

1

2 **Projection under pairwise distance control**

3 Hiba Alawieh ^a, Nicolas Wicker ^a and Christophe Biernacki ^a

4 ^aUniversité Lille 1, - UFR de Mathématiques, cité scientifique, 59655 Villeneuve d'Ascq,
5 France

6 **ARTICLE HISTORY**

7 Compiled March 1, 2020

8 **ABSTRACT**

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

18 **KEYWORDS**

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

21 **1. Introduction**

22 Several domains in science use data with large numbers of variables in their studies
23 such as in biology (Cheung 2012, Golub *et al.* 1999), chemistry (Svante *et al.* 1984),
24 geography (Van der Hilst *et al.* 2007) and finance (Jagannathan and Ma 2003). These
25 data can be viewed as a large matrix and extracting results from this 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, variable 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 PCA and MDS is that as the number of dimensions grows, important multi-dimensional relationships might not be visualized. Moreover, the quality of projection usually assessed by the percentage of variance (PCA case) that is conserved or by the stress factor (MDS case) is a global projection quality measure and does not give information about local quality.

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 (Jolliffe 1986). 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 structure in the data which

57 does not allow to give powerful results in case of nonlinear configurations. Therefore,
58 many methods have been developed to perform nonlinear projections by nonlinearizing
59 a linear dimensionality reduction or by using manifold learning methods.

60 The nonlinearization of linear dimensionality reduction is applied to extract nonlinear
61 principal components. Kernel PCA is one of the most popular methods in this domain,
62 which integrates a kernel function to determine principal components in different high-
63 dimensional space (Schölkopf 1998). Manifold learning methods are an approach to
64 construct a matrix using the neighborhood information and take a spectral decom-
65 position to find a nonlinear embedding (like Locally Linear Embedding LLE, Isomap
66 algorithm etc) (Lee and Verleysen 2007, Tenenbaum *et al.* 2000, Roweis and Saul
67 2000).

68 In this paper, we propose a new nonlinear projection method that projects the
69 points into a reduced space by using the pairwise distance between pairs of points and
70 by taking into account the projection quality of each point taken individually. Nonlin-
71 ear projection methods cited in the previous paragraph project the points in a feature
72 space which makes the distances between the projected points hard to be interpreted.
73 In our method, the distances between projected points are related to the initial dis-
74 tances between points, offering a way to easily interpret the distances observed in the
75 projection plane. This projection leads to a representation of the points as circles with
76 a different radius associated to each point. Henceforth, this method will be referred to
77 as “Projection under pairwise distance control”. Furthermore, visualization of data in
78 a reduced space is not the only objective of this method. It can serve as a dimension-
79 ality reduction method to reduce the number of variables by minimizing the sum of
80 the radii and to then determine the number of variables that can be kept.

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

87 2. Projection under pairwise distance control

88 Let us consider n points given by their pairwise distances denoted by d_{ij} for $i, j \in$
89 $\{1, \dots, n\}$. The objective is to project these points using distances into a reduced
90 space \mathbb{R}^q by introducing additional variables, called hereafter radii, that indicate the
91 extent to which the projection of each point is accurate. The local quality is then given
92 by the values of the radii. A good projection quality of point i is indicated by a small
93 radius value denoted by r_i . It is important to note that both units of d_{ij} 's and r_i 's are
94 identical, thus allowing for a direct comparison.

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

97 2.1. Overview of certain existing methods: PCA, KPCA and MDS

98 *Principal Component Analysis (PCA)*

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

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

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

109 **Kernel PCA (KPCA)**

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

$$114 \quad \begin{aligned} \Phi : \mathbb{R}^p &\rightarrow \mathcal{F} \\ Y &\rightarrow \Phi(Y) . \end{aligned}$$

115 The original data y_i is represented in the feature space as a function $\Phi(y_i) = k(y_i, \cdot)$,
 116 where $k(\cdot, \cdot)$ is a positive kernel. Similar to PCA, KPCA is based on finding the first
 117 q eigenvectors corresponding to the q largest eigenvalues λ_i of the Gram matrix $K =$
 118 $(k_{ij})_{i,j \in 1, \dots, n}$, where $k_{ij} = k(y_i, y_j) = \langle \Phi(y_i), \Phi(y_j) \rangle$ is a chosen positive kernel. Letting
 119 V_v , for $v = 1, \dots, q$, the eigenvectors in the feature space and $P_{\Phi(y_i)}$ the projection
 120 of $\Phi(y_i)$ onto the subspace V_1, \dots, V_q . The KPCA problem can be represented as a
 121 minimization problem with the following error:

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

$$122 \quad \text{where } P_{\Phi(y)} = \sum_{v=1}^q \langle \Phi(y), V_v \rangle V_v .$$

123 Furthermore, the most well-known and used measure applied to evaluate the pro-
 124 jection quality of points for PCA and KPCA is the squared cosine value. Squared
 125 cosine values cannot be interpreted at the same time as the distances in the projection
 126 because the cosine values do not have a specific unit. More precisely, the visualization
 127 of the projection in the reduced space using PCA and KPCA cannot simply be inter-
 128 preted in terms of original distances between the points. Indeed, in PCA, the cosine
 129 values do not provide a quantitative assessment of the error made when considering
 130 the distances between the projected points, even less in KPCA where the projected
 131 points are in the feature space so the term “distances” is not related to the distances
 132 between the points in the original space.

133 **Multidimensional Scaling (MDS)**

134 As with PCA, Multidimensional scaling (MDS) consists of finding a new data config-
 135 uration in a reduced space. The main difference between these two methods is that
 136 the input data in MDS is in the form of a similarity or dissimilarity matrix, called
 137 “proximity”, representing the proximity between pairs of objects. MDS are developed
 138 where the proximities behave like distances or not respectively (Borg and Groenen
 139 2005, Shepard 1962). The key idea of MDS is to perform dimensionality reduction in
 140 a way to approximate high-dimensional distances denoted by δ_{ij} the low-dimensional
 141 distances d_{ij} , where d_{ij} is equal to the distance between x_i and x_j , the coordinates of
 142 i and j in the reduced space. In his original paper on MDS (Kruskal 1964), Kruskal
 143 proposed the least-squares loss function denoted by “Stress” as follows

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

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

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

$$\mathcal{P}_{\text{MDS}} : \begin{cases} \min_{x_1, \dots, x_n \in \mathbb{R}^q, r_1, \dots, r_n \in \mathbb{R}^+} \sum_{i=1}^n r_i \\ \text{s.t.} \quad \sum_{i=1}^n r_i \geq \frac{1}{n-1} \sqrt{\frac{\sum_{1 \leq i < j \leq n} (d_{ij} - \|x_i - x_j\|)^2}{\sum_{1 \leq i < j \leq n} d_{ij}^2}}. \end{cases}$$

148 Note that the optimal solution of the MDS problem may not be unique (Kruskal and
 149 Wish 1978).

150 A criterion to determine the local projection quality has been proposed by Born
 151 and Groenen called Stress-per-point (*SPP*) (Borg and Groenen 2005). The *SPP* of

point i is given by:

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

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

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

Let x_1, \dots, x_n be the coordinates of the projected points in \mathbb{R}^p and $\|x_i - x_j\|$ 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 radius (r_i, r_j) . Figure 1 depicts this idea: for each point $i \in \{1, \dots, n\}$, the projection of i belongs to a sphere with center x_i and radius r_i such that for each couple $(i, j) \in \{1, \dots, n\}$ we have $\|x_i - x_j\| - (r_i + r_j) \leq d_{ij} \leq \|x_i - x_j\| + r_i + r_j$.

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 $1 \leq i < j \leq n$.

