Projection under pairwise distance controls

Hiba Alawieh, Nicolas Wicker, Christophe Biernacki

Projection under pairwise distance control

Hiba Alawieh [a], Nicolas Wicker [a] and Christophe Biernacki [a]

[a]Université Lille 1, - UFR de Mathématiques, cité scientifique, 59655 Villeneuve d'Ascq, France

ARTICLE HISTORY

Compiled December 11, 2019

ABSTRACT

Visualization of high dimensional and possibly complex data onto a low-dimensional space is often difficult. Several projection methods have been already proposed to display such high-dimensional structures on a lower-dimensional space, but the information lost is not always considered. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available at the scale of this reduced space. More specifically, this novel method allows for a straightforward visualization data in $\mathbb{R}^2$ with a simple reading of the approximation quality and thus provides a novel variant of dimensionality reduction.

KEYWORDS

Data visualization; dimensionality reduction; multidimensional scaling; principal component analysis; kernel principal component analysis.

1. Introduction

Several domains in science use data with large numbers of variables in their studies such as in biology (Cheung (2012), Golub *et al.* (1999)), chemistry (Svante *et al.* (1984)), geography (Van der Hilst *et al.* (2007)) and finance (Jagannathan and Ma (2003)). These data can be viewed as a large matrix and extracting results from this

Hiba Alawieh. Email: alawieh.hiba@gmail.com

type of matrix is often difficult and complicated. In such cases, it is desirable to reduce the number of dimensions of data by conserving as much information as possible from the given initial matrix.

Different types of multivariate data analysis methods have been developed to study these data such as dimensionality reduction, variables selection, cluster analysis and other methods. Typically, dimensionality reduction is used to summarize the data with variable selection used to choose the pertinent variables from the set of candidate variables and cluster analysis used to group the objects or variables. In our study, we focus on dimensionality reduction. Dimensionality reduction techniques can be used in different ways, to solely lower the dimensionality to prepare data for other treatments or for data visualization to provide a simple interpretation of the data in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Due to the difficulties faced by high dimensional data, many methods for data dimensionality reduction and data visualization have been proposed (Chan (2006); Chinchilli and Sen (1987); Dempster (1971); Keim and Kriegel (1996); Mardia *et al.* (1979)). Some of the most common methods include principal component analysis (PCA) (Jackson (1991)), multidimensional scaling (MDS) (Togerson (1958)), scatter plot matrix (Cleveland and McGill (1988)), parallel coordinates (Inselberg (1985)) and Sammon's mapping (Sammon (1969)). Scatter plot matrix and parallel coordinates methods are widely used to visualize multidimensional data sets. An issue with principal component analysis and multidimensional scaling is that as the number of dimensions grow, important multi-dimensional relationships might not be visualized. Moreover, the quality of projection assessed by the percentage of variance that is conserved or by the stress factor is a global projection quality measure and only takes into account what happens globally. Typically, it could be a good projection globally, if the percentage of variance obtained using PCA, for example, is large.

In some projection methods such as PCA, a local measure is defined to indicate the projection quality of each projected point taken individually. This local measure is evaluated by the squared cosine of the angle between the principal space and the vector of the point. A good representation in the projected space is hinted by high squared cosine values. This measure is useful in cases of linear projection, which happens in PCA, but cannot be applied in the case of nonlinear projection. Moreover, linear

dimensionality reduction misses important nonlinear structures in the data which does not allow to give powerful results in case of nonlinear configurations. Therefore, many methods have been developed to perform nonlinear projections by nonlinearizing a linear dimensionality reduction or by using manifold learning methods.

The nonlineariziation of linear dimensionality reduction is applied to extract nonlinear principal components. Kernel PCA is one of the most exciting methods in this domain, which integrates a kernel function to determine principal components in different high-dimensional space (Schölkopf (1998)). Manifold learning methods are an approach to construct a matrix using the neighborhood information and take a spectral decomposition to find a nonlinear embedding (like Locally Linear Embedding LLE, Isomap algorithm etc). (Lee and Verleysen (2007), Tenenbaum *et al.* (2000), Roweis and Saul (2000)).

In this paper, we propose a new nonlinear projection method that projects the points in a reduced space by using the pairwise distance between pairs of points and by taking into account the projection quality of each point taken individually. Nonlinear projection methods cited in the previous paragraph project the points in a feature space which makes the interpretation of distances between the projected points meaningless. In our method, the distances between projected points are related to the initial distances between points, offering a way to easily interpret the distances observed in the projection plane. This projection leads to a representation of the points as circles with a different radius associated to each point. Henceforth, this method will be referred to as "Projection under pairwise distance control". Furthermore, visualization of data in a reduced space is not the only objective of this method. It can serve as a dimensionality reduction method to reduce the number of variables by minimizing the sum of the radii and to then determine the number of variables that can be kept.

The main contribution of this study is to provide a simple data visualization in $\mathbb{R}^2$ with a straightforward interpretation and to provide a new variant of dimensionality reduction. Firstly, the new projection method is presented in Section 2. In Section 3, the algorithms used in solving the optimization problems related to this method are then illustrated. In Section 4 the application of this method to various real data sets is shown. Finally, the conlusions are drawn in Section 5.

2. Projection under pairwise distance control

Let us consider n points given by their pairwise distances denoted by d_{ij} for $i, j \in \{1, \ldots, n\}$. The objective is to project these points using distances into a reduced space $\mathbb{R}^q$ by introducing additional variables, called hereafter radii, that indicate the extent to which the projection of each point is accurate. The local quality is then given by the values of the radii. A good projection quality of point i is indicated by a small radius value denoted by r_i. It is important to note that both units of d_{ij}'s and r_i's are identical, thus allowing for a direct comparison.

Before presenting our method, an overview of principal component analysis, Kernel PCA and multidimensional scaling is given to highlight the significance of our method.

2.1. Overview of certain existing methods: PCA, KPCA and MDS

Principal Component Analysis (PCA)

The PCA method is the most used method for data visualization and dimensionality reduction. This method is a linear projection technique applied when the data is linearly separable. PCA can be stated as an optimization problem involving the squared Euclidean distances (Mardia *et al.* (1979)). This optimization problem is the following:

$$\mathcal{P}_{\text{PCA}} : \begin{cases} \min\limits_{A \in \mathcal{M}_{p\times q}} \sum\limits_{1 \le i < j \le n} |d_{ij}^2 - \|Ay_i - Ay_j\|^2| \\ s.t. \;\; \text{rank}(A) = m \\ \qquad AA^T = I_p, \end{cases}$$

where $y_i \in \mathbb{R}^p$ is the original coordinates vector of point i, d_{ij}^2 is the squared distance for couple (i, j) given by $\|y_i - y_j\|^2$ and A is the projection matrix of dimension $p \times q$ with q being the reduced space dimension. By its nature, PCA cannot take into account nonlinear structures, as it describes the data in terms of a linear subspace. To deal with nonlinearity, Kernel PCA, the reproducing kernel Hilbert space variant of PCA, can be used.

Kernel PCA (KPCA)

The idea behind KPCA is to perform PCA in a feature space denoted by $\mathcal{F}$, obtained by a nonlinear mapping of data from its original space into the feature space $\mathcal{F}$, where the low-dimensional latent structure is hopefully easier to discover (Schölkopf (1998)). The mapping function noted Φ is considered as:

$$\begin{array}{rccc} \Phi: & \mathbb{R}^p & \rightarrow & \mathcal{F} \\ & Y & \rightarrow & \Phi(Y)\ . \end{array}$$

The original data y_i is represented in the feature space as a function $\Phi(y_i) = k(y_i, .)$, where $k(., .)$ is a positive kernel. Similar to PCA, KPCA is based on finding the first q eigenvectors corresponding to the q largest eigenvalues λ_i of the Gram matrix $K = (k_{ij})_{ij\in 1,\ldots,n}$ where $k_{ij} = k(y_i, y_j) = \langle \Phi(y_i), \Phi(y_j) \rangle$ is a chosen positive kernel. Letting V_v, for $v = 1, \ldots, q$, are the eigenvectors in the feature space and $P_{\Phi(y_i)}$ is the projection of $\Phi(y_i)$ onto the subspace $V_1, \ldots, V_q$. The KPCA problem can be represented as a minimization problem with the following error:

$$\mathcal{E}_{\text{KPCA}}: \ \|\Phi(y) - P_{\Phi(y)}\|_2^2 \ ,$$

where $P_{\Phi(y)} = \sum_{v=1}^{q} \langle \Phi(y), V_v \rangle V_v$.

Furthermore, the only measure used to evaluate the projection quality of points for PCA and KPCA is the squared cosine value. Squared cosine values cannot be interpreted at the same time as the distances in the projection because the cosine values do not have a specific unit. More precisely, the visualization of the projection in the reduced space using PCA and KPCA cannot simply be interpreted in terms of original distances between the points. Indeed, in PCA, the cosine values do not provide a quantitative assessment of the error made when considering the distances between the projected points, all the more in KPCA where the projected points are in the feature space so the term "distances" is not related to the distances between the points in the original space.

Multidimensional Scaling (MDS)

As with PCA, Multidimensional scaling (MDS) consists of finding a new data configuration in a reduced space. The main difference between these two methods is that the input data in MDS is in the form of a similarity or dissimilarity matrix, called "proximity", representing the proximity between pairs of objects. The key idea of MDS is to perform dimensionality reduction in a way to approximate high-dimensional distances denoted by δ_{ij} the low-dimensional distances d_{ij}, where d_{ij} is equal to the distance between x_i and x_j, the coordinates of i and j in the reduced space. In the classic and simplest case of MDS, the least-squares loss function denoted by "Stress" is given as follows:

$$\text{Stress} = \sqrt{\sum_{1 \leq i < j \leq n} (d_{ij} - \|x_i - x_j\|)^2}.$$

By minimizing the Stress function, we find the best configuration of $(x_1, \ldots, x_n) \in \mathbb{R}^q$ such that the distances fit to the initial distances.

If we consider n variables as $r_1, \ldots, r_n \in \mathbb{R}^+$, the sum of which bounds the stress function, the optimization problem $\mathcal{P}_{\text{MDS}}$ can be equivalently rewritten as:

$$\mathcal{P}_{\text{MDS}} : \begin{cases} \min\limits_{r_1,\ldots,r_n \in \mathbb{R}^+} \sum\limits_{i=1}^{n} r_i \\ s.t. \quad \sum\limits_{i=1}^{n} r_i \geq \dfrac{1}{n-1} \sqrt{\sum\limits_{1 \leq i < j \leq n} (|d_{ij} - \|x_i - x_j\|)^2}. \end{cases}$$

A criterion to determine the local projection quality has been proposed by Born and Groenen in Borg and Groenen (2005) called Stress-per-point (SPP). The SPP of point i is given by:

$$SPP_i = \frac{\frac{\sum_{j=1, j \neq i}^{n} (d_{ij} - \|x_i - x_j\|)^2}{\sum_{j=1, j \neq i}^{n} d_{ij}^2}}{Stress},$$

with $Stress = \dfrac{\sum_{1\leq i<j\leq n}^{n} (d_{ij} - \|x_i - x_j\|)^2}{\sum_{1\leq i<j\leq n}^{n} d_{ij}^2}$.

Again, this is difficult to interpret directly on the projection as a distance error because the projected points are not in the same metric of initial data.

However, we can observe that the constraint on $\sum_{i=1}^{n} r_i$ can be modified to have a stronger control on each d_{ij} in the following way: $|d_{ij} - \|x_i - x_j\|| \leq r_i + r_j$ where x_i and x_j are the projected coordinates of points i and j.

Therefore, our objective is to propose a new nonlinear projection method that individually controls the projection of points and provides a graphical representation in the same metric as the original space with an error associated to each point.

