Projection under pairwise distance controls

Alawieh Hiba, Nicolas Wicker, Christophe Biernacki

Noname manuscript No.
(will be inserted by the editor)

Projection under pairwise distance control

Alawieh Hiba · Nicolas Wicker · Christophe Biernacki

Received: date / Accepted: date

Abstract Visualization of high-dimensional and possibly complex (non continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. 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 in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in $\mathbb{R}^2$ with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction.

Keywords Data visualization · dimension reduction · multidimensional scaling · principal component analysis.

1 Introduction

Several domains in science use data with large number of variables in their studies such as in biological [12,18], chemical [31], geographical [33], financial [22] studies and many others studies. These data can be viewed as a large matrix and extracting results from this type of matrix is often hard and complicate. 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 were developed to study these data as dimensionality reduction, variable selection, cluster analysis and others. These methods are directly related to the main goal of the researcher

H. Alawieh, N. Wicker, C. Biernacki
Université Lille 1, - UFR de Mathématiques, cité scientifique, 59655 Villeneuve d'Ascq, France
E-mail: alawieh.hiba@gmail.com

such as dimensionality reduction to summarize the data, variable selection to choose the pertinent variables from the set of candidate variables and cluster analysis to group the objects or variables. In our study, we are focus on dimension reduction.
Dimensionality reduction techniques can be used in different ways like data dimensionality reduction that projects the data form a high-dimensional space to a low-dimensional space or data visualization that provides a simple interpretation of the data in $\mathbb{R}^2$ or $\mathbb{R}^3$.
Many data dimensionality reduction and data visualization methods have been proposed to drop the difficulties associated to the high dimensional data [27, 17,13,25,9]. To quote a few, principal component analysis (PCA) [21], multidimensional scaling (MDS) [32], scatter plot matrix [14], parallel coordinates [20] and Sammon's mapping [28] are some of the known used methods.
Scatter plot matrix, parallel coordinates and Sammon's mapping methods are widely used to visualize multidimensional data sets. The first two methods have as inconvenience that when the number of dimensions grows, important dimensional relationships might not be visualized. Concerning Sammon's mapping method, the inconvenience is similar to that founded in PCA and MDS from the point of view of projection quality. Indeed, the quality of projection assessed by the percentage of variance that is conserved or by the stress factor is a global quality measure and takes only into account what happens globally but in some projection methods like 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 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 as happens in PCA but cannot be applied to the case of non-linear projection. ***Moreover, PCA will fail to give us a "good" representation in case of non-linear configurations therefore, Kernel PCA has been developed to extract non-linear principal components. Many disadvantages of Kernel PCA can be founded especially by comparing with other method. The storage of the data as dot products in the Gram matrix is too expensive since the size of kernel matrix increases quadratically with the number of data [5] . Additionally, in kernel PCA we can find the same local quality projection measure problem designed in PCA.***

In this paper, we propose a new non-linear 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. This projection leads to a representation of the points as circles with a different radius associated to each point. Henceforth, this method will be called "Projection under pairwise distance control". The main contributions of this study are to give a simple data visualization in $\mathbb{R}^2$ with a straightforward interpretation and provide a new variant of dimensionality reduction. First, the new projection method is presented in Section 2. Then, in Section 3, the algorithms used in the resolution of optimization problems related to this

method are illustrated. Next, Section 4 shows the application of this method to various real data sets. Finally, Section 5 concludes this work.

2 Projection under pairwise distance control

Let us consider n points given by their pairwise distance noted d_{ij} for $i, j \in \{1, \ldots, n\}$. The task here is to project these points using distances into a reduced space $\mathbb{R}^m$ by introducing additional variables, called hereafter radii, that indicate to which extent the projection of each point is accurate. The local quality is then given by the values of the radii. A good quality projection of point i is indicated by a small radius value noted r_i. It will be important to note that both units of d_{ij}'s and r_i's are identical, allowing direct comparison. Before developing our method, an overview of principal component analysis (PCA) is presented to highlight the interest of our method.

2.1 Principal Component Analysis (PCA)

PCA method is the most used method in the data visualization and dimensionality reduction. This method is a linear projection technique applied when the data is linearly separable. PCA problem can be stated as an optimization problem involving the squared Euclidean distances [27]. This optimization problem is the following:

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

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 is the reduced space dimension.

By construction, PCA cannot take into account non-linear structures, since it describes the data in terms of a linear subspace. Therefore, a new concept of PCA has been developed to lead with non-linear structures that is called Kernel PCA

2.2 Kernel PCA (KPCA)

The KPCA idea is to perform PCA in a feature space noted $\mathcal{F}$ produced by a non-linear mapping of data from its space into the feature space $\mathcal{F}$, where the low-dimensional latent structure is, hopefully, easier to discover. The mapping function noted Φ is considered as:

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

The original data y_i ***is then represented in the feature space as*** $\Phi(y_i)$***, where*** $\mathcal{F}$ ***and*** Φ ***are not known explicitly but obtained thanks to the "kernel trick". So, from the Gram matrix*** K ***given by*** $K = k_{ij}$ ***with*** $k_{ij} = K(y_i, y_j) = \langle \Phi(y_i), \Phi(y_j) \rangle$***, KPCA is based to find the first*** m ***eigenvectors corresponding to the largest eigenvalues*** λ_i ***of the Gram matrix*** K***. Let*** V_v ***for*** $v = 1, \ldots, m$ ***are the eigenvectors in the feature space and*** $P_{\Phi(y_i)}$ ***is the projection of*** $\Phi(y_i)$ ***onto the subspace*** $V_1, \ldots, V_m$***. KPCA problem can be represented as a minimization problem of the following error:***

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

with $P_{\Phi(y)} = \sum_{v=1}^{m} \langle \Phi(y); V_v \rangle V_v$

Furthermore, the only measures used to evaluate the projection quality of points are the squared cosines values which can only be used in the case of linear projection. Thus, the individual control of projection is no more guaranteed using non-linear projection method.

2.3 Multidimensional Scaling (MDS)

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

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

By minimizing the stress function, we find the best configuration of $(x_1, \ldots, x_n)$ such that the distances fit to the initial distances d_{ij}.
Now if we consider n variables $r_1, \ldots, r_n \in \mathbb{R}^+$, the sum of which bounds the stress function, the optimization problem $\mathcal{P}_{\text{MDS}}$ can be equivalently rewritten as:

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

Then, 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 projection coordinates of points i and j. Here the projection coordinates are not obtained necessarily by linear projection anymore.

Our new non-linear projection method that controls individually the projection of points is developed hereafter.

2.4 Our proposed method

Let $x_1, \ldots, x_n$ ***be the coordinates of the projected points in*** $\mathbb{R}^m$***. Radii are an important element of the paper introduced to assess how much the distance between two projected points*** (i, j) ***given by*** $\|x_i - x_j\|$ ***is far from given distance*** d_{ij}***. Indeed, radii*** (r_i, r_j) ***for couple*** (i, j) ***are small when*** $\|x_i - x_j\|$ ***is close to*** d_{ij}***. Figure*** 1 ***depicts this idea: for each point*** $P_i \in \{1, \ldots, n\}$***, the projected point of point*** P_i ***belongs to a sphere with center*** x_{P_i} ***and radius*** r_{P_i} ***such that for each couple*** $(i, j) \in \{1, \ldots, n\}$ ***we have*** $\|x_i - x_j\| - (r_i + r_j) \leq d_{ij} \leq \|x_i - x_j\| + r_i + r_j$***.*** This idea can be expressed by finding the value of radii that satisfy these two

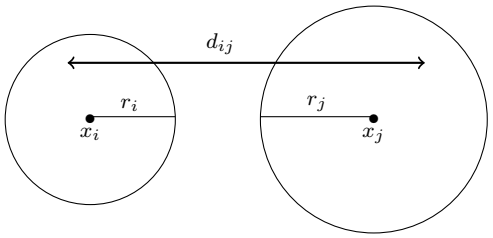


Fig. 1: Examples of radii for bounding the original distance d_{ij}

constraints:

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

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

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

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 problem can easily be solved in $(r_1, \ldots, r_n)$ using linear programming. This problem can be written as follows:

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

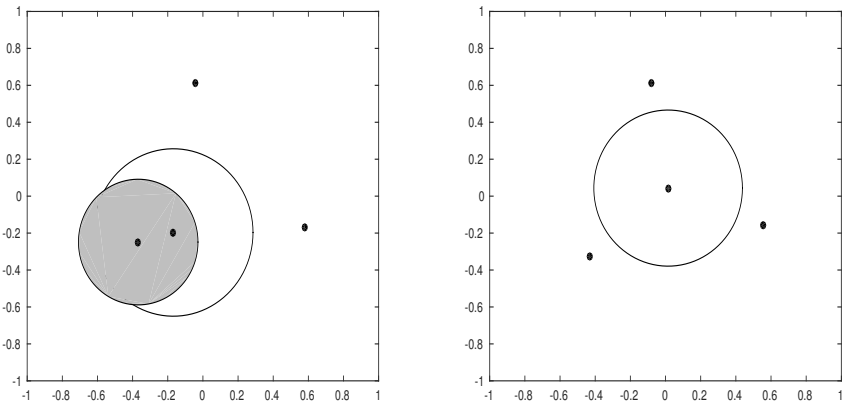
We can remark that a solution of $\mathcal{P}_r$ always exists. Indeed, to satisfy the constraints it is enough to increase all r_i. Besides, solving $\mathcal{P}_r$ with fixed coordinates $(x_1, \ldots, x_n)$ does not lead in general to the optimum of problem $\mathcal{P}_{r,x}$.

2.5 Visualization example

Let us apply our projection method to a simple example by taking a tetrahedron with all pairwise distance equal to 1. For problem $\mathcal{P}_r$, the coordinates x_i for $i = 1, \ldots, 4$ are obtained using multidimensional scaling. Using linear and non-linear optimization packages in Matlab respectively for problems $\mathcal{P}_r$ and $\mathcal{P}_{r,x}$ give a value of $\sum_{i=1}^{n} r_i$ equal to 0.7935 for problem $\mathcal{P}_r$ and 0.4226 for $\mathcal{P}_{r,x}$. Figure 2a corresponds to the first solution and Figure 2b corresponds to the second one. In Figures 2*a* and 2*b*, we depict circles with different radii. The circle color is related to the radius values, 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 considered as well projected points. Note that the points that are represented as points and not circles are very well projected, having radii almost equal to zero. In Figure 2*a*, half of the points is well projected whereas the other half have large radii indicating that they are not well projected. In Figure 2*b* just one circle appears marking that the projection quality using problem $\mathcal{P}_{r,x}$ is better than $\mathcal{P}_r$.