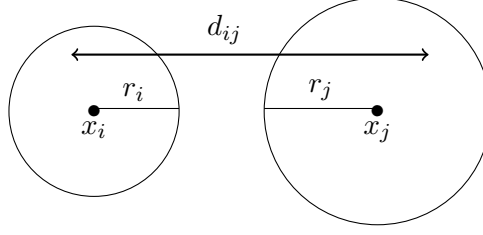


Figure 1.. Example of radii for bounding of the original distance d_{ij}

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

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

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

$$\mathcal{P}_r : \begin{cases} \min_{r_1, \dots, r_n \in \mathbb{R}^+} \sum_{i=1}^n r_i \\ \text{s.t. } |d_{ij} - \|x_i - x_j\|| \leq r_i + r_j, \text{ for } 1 \leq i < j \leq n \end{cases}$$

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

185 **$\mathcal{P}_{r,x}$ is a non-convex optimization problem:** For any dimension p , even with
 186 $p = 1$, note that the optimization problem $\mathcal{P}_{r,x}$ is not convex. Indeed, to easily illus-
 187 trate this fact, we take the function $g(x, y) = |d - \|x - y\||$ considering two solutions
 188 $(x_1, y_1) = (0, 2)$ and $(x_2, y_2) = (3, 1)$ with d equal to 2. Thus, we have $g(x_1, y_1) = 0$ and

189 $g(x_2, y_2) = 0$ but $g\left(\frac{x_1 + x_2}{2}, \frac{y_2 + y_2}{2}\right) = \left|d - \left\|\frac{x_1 + x_2}{2} - \frac{y_1 + y_2}{2}\right\|\right| = |2 - 0| = 2$
190 which is larger than $\frac{g(x_1, x_2) + g(y_1, y_2)}{2} = 0$ proving non convexity associated to this
191 sample design.

192 Many methods available in the literature propose different ways to solve such opti-
193 mization problems. Examples include: trust-region-reflective (Conn *et al.* 2000), which
194 chooses and computes an approximation of the objective function, and then chooses
195 and modifies the trust region and finally solves the trust-region subproblem; sequential
196 quadratic programming (SQP) which solves the optimization problem by addressing
197 a sequence of quadratic programming problems where the Lagrangian function is ap-
198 proximated by a quadratic function and the constraints are approximated by a linear
199 hyper-space (Boggs and Tolle 1995); the active-set method, which is composed of
200 two phases, wherein for the first phase (the feasibility phase) the objective function is
201 ignored while a feasible point is found for the constraints, and in the second phase (the
202 optimality phase) the objective function is minimized while feasibility is maintained
203 (Wong 2011, Cristofari *et al.* 2007). The choice of optimization method to use to
204 achieve optimality of the optimization problem is essential and depends on many fac-
205 tors such as the type of problem, desired quality of solution, time limit and availability
206 of the algorithm implementation etc. In fact, all of the methods cited above can be
207 used in optimizing problem $\mathcal{P}_{r,x}$ which is a constrained optimization problem having
208 inequality constraints and they are all available in MATLAB using the function “fmi-
209 con” for constrained nonlinear optimization problems. Having small radii is the main
210 constraint in our optimization problem, thus the objective is to obtain good solution
211 within a reasonable and practical timeframe. Therefore, a method that balances time
212 and quality of the solution is required.

213 **Another strategy of use: Dimensionality reduction** One of the main objectives
214 of high-dimensional data studies is to choose, from a large number of variables, those
215 that are important for understanding the underlying studied phenomena. In addition
216 to visualization, our aim can thus be to reduce the dimension rather than to visualize
217 data in \mathbb{R}^2 . Therefore, the proposed method can serve to reduce the number of variables
218 by taking into account the 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 $i = 1, \dots, 4$ 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.

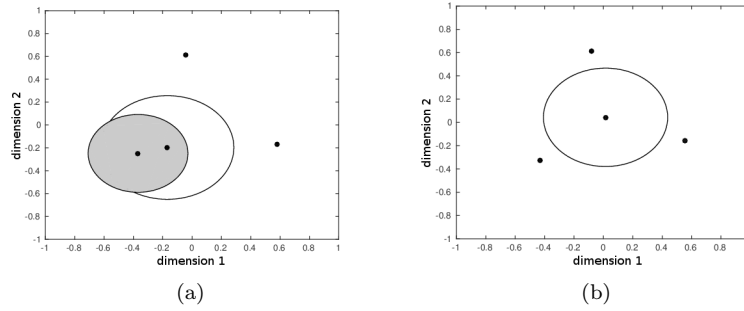


Figure 2.. Projected points after solving problem \mathcal{P}_r and problem $\mathcal{P}_{r,x}$. The x-axis and y-axis are dimension 1 and dimension 2, respectively. (a) and (b) show the projection obtained from the solution of problem \mathcal{P}_r using MDS and of problem $\mathcal{P}_{r,x}$, respectively.

229

This result is 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, the smaller the radius, the darker the 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 problem $\mathcal{P}_{r,x}$ is better than when using problem \mathcal{P}_r . In Figure 2a, half of the points are

239 well projected whereas the other half have large radii, indicating that they are not well
 240 projected. Moreover, it is worth noting that the three outer points all have radii equal
 241 to 0, which indicates that they are all perfectly placed with respect to one another.
 242 In Figure 2b, the distances between the three points that are very well projected
 243 are equal to the distances between these points in their original space ($d_{kl} = \|x_k -$
 244 $x_l\|$ where k and l are two very well projected points) whereas the distances from
 245 the badly projected points to the perfectly projected points are not yet conserved.
 246 Therefore, using the proposed method, we have succeeded in conserving half of the
 247 original distances in the new projection plane and the other half have been changed
 248 to fit the new configuration. If we now apply the proposed method to the distances
 249 obtained by MDS to find the radius of each projected point (Figure 2a), it can be noted
 250 that one distance is conserved as the original distance and the other five distances
 251 are changed which indicates that the proposed method projects the points well by
 252 conserving the distances between the points as much as possible.

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

258 **2.4. Connection with existing methods**

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

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

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

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

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

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

269 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. \square

270 3. Optimization tools for performing the proposed method

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

275 3.1. Initialization point for problem $\mathcal{P}_{r,x}$

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

278 **1- Initial point using a known projection method** The first possibility is to
 279 use the matrix obtained by PCA or another projection method. The choice of method
 280 must be based on the type of data. In this application, we use PCA for quantitative
 281 data and MDS for categorical and functional data.

282 **2- Initial point using squared distances** The optimization problem $\mathcal{P}_{r,x}$ can be
 283 changed by taking the squared distances between points instead of the distances.

284 Rewriting r_i^2 as R_i , the problem is changed into

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

285 This transformation is interesting because if the constraints of problem $\mathcal{P}_{R,x}$ are sat-
286 isfied, 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:

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

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

288 **3- Initial point using an improved solution of problem \mathcal{P}_r .** This strategy is
289 more involved. First, we need two properties that provide a way to improve the opti-
290 mization 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:*

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

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

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

the new optimization problem is given by:

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

Proposition 3.2. Let $r_1, \dots, r_n; x_1, \dots, x_n$ 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 $\exists \tilde{r}_1, \dots, \tilde{r}_n$ a solution of $\mathcal{P}_{r,a}$ such that $\sum_{i=1}^n \tilde{r}_i < \sum_{i=1}^n r_i$.

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: $X_{\mathcal{P}_{PCA/MDS}}$ 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 $R_{\mathcal{P}_{R,x}} = r_{\mathcal{P}_{R,x}}^2$ 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: $(X_{\mathcal{P}_{PCA/MDS}}, r_{\mathcal{P}_r})$.

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 $X_{\mathcal{P}_{PCA/MDS}}$, $X_{\mathcal{P}_{R,x}}$ and X_{imp} .

Choose the minimal solution obtained by these three different starting points.