2.2. Our proposal: Projection under pairwise distance control method

Letting $x_1, \ldots, x_n$ be the coordinates of the projected points in $\mathbb{R}^p$ and $\|x_i - x_j\|$ is the distance between two projected points (i, j). Radii are introduced in this paper to assess how far $\|x_i - x_j\|$ is from the given distance d_{ij}. Indeed, for the couple (i, j), we are aiming for a $\|x_i - x_j\|$ value close to d_{ij}, which should imply a small radii (r_i, r_j). Figure 1 depicts this idea: for each point $i \in \{1, \ldots, n\}$, the projection of i belongs to a sphere with center x_i and radius r_i such that for each couple $(i, j) \in \{1, \ldots, n\}$ we have $\|x_i - x_j\| - (r_i + r_j) \leq d_{ij} \leq \|x_i - x_j\| + r_i + r_j$.

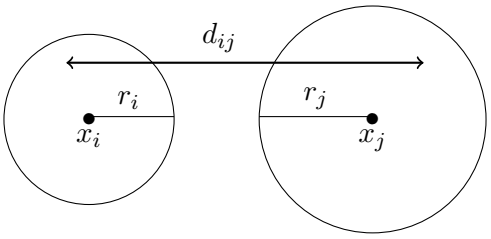


Figure 1.: Example of radii for bounding of the original distance d_{ij}

Radii for uncertainty metric: The idea presented above can be expressed by finding the value of radii that satisfy these two constraints:

- $\sum_{i=1}^{n} r_i$ is minimal.
- $d_{ij} \in [\|x_i - x_j\| - r_i - r_j; \|x_i - x_j\| + r_i + r_j]$, for $1 \leq i < j \leq n$.

The projection under pairwise distance control problem can be written as the following optimization problem:

$$\mathcal{P}_{r,x} : \begin{cases} \min\limits_{r_1,\ldots,r_n \in \mathbb{R}^+, x_1,\ldots,x_n \in \mathbb{R}^q} \sum\limits_{i=1}^{n} r_i \\ s.t \quad |d_{ij} - \|x_i - x_j\|| \leq r_i + r_j, \text{ for } 1 \leq i < j \leq n \end{cases}$$

Linear optimization program using fixed coordinates $(x_1, x_2, \ldots, x_n)$: Of course, by fixing the coordinates vectors x_i for all $i \in \{1, \ldots, n\}$ using principal component analysis or any other projection method, the optimization problem can easily be solved in $(r_1, \ldots, r_n)$ using linear programming. This problem can be written as follows:

$$\mathcal{P}_{r} : \begin{cases} \min\limits_{r_1,\ldots,r_n \in \mathbb{R}^+} \sum\limits_{i=1}^{n} r_i \\ s.t \quad |d_{ij} - \|x_i - x_j\|| \leq r_i + r_j, \text{ for } 1 \leq i < j \leq n \end{cases}$$

It should be noted that a solution for problem $\mathcal{P}_r$ always exists. Indeed, to satisfy the constraints it is sufficient to increase all r_i. Thus, for any method producing points in a reduced space as PCA for instance, we can compute the radii as a post-processing to assess the local quality of the projected points.

$P_{r,x}$ is a non-convex optimization problem: For any dimension p, even with $p = 1$, note that the optimization problem $P_{r,x}$ is not convex. Indeed, to easily illustrate this fact, we take 4 points with an arbitrary order indexed by i_1, i_2, i_3 and i_4 in $\mathbb{R}$ with respective coordinates $x_{i_1} = 0$, $x_{i_2} = 2$, $x_{i_3} = 3$ and $x_{i_4} = 1$. Note that distances $d_{i_1 i_2}$ and $d_{i_3 i_4}$ are both equal to 2.

Let us consider the function $g(x_i, x_j) = |d_{ij} - \|x_i - x_j\||$. Thus, we have $g(x_{i_1}, x_{i_2}) = 0$ and $g(x_{i_3}, x_{i_4}) = 0$ but $g(\frac{x_{i_1} + x_{i_3}}{2}, \frac{x_{i_2} + x_{i_4}}{2}) = |0 - 2| = 2$ which is larger than $\frac{g(x_{i_1}, x_{i_2}) + g(x_{i_3}, x_{i_4})}{2} = 0$ proving non convexity associated to this sample design.

In fact, problem $P_{r,x}$ is convex in one dimension only if $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$ are ordered. Indeed, let us consider $x_{i_4} \leq x_{i_3} \leq x_{i_2} \leq x_{i_1}$ so that $g(x_{i_1}, x_{i_2}) = |x_{i_1} - x_{i_2} - d_{i_1 i_2}|$ and $g(x_{i_3}, x_{i_4}) = |x_{i_3} - x_{i_4} - d_{i_3 i_4}|$ so that for any $\lambda, \mu \geq 0$:

$$\begin{aligned} g\left(\frac{\lambda x_{i_1} + \mu x_{i_3}}{\lambda + \mu}, \frac{\lambda x_{i_2} + \mu x_{i_4}}{\lambda + \mu}\right) &= \left|\frac{\lambda}{\lambda + \mu}(x_{i_1} - x_{i_2} - d_{i_1 i_2}) + \frac{\mu}{\lambda + \mu}(x_{i_3} - x_{i_4} - d_{i_3 i_4})\right| \\ &\leq \frac{\lambda}{\lambda + \mu} g(x_{i_1}, x_{i_2}) + \frac{\mu}{\lambda + \mu} g(x_{i_3}, x_{i_4}), \end{aligned}$$

which proves convexity. Therefore given an ordering, we have a convex optimization each time that can be solved exactly and the global optimum can be found by taking the minimum obtained for all permutations of $x_1, \ldots, x_n$. However, this only works in one dimension at a time; an approximate non-convex optimization is needed since we have multidimensional data.

Many methods available in the literature propose different ways to solve such optimization problems. Examples include: trust-region-reflective (Conn *et al.* (2000)), which chooses and computes an approximation of the objective function, and then chooses and modifies the trust region and finally solves the trust-region subproblem; sequential quadratic programming (SQP) which solves the optimization problem by addressing a sequence of quadratic programming problems where the Lagrangian function is approximated by a quadratic function and the constraints are approximated by a linear hyper-space (Boggs and Tolle (1995)); the active-set method, which is composed of two phases, wherein for the first phase (the feasibility phase) the objective function is ignored while a feasible point is found for the constraints, and in the second phase (the optimality phase) the objective function is minimized while feasibility is maintained (Wong (2011), Cristofari *et al.* (2007)). The choice of optimization method to use to achieve optimality of the optimization problem is essential and depends on many factors such as the type of problem, desired quality of solution, time limit and availability of the algorithm implementation etc. In fact, all of the methods cited above can be used in optimizing problem $\mathcal{P}_{r,x}$ which is a constrained optimization problem having inequalities constraints and they are all available in MATLAB using the function "fmicon" for constrained nonlinear optimization problems (since

the proposed method is implemented in MATLAB). Having small radii is the main constraint in our optimization problem thus, the objective is to obtain good solution within a reasonable and practical timeframe. Therefore, a method that balances time and quality of the solution is required.

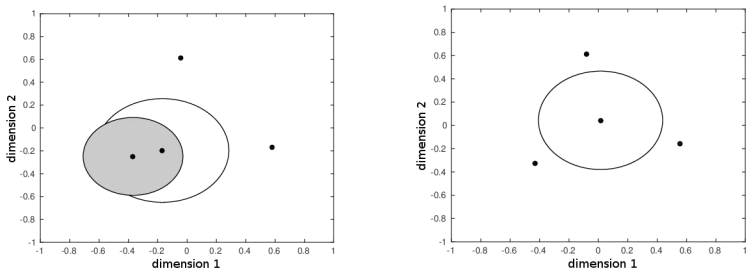
Another strategy of use: Dimensionality reduction One of the main objectives of high-dimensional data studies is to choose, from a large number of variables, those that are important for understanding the underlying studied phenomena. In addition to visualization, our aim can thus be to reduce the dimension rather than to visualize data in $\mathbb{R}^2$. Therefore, the proposed method can serve to reduce the number of variables by taking into account the minimal value of $\sum_{i=1}^{n} r_i$. Indeed, by solving the problem $\mathcal{P}_{r,x}$ using different dimension values, we can choose the dimension with respect to the local projection quality promoted in this study.

2.3. A toy example for illustrating our method

Let us apply the proposed projection method to a simple example by taking a tetrahedron with all pairwise distances equal to 1. For problem $\mathcal{P}_r$, the coordinates of points x_i for $i = 1, \ldots, 4$ are obtained using multidimensional scaling. The optimization was carried out using the MATLAB software with the optimization toolbox for linear and nonlinear optimization problem used for problems $\mathcal{P}_r$ and $\mathcal{P}_{r,x}$, respectively. The value of $\sum_{i=1}^{4} r_i$ is equal to 0.7935 for problem $\mathcal{P}_r$ and 0.4226 for $\mathcal{P}_{r,x}$. It is clear that problem $\mathcal{P}_{r,x}$ gives better solutions than problem $\mathcal{P}_r$ with smaller radii, which indicates better projection quality of points.

This result can be shown in Figure 2, which depicts the solution obtained using $\mathcal{P}_r$ and $\mathcal{P}_{r,x}$. In Figures 2a and 2b, the circles with different radii indicate the quality of projection for each point. The circle color is related to the radius value, the shades of gray lie between white and black in the descending direction of the radius values; the smaller the radius, the darker circle. The points that have circles with small radii are also considered as projected points. Note that the points represented as points and not as circles are very well projected, having radii almost equal to zero.

In Figure 2b, just one circle appears indicating that the projection quality using prob-



(a) (b)

Figure 2.: Projected points after solving problem $\mathcal{P}_r$ and problem $\mathcal{P}_{r,x}$. The x-axis and y-axis are dimension 1 and dimension 2, respectively. (a) and (b) show the projection obtained from the solution of problem $\mathcal{P}_r$ using MDS and of problem $\mathcal{P}_{r,x}$ respectively.

lem $\mathcal{P}_{r,x}$ is better than when using problem $\mathcal{P}_r$. In Figure 2a, half of the points are well projected whereas the other half have large radii, indicating that they are not well projected. Moreover, it is worth noting that the three outer points all have radii equal to 0, which indicates that they are all perfectly placed with respect to one another. In Figure 2b, the distances between the three points that are very well projected are equal to the distances between these points in their original space ($d_{kl} = \|x_k - x_l\|$ where k and l are two very well projected points) whereas the distances from the badly projected points to the perfectly projected points are not yet conserved. Therefore, using the proposed method, we have succeeded in conserving half of the original distances in the new projection plane and the other half have been changed to fit the new configuration. If we now apply the proposed method to the distances obtained by MDS to find the radius of each projected point (Figure 2a), it can be noted that one distance is conserved as the original distance and the other five distances are changed which indicates that the proposed method projects the points well by conserving the distances between the points as much as possible.

It is also important to note that, in general, our method is not only a nonlinear projection method with local quality measure, but it can act as a new tool to give the local quality of projection for the classical projection methods using the radii by solving problem $\mathcal{P}_r$. It can be used outside our method as post-processing of classical methods.

2.4. Connexion with existing methods

Multidimensional fitting (MDF) (Berge *et al.* (2010)) is a method that modifies the coordinates of a set of points in order to make the distances calculated on the modified coordinates similar to a given set of distances on the same set of points. The so-called "target matrix", the matrix that contains the point coordinates and "reference matrix" is the matrix that contains the given distances.

