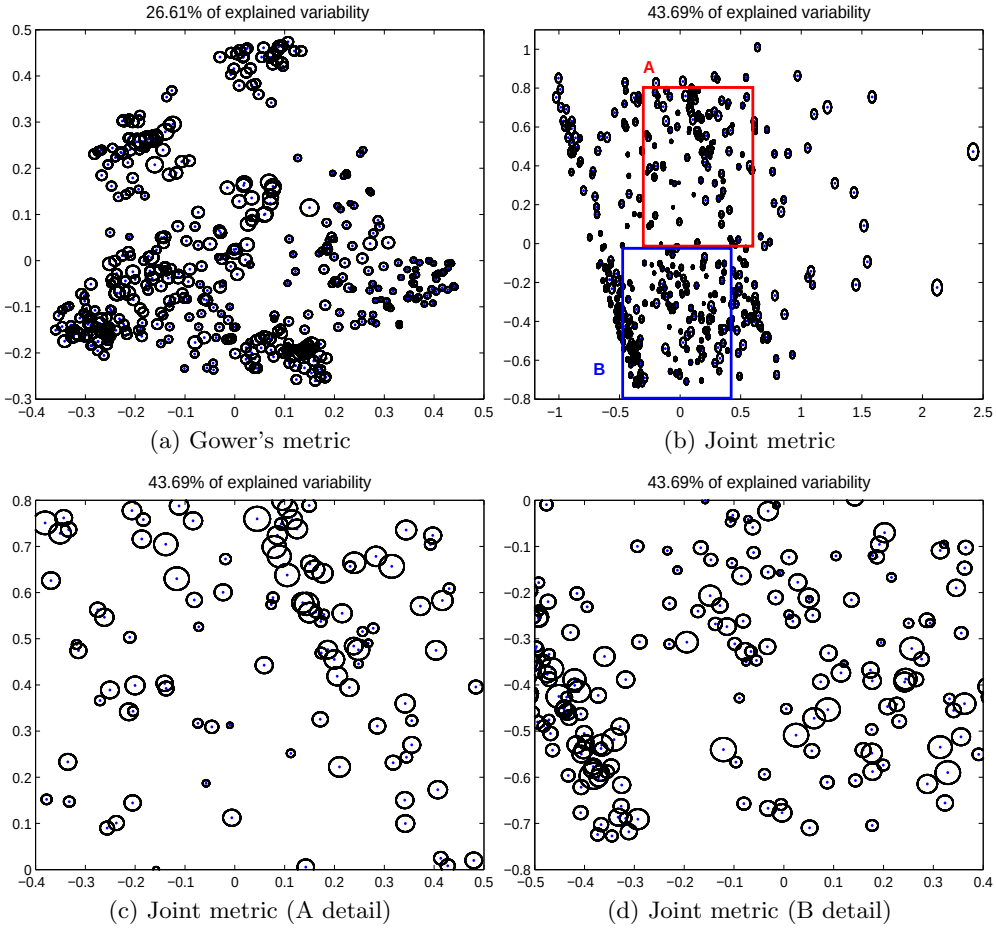


(a4) Expectations for their life. Gower's metric.



(b4) Housing problems. Joint metric.

Figure 3: Sensitivity of the MDS configurations. Original data set.



(a) Gower's metric (b) Joint metric

(c) Joint metric (A detail) (d) Joint metric (B detail)

the 95%-stability regions for the MDS configurations using Gower's metric (panel (a)), and the joint metric (panels (b)–(d)). The circles drawn are the result of the projection onto the plane of the smallest hyperspheres containing the 95% of the replicated points. The radius of each hypersphere is given by the squared root of the 95-th quantile of the Euclidean distances between the original coordinates of the point and the coordinates of its replicates. Notice that the axes scale is different in panels (a) and (b). Therefore, we include panels (c) and (d) for better comparison with panel (a). In panels (c) and (d) we can see few overlapping regions, indicating that individuals are better separated using the joint metric. On the other hand, the stability of the points is slightly greater with Gower's metric (see also Table 2, where we report some descriptive statistics for the circle radii).

Regarding to robustness, we implemented an experimental design to generate different scenarios of contamination according to the following factors: the percentage of contamination, the *P1-P4* profiles to be contaminated and the contamination type for mixed-type data. Here we only report some of the most representative outputs of the simulation study.