2.6 Link with other methods

Multidimensional fitting (MDF) [4] 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 given distances on the same set of points. We call "target



(a) (b)

Fig. 2: Projected points after solving $\mathcal{P}_r$ and $\mathcal{P}_{r,x}$. (a) shows the projection obtained from the solution of $\mathcal{P}_r$ using MDS and (b) shows that obtained from the solution of $\mathcal{P}_{r,x}$.

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

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

Property 1 Problem $\mathcal{P}_{r,x}$ is bounded below by $\dfrac{1}{n-1}\displaystyle\sum_{1\leq i<j\leq n} |d_{ij} - \|x_i - x_j\||$ where $x_1, \ldots, x_n$ is the optimum of 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, $\displaystyle\sum_{i=1}^{n} r_i \geq \frac{1}{n-1} \sum_{1\leq<i<j\leq n} |d_{ij} - \|x_i - x_j\||$, which concludes the proof.

3 Optimization tools

3.1 Lower Bound of the objective function of problem $\mathcal{P}_{r,x}$

Minimization problem $\mathcal{P}_{r,x}$ is too hard to be solved analytically. A way to assess how good a solution is, is to provide a lower bound on the objective function. Then, if the bound is close to the best found solution, we can conclude that this solution is fixed. Thus, in this section we present such a lower bound of problem $\mathcal{P}_{r,x}$.
Let $x_1, \cdots, x_n; r_1, \cdots, r_n$ a feasible solution of $\mathcal{P}_{r,x}$ and $M \in \mathbb{R}$ such that $M = \max_{(i,j)} \| x_i - x_j \|$. We consider three functions noted f, g, h depending on M as follows:

- $f(M) = \sqrt{\left(1 - \frac{n}{3}\right) M^2 + \frac{1}{n-1} \sum_{i<j} d_{ij}^2} - M.$
- $g(M) = |M - d_{max}|.$
- $h(M) = \min_{i<j} \max_{k<l; k,l \neq i,j} \min \left\{ L^1_{ijkl}; L^2_{ijkl} \right\}.$

where:

- $d_{max} = \max_{1 \leq i < j \leq n} \{d_{ij}\},$
- $L^1_{ijkl} = \max \left\{ \frac{d_{jk} - d_{kl}}{2}; \frac{d_{jl} - d_{kl}}{2}; |d_{ij} - M| \right\},$
- $L^2_{ijkl} = \max \left\{ \frac{d_{ik} - d_{kl}}{2}; \frac{d_{il} - d_{kl}}{2}; |d_{ij} - M| \right\}.$

Using results presented in appendix B, we can write:

$$\begin{cases} \sum_{i=1}^{n} r_i \geq f(M) \\ \sum_{i=1}^{n} r_i \geq g(M) \\ \sum_{i=1}^{n} r_i \geq h(M). \end{cases}$$

Consequently,

$$\sum_{i=1}^{n} r_i \geq \max\{f(M); g(M); h(M)\}. \quad (1)$$

The inequality (1) is true for all solutions of $\mathcal{P}_{r,x}$ particularly for the optimal solution. Thus:

$$\sum_{i=1}^{n} r_i^{opt} \geq \max\{f(M); g(M); h(M)\}. \quad (2)$$

Hence, the lower bound is given by:

$$\sum_{i=1}^{n} r_i^{opt} \geq \min_{M} \max\{f(M); g(M); h(M)\} \text{ for all feasible solutions.}$$

Given M, a lower bound of problem $\mathcal{P}_{r,x}$ is derived. Afterwards, a bound free of M is given by minimizing the bounds depending on M.
By applying this bound to the tetrahedron example, the three functions are plotted. The result is shown in Figure 3. The lower bound is equal to 0.1276 for $M = 1.1276$ and as we have seen the minimum obtained by solving $\mathcal{P}_{r,x}$ is equal to 0.42.

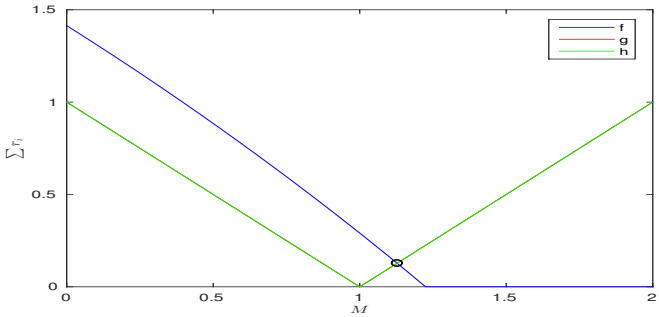


Fig. 3: The curves of the three functions f, g and h. Functions g and h are equal due to the fact that all distances are equal to 1 . The minimal intersection is given by the black circle for $M = 1.1276$ and $\sum_{i=1}^{n} r_i > 0.1276$.

3.2 Initialization point of problem $\mathcal{P}_{r,x}$

Different resolutions of problem $\mathcal{P}_{r,x}$ can be obtained using different initial values of matrix X. Three possible initial values can be used. The first of them is the matrix obtained by PCA or another projection method. In what follows, we present two other possibilities.

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

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

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

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

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

Initial point using improved solution of problem $\mathcal{P}_r$ First, we give two properties which provide a way to improve the optimization results of problem $\mathcal{P}_{r,x}$.

Property 2 Let us consider a point x_i such that for an index j, the following inequality is satisfied :

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

and the other inequalities involving i are not satisfied. Then, the corresponding solution can be improved by moving x_i along the line $x_j - x_i^*$ with x_i^* is the vector of coordinates for point j in order decrease r_i and $|d_{ij} - \|x_i - x_j\||$.

Proof The above condition means that x_i is rewritten $x_i + a(x_j - x_i)$ 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 $a > 0$ otherwise. Let us now consider the other inequalities corresponding to index pairs (i,k) with $k \neq j$. For each of them, either $\exists a \in [a_k^{'}, a_k^{''}]$ 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, if we take 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, decreasing so the objective function. Depending on whether a must be negative or positive, we take $a = a^{'}$ or $a = a^{''}$ respectively.

Another manner to improve the resolution of problem $\mathcal{P}_{r,x}$ is to effectuate 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:

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

Property 3 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 $a, \tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_n$ the optimal solution of $\mathcal{P}_{r,a}$. For the solution of $\mathcal{P}_{r,a}$, for each point i, we have a certain saturated constraint associated to point k noted $C_{ik(i)}$, otherwise it would not be an optimum. So, 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 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 $\sum_{i=1}^{n} \tilde{r}_i < \sum_{i=1}^{n} r_i$.

The new initial point is then given by using these two properties as follows:

- firstly, improve the solution of $\mathcal{P}_r$ using property 2 by solving $\mathcal{P}_{r,a}$.
- secondly, improve the solution of $\mathcal{P}_{r,a}$ using property 1.

3.3 Algorithm 1

Using the different initial values of matrix X presented above, we solve now problem $\mathcal{P}_{r,x}$. For this task, we introduce a new algorithm denoted algorithm 1 which gives the best solution that can be obtained using the different initial values cited above. This algorithm is consisted of two steps: initialization step and optimization step and it is presented as follows:

Algorithm 1

Input: D: distance matrix, N: number of iterations.
Initialization step
Project the points using PCA or MDS.
Solve $\mathcal{P}_r$ using an interior-point method. Obtained solution: $(X_{\mathcal{P}_r}, r_{\mathcal{P}_r})$.
Solve $\mathcal{P}_{R,x}$ using an active-set method and starting from the solution of $\mathcal{P}_r$ obtained at the previous step. Obtained solution: $(X_{\mathcal{P}_{R,x}}, R_{\mathcal{P}_{R,x}})$.
$X_0 \leftarrow X_{\mathcal{P}_{R,X}}$.
for $t = 1$ to N **do**
  Solve $\mathcal{P}_{r,a}$ starting from X_0 using an interior-point method.
  Improve the solution of $\mathcal{P}_{r,a}$ using property 1. Obtained solution: $(X^I_{\mathcal{P}_{r,a}}, r^I_{\mathcal{P}_{r,a}})$.
  $X_0 \leftarrow X^I_{\mathcal{P}_{r,a}}$.
end for
Optimization step
Optimize $\mathcal{P}_{r,x}$ using an active-set method and starting from X_0, $X_{\mathcal{P}_r}$ and $X_{\mathcal{P}_{R,x}}$.
Choose the minimal solution obtained by these three different starting points.

3.4 Algorithm 2

Problem $\mathcal{P}_{r,x}$ is a hard problem, so it is natural to resort to stochastic optimization methods. In the present case, Metropolis-Hastings algorithm [23] allows us to build a Markov chain with a desired stationary distribution. The only delicate part is the choice of the proposal distribution and the necessity to solve a $\mathcal{P}_r$ problem at each iteration. In details, this Metropolis-Hastings algorithm requires:

1- A *target distribution*:
The target distribution is related with the objective function of problem $\mathcal{P}_{r,x}$ and it is given by:

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

with E an application given by:

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

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 interesting results. It should be chosen in such a way that the proposal distribution gets close to the target distribution. The proposal distribution $q(X \to .)$ has been 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 $x_i^* = x_i - |c_{ij^{(i)}}| L_{ij^{(i)}}$.

  with $L_{ij^{(i)}} = \dfrac{x_i - x_j}{\|x_i - x_j\|}$.

3- A *linear optimization problem*:
For the matrix X generated in each iteration, we solve the linear optimization problem $\mathcal{P}_r$.

4 Numerical application

The presented projection method has been applied to different types of real data sets so as to illustrate its generality.

4.1 The data

Four real data sets are 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 [1] is a famous data set and is presented to show that the projection is as expected. This data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The four variables studied in this data set are: sepal length, sepal width, petal length and petal width (in cm). Car data set [29] is a data set studied in the book of Saporta (Table 17.1, page 428). This data set describes 18 cars according to various variables (cylinders, power, length, width, weight, speed).

The soybean data set [30] 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 (D3). The coffee data set is a time series data set used in chemometrics to classify food types. This kind of time series is seen in many applications in food safety and quality insurance. This data set is taken from *UCR time Series Classification and Clustering* website [10]. Coffea Arabica and Coffea Canephora variant Robusta are the two species of coffee bean which have acquired a worldwide economic importance and many methods have been developed to discriminate between these two species by chemical analysis [8].