Let us take $X = \{x_1|\cdots|x_n\}$, the target matrix of coordinates and $D = \{d_{ij}\}$, the reference matrix of distances. The objective function of MDF problem is given by:

$$\sum_{1\leq i<j\leq n} |d_{ij} - \|x_i - x_j\||.$$

Proposition 2.1. *Problem $\mathcal{P}_{r,x}$ is bounded from below by $\frac{1}{n-1}\sum_{1\leq i<j\leq n} |d_{ij}-\|x_i-x_j\||$ where $x_1, \ldots, x_n$ is the optimum for the associated MDF problem.*

Proof. By summing all the constraints of problem $\mathcal{P}_{r,x}$, we obtain:

$$\sum_{1\leq i<j\leq n} |d_{ij} - \|x_i - x_j\|| \leq \sum_{1\leq i<j\leq n} (r_i + r_j) = (n-1)\sum_{i=1}^{n} r_i.$$

So, $\sum_{i=1}^{n} r_i \geq \frac{1}{n-1}\sum_{1\leq<i<j\leq n} |d_{ij} - \|x_i - x_j\||$, which concludes the proof. □

3. Optimization tools for performing the proposed method

Problem $\mathcal{P}_{r,x}$ can be solved using different initialization points for the coordinate matrix X. In this section, we first discuss the different initialization points of the proposed optimization problem and then propose two algorithms to be used in our optimization.

3.1. Initialization point for problem $\mathcal{P}_{r,x}$

Different solutions of problem $\mathcal{P}_{r,x}$ can be obtained using different initial values of matrix X. We have considered three possibilities:

1- Initial point using a known projection method The first possibility is to use the matrix obtained by PCA or another projection method. The choice of method must be based on the type of data. In this application, we use PCA for quantitative data and MDS for categorical and functional data.

2- Initial point using squared distances The optimization problem $\mathcal{P}_{r,x}$ can be changed by taking the squared distances between points instead of the distances. Rewriting r_i^2 as R_i, the problem is changed into

$$\mathcal{P}_{R,x}: \begin{cases} \min\limits_{R_1,\ldots,R_n\in\mathbb{R}^+, x_1,\ldots,x_n\in\mathbb{R}^k} \sum\limits_{i=1}^{n} R_i \\ s.t. \;\; |d_{ij}^2 - \|x_i - x_j\|^2| \leq R_i + R_j, \quad \text{for } 1 \leq i < j \leq n. \end{cases}$$

This transformation is interesting because if the constraints of problem $\mathcal{P}_{R,x}$ are satisfied, the constraints of problem $\mathcal{P}_{r,x}$ will also be satisfied. Indeed,

$$|d_{ij}^2 - \|x_i - x_j\|^2| \leq R_i + R_j = r_i^2 + r_j^2.$$

If without loss of generality, $d_{ij} \geq \|x_i - x_j\|$, we obtain:

$$\begin{aligned} &(d_{ij} - \|x_i - x_j\|)\,(d_{ij} + \|x_i - x_j\|) \leq r_i^2 + r_j^2 \leq (r_i + r_j)^2 \Rightarrow \\ &|d_{ij} - \|x_i - x_j\||^2 \leq (r_i + r_j)^2 \Rightarrow |d_{ij} - \|x_i - x_j\|| \leq (r_i + r_j). \end{aligned}$$

In this way problem $\mathcal{P}_{R,x}$ can serve as an initial step in solving problem $\mathcal{P}_{r,x}$.

3- Initial point using an improved solution of problem $\mathcal{P}_r$ This strategy is more involved. First, we need two properties that provide a way to improve the optimization results of problem $\mathcal{P}_{r,x}$.

Proposition 3.1. *Let us consider a point x_i such that for an index j, the following inequality is saturated:*

$$|d_{ij} - \|x_i - x_j\|| \leq r_i + r_j,$$

and the other inequalities involving i are not saturated. The corresponding solution can then be improved by moving x_i along the line $x_j - x_i$ in order to decrease r_i and $|d_{ij} - \|x_i - x_j\||$.

Another manner to improve the resolution of problem $\mathcal{P}_{r,x}$ is to perform a scale change by multiplying the coordinates x_i, for $i = 1, \ldots, n$, by a constant $a \in \mathbb{R}$. Thus, the new optimization problem is given by:

$$\mathcal{P}_{r,a} : \begin{cases} \min\limits_{r_1,\ldots,r_n,a \in \mathbb{R}^+} \sum\limits_{i=1}^{n} r_i \\ s.t. \;\; |d_{ij} - a\|x_i - x_j\|| \leq r_i + r_j. \end{cases}$$

Proposition 3.2. *Let $r_1, \ldots, r_n; x_1, \ldots, x_n$ be a feasible solution of $\mathcal{P}_{r,x}$, if $\exists a$ such that $\eta(a) < \sum\limits_{i=1}^{n} r_i$ with $\eta(a) = \sum\limits_{1 \leq i < j \leq n} |d_{ij} - a\|x_i - x_j\||$, then $\exists\ \tilde{r}_1, \ldots, \tilde{r}_n$ a solution of $\mathcal{P}_{r,a}$ such that $\sum\limits_{i=1}^{n} \tilde{r}_i < \sum\limits_{i=1}^{n} r_i$.*

The new initial point called X_{imp}, is the improved solution given by using these two properties as follows:

- Firstly, improving the solution of problem $\mathcal{P}_r$ by solving problem $\mathcal{P}_{r,a}$ and using proposition 3.2.
- Secondly, improving the solution of problem $\mathcal{P}_{r,a}$ using proposition 3.1.

3.2. A deterministic strategy: Algorithm 1

As discussed, three possibilities of coordinate matrix X can be used as the initial point:

1- Coordinates given by PCA or MDS: $X_{\mathcal{P}_{\text{PCA/MDS}}}$ is the coordinate matrix obtained by applying PCA or MDS and $r_{_{\mathcal{P}_r}}$ is a vector that contains the radius of each

point obtained by solving $\mathcal{P}_r$.

2- Coordinates given by squared distances: $X_{\mathcal{P}_{R,x}}$ is the coordinate matrix obtained by solving problem $\mathcal{P}_{R,x}$ and $R_{\mathcal{P}_{R,x}} = r^2_{\mathcal{P}_{R,x}}$ is a vector that contains the squared radius for each point obtained by solving the subsequent $\mathcal{P}_{R,x}$ problem.

3- Coordinates given by improving the solution of problem $\mathcal{P}_r$: X_{imp} is the coordinate matrix obtained by improving the previous solution using Proposition 3.1 and r_{imp} is a vector that contains the radius of each point obtained after each iteration of solving problem $\mathcal{P}_{r,a}$

Finding these matrices requires solving the following optimization problems: $\mathcal{P}_r$, $\mathcal{P}_{R,x}$ and $\mathcal{P}_{r,a}$. Problems $\mathcal{P}_r$ and $\mathcal{P}_{r,a}$ are both constrained linear optimization problems that can be solved using interior-point or simplex algorithms, which are the most widely used algorithms for linear programming. The interior-point algorithm uses a primal-dual predictor-corrector algorithm and the simplex algorithm uses a systematic procedure for generating and testing candidate vertex solutions to a linear program (Murty (1983)). On the contrary, problem $\mathcal{P}_{R,x}$ is a nonlinear optimization problem that can be solved using one of the nonlinear optimization algorithms cited in Section 2.2. All these algorithms are available in MATLAB using the optimization toolbox and can be used for the corresponding problem.

To find the best solution of problem $\mathcal{P}_{r,x}$, we solve it with the three different initialization matrices described above. For this task, we define Algorithm 1 that gives the best solution using the different coordinate matrices. This algorithm consists of two steps, an initialization step and an optimization step. The initialization step offers three different coordinate matrices to be used in the optimization step as an initial point to quickly reach the best solution. During the optimization step, problem $\mathcal{P}_{r,x}$ is solved using one of the nonlinear optimization algorithms mentioned in Section 2.2, starting each time with one matrix of the three initial matrices already found.

Thus, for Algorithm 1, described below, the three different initialization matrices are tried and then the best one is chosen that gives the minimum value of $\sum_{i=1}^{n} r_i$.

Algorithm 1

Input: D: distance matrix, N: number of iterations.
Initialization step
Project the points using PCA or MDS.
Solve $\mathcal{P}_r$ using a linear optimization method. Obtained solution: $(X_{\mathcal{P}_{\text{PCA/MDS}}}, r_{\mathcal{P}_r})$.
Solve $\mathcal{P}_{R,x}$ using a nonlinear optimization method and starting from the solution of $\mathcal{P}_r$ obtained at the previous step. Obtained solution: $(X_{\mathcal{P}_{R,x}}, R_{\mathcal{P}_{R,x}})$.
$X_{imp} \leftarrow X_{\mathcal{P}_{R,X}}$.
for $t = 1$ to N **do**
  Solve $\mathcal{P}_{r,a}$ starting from X_{imp} using a linear optimization method.
  Improve the solution of $\mathcal{P}_{r,a}$. Obtained solution: (X_{imp},r_{imp}).
end for
Optimization step
Optimize $\mathcal{P}_{r,x}$ using a nonlinear optimization method and starting from $X_{\mathcal{P}_{\text{PCA/MDS}}}$, $X_{\mathcal{P}_{R,x}}$ and X_{imp}.
Choose the minimal solution obtained by these three different starting points.

3.3. A stochastic strategy: Algorithm 2

Problem $\mathcal{P}_{r,x}$ is a hard problem, thus it is natural to resort to stochastic optimization methods. In the present case, we resort to the Metropolis-Hastings algorithm (Johansen and Evers (2007)) which allows us to build a Markov chain with the desired stationary distribution. The challenging parts are the choice of the proposal distribution and the necessity to solve the problem $\mathcal{P}_r$ at each iteration. Specifically, the Metropolis-Hastings algorithm requires:

1- A *target distribution*:

The target distribution is related to the objective function of problem $\mathcal{P}_{r,x}$ and is given by:

$$\pi(x) \propto \exp\left(\frac{-E(x)}{T}\right),$$

where E is an application given by:

$$\begin{array}{rccl} E: & \mathbb{R}^n & \longmapsto & \mathbb{R} \\ & x = (x_1, \ldots, x_n) & \longmapsto & E(x) = \text{Solution of problem } \mathcal{P}_r \text{ with fixed } x. \end{array}$$

The variable T is the temperature parameter, to be fixed according to the value range of E.

2- A *proposal distribution*:

The choice of the proposal distribution is very important to obtain meaningful results. It should be chosen in such a way that the proposal distribution

approaches the target distribution. The proposal distribution $q(X \to .)$ is constructed as follows, giving priority to the selection of points involved in saturated constraints:

- For each point i, choose a point $j^{(i)}$ with probability equal to:

$$P_{j^{(i)}} = \frac{\lambda \exp\left(-\lambda(r_i + r_{j^{(i)}} - |d_{ij^{(i)}} - \|x_i - x_{j^{(i)}}\||)\right)}{\sum_{k=1,k\neq i}^{n} \lambda \exp\left(-\lambda(r_i + r_k - |d_{ik} - \|x_i - x_k\||)\right)}.$$

- Choose a constant $c_{ij^{(i)}}$ using Gaussian distribution $\mathcal{N}_k(0, \sigma)$.
- Generate a matrix X^* by moving each vector x_i of matrix X^{t-1} as follows:

  - If $d_{ij^{(i)}} - \|x_i - x_{j^{(i)}}\| > 0$ then $x_i^* = x_i + |c_{ij^{(i)}}| L_{ij^{(i)}}$.
  - else $x_i^* = x_i - |c_{ij^{(i)}}| L_{ij^{(i)}}$,

  where $L_{ij^{(i)}} = \dfrac{x_i - x_{j^{(i)}}}{\|x_i - x_{j^{(i)}}\|}$.

3- A *linear optimization problem*:

For the matrix X generated at each iteration, we solve the linear optimization problem $\mathcal{P}_r$.

Algorithm 1 and Algorithm 2 are both implemented in MATLAB and a code for each algorithm can be provided by the authors upon request.

4. Numerical applications

The projection method presented has been applied to different types of real data sets and also to a simulated data set to illustrate its practical interest.