337 3.3. A stochastic strategy: Algorithm 2

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

344 1- A target distribution:

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

$$347 \pi(x) \propto \exp\left(\frac{-E(x)}{T}\right),$$

where E is a function in \mathbb{R} given by:

$$E(x) = \sum_{i=1}^n r_i, \text{ where } \{r_1, \dots, r_n\} \text{ is the solution of problem } \mathcal{P}_r \text{ with fixed } x.$$

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

350 2- A proposal distribution:

351 The choice of the proposal distribution is very important to obtain mean-
 352 ingful results. It should be chosen in such a way that the proposal distribution
 353 approaches the target distribution. The proposal distribution $q(X \rightarrow \cdot)$ is con-
 354 structed as follows, giving priority to the selection of points involved in saturated
 355 constraints:

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

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

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

359

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

361 – else $x_i^* = x_i - |c_{ij^{(i)}}|L_{ij^{(i)}}$,

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

363 3- A linear optimization problem:

364 For the matrix X generated at each iteration, we solve the linear optimization
 365 problem \mathcal{P}_r and we choose finally the matrix X and the vector of radii which
 366 give the smallest value of $\sum_{i=1}^n r_i$.

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

369 4. Numerical applications

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

372 4.1. Experimental setup

373 In practice, we have tested the proposed method on different simulated and real data
 374 sets by solving the optimization problem $\mathcal{P}_{r,x}$ using Algorithm 1 in addition to the
 375 proposed Metropolis-Hastings algorithm (Algorithm 2). A distance matrix is required
 376 each time. For quantitative data, the Euclidean distance between points $y_i \in \mathbb{R}^p$,
 377 for $i = 1, \dots, n$, is computed by the known formula $d_{ij} = \sqrt{\sum_{k=1}^p (y_{ik} - y_{jk})^2}$. For
 378 categorical data, the distance between two points (i, j) is given through the Eskin
 379 similarity measure (Boriah *et al.* 2008) computed by the formula $p_{ij} = \sum_{t=1}^Q w_t p_{ij}^t$

380 where $p_{ij}^t = \begin{cases} 1 & \text{if } i^t = j^t \\ \frac{n_t^2}{n_t^2 + 2} & \text{else} \end{cases}$, p_{ij}^t is the per-attribute Eskin similarity between
 381 two values for the categorical attribute indexed by t , w_t is the weight associated to
 382 the attribute t called w_t which is defined by: $w_t = \frac{1}{Q}$, Q is the number of attributes
 383 and n_t is the number of values taken by each attribute. Then, the distances can be

384 obtained by the standard transformation formula (Du Toit *et al.* 1986) converting
385 similarities to distances: $d_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$.

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

395 For problems \mathcal{P}_r and $\mathcal{P}_{r,a}$, we first applied PCA for quantitative data and MDS for
396 categorical and functional data; a linear programming package, called “linprog” which
397 solves linear programming problems, was then used to solve the optimization problems
398 with an interior-point algorithm. Problems $\mathcal{P}_{r,x}$ and $\mathcal{P}_{R,x}$ are nonlinear optimization
399 problems; therefore, we used a nonlinear programming package, called “fmincon” which
400 finds minimum of constrained nonlinear multi-variable function, to solve them. The
401 algorithms cited in Section 2.2 can be used here, but we recommend to use the active-
402 set algorithm. Algorithm 2 can provide a good solution if the parameters λ , σ and
403 T are chosen adequately. For instance, λ should be such that the points belonging
404 to unsaturated constraints are chosen with small probabilities. Therefore, we took it
405 equal to 100. For the other parameters σ and T , we took their values in the range
406 from 0.01 to 100. The choice of these numbers is taken after trying different values of
407 σ and T in order to have the best solution that gives a minimal value of $\sum_{i=1}^n r_i$.

408 Moreover, the visualization of the projection of each point i in \mathbb{R}^2 is represented
409 as a circle having x_i as the center and r_i as the radius in a two-dimensional space,
410 where the horizontal and vertical axes correspond to the first and the second dimension
411 associated to the projection in \mathbb{R}^2 , respectively. The projected point belongs to this
412 circle and this is the specificity of our method. For each data set, the circles obtained
413 for each point after solving the optimization problem $\mathcal{P}_{r,x}$ are shown. To compare the

414 projection quality of our representation with that obtained by PCA and KPCA, we
 415 used the squared cosine values as projection quality, and for MDS, the Stress-per-
 416 point (*SPP*) (Borg and Groenen 2005). Indeed, for PCA and KPCA, we plotted
 417 the projected points indexed by their squared cosine values and for MDS, we used
 418 the smacof package in R to compute the stress-per-point and to plot the bubble plot
 419 represented the stress-per-point.

420 4.2. A simulation study

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

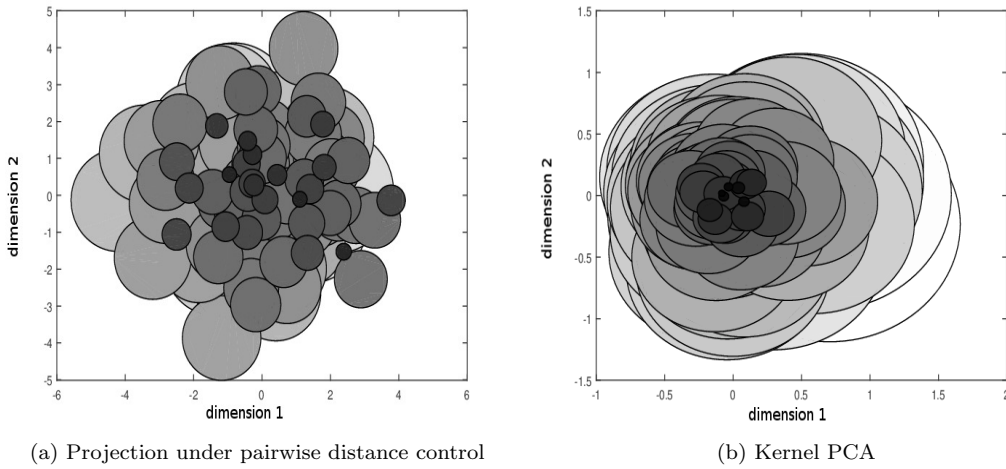


Figure 3.. Projection of the simulated data using the proposed method (a) and Kernel PCA (b). The points that have circles with small radii are considered as well projected points.

426

427 Figure 3 shows the results of the projection of the simulated data using the proposed
 428 method and KPCA. By comparing Figure 3a and Figure 3b, it can be shown that
 429 the projection quality of points using KPCA is somehow dependent on the position
 430 of the points in the reduced space. Indeed, the projection is likely to give better local

431 projection quality if the projected point is located near to the center $(0,0)$. On the
432 contrary, this is less visible for the proposed method.
433 This simulated data illustrates the originality and the efficiency of the proposed
434 method in giving a good local projection quality.

435 **4.3. Introducing the real data sets**

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

- 437 • Quantitative data: Iris and car data sets.
- 438 • Categorical data: Soybean data set.
- 439 • Functional data: Coffee data set.

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

447 The soybean data set (Stepp 1984) from *UCI Machine Learning Repository* charac-
448 terizes 47 soybean disease case histories defined over 35 attributes. Each observation is
449 identified by one of the 4 diseases: Diaporthe Stem Canker (D1), Charcoal Rot (D2),
450 Rhizoctonia Root Rot (D3) and Phytophthora Rot (D4).

451 The coffee data set is a time series data set used in chemometrics to classify food
452 types. It is a functional data set where 56 samples of coffee are available with 286
453 timestamps for each sample (as a result of spectroscopic analysis). This kind of time
454 series is common in many applications in food safety and quality assurance and was
455 taken from the *UCR time Series Classification and Clustering* website (Chen *et al.*
456 2015). *Coffea Arabica* and *Coffea Canephora* variant *Robusta* are the two species of
457 coffee bean that have acquired a worldwide economic importance, and many methods
458 have been developed to discriminate between these two species by chemical analysis
459 (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 solution of Algorithm 1. 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 computing time. In fact, the important decrease in the value of $\sum_{i=1}^n r_i$ occurred in the first 500 iterations through 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 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.