4.2 Experimental setup

In practice, we have tested our method on the different data sets by solving the optimization problem $\mathcal{P}_{r,x}$ using algorithm 1 and also the proposed Metropolis-Hastings algorithm (algorithm 2). Each time, a distance matrix is required. For the quantitative data, we compute the Euclidean distance between points y_i, for $i = 1, \ldots, n$, 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 [7] computed by the formula $p_{ij} = \sum_{t=1}^{Q} w_t p_{ij}^t$ where $p_{ij}^t = \begin{cases} 1 & \text{if } i^t = j^t \\ \frac{n_k^2}{n_k^2 + 2} & \text{else} \end{cases}$, p_{ij}^t is the per-attribute Eskin dissimilarity between two values for the categorical attribute indexed by t, w_t is the weight assigned to the attribute t, Q is the number of attributes and n_t is the number of values taken by each attribute. Then, using the formula which transforms the dissimilarity into similarity: $p_{ij} = 1 - s_{ij}$, the distances can be given by the standard transformation formula from similarity to distance: $d_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$. On top of that, to compute the distances between the curves of functional data, we have chosen a measure of proximity similar to that studied in [19]. In this article, the authors develop a proper classification designed to distinguish the grouping structure by using a functional k-means clustering procedure with three sorts of distances. So, in our work we choose one of these three proximity measures forasmuch their results are similar. Thus, 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.

To solve the different optimization problems, we have used the optimization toolbox in MATLAB. For problems $\mathcal{P}_r$ and $\mathcal{P}_{r,a}$, we apply firstly PCA – for quantitative data – or MDS – for categorical and functional data – and then a

linear programming package is used to solve the optimization problems using an interior-point algorithm. Problems $\mathcal{P}_{r,x}$ and $\mathcal{P}_{R,x}$ are non-linear optimization problems, therefore we use a non-linear programming package to solve it selecting the active-set algorithm to obtain the best values of $(x_1, \ldots, x_n)$ and $(r_1, \ldots, r_n)$. This iterative algorithm is composed of two phases. In the first phase (the feasibility phase), the objective function is ignored while a feasible point is found for the constraints, in the second phase (the optimality phase), the objective function is minimized while feasibility is maintained [34].

Our proposed Metropolis-Hastings algorithm can provide a good solution if 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 take it equal to 100. For the other parameters σ and T, we take their values respectively in a range from 0.01 and 100.

As we have mentioned in the section of visualization, the visualization of the projection of each point i in $\mathbb{R}^2$ is presented as a circle having x_i as center and r_i as radius so as the projected point belongs to this circle and this is the specificity of our method. For each data set, we show the circles obtained for each point after resolution of optimization problem $\mathcal{P}_{r,x}$. ***To compare the projection quality of our representation with that obtained by PCA and KPCA, we use the squared cosine values as projection quality. Moreover, the comparison of the projection quality of our representation with MDS is done using Stress-per-point*** (SPP) ***proposed by Born and Groenen in [11]. The*** SPP ***of the point*** i ***is given by:***

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

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

Furthermore, the lower bound defined in section 3.1 is each time computed.

4.3 Results

4.3.1 Visualization data in $\mathbb{R}^2$

The optimization results for these four data sets are given in Table 1. For each data, we give the algorithm 1 and Metropolis-Hastings results with which initial starting point is used in algorithm 1. The lower bound value for each data set is also given in this table, we observe that in one case (cars), this lower bound indicates that the found solution is not far from the optimum but in the other cases, it seems that the lower bound while providing a good starting point can be improved.

Figures 5 and 7 depict the results of projection under pairwise distance control for quantitative data. This projection is compared

Table 1: Optimization solution of problem $\mathcal{P}_{r,x}$ for different data sets.

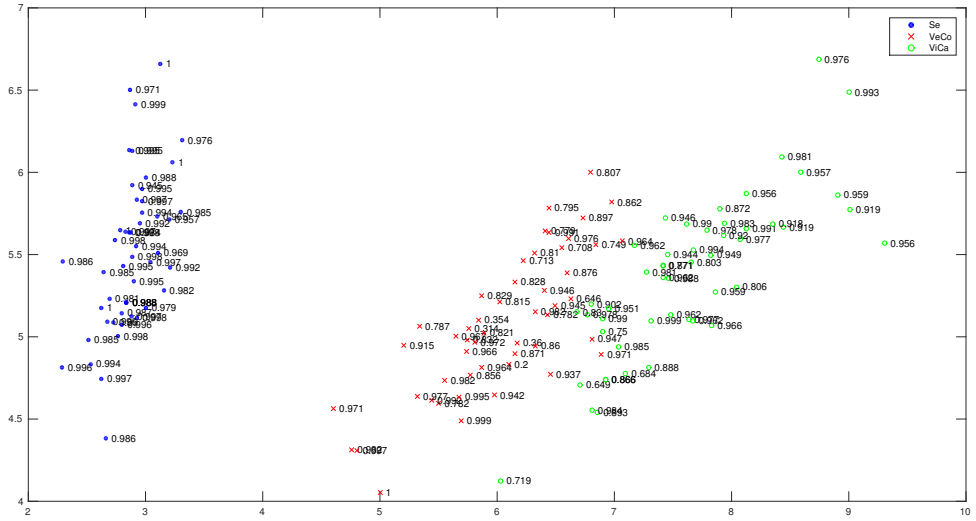
	$\sum r_i^{\text{Algo 1}}$	$\sum r_i^{\text{MH}}$	Lower Bound
Iris	16.19	17.2	1.07
Cars	3.27	3.35	1.21
Soybean	3.98	3.93	0.29
Coffee	21.68	21.97	0.89

with the projection given by PCA, KPCA and MDS. For PCA and KPCA, we have plotted the projection of the points indexed by their squared cosine values. For MDS, we have used the smacof package in R to compute the stress-per-point and to plot the bubble plot represented the stress-per-point.

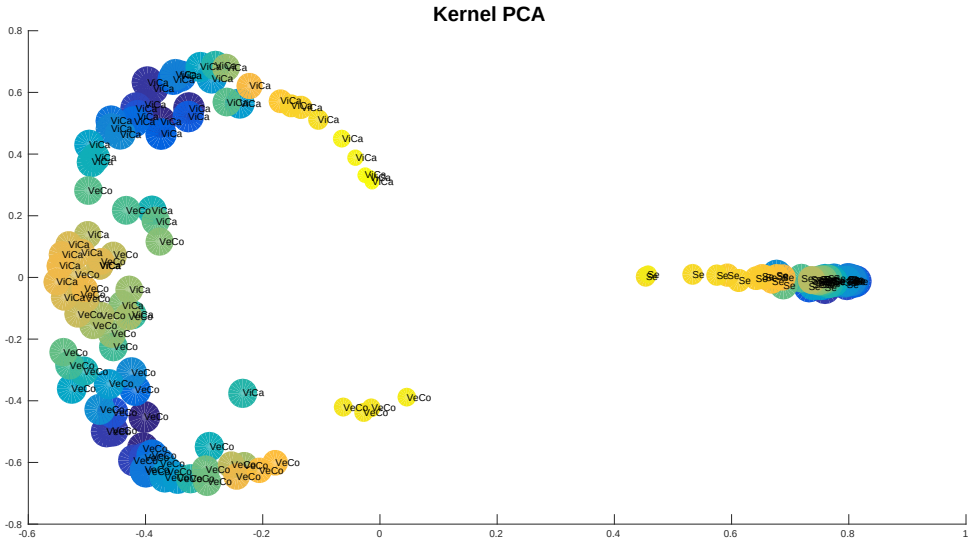
In the projection of Iris data set showed in Figure 5, it is interesting to remark that appealingly two areas are well separated. This corresponds to the well-known fact that Iris versicolor and virginica are close whereas the species Iris setosa are more distant.

Concerning car data set, the projection of points using projection under pairwise distance control is given in Figure 7. The expensive cars as "Audi 100", "Alfetta-1.66", "Dastun-200L", "Renault 30" are well-separated from the low-standard cars as "Lada-1300", "Toyota Corolla", "Citroen GS Club", "Simca 1300". We remark that the expensive cars are located on the right and low-standard one are located on the left.

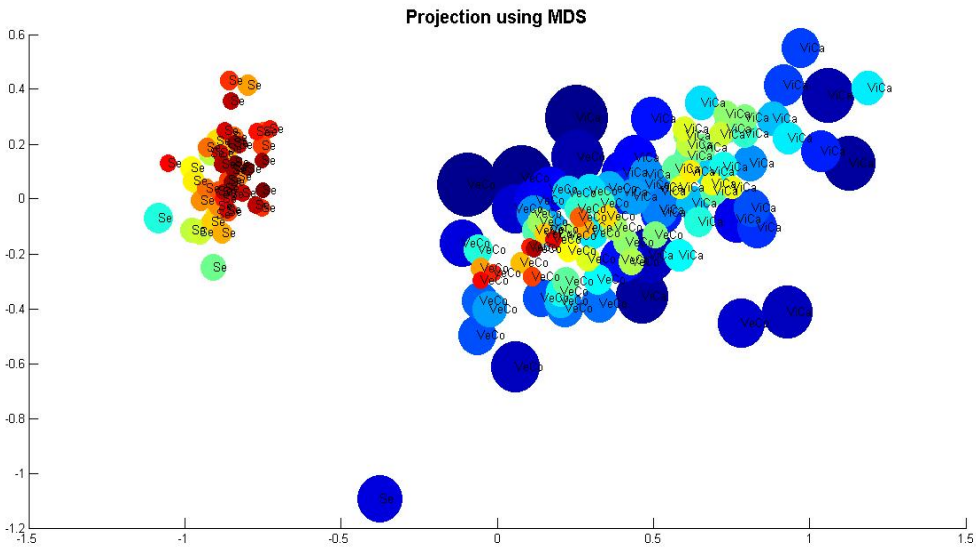
For these two data sets, we want to compare the projection quality for each method presented in Figures 4 and 6 with the projection quality obtained using projection under pairwise distance contral in Figures 5 and 7. For PCA, we can say that our method projected the points without giving any importance to any group. Indeed, Figure 4a depicts a group with small values of quality measure and a group with high values of quality measure whereas the radii obtained by projection under pairwise distance control method are distributed in an equivalent way. Additionally, from Figure 7, we can assert that the projected points obtained using projection under pairwise distance control method are well separated as there is any intersection between the circle. For KPCA,we have plotted the squared cosine as a circle to make the representation more clear especially for iris data set as the Iris setosa specie are projected next to each other. From Figure 4b we can conclude that in each category, the points which have close quality values are located side by side. The same conclusion is appeared for cars data set in Figure 6b, we can see that the point with navy circle are located almost in the same Y axis coordinates similarly for the red circles. So the local quality in KPCA dependent on the place on the points. Additionally, by comparing the projection of our method with those obtained in MDS, we can find the same results appear in PCA for Iris data set, the points



(a)



(b)



(c)

Fig. 4: Projection of Iris data set using PCA, KPCA and MDS respectively.