Table 2: 90-th and 95-th quantiles for the circle radii. Original data set.

	Gower	joint (robust)
90th-quantile	0.0078	0.0126
95th-quantile	0.0100	0.0183

We contaminate the original data set with a 5% of outliers in order to study the robustness of both MDS configurations. We consider 22 individuals either from *P2* or *P4* profiles and modify some of their characteristics in a contradictory way, simulating individuals that would rarely be participants of the social programs described in Section 4.1. In this way, the four contaminated data sets are constructed according to multivariate contamination patterns. The contaminations reported are:

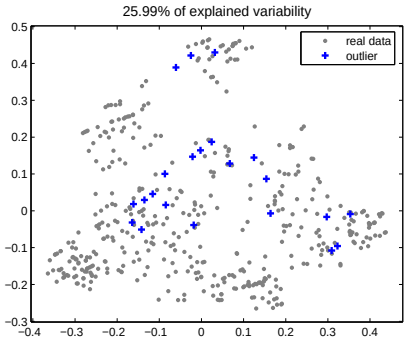
> *Contamination 1.* Participants from the Aged People social program, 70 year-old (in mean) women, that had never worked, but with 2000 euros monthly income (in mean).
>
> *Contamination 2.* Participants from the Immigrants and Refugees social program, 25 year-old (in mean) women, that had never worked, but with 2000 euros monthly income (in mean).
>
> *Contamination 3.* Participants from the Immigrants and Refugees social program, over 95 year-old, unemployed, with 400 monthly income (in mean).
>
> *Contamination 4.* Participants from the Aged People social program, under 16 year-old, that had never worked, with a monthly income greater than 1000 euros.

We start by computing the MDS configurations for the four contaminated data sets, both using Gower's and the joint metric. They are depicted in Figure 4, where we can see that these groups of rare observations are quite well located in the MDS configuration using the join metric. As before, the percentage of variability explained with the joint metric is greater than with Gower's. For each contaminated data set and each MDS configuration we carry out a leave-one-out crossvalidation procedure in order to compute the 95%-stability regions, analogously as we have done for the original data set at the beginning of this section. Table 3 contains some descriptive statistics for the radii of these regions. Although these radii show similar values for both configurations, it is important to note that joint metric is more powerful allocating outliers and explaining the variability of the data than the classical alternative that uses Gower's metric. In general, from Figure 4 and Table 3 we may conclude that, in case of having mixed-type data with more than one continuous variable, MDS configurations using the joint metric are preferable to those that use Gower's metric.

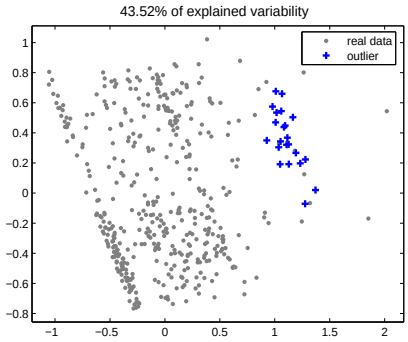
5 Outliers identification

Quite often, in the context of multivariate analysis, an outlier is an observation that is located far (in one or more directions) from the rest of the individuals. Our

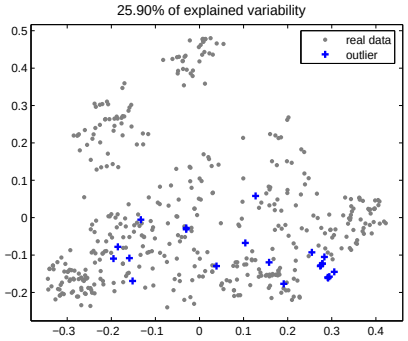
Figure 4: MDS configurations for the contaminated data sets. Outlier location.



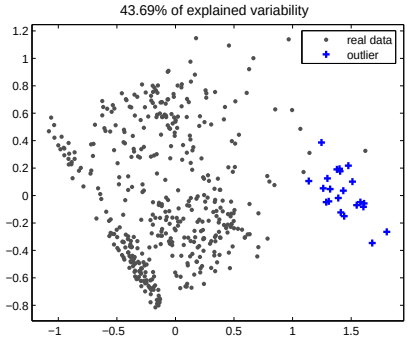
(a1) Contamination 1. Gower's metric.