Referring to the original data, the Iris data set contains three classes corresponding

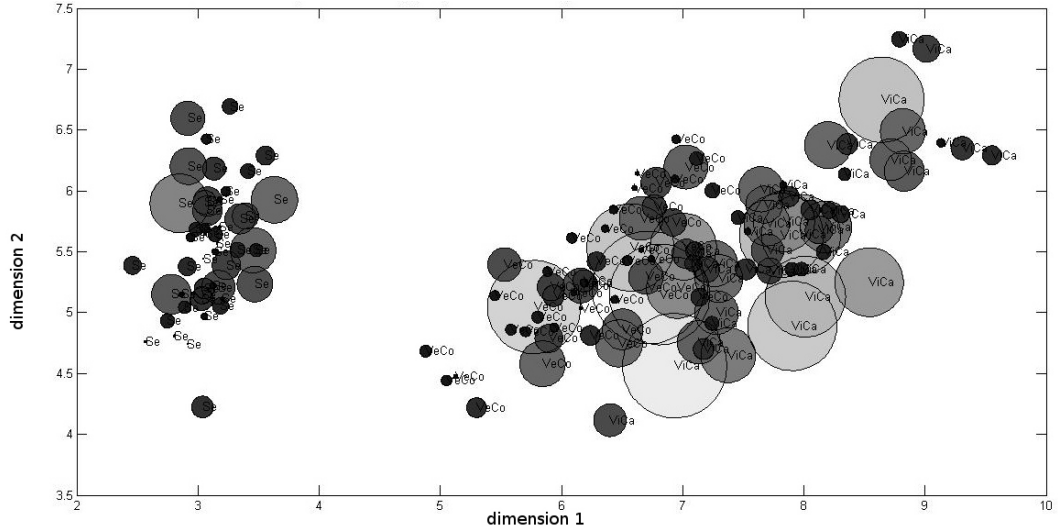


Figure 4.. Projection of the Iris data set using projection under pairwise distance control method. Two well separated groups can be observed. The points that have circles with small radii are considered as well projected points.

481 to the three types of Iris plants and one class is linearly separable from the other two
 482 classes. This result clearly appears in our projection.

483 Moreover, we have compared the local projection quality of PCA, KPCA and MDS
 484 with the local projection quality obtained using projection under pairwise distance
 485 control. By comparing the projection of PCA with the projection of our method for
 486 the Iris data set given respectively in Figures 5 and 4, we can say that our method
 487 projected the points without giving any importance to any group. Figure 5 depicts a
 488 group with small values of the quality measure and another group with high values
 489 of quality measure, whereas the radii obtained by projection under pairwise distance
 490 control method are distributed in an equivalent way.

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

495 Furthermore, by comparing the proposed projection method with the one obtained by
 496 MDS, it can be concluded that, as is the case when using PCA, the points in Figure
 497 6b are projected by giving more importance to the Iris setosa group. Indeed, almost
 498 all the red circles (indicating a very good projection) are assigned to the Iris setosa

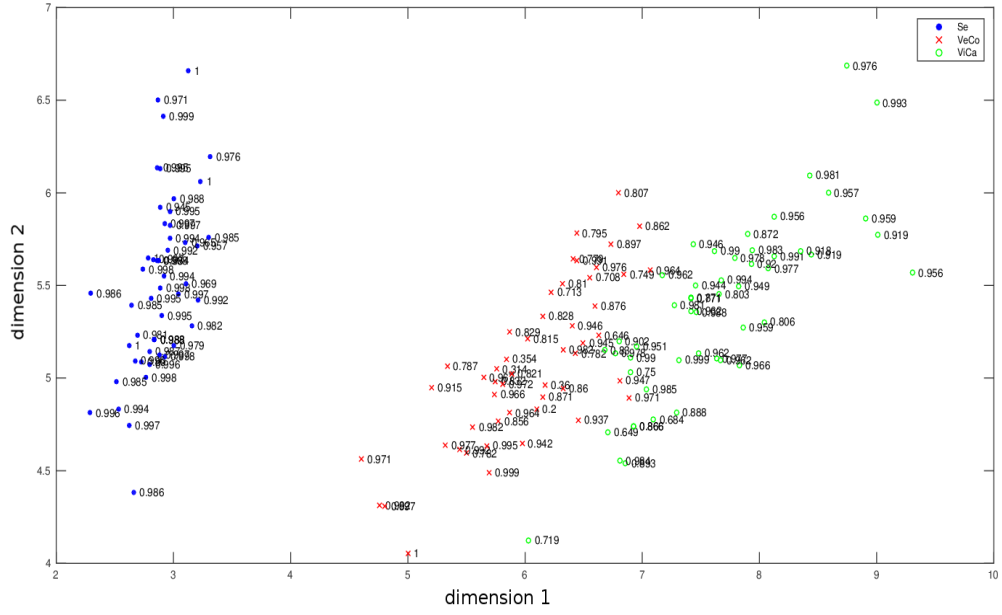


Figure 5.. Projection of the Iris data set using PCA. The values of local projection quality are given for each instance. The values of projection quality for Iris setosa species (Se) vary between 0.97 and 1 indicating then a very good projection quality whereas for Iris versicolor (VeCo) and Iris virginica (ViCa) species, the values of projection quality vary between 0.2 and 0.99 indicating a very large variability in the projection quality.

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 are not in a metric compatible to the initial distances, whereas in our method we have conserved the metric of the initial distances.

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 are well separated as there are no circle intersections. 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

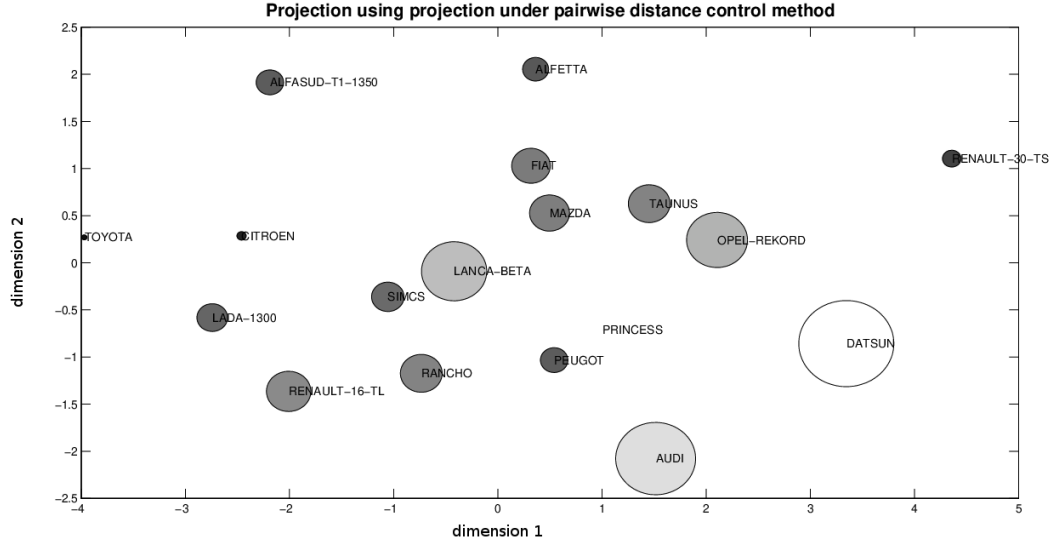


Figure 7.. Projection of the car data set using projection under pairwise distance control.