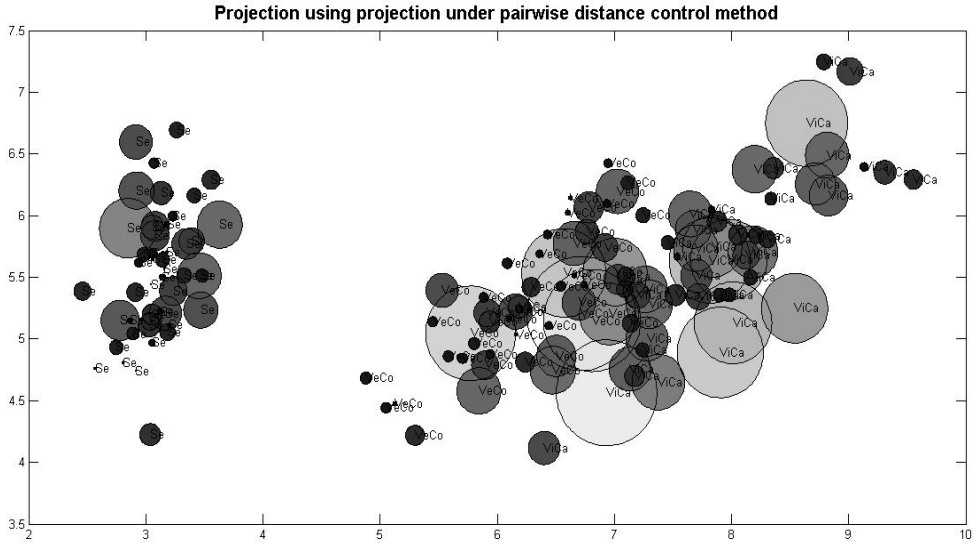
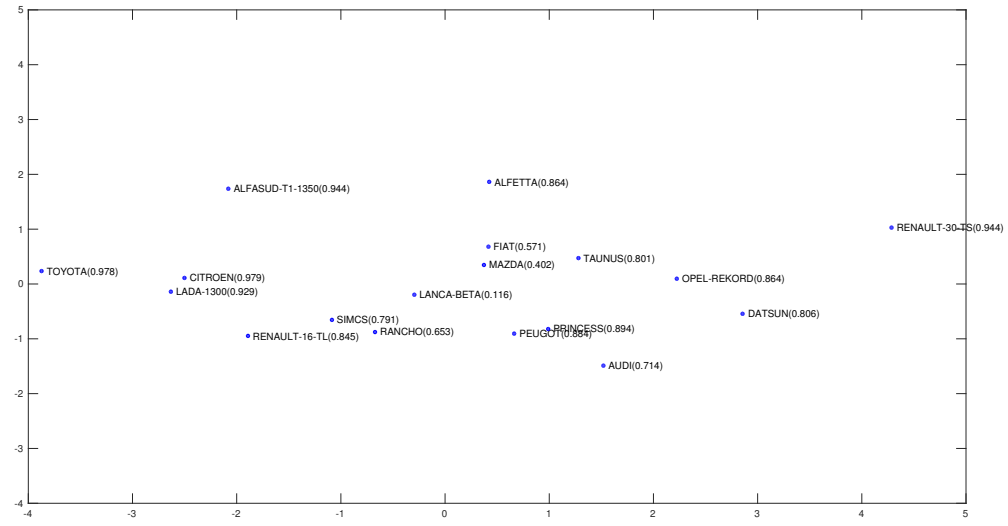


Fig. 5: Projection of Iris data set using projection under pairwise distance control methods respectively. Two well separated groups can be observed.

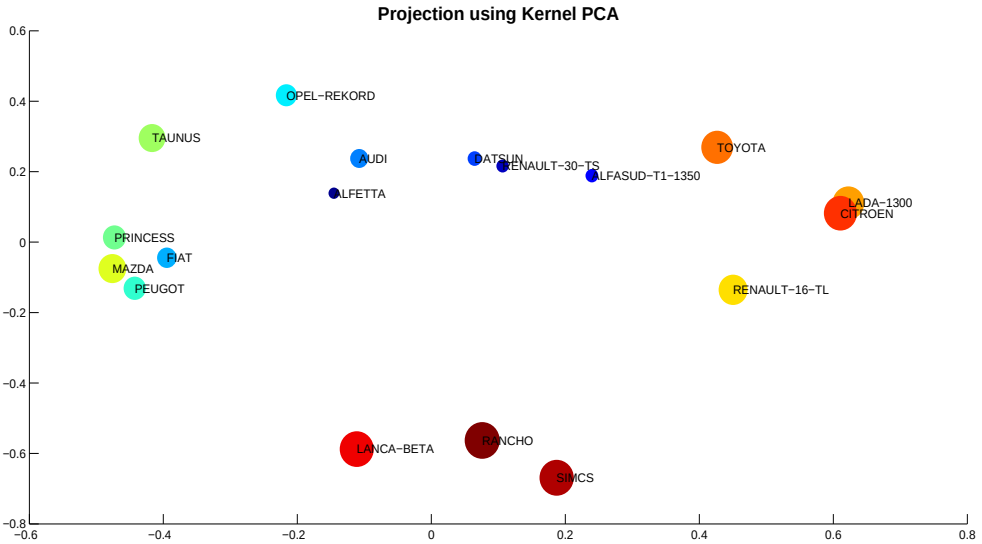
are projected by giving an importance to Iris setosa group. Indeed, almost all the red circles (indicating the very well projection) are given to the Iris setosa specie.

Moreover, the pairwise distances are significant in our method and give an interpretation on the position 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 can not be interpreted as distances in PCA and KPCA and MDS. This is the particular strength of our method. Hence, projection under pairwise distance control suggests an absolute interpretation whereas the other methods give a relative one. From this, we can conclude from Figure 7 that there is a big difference between the two cars "Toyota" and "Renault 3" as the distances between this two cars is very important. Conversely, the distance between "Lada1300" and "Citroen" cars are small indicating then the closeness of these two cars. Note here that these two cars are very well projected leading to a very good interpretation.

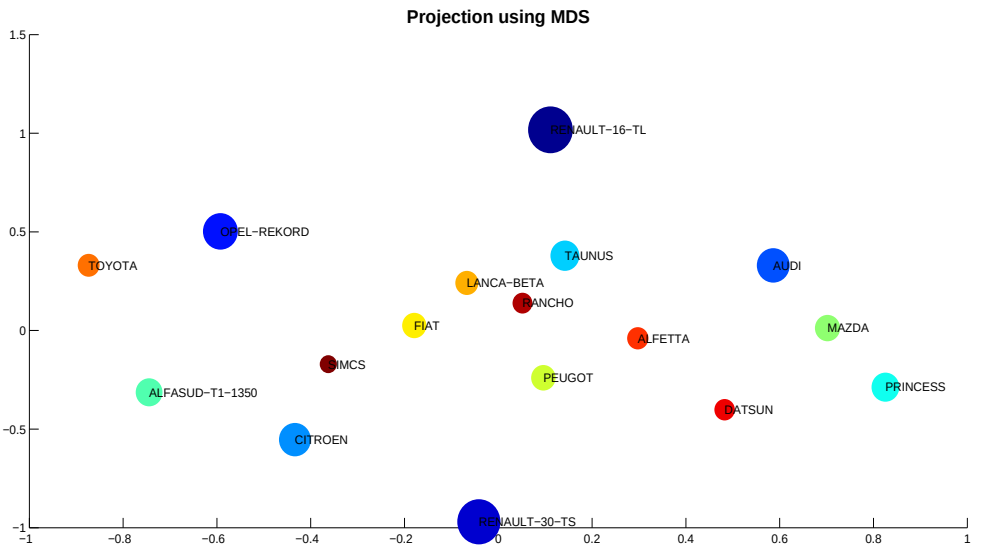
For the qualitative and functional data sets, it is necessary to verify that the matrix B obtained by MDS method is semi-definite positive to use the squared cosine as quality measure because the starting point of optimization is obtained from MDS. After that, in case of positiveness of matrix B, we can calculate the quality measure. In the projection of the soybean data set, four classes have been shown in Figure 8 and each class contains the diseases number of the class. But basically, the whole set of points can be divided in



(a)



(b)



(c)

Fig. 6: Projection of car data set using PCA, KPCA and MDS. For PCA, the

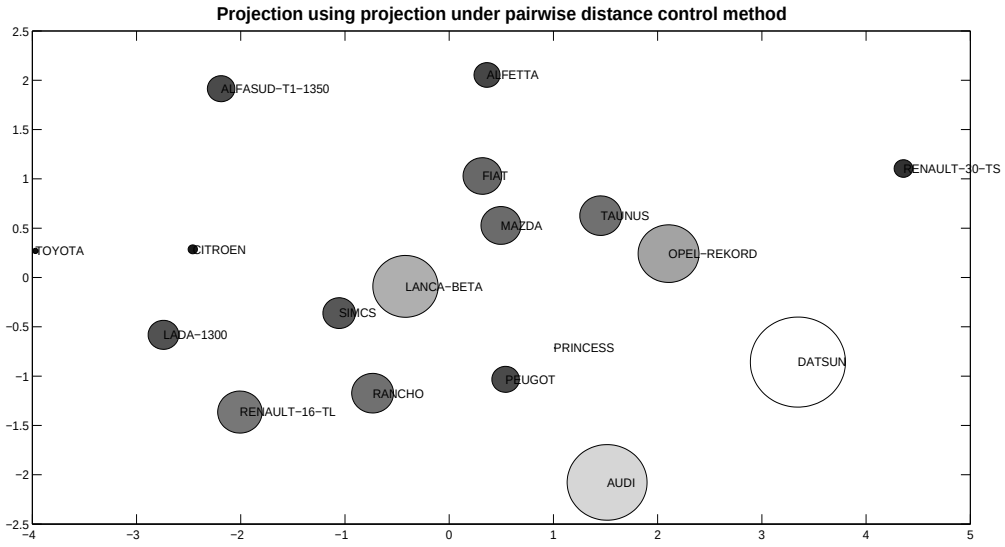


Fig. 7: Projection of car data set using projection under pairwise distance control.

two large classes. Indeed, It is clear that class 2 is well separated from the others classes as there is no intersection between the circles of class 2 and the circles of others classes. Moreover, class 1 can be considered as well separated class from classes 3 and 4 if we do not take into account the point D_3^*. Classes 3 and 4 are not at all well separated as we can exhibit that there are different intersections between the circles of these two classes. This result is figured in [30] which lists the value "normal" for the first two classes and "irrelevant" for the later two classes. The comparison of projection under pairwise distance control result with **PCA and KPCA** is not possible for this data set because the matrix B is not semi-definite positive.