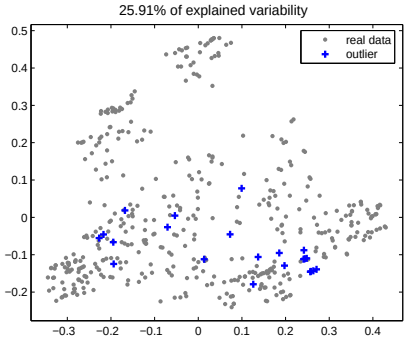
(b1) Contamination 1. Joint metric.



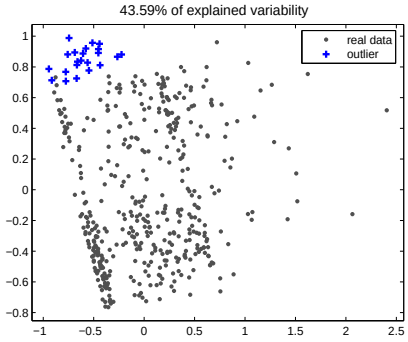
(a2) Contamination 2. Gower's metric.



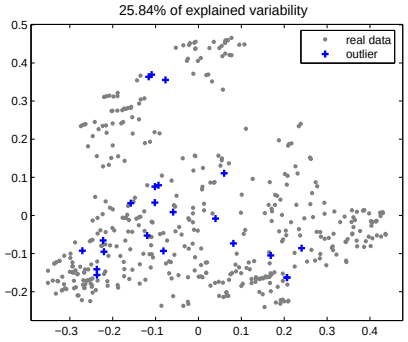
(b2) Contamination 2. Joint metric.



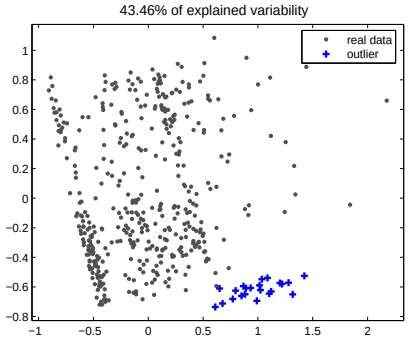
(a3) Contamination 3. Gower's metric.



(b3) Contamination 3. Joint metric.



(a4) Contamination 4. Gower's metric.



(b4) Contamination 4. Joint metric.

Table 3: 90-th and 95-th quantiles for the circle radii. Contaminated data sets.

	Contamination 1		*Contamination 2*	
	Gower	joint(robust)	Gower	joint(robust)
90th-quantile	0.0103	0.0093	0.0080	0.0122
95th-quantile	0.0142	0.0130	0.0101	0.0167

	Contamination 3		*Contamination 4*	
	Gower	joint(robust)	Gower	joint(robust)
90th-quantile	0.0080	0.0140	0.0081	0.0102
95th-quantile	0.0102	0.0264	0.0103	0.0161

contribution in this section is to propose a new test statistic related to a distance-based proximity function for detecting multivariate outliers in the context of mixed-type data. To our knowledge, this offers an original contribution in this framework. We start by defining the test statistic and, due to the difficulties to find its exact distribution (which depends on the metric selection), we obtain the approximate distribution by nonparametric procedures. To study the effectiveness of the test we apply it to the artificially contaminated data sets described in Section 4.3.

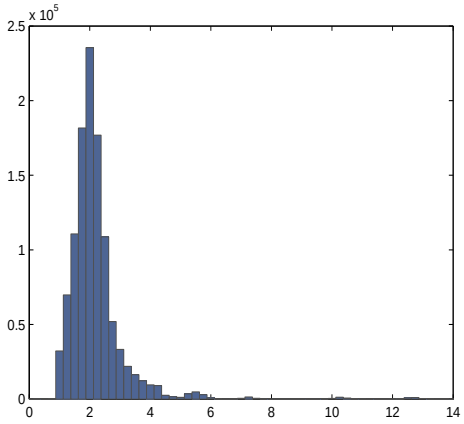
5.1 Definition of the test statistic

Let $\mathbf{D} = (d(\mathbf{z}_i, \mathbf{z}_j))_{1 \leq i,j \leq n}$ be a $n \times n$ distance (or dissimilarity) matrix computed on a set of n p-dimensional vectors $\{\mathbf{z}_i, 1 \leq i \leq n\}$.
Given a new individual $\mathbf{z}_0 \in \mathbb{R}^p$, Cuadras, Fortiana, and Oliva (1997) define the *distance-based proximity* function of $\mathbf{z}_0$ to the previous set of n individuals as:

$$\phi(\mathbf{z}_0) = \frac{1}{n} \sum_{i=1}^{n} d^2(\mathbf{z}_0, \mathbf{z}_i) - V_D, \tag{6}$$

where V_D is the geometric variability of $\mathbf{D}$ defined in (3). These authors studied the properties of (6) and used it to allocate a new individual to one of a given set of populations. Moreover, they proved that, the distance-based proximity function is a Matusita (1956) rule, whenever an Euclidean configuration exists. This concept was also used in Boj, Claramunt, Grané, and Fortiana (2009) in the context of distance-based prediction.
The new proposal is to use the proximity function defined by (6) to identify multivariate outliers in the context of mixed-type data. For this purpose it would be very useful the availability of the probability distribution of (6) under the null hypothesis that there are no outliers. However, in the context of mixed-type data, we think that an analytical derivation of such a distribution is an exceedingly complex task for any realistic situation. Note that the distribution of (6) may depend on more than one distance function (in fact, this is the case when using the joint metric). Therefore, we propose to estimate the probability distribution of the proximity function (6) through an iid-bootstrap procedure with B resamples of the original data set. For each resample, we get n values of the proximity function. One possibility for setting the cut-off value is to select the 95-th percentile, that can be computed on Bn values.

Figure 5: Histogram and descriptive statistics for the proximity function computed on $B = 2500$ resamples of the original data set.



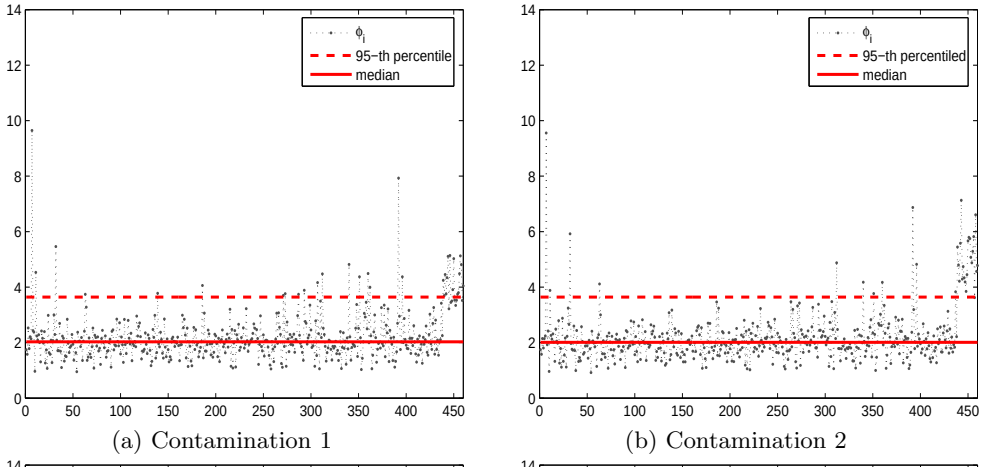
mean	2.1971
std. deviation	0.9883
median	2.0386
MAD	0.5600
90-th percentile	2.9992
95-th percentile	**3.6430**
97.5-th percentile	4.2954
99-th percentile	5.6458

We also propose the 90-th percentile or even robust alternatives based on the median to be used as threshold values, which may result in less conservative tests.
Although in the application that follows the distance matrix $\mathbf{D}$ is obtained using the joint metric (5), the procedure is general enough to be extended in several directions. For example, one can use robust distances in the construction the joint metric, or eventually, to use weighted approaches for the construction of the joint metric, which are beyond the scope of in this paper.