4.1. Experimental setup

In practice, we have tested the proposed method on the different simulated and real data sets by solving the optimization problem $\mathcal{P}_{r,x}$ using Algorithm 1 in addition to the proposed Metropolis-Hastings algorithm (Algorithm 2). A distance matrix is required each time. For the quantitative data, the Euclidean distance between points

$y_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, is computed by the known formula $d_{ij} = \sqrt{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2}$. For categorical data, the distance between two soybean diseases (i, j) is given through Eskin dissimilarity (or proximity) measure (Boriah *et al.* (2008)) computed by the formula $p_{ij} = \sum_{t=1}^{Q} w_t p_{ij}^t$ where $p_{ij}^t = \begin{cases} 1 & \text{if } i^t = j^t \\ \dfrac{n_t^2}{n_t^2 + 2} & \text{else} \end{cases}$, p_{ij}^t is the per-attribute Eskin dissimilarity between two values for the categorical attribute indexed by t, w_t is the weight associated to the attribute t called w_t which is defined by: $w_t = \dfrac{1}{Q}$, Q is the number of attributes and n_t is the number of values taken by each attribute. Then, using the following formula that transforms dissimilarities into similarities: $s_{ij} = 1 - p_{ij}$, the distances can be obtained by the standard transformation formula converting similarities to distances: $d_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$.

In addition, to compute the distances between the curves of functional data, we have chosen a measure of proximity similar to that studied by Ieva *et al.* (2012). In their paper, the authors develop a proper classification designed to distinguish the grouping structures by using a functional k-means clustering procedure with three sorts of distances. For our work we chose one of these three proximity measures as their results are similar. The proximity measure chosen between two curves F_i and F_j is the following: $d_0(F_i, F_j) = \sqrt{\int_{\mathcal{T}} (F_i^0(t) - F_j^0(t))^2 dt}$. This measure is calculated using the function *metric.lp()* of the *fda.usc* package for the **R** software (Febrero-Bande and Oviedo de la Fuente (2011)).

To solve the different optimization problems presented in Algorithm 1, we used the optimization toolbox available in MATLAB. For problems $\mathcal{P}_r$ and $\mathcal{P}_{r,a}$, we first applied PCA for quantitative data and MDS for categorical and functional data; a linear programming package was then used to solve the optimization problems with an interior-point algorithm. Problems $\mathcal{P}_{r,x}$ and $\mathcal{P}_{R,x}$ are nonlinear optimization problems; therefore, we used a nonlinear programming package to solve them. The algorithms cited in Section 2.2 can be used here but we recommend to use the active-set algorithm. Indeed, to choose the best algorithm in our optimization problems, we tried the different algorithms and chose the algorithm that gives the smallest value of $\sum_{i=1}^{n} r_i$ in the shortest time compared to the other algorithms.

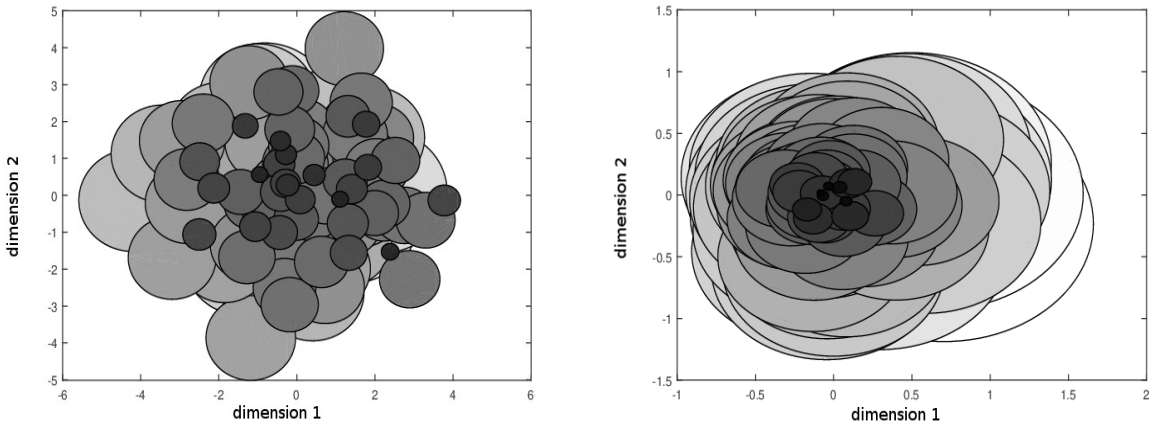
Algorithm 2 can provide a good solution if the parameters λ, σ and T are chosen adequately. For instance, λ should be such that the points belonging to unsaturated constraints are chosen with small probabilities. Therefore, we took it equal to 100. For the other parameters σ and T, we took their values in the range from 0.01 to 100.

Moreover, the visualization of the projection of each point i in $\mathbb{R}^2$ is represented as a circle having x_i as the center and r_i as the radius in a two-dimensional space, where the horizontal and vertical axes correspond to the first and the second dimension associated to the projection in $\mathbb{R}^2$, respectively. The projected point belongs to this circle and this is the specificity of our method. For each data set, the circles obtained for each point after solving the optimization problem $\mathcal{P}_{r,x}$ are shown. To compare the projection quality of our representation with that obtained by PCA and KPCA, we used the squared cosine values as projection quality, and for MDS, the Stress-per-point (SPP). Indeed, for PCA and KPCA, we plotted the projected points indexed by their squared cosine values and for MDS, we used the smacof package in R to compute the stress-per-point and to plot the bubble plot represented the stress-per-point.

4.2. A simulation study

To evaluate the performance of projection under pairwise distance control method, we conducted a simulation study. We generated 100 random samples of y_i from a 5-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix I, the identity matrix, and we calculated the Euclidean distances between pairs (y_i,y_j) for $1 \leq i < j \leq n$. The projection result was compared with those obtained by KPCA.

Figure 3 shows the results of the projection of the simulated data using the proposed method and KPCA. By comparing Figure 3a and Figure 3b, it can be shown that the projection quality of points using KPCA is somehow dependent on the position of the points in the reduced space. Indeed, the projection is likely to give better local projection quality if the projected point is located near to the center $(0, 0)$. On the contrary, the proposed method gives local projection quality without giving any importance to the position of the points in the reduced space. This result can also be shown in the real data sets.



(a) Projection under pairwise distance control (b) Kernel PCA

Figure 3.: Projection of the simulated data using the proposed method (a) and Kernel PCA (b).

This simulated data illustrates the originality and the efficiency of the proposed method in giving a good local projection quality.

4.3. Introducing the real data sets

Four real data sets were used and divided into three categories:

- Quantitative data: Iris and car data sets.
- Categorical data: Soybean data set.
- Functional data: Coffee data set.

The Iris data set (Anderson (1935)) is a famous data set and is presented to show that the projection works as expected. This data set contains 3 classes of 50 instances each, where each class refers to a species of Irises. The four variables studied in this data set are: sepal length, sepal width, petal length and petal width (in cm). The car data set (Saporta (2006)) is a data set studied in the book by Saporta (Table 17.1, page 428). This data set describes 18 cars according to various variables (cylinders, power, length, width, weight and speed).

The soybean data set (Stepp (1984)) from *UCI Machine Learning Repository* characterizes 47 soybean disease case histories defined over 35 attributes. Each observation is identified by one of the 4 diseases: Diaporthe Stem Canker (D1), Charcoal Rot (D2),

Rhizoctonia Root Rot (D3) and Phytophthora Rot (D4).

The coffee data set is a time series data set used in chemometrics to classify food types. It is a functional data set where 56 samples of coffee are available with 286 timestamps for each sample (as a result of spectroscopic analysis). This kind of time series is common in many applications in food safety and quality assurance and was taken from the *UCR time Series Classification and Clustering* website (Chen *et al.* (2015)). Coffea Arabica and Coffea Canephora variant Robusta are the two species of coffee bean that have acquired a worldwide economic importance, and many methods have been developed to discriminate between these two species by chemical analysis (Briandet *et al.* (1996)).

4.4. Results from the real data sets

4.4.1. Data visualization in $\mathbb{R}^2$

The optimization results for these four data sets are given in Table 1. For each data, the sum of radii $\sum_{i=1}^{n} r_i$ obtained using Algorithm 1 and Algorithm 2 is provided.

Table 1.: Solution of problem $\mathcal{P}_{r,x}$ for data sets using Algorithm 1 and Algorithm 2.

	$\sum_{i=1}^{n} r_i$	
	Algorithm 1	Algorithm 2
Iris	16.19	17.2
Cars	3.27	3.35
Soybean	3.98	3.93
Coffee	21.68	21.97

Based on Table 1, the solutions of Algorithm 2 for the different data sets are shown to be very close to those obtained using Algorithm 1. Thus, the radii obtained are estimated to be close to the optimum. Moreover, it is interesting to note here that the number of iterations N in Algorithm 1 has an important role in finding the minimal value of $\sum_{i=1}^{n} r_i$ for problem $\mathcal{P}_{r,a}$ and then for problem $\mathcal{P}_{r,x}$ and also to reduce the computer speed time. In fact, the important decrease in the value of $\sum_{i=1}^{n} r_i$ occurred in the first 500 iterations through the loop of 1000 iterations, and then a small decrease

occurred after 500 iterations. This small decrease in value of $\sum_{i=1}^{n} r_i$ after 500 iterations shows that a size of 500 iterations can be a good choice for the Algorithm 1 since all the studied data sets are concerned. Indeed, this result can be observed for all data sets presented in our application with approximately 500 iterations.

Iris data set: Figure 4 depicts the result of projection under pairwise distance control for the Iris data set. In the projection of the Iris data set shown in Figure 4, it is interesting to note that the two areas are well separated. This corresponds to the well-known fact that Iris versicolor and virginica are close whereas the species Iris setosa are more distant.

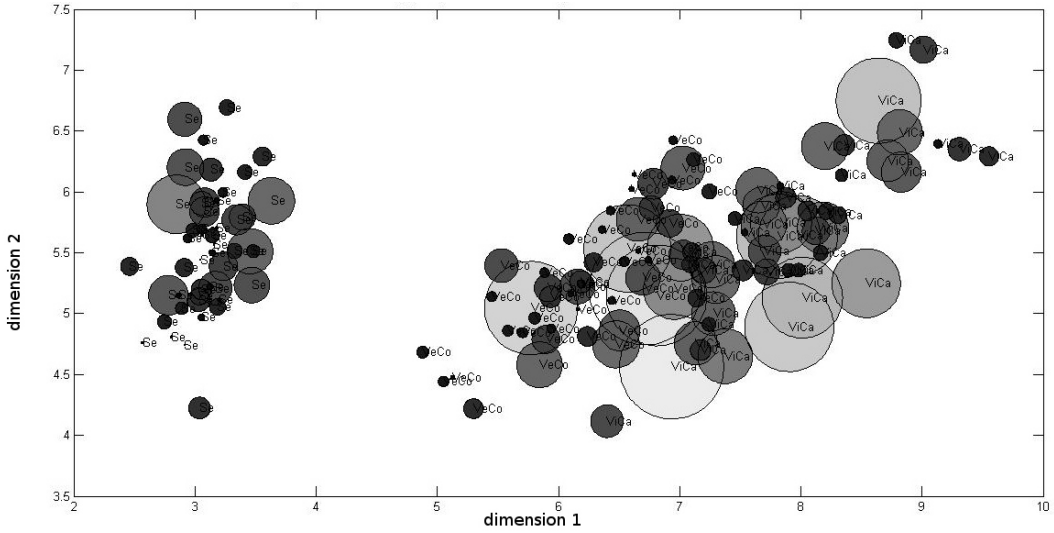


Figure 4.: Projection of the Iris data set using projection under pairwise distance control method. Two well separated groups can be observed.

Referring to the original data, the Iris data set contains three classes corresponding to the three types of Iris plants and one class is linearly separable from the other two classes. This result clearly appears in our projection.

Moreover, we have compared the local projection quality of PCA, KPCA and MDS with the local projection quality obtained using projection under pairwise distance control. By comparing the projection of PCA with the projection of our method for the Iris data set given respectively in Figures 5 and 4, we can say that our method projected the points without giving any importance to any group. Indeed, Figure 5 depicts a group with small values of quality measure and another group with high

values of quality measure, whereas the radii obtained by projection under pairwise distance control method are distributed in an equivalent way.