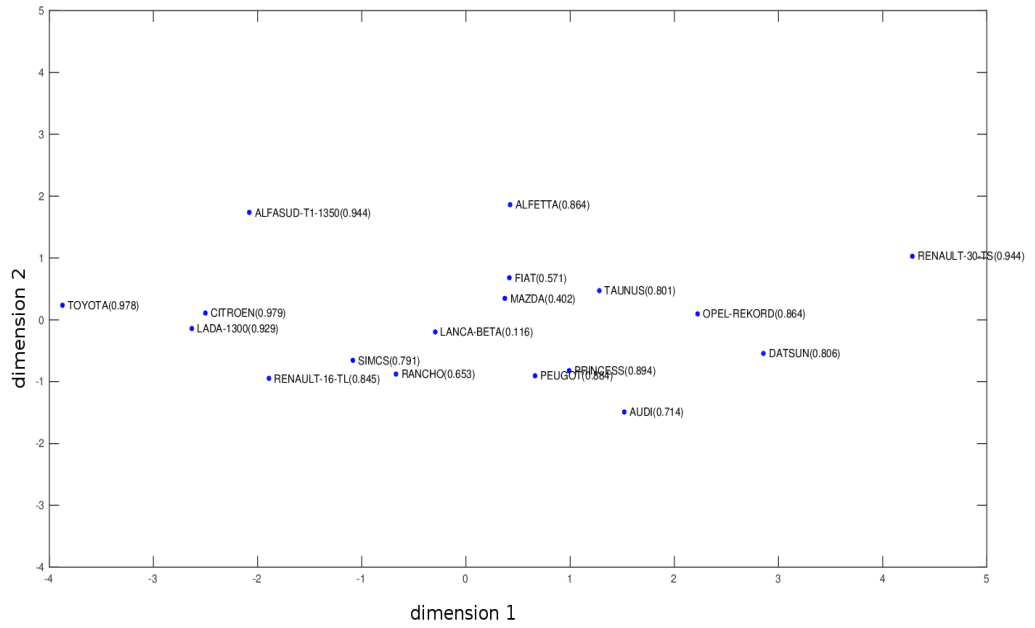


Figure 8.. Projection of the cars data set using PCA. The values of local projection quality are given for each car.

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

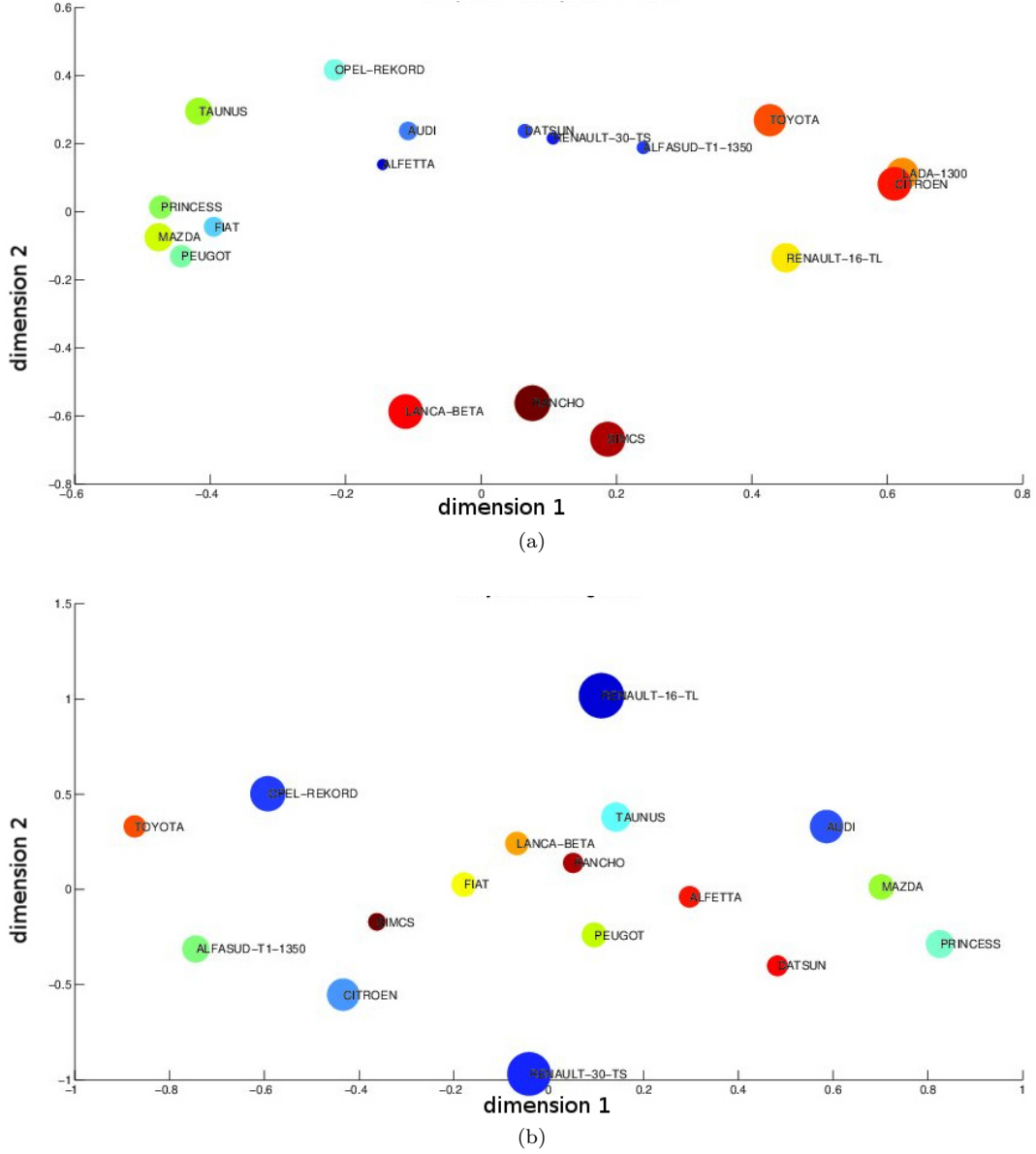


Figure 9.. Projection of the car data set using KPCA (a) and MDS (b).

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 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

527 significant. Conversely, the distance between the “Lada1300” and “Citroen” is small,
528 thus indicating the closeness of these two cars. Note that these two cars are very well
529 projected, resulting in a very good interpretation of the distance between them.

530 Therefore, radii are meaningful in our method and give an interpretation about the
531 distances between points whereas the distances between the projected points using
532 PCA, KPCA and MDS are not interpretable. This is a particular strength of our
533 method. Projection under pairwise distance control suggests an absolute interpretation
534 whereas the other methods provide a relative one.

535 For the qualitative and functional data sets and using MDS, recall the definition of
536 the Gram matrix called B which is equal to $X'X$ where X is the coordinate matrix in
537 the reduced space. Thus, it is necessary to verify that the matrix B obtained by the
538 MDS method is semi-definite positive to use the squared cosine as the quality measure
539 because the starting point of optimization is obtained from MDS. After this, in case
540 of positiveness of matrix B , the quality measure can be calculated.

541 **Soybean data set:** In the projection of the soybean data set, four classes are shown
542 in Figure 10 and each class contains the disease number of the class. The whole set of
543 points can however be divided in two large classes. Indeed, it is clear that Class 2 is
544 well separated from the other classes as there is no intersection between the circles of
545 Class 2 and the circles of other classes. Moreover, Class 1 can be considered as well
546 separated class from Classes 3 and 4 if the largest circle D_3 is not taken into account.
547 Classes 3 and 4 are not well separated at all, as there are different intersections between
548 the circles of these two classes. This result is shown in Stepp (1984) which labels the
549 first two classes as “normal” and the latter two classes as “irrelevant”. A comparison
550 of results from projection under pairwise distance control with PCA and KPCA is not
551 possible for this data set because the matrix B is not semi-definite positive. Regarding
552 Figure 11, it is clear that Class 4 exhibits the worst projection quality, whereas Classes
553 1 and 2 show better projection quality. Therefore, it is possible to draw the same
554 conclusion for the Iris and car data sets when using MDS as a projection method, the
555 projection quality of points is dependent on the class of the points.