5.2 Application

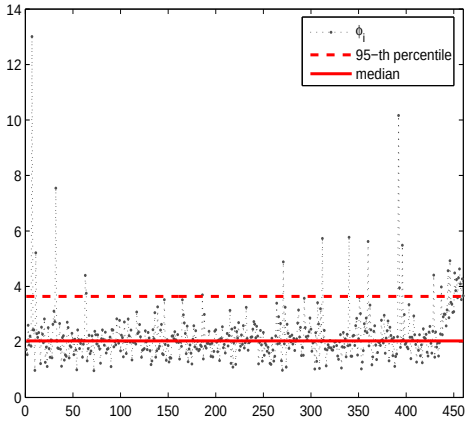
Since we are interested in detecting outlier observations, in the construction of the joint metric (5) we start by considering the usual Mahalanobis' distance for continuous variables. Nevertheless, one can also use a robust version as that described in Section 3.2. Although we know that there are multivariate outliers in the original data set presented in Section 4, we consider more challenging to work with the four contaminated data sets introduced in Section 4.3 in order to check the effectiveness of our proposal in front of different types of outliers.
To this end, we take $B = 2500$ resamples of the original data set (of size $n = 438$) and estimate the probability distribution of (6) through an iid-bootstrap. Figure 5 contains the histogram and some descriptive statistics of this proximity function computed on Bn values. Hereafter, we use the value 3.6430, the 95-th percentile (in bold in Figure 5) as the threshold value for the outlier identification test.
To evaluate the performance of the test, we compute the proximity function for the four contaminated data sets described in Section 3.2. The results are reported in Figure 6, where jointly with the proximity function values, we depict the the proposed threshold (red dashed line) and the median (red solid line) reproduced in Figure 5. In all panels there are several observations with high proximity function values, half of them located at the end of the sample, which correspond to the 22 artificially added observations. These findings are reinforced by the results contained in Table 4. The

Figure 6: Proximity function values for *Contamination 1–4* data sets.

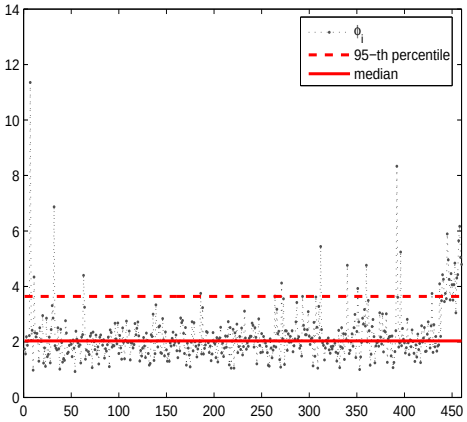


(a) Contamination 1

(b) Contamination 2



(c) Contamination 3



(d) Contamination 4

first column of Table 4 lists the number of outliers detected on the four contaminated sets. Since we know that 22 of them are artificially added points, we may now identify the outlying points in the original data set. Moreover, analyzing the characteristics of each type of contamination we can discover the features of each identified outlier in the original data set.

Additionally, there are many other thresholds that can be used. For instance, a widely used method in univariate distributions is the one that considers an observation to be an outlier if it departs from the median (in absolute value) more than 4.5 times the MAD (Median of Absolute Deviations to the median). We compare the effectiveness of both thresholds in the third and fourth columns of Table 4, where we can see that our proposal (the 95-th percentile of the proximity function distribution computed on B resamples of the original data set) overperforms the $Me + 4.5MAD$ threshold.

Table 4: Number of outlier detections in *Contamination 1–4* data sets.

	number of detections in the whole data set		number of detections in the set of 22 artificially added points	
	threshold	$Me + 4.5\,MAD$	threshold	$Me + 4.5\,MAD$
Contamination 1	33	9	15 (68.2%)	5 (22.7%)
Contamination 2	32	17	21 (95.5%)	12 (54.5%)
Contamination 3	25	10	10 (45.5%)	1 (4.5%)
Contamination 4	31	12	16 (72.2%)	7 (31.8%)

Me and MAD are estimated on each contaminated data set.

6 Conclusions

Assessing sensitivity of MDS configurations we found that our proposal overperforms the classical one that uses Gower's metric.

The presence of outliers in survey data is a relevant problem when one is interested in data visualization through MDS techniques. Mainly, this problem comes out from the mixed-type nature of the data. To our knowledge there are only few attempts to tackle this problem. This paper contributes on this direction by presenting a systematic approach to sensitivity and robustness of MDS configurations computed on mixed-type data, in particular when variables measured at continuous level as well as categorical variables are considered.