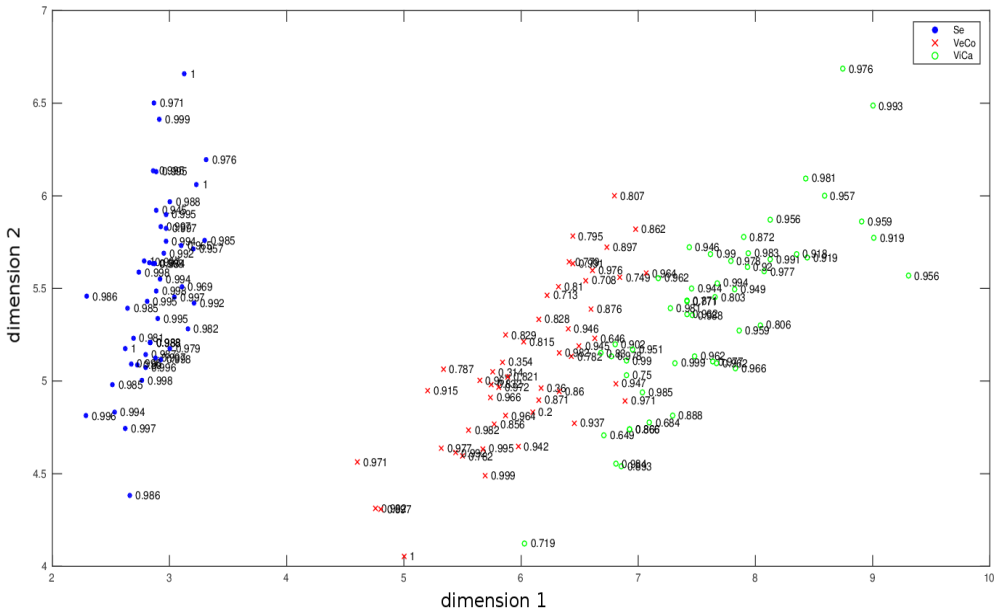
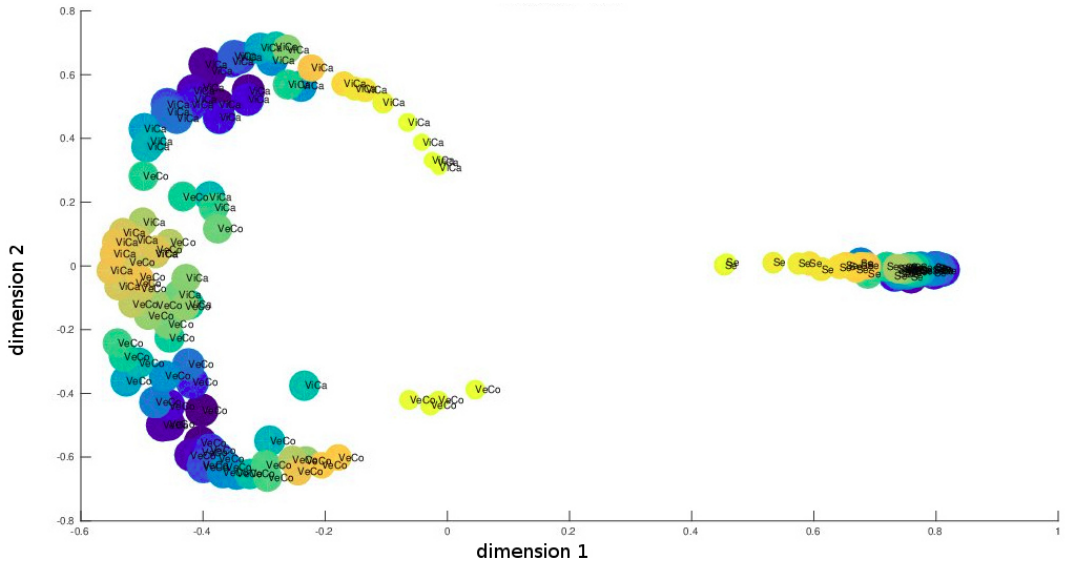


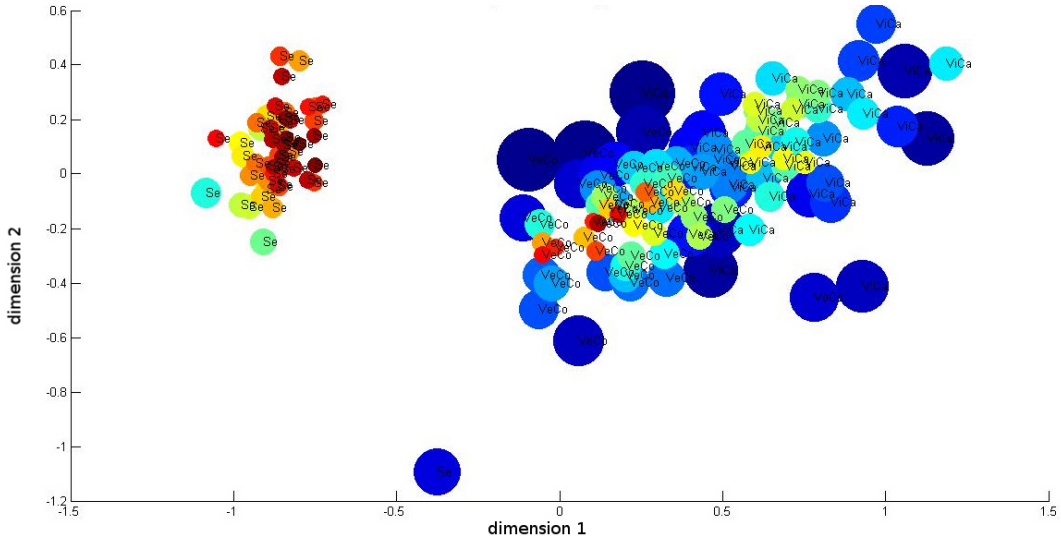
Figure 5.: Projection of the Iris data set using PCA.

For KPCA, we plotted the squared cosine values as circles to make the representation clearer, especially for the Iris data set as the Iris setosa species are projected next to each other. From Figure 6a, we can conclude that in each category, the points that have close quality values are located side by side.

Furthermore, by comparing the proposed projection method with the one obtained by MDS, it can be concluded that, as is the case when using PCA, the points in Figure 6b are projected by giving more importance to the Iris setosa group. Indeed, almost all the red circles (indicating a very good projection) are assigned to the Iris setosa species. Moreover, the comparison of the position of points in the reduced space in terms of distance between points cannot be viewed in this classical method as the points in the reduced space is not in the same metric of the initial distances, whereas in our method we have conserved the metric of the initial distances.



(a)



(b)

Figure 6.: Projection of the Iris data set using KPCA (a) and MDS (b). The color convention is as follows: the darker the red color of a particular disk, the better the projection. Inversely, the darker the blue color of a particular disk, the worse the projection.

Cars data set: The projection of points using projection under pairwise distance control for the car data set is shown in Figure 7. The expensive cars, the "Audi 100", "Alfetta-1.66", "Datsun-200L" and "Renault 30" are well-separated from the low-standard cars, the "Lada-1300", "Toyota Corolla", "Citroen GS Club" and "Simca 1300". Moreover, we can assert that the projected points obtained using projection

under pairwise distance control method are well separated as there are no circle intersections.

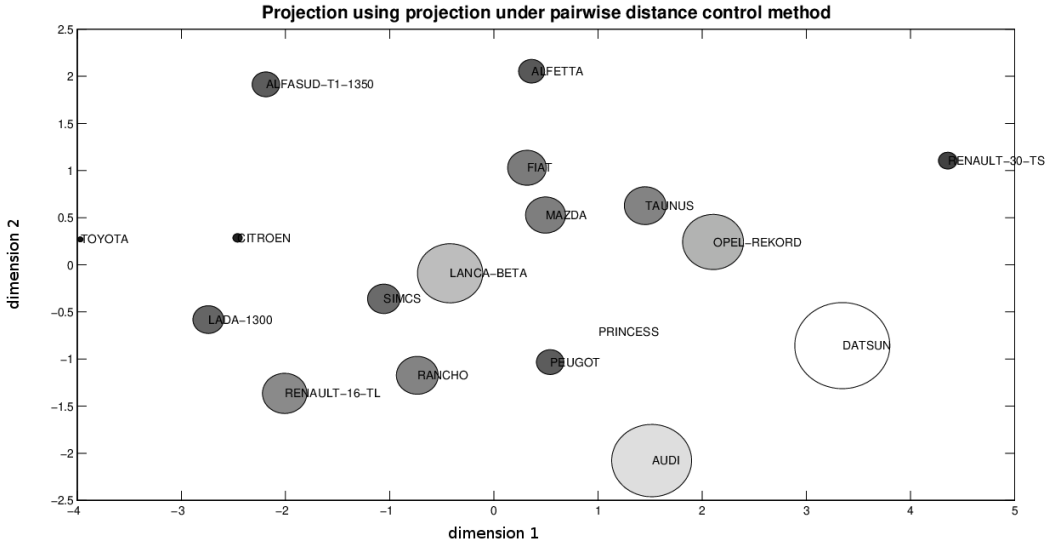


Figure 7.: Projection of the car data set using projection under pairwise distance control.

By comparing our projection with the projection obtained using PCA presented in Figure 8, it can be shown that in the projection of PCA, there is a group with small values of quality measure located at the center, which corresponds to the cars: Lanca-Beta, Mazda, Fiat, Simcs and Rancho, and a group with high values of quality measure located far from the center. Thus, as shown for the Iris data set, projection under pairwise distance control method projects the points without giving any importance to the position of the points in the reduced space.

Regarding KPCA, we can see in Figure 9a that the points with navy circles are almost all located almost around the same y-axis coordinates and the same applies for the red circles. So the local quality for KPCA is dependent on the position of the points. It can also be noticed that the cars Princess, Mazda, Fiat and Peugeot located in the same area with small circles. Therefore, the only conclusion that we arrive at is in relation to the size of the circles and to the quality of the projected points. However, it is not possible to conclude anything about the closeness of these 4 points as the distances here are in the feature space and are not related to the original space. In Figure 7, we can however conclude that the two cars, the Mazda and Fiat, are