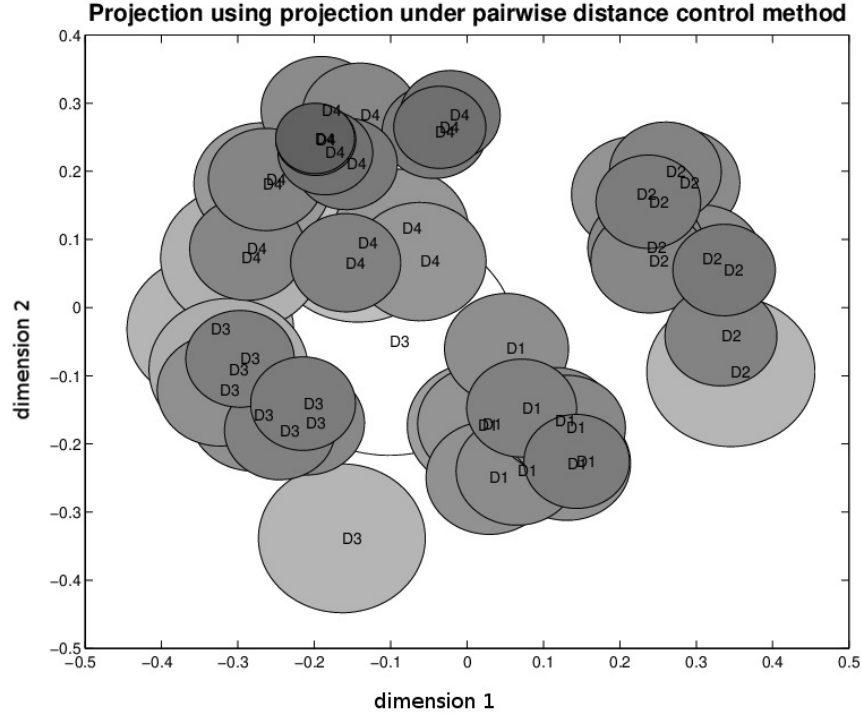


Figure 10.. Projection under pairwise distance control for the soybean data set. Four groups are presented, indexed by D1, D2, D3 and D4.

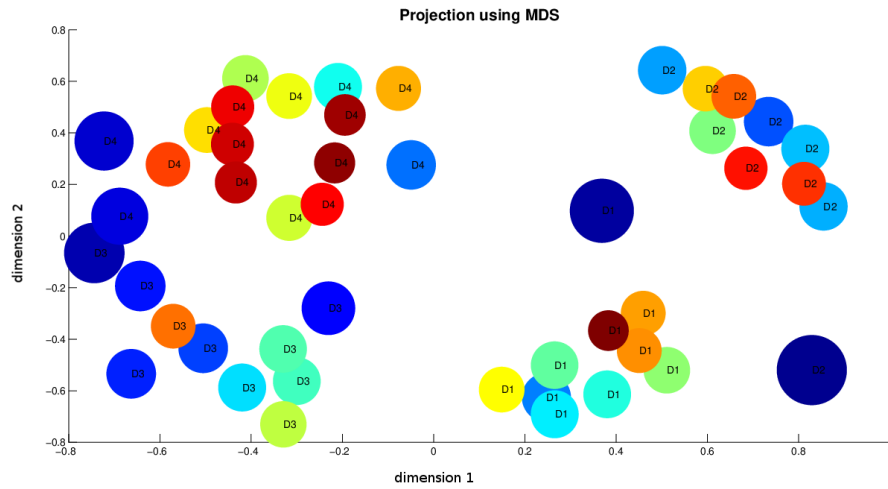


Figure 11.. MDS for the soybean data set. Four groups are presented, indexed by D1, D2, D3 and D4.

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

559 clearly seen in Figures 12 and 13

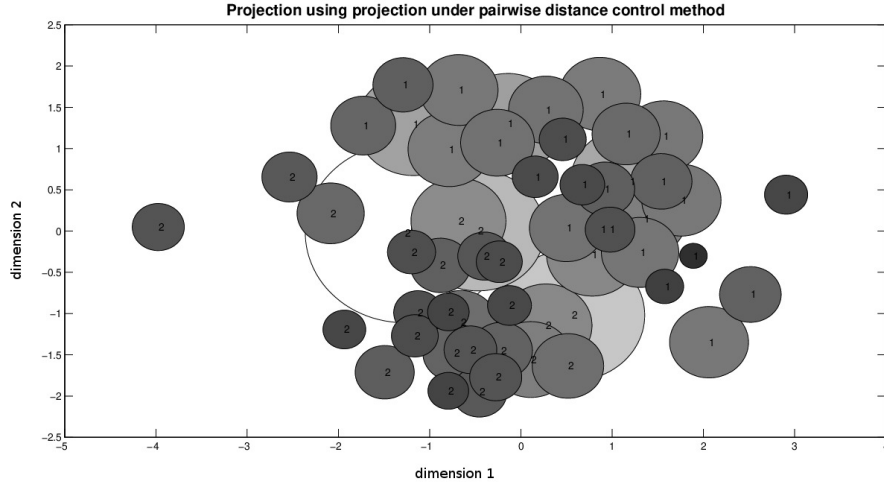
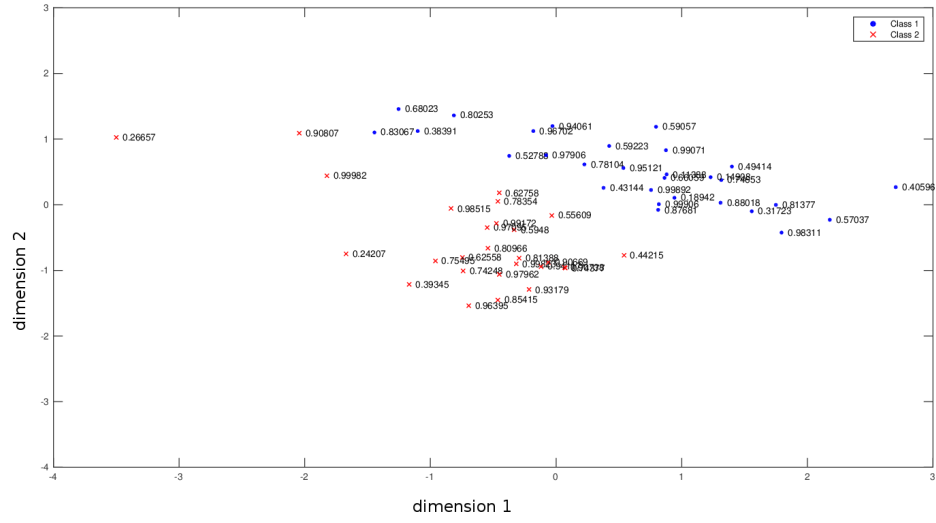


Figure 12.. Projection of the coffee data set using projection under pairwise distance control. Two clusters, indexed 1 and 2, indicate the Arabica and Robusta classes respectively.

