Multidimensional Projection with Radial Basis Function and Control Points Selection

Elisa Amorim[1*] Emilio Vital Brazil[1*] Luis Gustavo Nonato[2†] Faramarz Samavati[1*]
Mario Costa Sousa[1*]
[1]University of Calgary - Canada, [2]University of São Paulo - Brazil

ABSTRACT

Multidimensional projection techniques provide an appealing approach for multivariate data analysis, for their ability to translate high-dimensional data into a low-dimensional representation that preserves neighborhood information. In recent years, pushed by the ever increasing data complexity in many areas, numerous advances in such techniques have been observed, primarily in terms of computational efficiency and support for interactive applications. Both these achievements were made possible due to the introduction of the concept of control points, which are used in many different multidimensional projection techniques. However, little attention has been drawn towards the process of control points selection. In this work we propose a novel multidimensional projection technique based on radial basis functions (RBF). Our method uses RBF to create a function that maps the data into a low-dimensional space by interpolating the previously calculated position of control points. We also present a built-in method for the control points selection based on "forward-selection" and "Orthogonal Least Squares" techniques. We demonstrate that the proposed selection process allows our technique to work with only a few control points while retaining the projection quality and avoiding redundant control points.

Keywords: High-Dimensional Data, Dimensionality Reduction, Multidimensional Projection, Interpolation with Radial Basis Function.

1 INTRODUCTION

Understanding the underlying structure of multidimensional data sets is a fundamental requirement in many scientific and real-world applications. Consequently, visualization tools have long become indispensable for multidimensional data analysis, as they present significant aspects of the data in a comprehensible manner to the end user [15]. In particular, multidimensional projection (MP) techniques have gained popularity for their ability to convey dissimilarity information in a straightforward manner.

The goal of multidimensional projection techniques is to find a low-dimensional representation (usually 2D) for the data, where original dissimilarities between pairs of instances are reflected as their Euclidean distance in the low-dimensional space. The resulted 2-dimensional projection can be interpreted by the user, who is able to visually infer classifications and recognize possible patterns in the data set. MP has been successfully used as a visual analysis tool in applications such as feature exploration in multivariate scalar fields [17], vector field visualization [9], text mining [7, 20], finances [12] and psychology [3], to name just a few.

Ever increasing data complexity has driven the development of more robust MP methods and algorithms that are capable of dealing with massive data sets efficiently. Consequently, MP has experienced various improvements, both in terms of mapping quality [16] and computational efficiency [14, 22, 18]. An important contribution to these advancements was the work of Pekalska et al. [24], who proposed the use of a subset of samples to speed up the projection process. Since Pekalska's work, several other techniques have used this concept. More recently, such a subset of samples, called *control points*, has had its functionality expanded to work as an interface between user input and projection results. The manipulation of such control points in the projection space has been suggested as a means for incorporating user-knowledge into the projection process, as can be seen in Joia et al. [18].

Even though the control points paradigm is gaining strength and being incorporated in most of the new techniques, the selection of instances that should be used as control points has not been thoroughly investigated. The set of control points has a direct impact in the quality of the final results. However, most techniques suggest a random or clustering-based approach for control points selection. The former alternative is naive and the latter, besides being computationally expensive, requires the user to specify the number of clusters used to divide the data set, which may not be an obvious choice. Ideally, one should have a technique that finds a good set of points that represents the original data set in such a way that improves the projection approximation without requiring many user-defined parameters.

In this paper we propose a novel multidimensional projection technique, built upon the *radial basis function* (RBF) theory. RBF presents a well-established mathematical formulation, which has been used in diverse approximation applications [4]. In general terms, RBF constructs a function that interpolates given sample points and their outputs. The interpolation function is formed by a linear combination of radial basis functions, which are real-valued functions whose value depend only on the distance from a point to the RBF center. In this work, we apply RBF to find an interpolation function that respects the low-dimensional position of previously projected control points, and use this function to approximate the projection of the remaining instances. This method provides an explicit mapping from high to low dimension, and allows one to incorporate new data in real time with little computational effort. The technique introduced by Pekalska et al. [24] is a particular case of RBF. With the proposed framework, we generalize this traditional method and improve it by reducing the number of required control points.

An advantage of the proposed RBF formulation is the existence of different works dedicated to center selection. One of these approaches is based on the solution of Orthogonal Least Squares problems to select a subset of samples that satisfactorily represents the data set [6]. We incorporated this technique into our Multidimensional Projection framework as a means to perform control points selection and to improve the final projection results. This approach automatically determines a good number of control points; thus, the user is not required to provide this important parameter. Furthermore, the proposed technique can produce good-quality results with a reduced number of control points, which improves efficiency as well as favors user interactivity [18].

*e-mail: {epdamori,evbrazil,samavati,smcosta}@ucalgary.ca
†email:gnonato@icmc.usp.br

Another advantage of the proposed technique is that it does not require the original data set to be embedded in a Cartesian space. This restriction was introduced in most of the recent MP methods, but is not part of traditional techniques, such as MDS and Sammon's mapping. Moreover, our method does not require the full dissimilarity matrix between all pairs of instances, only those pertaining to the control points. This makes the proposed technique competitive to the most recent MP methods in terms of performance.