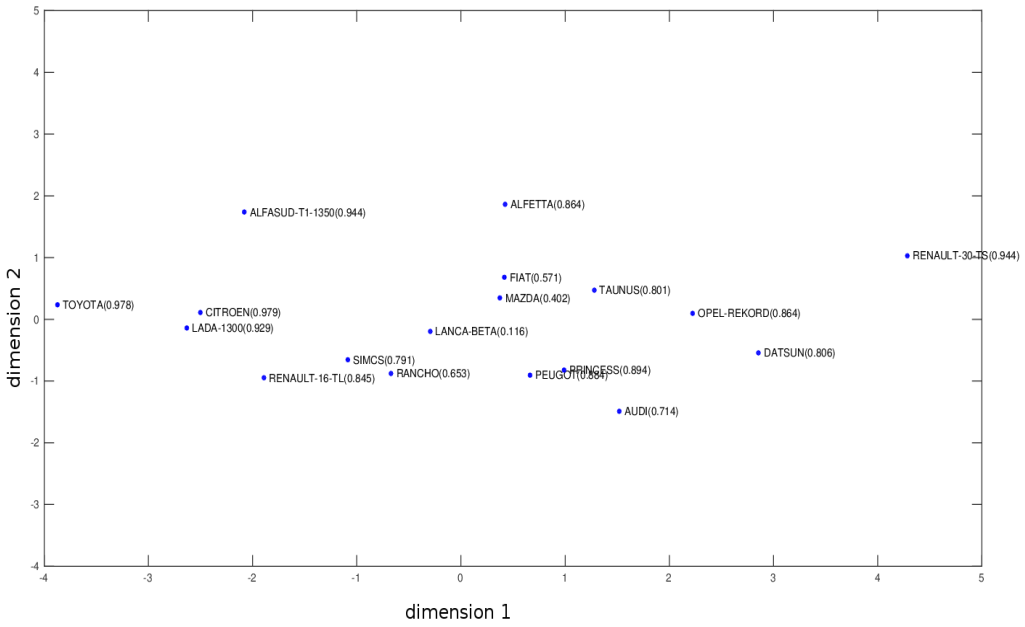
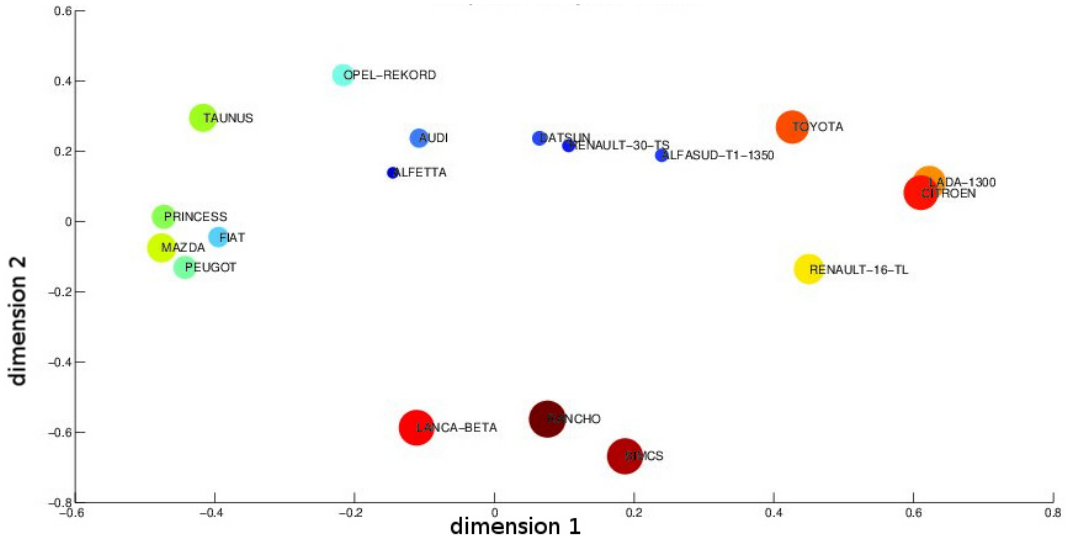


Figure 8.: Projection of the cars data set using PCA.

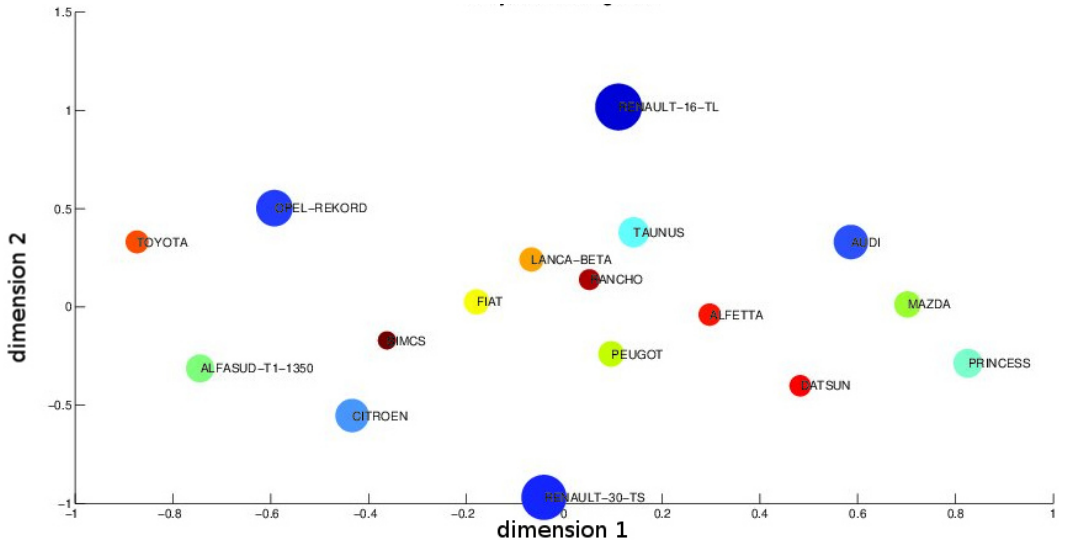
well projected in the reduced space, and they have similar characteristics as these two cars are close. The same conclusion can be made for the Peugeot and Princess cars. From this, it is possible to conclude that there is a large difference between the two cars, the "Toyota" and "Renault 3" as the distance between these two cars is significant. Conversely, the distance between the "Lada1300" and "Citroen" is small, thus indicating the closeness of these two cars. Note that these two cars are very well projected, resulting in a very good interpretation of the distance between them.

Therefore, the pairwise distances are meaningful in our method and give an interpretation about the distances between points whereas the distances between the projected points using PCA, KPCA and MDS are not interpretable as the cosine values and the Stress-per-point cannot be interpreted as distances. This is a particular strength of our method. Projection under pairwise distance control suggests an absolute interpretation whereas the other methods provide a relative one.

For the qualitative and functional data sets and using MDS, recall the definition of the Gram matrix called B which is equal to $X'X$ where X is the coordinate matrix in



(a)



(b)

Figure 9.: Projection of the car data set using KPCA (a) and MDS (b).

the reduced space. Thus, it is necessary to verify that the matrix B obtained by the MDS method is semi-definite positive to use the squared cosine as the quality measure because the starting point of optimization is obtained from MDS. After this, in case of positiveness of matrix B, the quality measure can be calculated.

Soybean data set: In the projection of the soybean data set, four classes are shown in Figure 10 and each class contains the disease number of the class. The whole set of points can however be divided in two large classes. Indeed, it is clear that Class 2 is

559 well separated from the other classes as there is no intersection between the circles of
560 Class 2 and the circles of other classes. Moreover, Class 1 can be considered as well
separated class from Classes 3 and 4 if the largest circle D_3 is not taken into account.

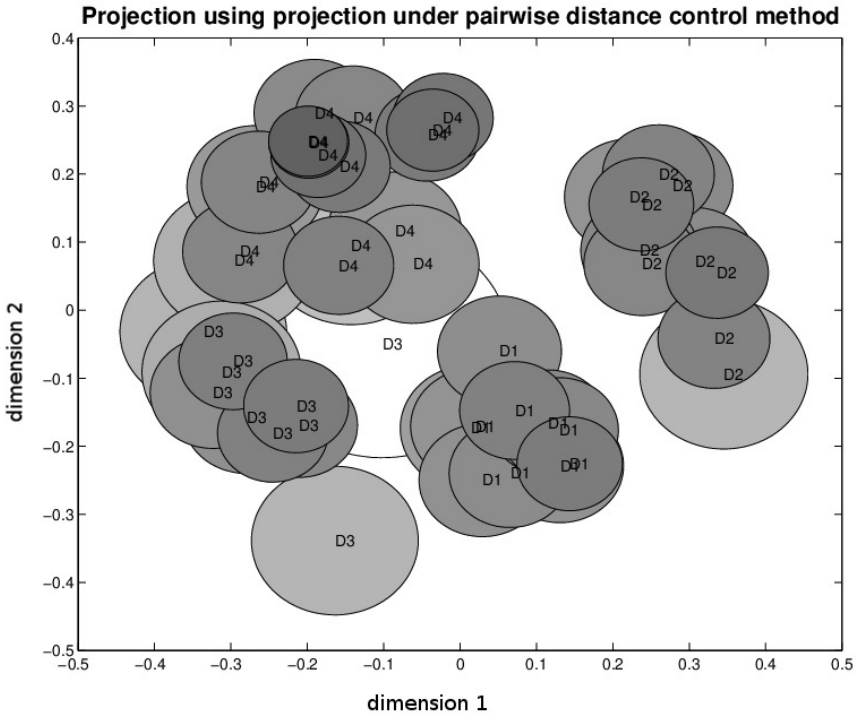


Figure 10.: Projection under pairwise distance control for the soybean data set. Four groups are presented, indexed by D1, D2, D3 and D4.

561

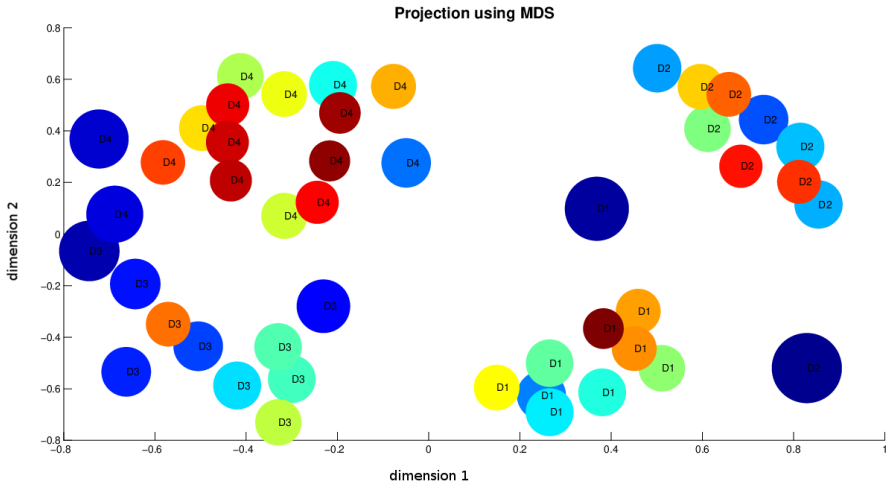


Figure 11.: MDS for the soybean data set. Four groups are presented, indexed by D1, D2, D3 and D4.

Classes 3 and 4 are not well separated at all, as there are different intersections between the circles of these two classes. This result is shown in Stepp (1984) which labels the first two classes as "normal" and the latter two classes as "irrelevant". A comparison of results from projection under pairwise distance control with PCA and KPCA is not possible for this data set because the matrix B is not semi-definite positive. Regarding Figure 11, it is clear that Class 4 exhibits the worst projection quality, whereas Classes 1 and 2 show better projection quality. Therefore, it is possible to draw the same conclusion for the Iris and car data sets when using MDS as a projection method, the projection quality of points is dependent on the class of the points.