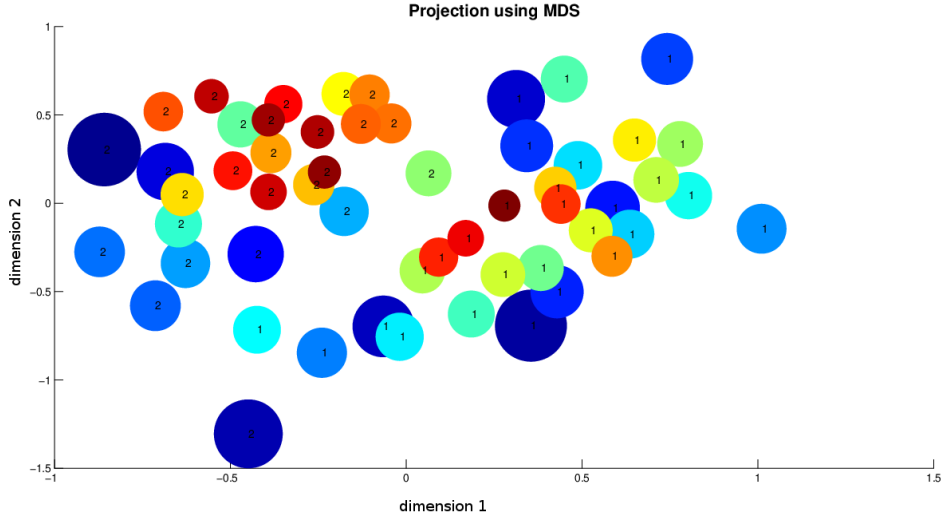
560 In Figure 12, we show that we have succeeded in differentiating the Arabica from
 561 Robusta coffee. These two classes are clearly presented, the first class indexed by
 562 number 1, corresponding to Arabica coffee, and the second one indexed by number 2,
 563 corresponding to Robusta coffee. These classes are not well separated in comparison
 564 with the results of quantitative data, since there are many intersections. Therefore,
 565 the representation of the points as circles and not as simple points provides more
 566 information about the real point classes and shows the points that are at risk of being
 567 misplaced in a particular class.

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

573 Additionally, Algorithm 2 was applied to these data sets. The trace plots of the
 574 optimization problem $\mathcal{P}_{r,x}$ are shown in Figure 14 after 5000 iterations. It is important
 575 to note that the value of the sum of radii $\sum_{i=1}^n r_i$ decreases rapidly in the first iterations
 576 and stays roughly constant after 1000 iterations for the different data sets, with the



(a)



(b)

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

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.

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

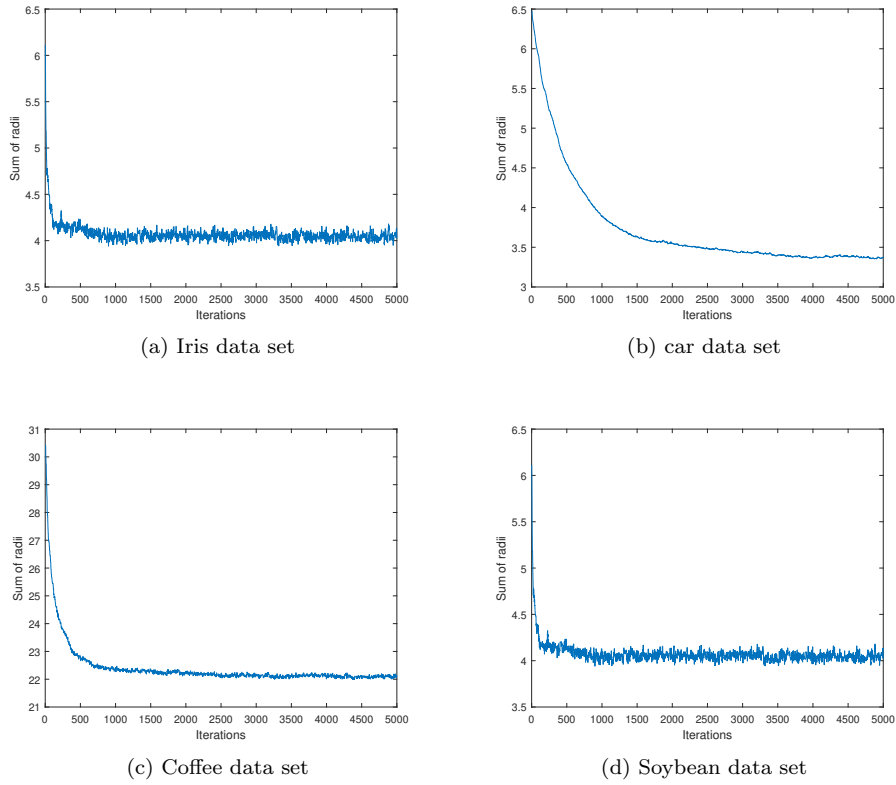


Figure 14.. Trace plots of Metropolis Hastings for different data sets. The x-axis corresponds to the iteration number and the y-axis to the value of $\sum_{i=1}^n r_i$.

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 |

590 4.4.2. Dimensionality reduction results

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

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

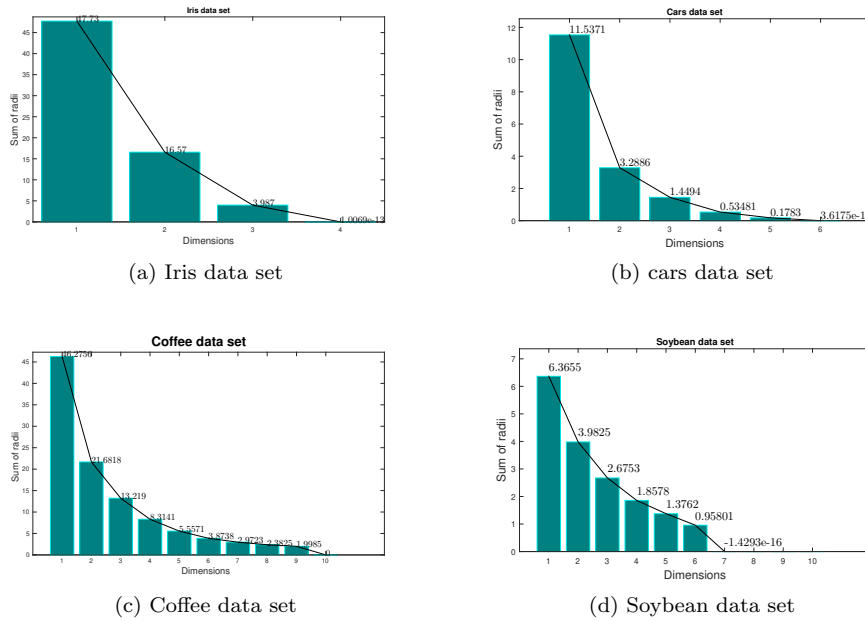


Figure 15.. Scree plots of $\sum_{i=1}^n r_i$ for different dimensions for the four data sets.

605 Regarding the quantitative data sets (Iris and car), if the main objective of the
606 user is to obtain a very good projection quality, then a choice of three components

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

613 A comparison of our results with the existing results shows a coherence between
614 them. For the Iris data set, Chiu (1996) and Liu and Setiono (1995) concluded
615 that the number of variables could be reduced to 2 as the petal length and petal
616 width variables are the most important variables from all the variables. For the car
617 data set, Saporta (2006) (Table 7.4.1 page 178) noticed that the conservation of two
618 dimensions led to the explanation of 88% of inertia, where the inertia term reflects
619 the importance of a component. Therefore, these results seem very similar to our
620 results, with the important decrease located between dimensions 1 and 2. The other
621 reductions are negligible for these two data sets. A selection of variables was studied on
622 time series coffee data set by Andrews and McNicholas (2014). Using several analysis
623 methods, the number of selected variables ranged between 2 and 13. This result is also
624 seen using our method, a number of reduced variables taken between 2 and 9 gives a
625 good projection. Regarding the soybean data set, Dela Cruz shows in his paper Dela
626 Cruz (2015) that the 35 attributes can be reduced to 15. With our method, we have
627 succeeded in reducing the attributes to 6 by having a very good projection of points.
628 Hence, the results presented confirm that the dimension nonlinearly can be reduced
629 while assessing a reasonable number of dimensions at the same time.