A primary motivation of this work was to extend a previous study on a real data set coming from a panel-type analysis designed to assess the economic crisis impact on Spanish people who were in situations of greater social fragility. Due to the mixed-type structure of the data we were unable to identify possible outliers from the MDS configurations with the help of the available techniques.

Through leave-one-out crossvalidation procedures we compare the performance, in terms of sensitivity and robustness, of two MDS configurations obtained using Gower's similarity coefficient versus a joint metric computed via related metric scaling, a technique that combines different distance matrices avoiding redundant information. We illustrate these methodologies on a real data set and on four contaminations of it (with 5% of outliers). In the case of having mixed-type data with more than one continuous variable, we can conclude that with MDS configurations obtained via related metric scaling (joint metric), firstly, individuals are better separated and, secondly, contaminated groups of observations (outliers) are easier to locate. This last finding is reinforced with the distance-based proximity function, that we propose as a test statistic for detecting multivariate outliers in the context of mixed-type data.

The application of several robust alternatives to the construction of the joint metric as well as a possible robustification of the proximity function (6) are lines left for further research.

Acknowledgements

Special thanks to J. Fortiana for his helpful comments concerning to the bootstrap distribution of the proximity function.

References

Boj, E., M. M. Claramunt, A. Grané, and J. Fortiana (2009). Projection error term in Gower's interpolation. *Journal of Statistical Planning and Inference 139*, 1867–1878.

Borg, I. and P. J. F. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications* (second ed. ed.). New York: Springer.

Cox, T. F. and M. A. A. Cox (2000). *Multidimensional Scaling* (second ed. ed.). London, Boca Raton, FL.: Chapman & Hall.

Cuadras, C. M. and J. Fortiana (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis 52*, 1–14.

Cuadras, C. M. and J. Fortiana (1998). Visualizing categorical data with related metric scaling. In J. Blasius and M. Greenacre (Eds.), *Visualization of Categorical Data*, London. Academic Press.

Cuadras, C. M., J. Fortiana, and F. Oliva (1997). The proximity of an individual to a population with applications to discriminant analysis. *Journal of Classification 14*, 117–136.

DeLeeuw, J. and J. Meulman (1986). A special jackknife for multidimensional scaling. *Journal of Classification 3*, 97–112.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons.

González, J., D. Peña, and R. Romera (2009). A robust partial least squares regression method with applications. *Journal of Chemometrics 23*, 78–90.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika 55*, 582–585.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 857–874.

Gower, J. C. (1992). Generalized biplots. *Biometrika 79*, 475–493.

Gower, J. C. and D. Hand (1996). *Biplots*. London, UK: Chapman & Hall.

Hubert, M., P. J. Rousseeuw, and K. Vanden Branden (2005). Robpca: A new approach to robust principal component analysis. *Technometrics 47*, 64–79.

Krzanowski, W. J. (1994). Ordination in the presence of of group structure for general multivariate data. *Journal of Classification 11*, 195–207.

Krzanowski, W. J. (2006). Sensitivity in Metric Scaling and Analysis of Distance. *Biometrics 62*, 239–244.

Krzanowski, W. J. and F. H. C. Marriott (1994). *Multivariate Analysis. Part 1*, Volume Distributions, ordination and inference. London: Edward Arnold.

Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. *Annals of the Institute of Statistical Mathematics 8*, 67–77.

Peña, D. and F. J. Prieto (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics 16*, 228–254.

Piccarella, R. and O. Lior (2007). Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society A 173*, 165–184.

Ramsay, J. O. (1980). Joint analysis of direct ratings, pairwise preferences and dissimilarities. *Psychometrika 45*, 149–165.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society A 145*, 285–312.

Rao, C. R. (1982a). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology 21*, 24–43.

Rao, C. R. (1982b). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā. The Indian Journal of Statistics, Series A 44*, 1–22.

Solaro, N. (2010). Sensitivity analisys and robust approach in multidimensional scaling: an evaluation of customer satisfaction. *Quality Technology & Quantitative Management 7*, 169–184.

Tuckey, J. W. (1960). A survey of sampling from contaminated distributions. In I. O. et al. (Ed.), *Contributions to Probability and Statistics*, pp. 448–485. Standford University Press.