Coffee data set: The coffee data set has been studied in several articles (Briandet *et al.* (1996), Bagnall *el al.* (2012)) and different classification methods have shown the different groups contained in this data set. The grouping structure obtained can be clearly seen in Figures 12 and 13

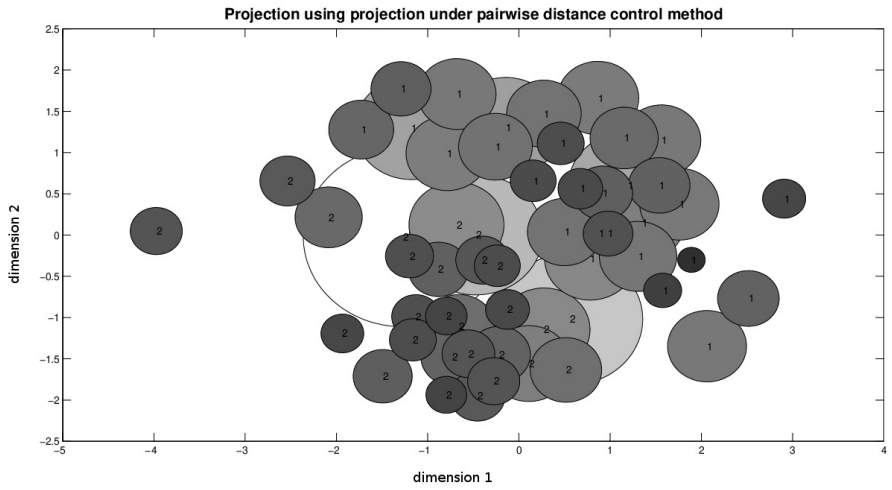
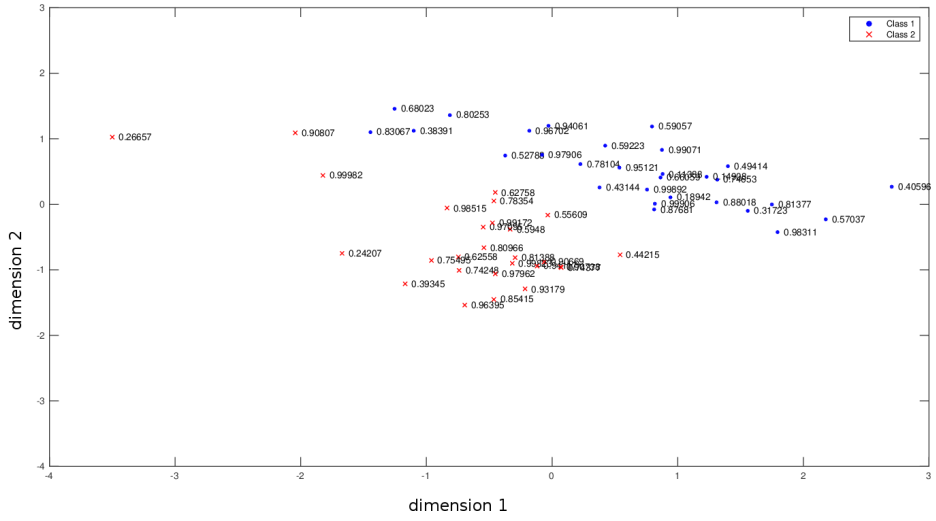


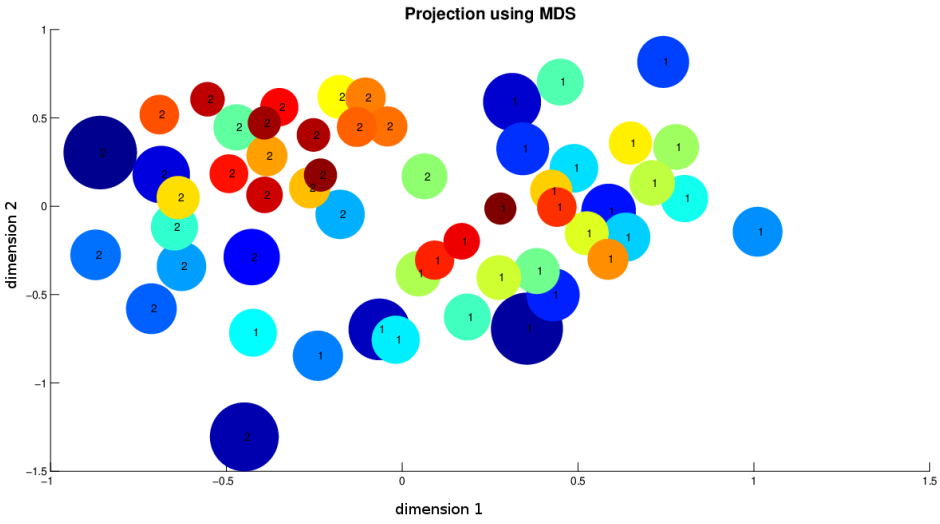
Figure 12.: Projection of the coffee data set using projection under pairwise distance control. Two clusters, indexed 1 and 2, indicate the Arabica and Robusta classes respectively.

In Figure 12, we show that we have succeeded in differentiating the Arabica from Robusta coffee. These two classes are clearly presented, the first class indexed by number 1, corresponding to Arabica coffee, and the second one indexed by number 2, corresponding to Robusta coffee. These classes are not well separated in comparison with the results of quantitative data, since there are many intersections. Therefore,

the representation of the points as circles and not as simple points provides more information about the real point classes and shows the points that are at risk of being misplaced in a particular class.



(a)



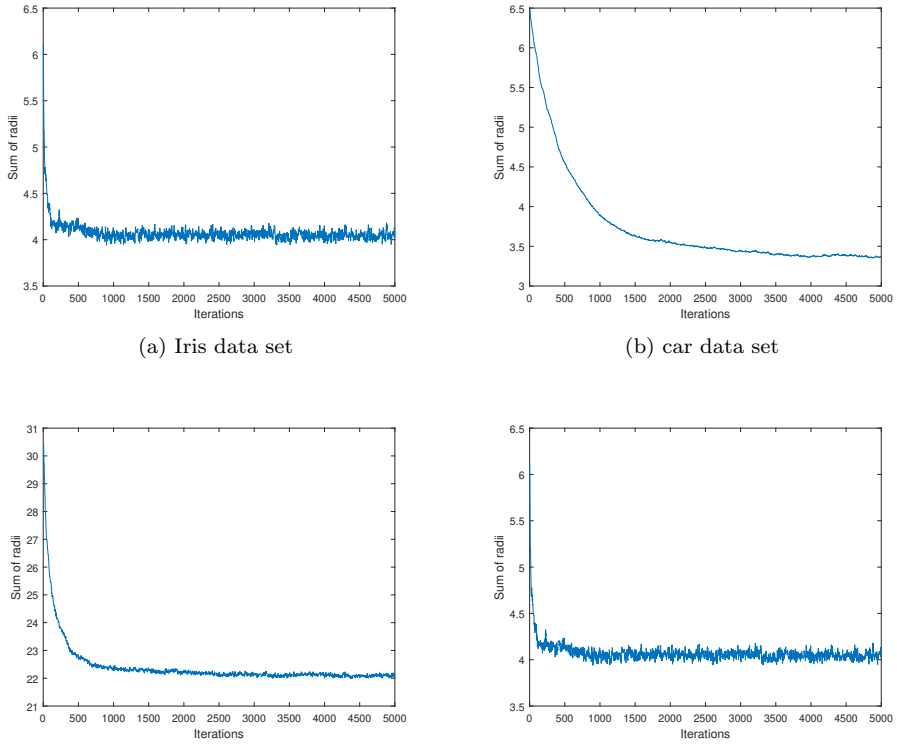
(b)

Figure 13.: Projection of coffee data set using PCA and MDS.

Figures 13a and 13b show the projection quality using PCA and MDS respectively. As all the eigenvalues of matrix B are positive, we can compute the quality measure given by PCA. Comparing the projection quality of PCA and projection under pairwise distance control provided by Figures 13a and 12, respectively, it can be seen that the

quality of projection of the set of points is quite steady.

Additionally, Algorithm 2 was applied to these data sets. The trace plots of the optimization problem $\mathcal{P}_{r,x}$ are shown in Figure 14 after 5000 iterations. It is important to note that the value of the sum of radii $\sum_{i=1}^{n} r_i$ decreases rapidly in the first iterations and stays roughly constant after 1000 iterations for the different data sets, with the exception of the car data sets. Thus, we can decrease the number of iterations from 5000 to almost 2000, or even 1000, in order to reduce the speed time.



(a) Iris data set

(b) car data set

(c) Coffee data set

(d) Soybean data set

Figure 14.: Trace plots of Metropolis Hastings for different data sets. The x-axis corresponds to the iteration number and the y-axis to the value of $\sum_{i=1}^{n} r_i$.

Finally, the computer speed time of the proposed method is compared with that using the classical methods. Table 2 shows the computer speed time for the four data sets using PCA, KPCA, MDS, Algorithm 1 and Algorithm 2. It is clear that our method takes more time than the existing methods. However, Algorithms 1 and 2 are expected to significantly increased by using the C++ programming language (instead of MATLAB currently) to produce more efficient code. In addition, by comparing the

computer speed time of the two algorithms and by referring to Table 1, the solutions obtained using Algorithm 1 and Algorithm 2 are very close, which indicates that Algorithm 2 can be used instead of Algorithm 1 to obtain a better solution faster (between two and four times faster). Thus, Algorithm 2 (Metropolis Hastings algorithm) is recommended for use as it takes less time.

Table 2.: Computer speed time (in seconds) using different methods for the four data sets

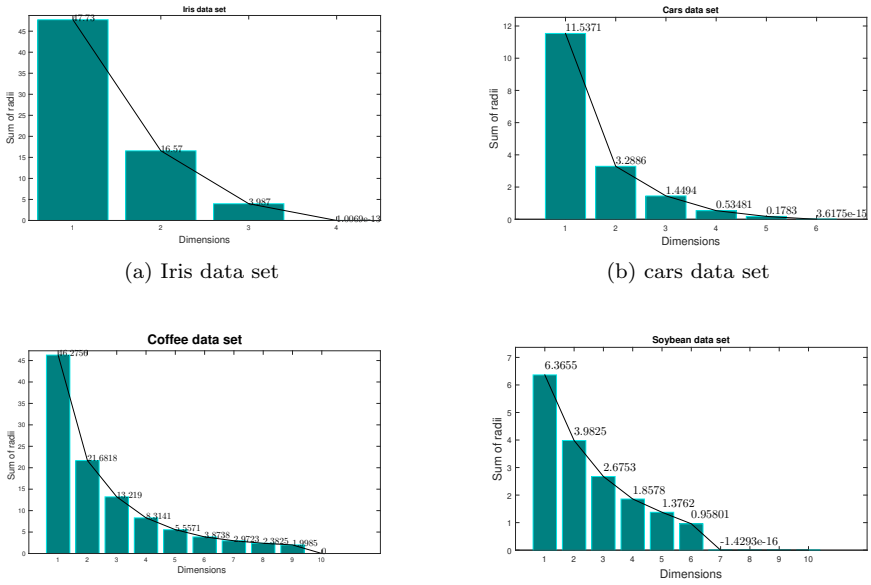
Computer speed time (sec.)					
	PCA	KPCA	MDS	Algo 1	Algo 2
Iris	3.61	5.21	5.54	1124	600
Cars	2.70	4.17	4.62	671	300
Soybean	–	–	2.65	2036	698
Coffee	3.68	–	3.18	1968	589

4.4.2. Dimensionality reduction results

Our method can also be directly used to reduce the dimensionality of data (possibly using it beyond visualization in $\mathbb{R}^2$). This only requires solving problem $\mathcal{P}_{r,x}$ using different dimension values. In Figure 15, the values of $\sum_{i=1}^{n} r_i$ were plotted as a guide for choosing the reduced number of variables. This figure shows the values of $\sum_{i=1}^{n} r_i$ for the different data sets using different dimensions. It is clear that the value of $\sum_{i=1}^{n} r_i$ decreases when the dimension increases.

The main problem, which is widely posed in dimensionality reduction methods, is the determination of the number of components that need to be kept. Many methods have been discussed in the literature (Besse (1992); Jollife (1986)) to determine the dimension of the reduced space, relying on different strategies related to a good explanation or a good prediction. Thus, with our method the choice of the reduced space dimension is related to the local projection quality of points and how much the user is interested in the projection quality of points.

Regarding the quantitative data sets (Iris and car), if the main objective of the user is to obtain a very good projection quality, then a choice of three components against four, for Iris data set and six for the car data set can be a good choice, as the



(a) Iris data set

(b) cars data set

(c) Coffee data set

(d) Soybean data set

Figure 15.: Scree plots of $\sum_{i=1}^{n} r_i$ for different dimensions for the four data sets.

value of $\sum_{i=1}^{n} r_i$ is small and there is not a large difference between this value and the values for higher dimensions. For the coffee data set, a dimensionality reduction from 56 sample time series down to 6 simple extracted features is considered as a good choice. As for the soybean data set, a reduced space dimension equal to 4 dimensions can be considered as an appropriate reduced space.