By comparing our projection with the projection obtained by MDS, we can see that the 4 ***groups of soybean are well separated using MDS than our method and the bubble are small than our circles. So we can conclude that our method detect more the misplaced points in the groups and the points which are not well projected.***

The coffee data set has been studied in several articles ([8,3]) and different classification methods have shown the different groups contained in this data set using our method and the other methods. We can see clearly in Figures 10 and 11 the grouping structure that is obtained.

In Figure 11, we show that we have succeeded in differentiating the Arabica from Robusta coffee. These two classes are clearly presented, the first class indexed by 1 corresponding to Arabica coffee and the second one indexed by 2

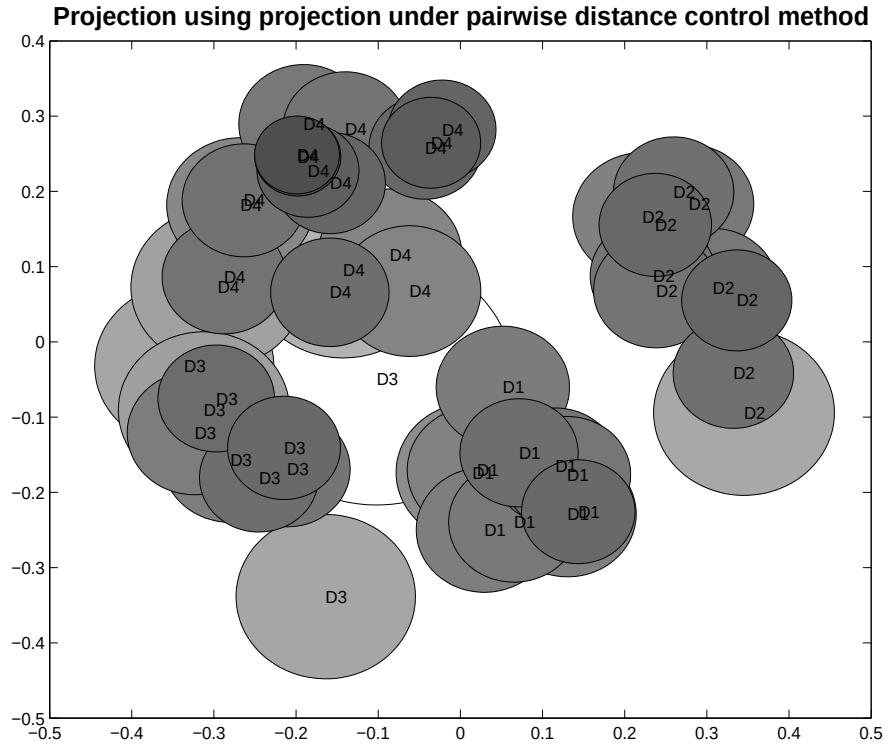


Fig. 8: Projection under pairwise distance control for soybean dat set. Four groups are presented, indexed by D1, D2, D3 and D4.

corresponding to Robusta coffee. These classes are not well separated by comparing with the results of quantitative data, since there are many intersections. Therefore, the representation of the points as circles and not as coordinates points gives more information about the real class of points and shows the points who have the possibility to be misplaced in a class.

Figure 10a and 10b show the projection quality using PCA and MDS respectively. As all the eigenvalues of matrix B ***are positive, so we can compute the quality measure given by PCA. Comparing the projection quality of PCA and projection under pairwise distance control given respectively by Figures 10a and 11, we can observe that the quality of projection of the set of points is pretty steady.***

The comparison with KPCA projection is not allowed because all the local quality projection obtained using KPCA is equal to 1.

Additionally, Metropolis-Hastings has been applied to these data sets. The trace plots of the optimization problem $\mathcal{P}_{r,x}$ are shown in Figure 12 after

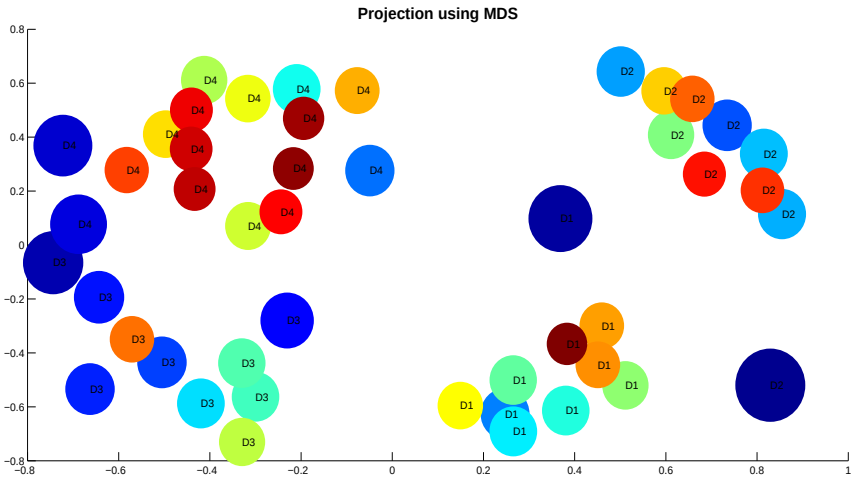
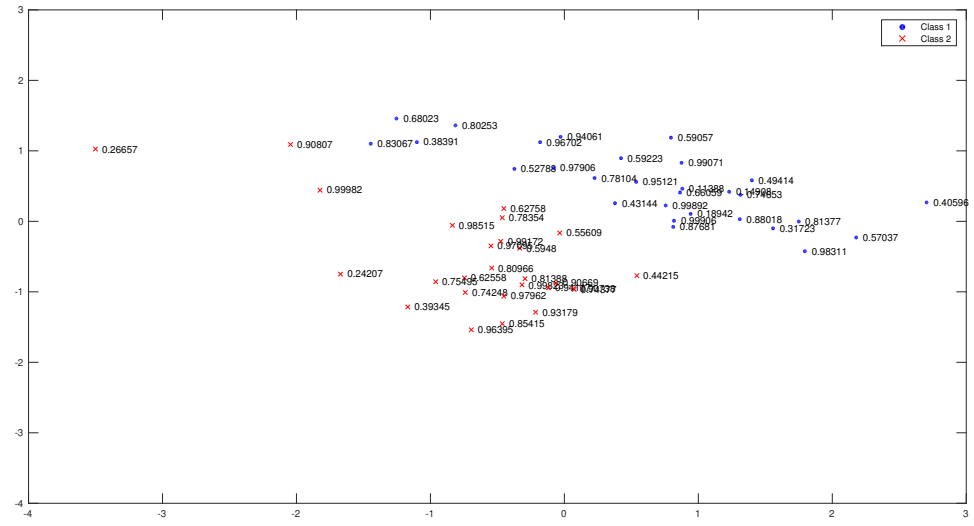


Fig. 9: MDS for soybean dat set. Four groups are presented, indexed by D1, D2, D3 and D4.

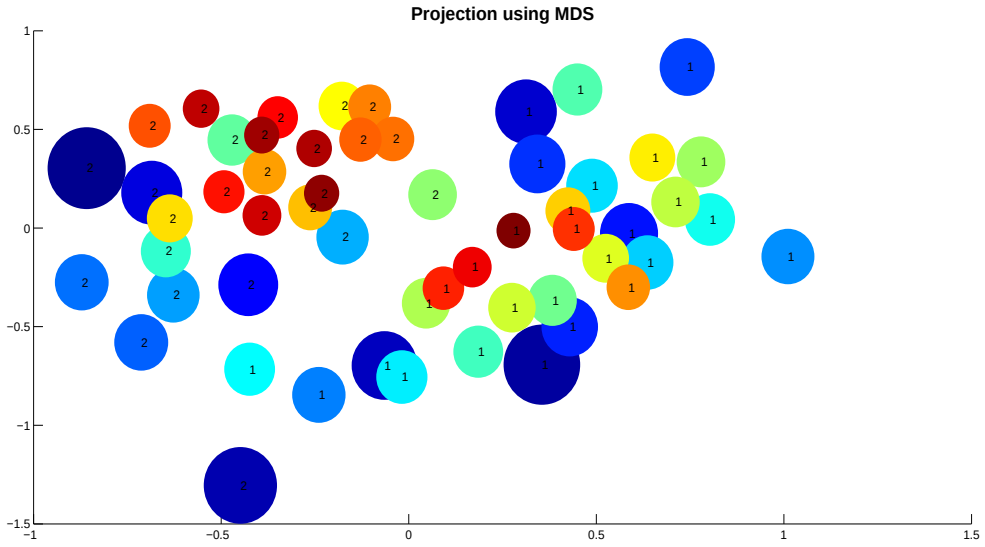
5000 iterations. Returning to Table 1, we can exhibit that Metropolis-Hastings algorithm solutions are very close to those obtained using the optimization package of Matlab and reciprocally. Thus, the obtained radii should be close to the optimum.
Finally, we present the lower bound computed from the three functions described in Section 3. The lower bound is taken by the minimal intersection of these functions. Returning to Table 1, it is clear that the value of the bound is small compared to the value of the solution obtained by algorithm 1 and Metropolis-Hastings. Thus, this bound while providing a good starting point, should be improved Note that this bound for tetrahedron example gives also good results as algorithm 1 provides a solution three times smaller than the bound.