630 5. Conclusion

631 The purpose of this paper was to outline a new nonlinear projection method based
632 on a new local measure of projection quality. Of course, in some projection methods,
633 a local measure is given but this measure cannot be applied unless in cases of linear
634 projections, and even then it is not efficiency for graphical representation.
635 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.
- Andrews, J. L. and McNicholas, P. D. (2014). Variable Selection for Clustering and Classification. *Journal of Classification* 31:136-153.
- Bagnall, A., Davis, L., Hills, J., and Lines, J. (2012). Transformation Based Ensembles for Time Series Classification. *Proceedings of the 12th SIAM International Conference on Data Mining* 307–319.
- Berge, C., Froloff, N., Kalathur, RK., Maumy, M., Poch, O., Raffelsberger, W. and Wicker, N. (2010). Multidimensional fitting for multivariate data analysis. *Journal of Computational Biology* 17:723–732.
- Besse, P(1992). PCA stability and choice of dimensionality. *Statistics & Probability Letters* 13:405-410.
- Boggs, P. T. and Tolle, J. W. (1995). Sequential quadratic programming. *Acta Numer* 4:1–51.
- Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.) New York: Springer-Verlag.

- 666 Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity Measures for Categorical Data:
667 A Comparative Evaluation. *Proceedings of the SIAM International Conference on Data*
668 *Mining*.
- 669 Briandet, R., Kemsley, E. K., and Wilson, R. H. (1996). Discrimination of arabica and robusta
670 in instant coffee by fourier transform infrared spectroscopy and chemometrics. *Journal of*
671 *Agricultural and Food Chemistry* 44(1):170–174.
- 672 Chan, W. W-Y. (2006). A survey on multivariate data visualization in Science and technology.
673 *Department of Computer Science and Engineering Hong Kong, University of Science and*
674 *Technology* 8(6):1–29.
- 675 Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. (2015). *The*
676 *UCR Time Series Classification Archive*. www.cs.ucr.edu/~eamonn/time_series_data/.
- 677 Cheung, L. W. (2012). Classification approaches for microarray gene expression data analysis.
678 *Methods in Molecular Biology* 802:73–85.
- 679 Chinchilli, V. M. and Sen, P. K. (1987). Multivariate Data Analysis: Its Methods. *Chemomet-*
680 *rics and Intelligent Laboratory Systems* 2:29–36.
- 681 Cleveland, W. S. and McGill, M. E. (1988). Dynamic Graphics for Statistics. *Wadsworth and*
682 *Brooks/Cole*, Pacific Grove, Canada.
- 683 Chiu, S. L. (1996). Method and Software for Extracting Fuzzy Classification Rules by Sub-
684 tractive Clustering. *Proceedings of North American Fuzzy Information Processing Society*
685 *Conference*.
- 686 Cristofari, A., De Santis, M., Lucidi, S. and Rinaldi, F. (2007). A Two-Stage Active-Set Algo-
687 rithm for Bound-Constrained Optimization. *J. Optim. Theory Appl.* 172(2):369–401.
- 688 Conn, De A. R., Gould, N. I. M. and Toint, Ph. L. (2000). Trust Region Methods, SIAM.
- 689 Dela Cruz, G. B. (2015). Comparative Study of Data Mining Classification Techniques over
690 Soybean Disease by Implementing PCA-GA. *International Journal of Engineering Research*
691 *and General Science* 3(5):6–11.
- 692 Dempster, A. P. (1971). An overview of multivariate data analysis. *Journal of Multivariate*
693 *Analysis* 1(3):316–346.
- 694 Du Toit, S. H. C., Steyn, A. G. W., and Stumpf, R. H. (1986). Graphical Exploratory Data
695 Analysis, Chapter 5: Cluster Analysis, *Springer-Verlag* p. 79.
- 696 Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Collier,
697 H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999).
698 Molecular classification of cancer: class discovery and class prediction by gene expression

699 monitoring. *Science* 286:531–537.

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

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

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

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

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

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

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

713 Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*
714 29: 115-129.

715 Kruskal, J. B., Wish, M. (1978). Multidimensional Scaling. Series: Quantitative applications
716 in social Sciences. Sage university.

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

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

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

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

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

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

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

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

730 Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling withan unknown
731 distance function. *Psychometrika* 27: 125-139 & 219-246.

- 732 Schölkopf, B. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural*
733 *Computation* 10(5): 1299–1319.
- 734 Stepp, R. (1984). Conjunctive conceptual clustering. Doctoral dissertation, department of com-
735 puter science, university of Illinois, Urbana-Champaign, IL.
- 736 Svante, W., C. Albano, W. J. DunnIII, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E.
737 Johansson, W. Lindberg , M. Sjostrom. (1984). Multivariate Data Analysis in Chemistry.
738 *Chemometrics* 138:17–95.
- 739 Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for
740 nonlinear dimensionality reduction. *science*, 290(5500):2319-2323.
- 741 Togerson, W. S. (1958). Theory and methods of scaling, New York: Wiley.
- 742 Van der Hilst, R., de Hoop, M., Wang, P., Shim, S.-H., Ma, P. and Tenorio, L. (2007).
743 Seismo-stratigraphy and thermal structure of earth’s core-mantle boundary region. *Science*
744 315:1813–1817.
- 745 Wong, E. (2011). Active-Set Methods for Quadratic Programming. Ph.D. thesis, university of
746 California, San Diego.

747

Appendix

748

Proof of proposition 3.1

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

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

749

and the other inequalities involving i are not saturated. Then, the corresponding solu-

750

tion can be improved by moving x_i along the line $x_j - x_i$ in order to decrease r_i and

751

$|d_{ij} - \|x_i - x_j\||$.

752

Proof. The above condition means that x_i is rewritten as $x_i + a(x_j - x_i)$ with $a \in \mathbb{R}$

753

and we look for a such that $|d_{ij} - \|x_i + a(x_j - x_i) - x_j\|| < r_i + r_j$. In particular $a \leq 0$

754

if $d_{ij} - \|x_i - x_j\| \geq 0$ and is otherwise > 0 . Let us now consider the other inequalities

755

corresponding to index pairs (i, k) with $k \neq j$. For each of them, $\exists a \in [a'_k, a''_k]$ with

756

$a'_k < 0$ and $a''_k > 0$ such that

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

757

as these constraints are unsaturated. Finally, taking a different from 0 in $[a', a'']$ with

758

$a' = \max_k a'_k$ and $a'' = \min_k a''_k$, all constraints involving i get unsaturated so that r_i

759

can be decreased, thereby decreasing the objective function. Depending on whether a

760

must be negative or positive, we take $a = a'$ or $a = a''$ respectively.

761

□

762 **Proof of proposition 3.2**

763 Let $r_1, \dots, r_n; x_1, \dots, x_n$ be a feasible solution of $\mathcal{P}_{r,x}$, if $\exists a$ such that $\eta(a) < \sum_{i=1}^n r_i$
 764 with $\eta(a) = \sum_{1 \leq i < j \leq n} |d_{ij} - a\|x_i - x_j|||$, then $\exists \tilde{r}_1, \dots, \tilde{r}_n$ a solution of $\mathcal{P}_{r,a}$ such that
 765 $\sum_{i=1}^n \tilde{r}_i < \sum_{i=1}^n r_i$.

766 **Proof.** Let us consider $r_1, \dots, r_n; x_1, \dots, x_n$ a feasible solution of problem $\mathcal{P}_{r,x}$ and
 767 $a, \tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_n$ a solution of $\mathcal{P}_{r,a}$ where a is kept constant. For the solution of $\mathcal{P}_{r,a}$, for
 768 each point i , we have a certain saturated constraint associated to point k denoted by
 769 $C_{ik(i)}$, otherwise we can easily saturate it using proposition 3.1. Thus, we have:

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

770 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 $i =$
 771 $1, \dots, n$, we obtain:

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

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

773 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 $\mathcal{P}_{r,a}$
 774 such that $\sum_{i=1}^n \tilde{r}_i < \sum_{i=1}^n r_i$. □