A comparison of our results with the existing results shows a coherence between them. For the Iris data set, Chiu (1996) and Liu and Setiono (1995) concluded that the number of variables could be reduced to 2 as the petal length and petal width variables are the most important variables from all the variables. For the car data set, Saporta (2006) (Table 7.4.1 page 178) noticed that the conservation of two dimensions led to the explanation of 88% of inertia, where the inertia term reflects the importance of a component. Theredore, these results seem very similar to our results, with the important decrease located between dimensions 1 and 2. The other reductions are negligible for these two data sets. A selection of variables was studied on time series coffee data set by Andrews and McNicholas (2014). Using several analysis methods, the number of selected variables ranged between 2 and 13. This result is also seen using our method, a number of reduced variables taken between 2 and 9 gives a

good projection. Regarding the soybean data set, Dela Cruz shows in his paper Dela Cruz (2015) that the 35 attributes can be reduced to 15. With our method, we have succeeded in reducing the attributes to 6 by having a very good projection of points. Hence, the results presented confirm that the dimension nonlinearly can be reduced while assessing a reasonable number of dimensions at the same time.

5. Conclusion

The purpose of this paper was to outline a new nonlinear projection method based on a new local measure of projection quality. Of course, in some projection methods, a local measure is given but this measure cannot be applied unless in cases of linear projections, and even then it is not suitable for graphical representation.

The quality of projection is given here by additional variables called radii, which enable bound on the original distances to be obtained. We have also shown that the idea can be written as an optimization problem in order to minimize the sum of the radii under some constraints. As the solution of this problem cannot be obtained exactly, we developed a stochastic optimization method.

This method has several advantages. Firstly, it is a nonlinear projection method that takes into account the projection quality of each point individually. Secondly, the distances between projected points are related to the initial distances between points offering a way to easily interpret the distances observed in the projection plane. The projection quality of each point can even then be used outside our method, as a post-processing of PCA or MDS for example. Finally, it appears to be efficient in terms of dimensionality reduction for the selection of the dimension of the reduced space based on the local quality of projection.

As perspectives, a lower bound for the optimization problem is needed and this radii approach could also be applied to other methods.

References

Anderson, E. (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society* 59:2–5.

667 Andrews, J. L. and McNicholas, P. D. (2014). Variable Selection for Clustering and Classifi-
668 cation. *Journal of Classification* 31:136-153.
669 Bagnall, A., Davis, L., Hills, J., and Lines, J. (2012). Transformation Based Ensembles for
670 Time Series Classification. *Proceedings of the 12th SIAM International Conference on Data*
671 *Mining* 307–319.
672 Berge, C., Froloff, N., Kalathur, RK., Maumy, M., Poch, O., Raffelsberger, W. and Wicker, N.
673 (2010). Multidimensional fitting for multivariate data analysis. *Journal of Computational*
674 *Biology* 17:723–732.
675 Besse, P(1992). PCA stability and choice of dimensionality. *Statistics & Probability Letters*
676 13:405-410.
677 Boggs, P. T. and Tolle, J. W. (1995). Sequential quadratic programming. *Acta Numer* 4:1–51.
678 Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*
679 (2nd ed.) New York: Springer-Verlag.
680 Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity Measures for Categorical Data:
681 A Comparative Evaluation. *Proceedings of the SIAM International Conference on Data*
682 *Mining.*
683 Briandet, R., Kemsley, E. K., and Wilson, R. H. (1996). Discrimination of arabica and robusta
684 in instant coffee by fourier transform infrared spectroscopy and chemometrics. *Journal of*
685 *Agricultural and Food Chemistry* 44(1):170–174.
686 Chan, W. W-Y. (2006). A survey on multivariate data visualization in Science and technology.
687 *Department of Computer Science and Engineering Hong Kong, University of Science and*
688 *Technology* 8(6):1–29.
689 Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. (2015). *The*
690 *UCR Time Series Classification Archive.* `www.cs.ucr.edu/~eamonn/time_series_data/`.
691 Cheung, L. W. (2012). Classification approaches for microarray gene expression data analysis.
692 *Methods in Molecular Biology* 802:73–85.
693 Chinchilli, V. M. and Sen, P. K. (1987). Multivariate Data Analysis: Its Methods. *Chemomet-*
694 *rics and Intelligent Laboratory Systems* 2:29–36.
695 Cleveland, W. S. and McGill, M. E. (1988). Dynamic Graphics for Statistics. *Wadsworth and*
696 *Brooks/Cole*, Pacific Grove, Canada.
697 Chiu, S. L. (1996). Method and Software for Extracting Fuzzy Classification Rules by Sub-
698 tractive Clustering. *Proceedings of North American Fuzzy Information Processing Society*
699 *Conference.*

Cristofari, A., De Santis, M., Lucidi, S. and Rinaldi, F. (2007). A Two-Stage Active-Set Algorithm for Bound-Constrained Optimization. *J. Optim. Theory Appl.* 172(2):369–401.

Conn, De A. R., Gould, N. I. M. and Toint, Ph. L. (2000). Trust Region Methods, SIAM.

Dela Cruz, G. B. (2015). Comparative Study of Data Mining Classification Techniques over Soybean Disease by Implementing PCA-GA. *International Journal of Engineering Research and General Science* 3(5):6–11.

Dempster, A. P. (1971). An overview of multivariate data analysis. *Journal of Multivariate Analysis* 1(3):316–346.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

Ieva, F., Paganoni, A.M., Pigoli, D., and Vitelli., V. (2012). Multivariate functional clustering for the analysis of ECG curves morphology, *Journal of the Royal Statistical Society. Applied Statistics, series C* 62(3):401–418.

Inselberg, A. (1985). The Plane with Parallel Coordinates. *Special Issue on Computational Geometry, The Visual Computer* 1:69–91.

Jackson, J. (1991). A Users Guide to Principal Components, *John Wiley & Sons, New York*.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: why imposing the wrong constraints helps. *The Journal of Finance* 58:1651–1683.

Johansen, A. M. and Evers, L. (2007). *Monte Carlo Methods*. Department of Mathematics, University of Bristol.

Jollife, I. T. (1986). Principal Component Analysis, Springer, New York

Keim, D. A. and Kriegel, H. P. (1996). Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on Knowledge and Data Engineering* 8(6):923-938.

Lee, J. A. and Verleysen, M. (2007). Nonlinear Dimensionality Reduction. Springer.

Liu, H. and Setiono, R. (1995). Chi2: feature selection and discretization of numeric attributes. *Proceedings Seventh International Conference on Tools with Artificial Intelligence*.

Febrero-Bande, M., Oviedo de la Fuente, M. (2011). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of statistical software* 51(4).

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). Multivariate analysis, *Academic Press*, London.

Murty, K. G. (1983). Linear programming. John Wiley & Sons, New York.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* 290(5500):2323–2326..

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* 18(5):401–409.

Saporta, G. (2006). Probabilités, analyse des données et statistique. *Technip*.

Schölkopf, B. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 10(5): 1299–1319.

Stepp, R. (1984). Conjunctive conceptual clustering. Doctoral dissertation, department of computer science, university of Illinois, Urbana-Champaign, IL.

Svante, W., C. Albano, W. J. DunnIII, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg , M. Sjostrom. (1984). Multivariate Data Analysis in Chemistry. *Chemometrics* 138:17–95.

Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319-2323.

Togerson, W. S. (1958). Theory and methods of scaling, New York: Wiley.

Van der Hilst, R., de Hoop, M., Wang, P., Shim, S.-H., Ma, P. and Tenorio, L. (2007). Seismo-stratigraphy and thermal structure of earth's core-mantle boundary region. *Science* 315:1813–1817.

Wong, E. (2011). Active-Set Methods for Quadratic Programming. Ph.D. thesis, university of California, San Diego.

Appendix

Proof of proposition 3.1

Let us consider a point x_i such that for an index j, the following inequality is saturated:

$$|d_{ij} - \|x_i - x_j\|| \leq r_i + r_j,$$

and the other inequalities involving i are not saturated. Then, the corresponding solution can be improved by moving x_i along the line $x_j - x_i$ in order to decrease r_i and $|d_{ij} - \|x_i - x_j\||$.

Proof. The above condition means that x_i is rewritten as $x_i + a(x_j - x_i)$ with $a \in \mathbb{R}$ and we look for a such that $|d_{ij} - \|x_i + a(x_j - x_i) - x_j\|| < r_i + r_j$. In particular $a \leq 0$ if $d_{ij} - \|x_i - x_j\| \geq 0$ and is otherwise > 0. Let us now consider the other inequalities corresponding to index pairs (i, k) with $k \neq j$. For each of them, $\exists a \in [a_k^{'}, a_k^{''}]$ with $a_k^{'} < 0$ and $a_k^{''} > 0$ such that

$$|d_{ij} - \|x_i + a(x_j - x_i) - x_j\|| \leq r_i + r_j,$$

as these constraints are unsaturated. Finally, taking a different from 0 in $[a^{'}, a^{''}]$ with $a^{'} = \max_k a_k^{'}$ and $a^{''} = \min_k a_k^{''}$, all constraints involving i get unsaturated so that r_i can be decreased, thereby decreasing the objective function. Depending on whether a must be negative or positive, we take $a = a^{'}$ or $a = a^{''}$ respectively.

□

Proof of proposition 3.2

Let $r_1, \ldots, r_n; x_1, \ldots, x_n$ *be a feasible solution of* $\mathcal{P}_{r,x}$*, if* $\exists a$ *such that* $\eta(a) < \sum_{i=1}^{n} r_i$ *with* $\eta(a) = \sum_{1 \leq i < j \leq n} |d_{ij} - a\|x_i - x_j\||$*, then* $\exists\ \tilde{r}_1, \ldots, \tilde{r}_n$ *a solution of* $\mathcal{P}_{r,a}$ *such that* $\sum_{i=1}^{n} \tilde{r}_i < \sum_{i=1}^{n} r_i$.

Proof. Let us consider $r_1, \ldots, r_n; x_1, \ldots, x_n$ a feasible solution of problem $\mathcal{P}_{r,x}$ and $a, \tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_n$ a solution of $\mathcal{P}_{r,a}$ where a is kept constant. For the solution of $\mathcal{P}_{r,a}$, for each point i, we have a certain saturated constraint associated to point k denoted by $C_{ik(i)}$, otherwise we can easily saturate it using proposition 3.1. Thus, we have:

$$
\begin{gathered}
|d_{i1} - a\|x_i - x_1\|| \leq \tilde{r}_i + \tilde{r}_1 \\
\vdots \\
|d_{ik(i)} - a\|x_i - x_{k(i)}\|| = \tilde{r}_i + \tilde{r}_{k(i)} \\
\vdots \\
|d_{ij} - a\|x_i - x_j\|| \leq \tilde{r}_i + \tilde{r}_j \\
\vdots \\
|d_{in} - a\|x_i - x_n\|| \leq \tilde{r}_i + \tilde{r}_n.
\end{gathered}
$$

Then, $|d_{ik(i)} - a\|x_i - x_{k(i)}\|| = \tilde{r}_i + \tilde{r}_{k(i)} \geq \tilde{r}_i$. By summing for all points i, for $i = 1, \ldots, n$, we obtain:

$$
\sum_{i=1}^{n} |d_{ik(i)} - a\|x_i - x_{k(i)}\|| \geq \sum_{i=1}^{n} \tilde{r}_i.
$$

Thus, $\sum_{1 \leq i < j \leq n} |d_{ij} - a\|x_i - x_j\|| \geq \sum_{i=1}^{n} |d_{ik(i)} - a\|x_i - x_{k(i)}\|| \geq \sum_{i=1}^{n} \tilde{r}_i$.

Note $\eta(a) = \sum_{1 \leq i < j \leq n} |d_{ij} - a\|x_i - x_j\||$, then if $\eta(a) < \sum_{i=1}^{n} r_i$ there is a solution of $\mathcal{P}_{r,a}$ such that $\sum_{i=1}^{n} \tilde{r}_i < \sum_{i=1}^{n} r_i$. □