4.3.2 Dimensionality reduction results

One of high-dimensional data studies objectives is to choose from a large number of variables those which are important for understanding the underlying phenomena of study. So, the aim will be to reduce the dimension rather than to visualize data in $\mathbb{R}^2$. So, our method can also serve to reduce the number of variables by taking into account the minimal value of $\sum_{i=1}^{n} r_i$.
Here, we have solved the problem $\mathcal{P}_{r,x}$ using the different possible dimension values. We have plotted in Figure 13 the values of $\sum_{i=1}^{n} r_i$ 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 to see that the value of $\sum_{i=1}^{n} r_i$ decreases when the dimension increases.



(a)



(b)

Fig. 10: Projection of coffee data set using PCA and MDS.

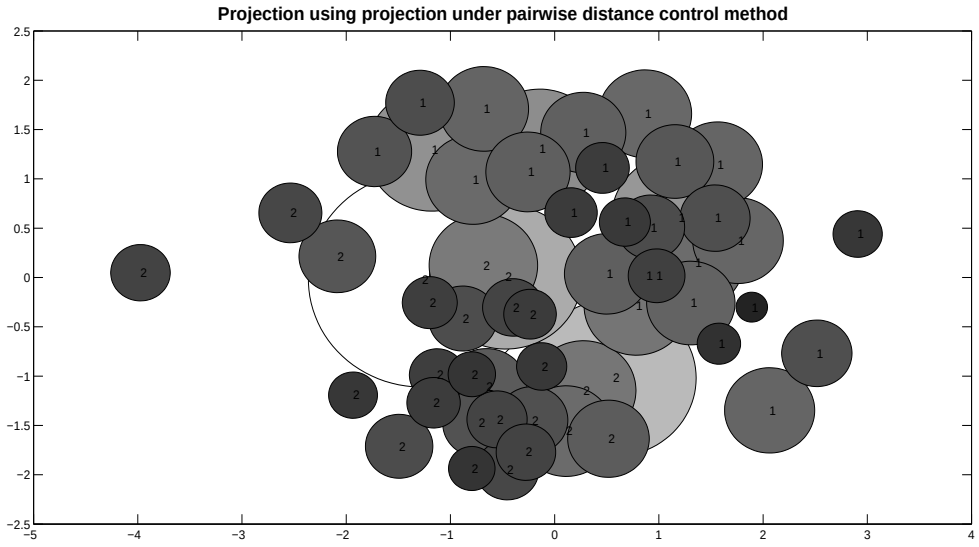
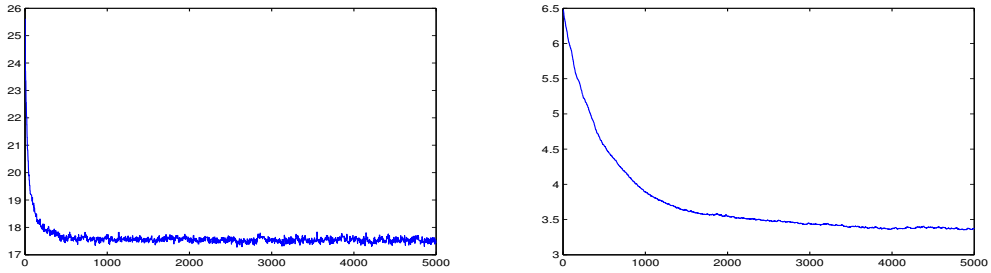


Fig. 11: Projection of coffee data set using projection under pairwise distance control. Two clusters indexed 1 and 2 indicate respectively Arabica and Robusta classes.

The main problem which is widely posed in dimension reduction methods is the determination of the number of components that are needed to be retained. Many methods have been discussed in the literature [24,6] to determine the dimension of reduced space relying on different strategies related to the good explanation or the good prediction. So, with our method the choice of the reduced space dimension is related to the locally projection quality of points and how much the user is interested by the projection quality of points.

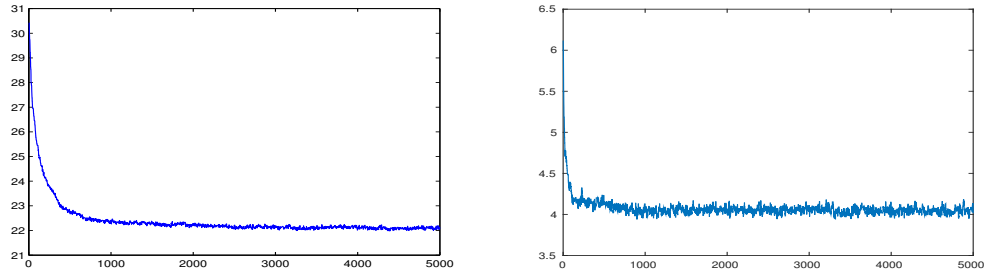
Concerning the quantitative data sets (Iris and car), if the main objective of the user is to obtain a very good projection quality then a choice of three components against 4 for iris and 6 for cars can be a good choice as the value of $\sum_{i=1}^{n} r_i$ is small and there is not a big difference between this value for this dimension and the values of the higher dimensions. For coffee data set, a dimensionality reduction from 56 sample time series down to 6 simple extracted features is considered as a good choice. The same idea can be seen for soybean data set, a reduced space dimension equal to 4 can be considered as efficient reduced space.

Moreover, a comparison of our results with the existent results shows a coherence between them. For Iris data set, [15] and [26] conclude that the number of variables can be reduced to 2 as the petal length and petal width variables are the most important variables among all variables. Similarly, this result can be seen for car data set, Saporta in his book [29] (Table 7.4.1 page 178) notices



(a) Iris data set

(b) car data set



(c) Coffee data set

(d) Soybean data set

Fig. 12: Trace plots of Metropolis Hastings for different data sets. The x-axis corresponds to the iteration number and y-axis to the value of $\sum_{i=1}^{n} r_i$.

that the conservation of two dimensions leads to the explanation of 88% of inertia **where the term inertia reflects the importance of a component**. So, these results seem very similar to our results, the important decrease is located between dimensions 1 and 2. The other decreases are negligible for these two data sets. Selection variables is studied on time series coffee data set in [2]. Using several analysis methods, the number of selected variables ranges 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 quality projection of the points. Concerning soybean data set, Dela Cruz shows in his paper [16] that the 35 attributes can be reduced to 15 and here with our method, we have succeeded to reduce the attributes to 6 by having a very good projection. Hence, the presented results confirm that we can reduce the dimension nonlinearly and still keep a way of assessing as reasonable number of dimensions. and that is efficient as a dimensionality reduction method.

4.4 Advantages of projection under pairwise distance control method

As we have seen, our presented method has several advantages. To summarize:

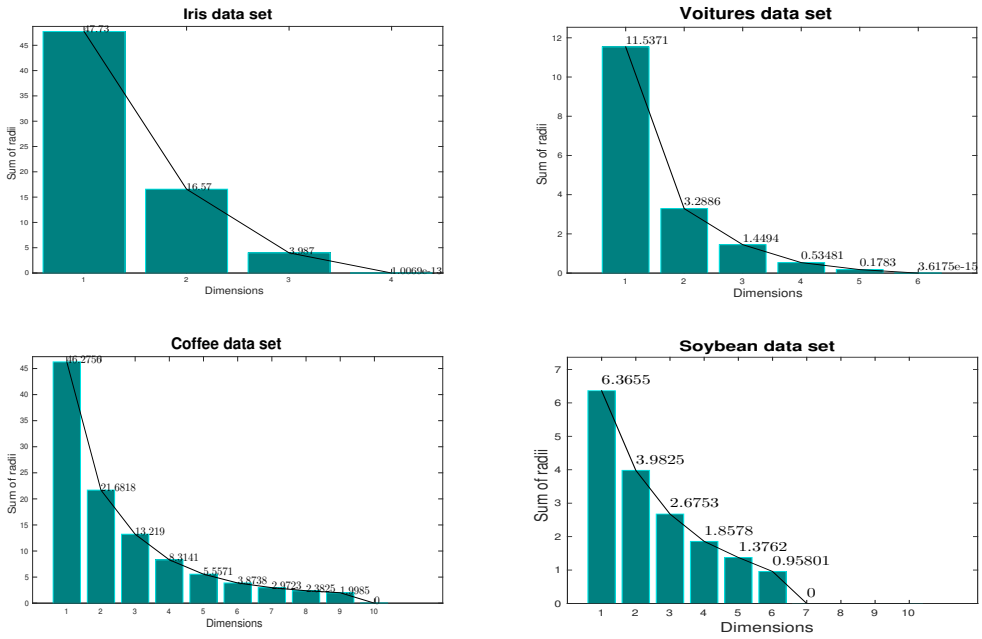


Fig. 13: The scree plot of $\sum_{i=1}^{n} r_i$ for different dimensions for the four data sets.

firstly, it is a non-linear projection method which 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 interpret easily the distances observed in the projection plane. Thirdly, the quality distribution between the points seems to be evenly distributed.

5 Conclusion

The purpose of this article was to outline a new non-linear 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 to give a bound on the original distances. We have 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 have developed different algorithms and proposed a lower bound for the objective function. As such, the method described here

needs further research to improve the lower bound in order to assess how close the algorithms are from the minimum.

References

1. Anderson E, The Irises of the Gaspé Peninsula, Bull. Am. Iris Soc., 59, 2–5 (1935)
2. Andrews, J. L. and McNicholas, P.D., Variable Selection for Clustering and Classification, Journal of Classification, 31, 136-153 (2014)
3. Bagnall, A., Davis, L., Hills, J., and Lines, J., Transformation Based Ensembles for Time Series Classification, Proceedings of the 12th SIAM International Conference on Data Mining, 307–319 (2012)
4. Berge, C., Froloff, N., Kalathur, RK., Maumy, M., Poch, O., Raffelsberger, W. and Wicker, N., Multidimensional fitting for multivariate data analysis. Journal of Computational Biology, 17, 723–732 (2010)
5. Xiaofang, L., Chun, Y., Greedy kernel PCA for training data reduction and nonlinear feature extraction in classification. Proceedings of SPIE - The International Society for Optical Engineering, 2009.
6. Besse, P., PCA stability and choice of dimensionality, Statistics & Probability Letters, 13, 405-410, (1992)
7. Boriah, S., Chandola, V., and Kumar, V., Similarity Measures for Categorical Data: A Comparative Evaluation, Proceedings of the SIAM International Conference on Data Mining (2008)
8. Briandet, R., Kemsley, E. K., and Wilson, R. H., Discrimination of arabica and robusta in instant coffee by fourier transform infrared spectroscopy and chemometrics, J. Agric. Food Chem., 44 (1), 170–174 (1996)
9. Chan, W. W-Y., A survey on multivariate data visualization in Science and technology, Department of Computer Science and Engineering Hong Kong, University of Science and Technology, 8(6), 1–29 (2006)
10. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G., The UCR Time Series Classification Archive, `www.cs.ucr.edu/~eamonn/time_series_data/` (2015).
11. Borg, I., Groenen, P. Modern Multidimensional Scaling: Theory and Applications (2nd ed.). New York: Springer-Verlag (2005).
12. Cheung, L.W., Classification approaches for microarray gene expression data analysis, Methods Mol Biol., 802, 73–85 (2012)
13. Chinchilli, V.M., Sen, P.K., Multivariate Data Analysis: Its Methods, Chemometrics and Intelligent Laboratory Systems, 2, 29–36 (1987)
14. Cleveland, W.S., McGill, M.E., Dynamic Graphics for Statistics, Wadsworth and Brooks/Cole, Pacific Grove, CA, (1988)
15. Chiu, S. L., Method and Software for Extracting Fuzzy Classification Rules by Subtractive Clustering, In Proceedings of North American Fuzzy Information Processing Society Conference (1996)
16. Dela Cruz, G. B., 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 (2015)
17. Dempster, A.P., An overview of multivariate data analysis, Journal of Multivariate Analysis, 1(3), 316–346, (1971)
18. 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., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 286, 531–537 (1999)
19. Ieva, F., Paganoni, A.M., Pigoli, D., and Vitelli., V., Multivariate functional clustering for the analysis of ECG curves morphology, Journal of the Royal Statistical Society, Applied Statistics, series C., 62(3), 401–418 (2012)
20. Inselberg, A., The Plane with Parallel Coordinates, Special Issue on Computational Geometry, The Visual Computer, 1, 69–91 (1985)

21. Jackson, J., A Users Guide to Principal Components, John Wiley & Sons, New York (1991)
22. Jagannathan, R. and Ma, T., Risk reduction in large portfolios: why imposing the wrong constraints helps, J. Finan., 58, 1651–1683 (2003)
23. Johansen, A. M. and Evers, L. Monte Carlo Methods, Department of Mathematics, University of Bristol (2007)
24. Jollife, I.T., Principal Component Analysis, Springer, New York (1986)
25. Keim, D. A. and Kriegel, H. P., Visualization Techniques for Mining Large Databases: A Comparison, IEEE Transactions on Knowledge and Data Engineering, 8(6), 923-938 (1996)
26. Liu, H., Setiono, R., Chi2: feature selection and discretization of numeric attributes, In Proceedings., Seventh International Conference on Tools with Artificial Intelligence (TAT'95) (1995)
27. Mardia, K.V., Kent, J.T. and Bibby, J.M., Multivariate analysis, Academic Press, London (1979)
28. Sammon, J., A non-linear mapping for data structure analysis, IEEE Transactions on Computers, 18(5), 401–409 (1969)
29. Saporta, G., Probabilités, analyse des données et statistique, Technip (2006)
30. Stepp, R., Conjunctive conceptual clustering, Doctoral dissertation, department of computer science, university of Illinois, Urbana-Champaign, IL (1984)
31. Svante W., C. Albano, W. J. DunnIII, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg , M. Sjostrom., Multivariate Data Analysis in Chemistry, Chemometrics, 138, 17–95 (1984)
32. Togerson, W.S., Theory and methods of scaling, New York: Wiley (1958)
33. Van der Hilst, R., de Hoop, M., Wang, P., Shim, S.-H., Ma, P. and Tenorio, L., Seismostratigraphy and thermal structure of earth's core-mantle boundary region, Science, 315, 1813–1817 (2007)
34. Wong, E., Active-Set Methods for Quadratic Programming, Ph.D. thesis, university of California, San Diego (2011)

Appendices

Appendix A: Two lemmas

Let (C) be a circle with center O and radius r. Let consider n points with coordinates $x_1, \cdots, x_n$ such that for all $i = 1, \ldots, n$, $\|x_i - O\| \leq r$ and having g as center of gravity. This hypothesis is used in Lemma 1 and Corollary 1.

Lemma 1 *For all points $x_1, \ldots, x_n$, we have:*

$$\|x_i - O\| = r \text{ when } \sum_{i=1}^{n} \|x_i - g\|^2 \text{ is maximum .}$$

Proof Let A and B two points among the n points such that A is inside the circle (C) and B belonging (C). The point A and B have as coordinates (a_1, a_2) and $(r, 0)$ respectively. We want to show that by moving B by small movements along the circle, we can approach the point A to the circle border increasing thus the inertia $\sum_{i=1}^{n} \|x_i - g\|^2$.
We note B', A' the new positions after movements of B and A respectively. Let θ be the angle between (OB) and (OB'), u_θ the displacement of B and (a'_1, a'_2) the coordinate of point A'.
Approaching A to the circle border requires the opposite movements of A and B with equal length. This constraint is necessary to keep the center in the same position. We distinguish two cases:

1- A having $a_2 < 0$.
2- A having $a_2 > 0$.

The two cases are illustrated in Figure 14.

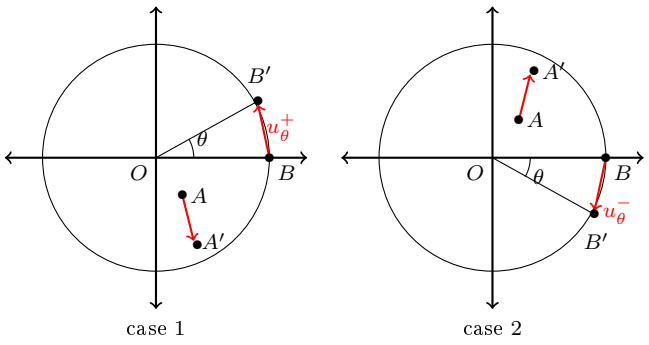


Fig. 14: Representation of movements of points A and B in cases 1 and 2.

For case 1, A in the lower half of the circle requires that B moves positively with angle $\theta \in [0; \frac{\pi}{2}]$. In this case, the vector u_θ is given by: $u_\theta^+ = \begin{pmatrix} \cos\theta - 1 \\ \sin\theta \end{pmatrix}$ and the inner product $\langle u_\theta^+, AA' \rangle$ is given by:

$$\langle u_\theta^+, AA' \rangle = (a'_1 - a_1)(\cos\theta - 1) + (a'_2 - a_2)\sin\theta \cdot \tag{3}$$

Here, we have $a'_1 \geq a_1$ and $a'_2 \leq a_2$ that imply $a'_1 - a_1 \geq 0$ and $a'_2 - a_2 \leq 0$. Moreover, we have $0 \leq \cos\theta \leq 1$ and $\sin\theta \geq 0$ that give $\langle u_\theta^+, AA' \rangle < 0$.

For case 2, A in the upper half of the circle requires that B moves negatively with angle $\theta \in [3\frac{\pi}{2}, 2\pi]$. In this case, the vector u_θ is given by: $u_\theta^- = \begin{pmatrix} \cos\theta - 1 \\ -\sin\theta \end{pmatrix}$ and the inner product is given by:

$$\langle u_\theta^-, AA' \rangle = (a_1' - a_1)(\cos\theta - 1) - (a_2' - a_2)\sin\theta \tag{4}$$

Here, we have $a_1' \geq a_1$ and $a_2' \leq a_2$ that imply $a_1' - a_1 \geq 0$ and $a_2' - a_2 \leq 0$. Moreover, we have $0 \leq \cos\theta \leq 1$ and $\sin\theta \leq 0$ that give $\langle u_\theta^-, AA' \rangle < 0$.

Corollary 1 *The center of gravity g of $x_1, \ldots, x_n$ is the center of circle (C) i.e. $O = g$ when $\sum_{i=1}^n \|x_i - g\|^2$ is maximum.*

Proof We have:

$$\begin{aligned}
\sum_{i=1}^n \|x_i - g\|^2 &= \sum_{i=1}^n \|x_i - O + O - g\|^2 \\
&= \sum_{i=1}^n \|x_i - O\|^2 + \sum_{i=1}^n \|O - g\|^2 + 2\sum_{i=1}^n (x_i - O)'(O - g) \\
&= \sum_{i=1}^n \|x_i - O\|^2 + \sum_{i=1}^n \|O - g\|^2 + 2n(g - O)'(O - g) \\
&= \sum_{i=1}^n \|x_i - O\|^2 + n\|O - g\|^2 - 2n\|O - g\|^2 \\
&= \sum_{i=1}^n \|x_i - O\|^2 - n\|O - g\|^2
\end{aligned}$$

All the points belong to the circle (C) as a result of Lemma 1. So, $\|x_i - O\|^2$ is fix and equal to r^2. Thus, maximizing inertia $\sum_{i=1}^n \|x_i - g\|^2$ amounts to minimize $\|O - g\|^2$. Then, the minimum of $\|O - g\|^2$ is zero so that $O = g$.

Lemma 2 *If $\sum_{i=1}^n \| x_i - g \|^2$ is maximum for points $x_1, \ldots, x_n$ under constraints, for all couple (i, j), $\|x_i - x_j\| \leq M$, then an upper bound of $\sum_{i=1}^n \| x_i - g \|^2$ is given by: $\frac{nM^2}{3}$.*

Proof Let (C) be the smallest circle containing the n points with coordinates $x_1, \ldots, x_n$. We consider three points noted A, B and C among the n points and having x_A, x_B and x_C as coordinates . We suppose that A, B and C belong to the circle (C) and the distance between B and C is equal to M *i.e.* $\|x_B - x_C\| = M$. By hypothesis, we have $\|x_A - x_B\| \leq M$. We note θ the angle between (AB) and (BC). Figure 15 illustrates the situation.
If $\theta > \frac{\pi}{3}$ then $\|x_A - x_B\| \leq M$ and $\|x_A - x_C\| > M$, by reversing the role of x_B and x_C we get $\theta = \frac{\pi}{3}$. So, $A = A'$ (A' is at the position indicated in figure 15) and then $r = \frac{M}{\sqrt{3}}$.
Now, if we consider the n points $y_1, \ldots, y_n$ and using Lemma 1 and Corollary 1, the maximum of the inertia $\sum_{i=1}^n \| y_i - g \|^2$, under the constraint that $y_1, \ldots, y_n$ are inside (C) is equal to $nr^2 = n\frac{M^2}{3}$. Thus, the maximum of $\sum_{i=1}^n \| x_i - g \|^2$ is upper bounded by $n\frac{M^2}{3}$.

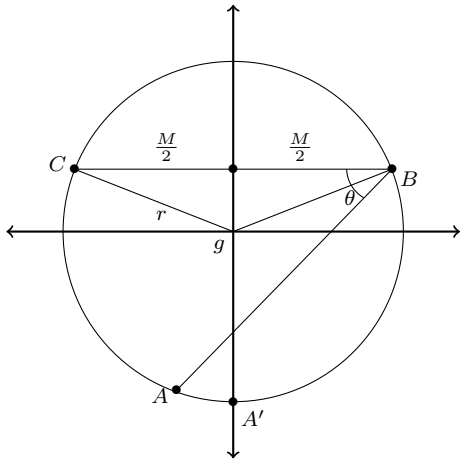


Fig. 15: Representation of the points on the circle.

Appendix B: Three functions used to compute the lower bound of $\sum_{i=1}^{n} r_i$

Recall optimization problem $\mathcal{P}_{r,x}$:

$$\begin{cases} \min\limits_{r_1,\ldots,r_n\in\mathbb{R}^+,x_1,\ldots,x_n\in\mathbb{R}^k} \sum\limits_{i=1}^{n} r_i \\ s.t \;\; d_{ij} - \parallel x_i - x_j \parallel \leq r_i + r_j \\ \qquad \parallel x_i - x_j \parallel - d_{ij} \leq r_i + r_j \end{cases}$$

Appendix B.1: Function $f(M)$

Using the first constraint, we have:

$$\begin{aligned} d_{ij} &\leq \parallel x_i - x_j \parallel + r_i + r_j \\ d_{ij}^2 &\leq (\parallel x_i - x_j \parallel + r_i + r_j)^2 \\ \sum_{i<j} d_{ij}^2 &\leq \sum_{i<j} \parallel x_i - x_j \parallel^2 + \sum_{i<j} (r_i + r_j)^2 + 2\sum_{i<j} (\parallel x_i - x_j \parallel)(r_i + r_j) \\ \sum_{i<j} d_{ij}^2 &\leq \sum_{i<j} \parallel x_i - x_j \parallel^2 + \sum_{i<j} (r_i + r_j)^2 + 2M \sum_{i<j} (r_i + r_j) \text{ as } \parallel x_i - x_j \parallel \leq M \cdot \end{aligned} \quad (5)$$

Let g be the center of gravity of the projected points $x_1, \ldots, x_n$, so:

$$\begin{aligned} \parallel x_i - x_j \parallel &= \parallel x_i - g - x_j + g \parallel \\ \parallel x_i - x_j \parallel^2 &= \parallel x_i - g \parallel^2 + \parallel x_j - g \parallel^2 + 2(x_i - g)'(x_j - g) \\ \sum_{i<j} \parallel x_i - x_j \parallel^2 &= \sum_{i<j} \left(\parallel x_i - g \parallel^2 + \parallel x_j - g \parallel^2 \right) + 2\sum_{i<j} (x_i - g)'(x_j - g) \cdot \end{aligned}$$

As $\sum_{i<j}(x_i-g)'(x_j-g)=0$ then:

$$\sum_{i<j}\parallel x_i-x_j\parallel^2=(n-1)\sum_{i=1}^{n}\parallel x_i-g\parallel^2\cdot \tag{6}$$

Replacing equation (6) in (5) gives :

$$(n-1)\sum_{i=1}^{n}\parallel x_i-g\parallel^2+\sum_{i<j}(r_i+r_j)^2+2M\sum_{i<j}(r_i+r_j)-\sum_{i<j}d_{ij}^2\geq 0\cdot \tag{7}$$

The quantity $\sum_{i=1}^{n}\parallel x_i-g\parallel^2$ is the inertia of the projected points $x_1,\ldots,x_n$. As long as we want to conserve the initial information, the inertia must be maximal under constraints $\|x_i-x_j\|\leq M$ for all $1\leq i<j\leq n$.
Recalling equation (7) and by using Lemma 2, we obtain:

$$\frac{n(n-1)}{3}M^2+\sum_{i<j}(r_i+r_j)^2+2M\sum_{i<j}(r_i+r_j)-\sum_{i<j}d_{ij}^2\geq 0$$

$$\frac{n(n-1)}{3}M^2+(n-1)\left(\sum_{i=1}^{n}r_i\right)^2+2(n-1)M\sum_{i=1}^{n}r_i-\sum_{i<j}d_{ij}^2\geq 0$$

$$\left(\sum_{i=1}^{n}r_i\right)^2+2M\left(\sum_{i=1}^{n}r_i\right)+\frac{n}{3}M^2-\frac{1}{n-1}\sum_{i<j}d_{ij}^2\geq 0\cdot \tag{8}$$

The discriminant of equation (8) is given by: $\Delta=4\left(1-\frac{n}{3}\right)M^2+\frac{4}{n-1}\sum_{i<j}d_{ij}^2$ and as $r_i\geq 0$, $\forall i=1,\ldots,n$ we get:

$$\sum_{i=1}^{n}r_i\geq\sqrt{\left(1-\frac{n}{3}\right)M^2+\frac{1}{n-1}\sum_{i<j}d_{ij}^2}-M\cdot$$

We note $f(M)=\sqrt{\left(1-\frac{n}{3}\right)M^2+\frac{1}{n-1}\sum_{i<j}d_{ij}^2}-M\cdot$

Appendix B.2: Function $g(M)$

Two situations are possible:

1. $\exists(i',j')$ such that $\parallel x_{i'}-x_{j'}\parallel=M$, that gives:

   $$r_{i'}+r_{j'}\geq\parallel x_{i'}-x_{j'}\parallel-d_{i'j'}\geq M-d_{i'j'}\geq M-d_{max}\cdot$$

   As $\sum_{i=1}^{n}r_i\geq r_{i'}+r_{j'}$, we obtain:

   $$\sum_{i=1}^{n}r_i\geq M-d_{max}$$

2. $\exists(i^*,j^*)$ such that $d_{i^*j^*}=d_{max}$, that gives:

   $$r_{i^*}+r_{j^*}\ \geq d_{i^*j^*}-\parallel x_{i^*}-x_{j^*}\parallel\geq d_{max}-M\cdot$$

   Then, we obtain:

   $$\sum_{i=1}^{n}r_i\geq d_{max}-M\cdot$$

Hence:

$$\sum_{i=1}^{n} r_i \geq |M - d_{max}|. \tag{9}$$

We note $g(M) = |M - d_{max}|$.

Appendix B.3: Function $h(M)$

Let us consider four distinct points i, j, k and l. We suppose that there is a couple (i, j) such that $\|x_i - x_j\| = M$ and one of their coordinates is equal to zero ($x_i = 0$ or $x_j = 0$). We distinguish two cases:

1. $x_i = 0$.
2. $x_j = 0$.

Case 1: For $x_i = 0$, we take $x_j = \alpha x_k + \beta x_l$ with $\alpha, \beta \in [0, 1]$. The constraints related to these four points are the following:

$$\begin{cases} |\|x_j\| - d_{ij}| \leq r_i + r_j & \text{(C1)} \\ |\|x_k\| - d_{ik}| \leq r_i + r_k & \text{(C2)} \\ |\|x_l\| - d_{il}| \leq r_i + r_l & \text{(C3)} \\ |\|x_j - x_k\| - d_{jk}| \leq r_j + r_k & \text{(C4)} \\ |\|x_j - x_l\| - d_{jl}| \leq r_j + r_l & \text{(C5)} \\ |\|x_k - x_l\| - d_{kl}| \leq r_k + r_l & \text{(C6)} \end{cases}$$

Firstly, using constraints (C4) and (C6) we obtain:

$$2\sum_{t=1}^{n} r_t \geq d_{jk} - d_{kl} + \|x_k - x_l\| - \|x_j - x_k\|.$$

Additionally, as $x_j = \alpha x_k + \beta x_l$ with $\alpha, \beta \in [0, 1]$, then

$$\|x_k - x_j\| = \|x_k - \alpha x_k - \beta x_l\| = \|(1 - \alpha)x_k - \beta x_l\| \leq \|x_k - x_l\|,$$

which gives

$$\sum_{t=1}^{n} r_t \geq \frac{d_{jk} - d_{kl}}{2}. \tag{10}$$

Secondly, using constraints (C5) and (C6) we obtain:

$$2\sum_{t=1}^{n} r_t \geq d_{jl} - d_{kl} + \|x_k - x_l\| - \|x_j - x_l\|.$$

As $\|x_l - x_j\| \leq \|x_k - x_l\|$ we obtain:

$$\sum_{t=1}^{n} r_t \geq \frac{d_{jl} - d_{kl}}{2}. \tag{11}$$

Thirdly, constraint (C1) and initial hypothesis $\|x_i - x_j\| = M$ lead to:

$$\sum_{t=1}^{n} r_t \geq |d_{ij} - M|. \tag{12}$$

Consequently, equations (10), (11) and (12) involve:

$$\sum_{t=1}^{n} r_t \geq \max\left\{\frac{d_{jk} - d_{kl}}{2}; \frac{d_{jl} - d_{kl}}{2}; |d_{ij} - M|\right\} \text{ denoted } L^1_{ijkl}$$

Case 2: For $x_j = 0$, we take $x_i = \alpha x_k + \beta x_l$ with $\alpha, \beta \in [0,1]$. By analogy with case 1, we obtain:

$$\sum_{t=1}^{n} r_t \geq \max\left\{\frac{d_{ik} - d_{kl}}{2}; \frac{d_{il} - d_{kl}}{2}; |d_{ij} - M|\right\} \text{ denoted } L^2_{ijkl}.$$

Due to the choice of one case among cases 1 and 2, we take the minimum of L^1_{ijkl} and L^2_{ijkl}. Thus:

$$\sum_{t=1}^{n} r_t \geq \min\left\{L^1_{ijkl}; L^2_{ijkl}\right\}.$$

Moreover, for a given i, j, this inequality is verified. So that:

$$\sum_{t=1}^{n} r_t \geq \min_{i<j} \max_{k<l;k,l\neq i,j} \min\left\{L^1_{ijkl}; L^2_{ijkl}\right\}.$$

We note $h(M) = \min_{i<j} \max_{k<l;k,l\neq i,j} \min\left\{L^1_{ijkl}; L^2_{ijkl}\right\}$.