We can summarize the main contributions of this work as follows:

- Radial Basis Function MP technique: A novel multidimensional projection technique based on Radial Basis Function interpolation theory. The proposed technique does not present the drawback of requiring a Cartesian representation of the data; it is computationally efficient and works well with a low number of control points.
- Selection of control points based on the Orthogonal Least Squares problem methodology: The decision of which control points to use is not determined randomly, but based on a deterministic algorithm that selects those instances that better explain the entire data set.

To the best of our knowledge, the proposed mechanism is the first to incorporate a quality measure for control points selection. In fact, measuring the quality of control points is a problem not properly tackled until now.

This paper is organized as follows: Section 2 presents an overview of MP techniques and contextualizes the advantages of our RBF approach; Section 3 details the mathematical formulation of our technique; Section 4 presents a detailed description of the use of Orthogonal Least Squares for the selection of control points. Section 5 contains the evaluation of the RBF technique and a comparison with other methods; we conclude and discuss about future work directions in Section 6.

2 Related work

Multidimensional projection has long been used for visual analysis of multidimensional data. In this section we present an overview of the main MP techniques. We discuss linear and nonlinear approaches for low-dimensional mapping in Section 2.1; and we outline the methods that make use of a subset of samples (control points) to speed the projection process in Section 2.2.

2.1 Linear and nonlinear techniques

The classification of linear and nonlinear accounts for the kind of transformation applied to the instances of the original data set. Linear mappings are designed to operate when the submanifold is embedded linearly, or almost linearly in the observation space [10]. However, they cannot capture nonlinear relationships between data instances, which are usually accomplished by nonlinear transformations. Some examples of linear projection methods are principal component analysis (PCA) [19] and classic multidimensional scaling [8].

Least-squares scaling methods, such as the Sammon's mapping [26], are examples of nonlinear mappings. Generally, nonlinear techniques attempt to minimize a function of the information loss caused by the projection. Such a function measures the error between dissimilarities in the original and projected spaces. Sammon's mapping, and many others that derive from it, applies a steepest-descent procedure to solve the optimization problem. One of the disadvantages of least-squares techniques is that gradient-based techniques do not guarantee convergence to the global minimum of the function. Consequently, the final projection layout may not be a good representation of the original data set. Roweis and Saul [25] proposed a method called Locally Linear Embedding (LLE) that uses local information to achieve an optimization without local minima. Other examples of nonlinear projection methods are Curvilinear Component Analysis (CCA) [13], which presents a variation on the loss function proposed by Sammon; Isomap [28], proposed by de Silva and Tenenbaum, that uses the geodesic distance information to compute dissimilarities; and Least Square Projection (LSP), introduced by Paulovich et al. [21].

Some recent methods propose the combination of linear and nonlinear transformations. For instance, Part-linear multidimensional projection (PLMP) [22] and LAMP [18] make use of a subset of samples initially positioned in the projection space through a nonlinear technique. The remaining instances are subjected to a linear transformation based on the final position of these samples. The aforementioned LLE [25] also combines both, linear and nonlinear approaches, by computing some weights and vectors linearly, but the overall process is nonlinear. The proposed technique based on RBF is a non-linear approximation that is able to approximate both linear and non-linear methods.

2.2 Control Points

In the context of multidimensional projection, control points are a subset of the original dataset that are positioned in the low-dimensional space in a preprocessing step with a global technique, such as MDS, Sammon's mapping, PCA, etc. The information from high- to low-dimensions calculated for this subset of points is used to approximate the final position of the remaining instances of the data set. The primary motivation behind the use of control points is to make the projection technique more efficient in terms of computational time, since global methods can be computationally expensive.

Pekalska et al. [24] were the first to use control points in multidimensional projection. The goal of their work was solely to speed up the Sammon's mapping algorithm while still preserving the projection quality. With this in mind, they proposed the application of Sammon's mapping only to a subset of points, followed by the computation of a linear transformation that would respect the high to low dimensional mapping of this subset. The remaining points are later projected using this linear transformation. One of the drawbacks of this technique is that, in order to provide a good approximation, the number of control points needs to be large. In fact, the authors recommend that 50% of the total number of instances of the data set should be used as control points. Using the technique proposed in this paper, we are able to reduce the number of control points, and still get good-quality mappings.

Since Pekalska *et al.*'s work, a variety of control-points-based techniques have been proposed: L-Isomap [10], L-MDS [11], LSP [21], PLMP [22], LAMP [18] and PLP [23] are a few examples of such methods. Even though control points are central to the aforementioned techniques, most of these works do not approach the problem of how to effectively select control points. Pekalska, L-Isomap, L-MDS, PLMP and LAMP suggest to select control points randomly, while LSP and PLP make use of clustering techniques to divide the data set into regions, and select one or more representative instances of each region as control points. The random approach is far from ideal, as the final set of control points may not be a good representation of the entire data set and may result in poor-quality mappings. On the other hand, the selection through clustering can become an issue in terms of computational time. Moreover, both approaches present the drawback of requiring the user to determine the number of control points, which may not be an obvious parameter to choose.

Besides these limitations, with the exception of LAMP, these techniques may require a large number of control points to maintain the projection quality. For example, PLP, PLMP and LSP suggest $\sqrt{n}$ control points, where n is the number of instances in the origi-

nal data set. A large number of control points create a less efficient high to low-dimensional mapping, and is not ideal for applications with user intervention, as pointed out by Joia et al. [18].

The proposed RBF projection also makes use of control points, however we present a built-in approach to perform control points selection efficiently. This approach is based on the "Regularized Orthogonal Least Squares" problem and performs a forward-selection of instances in which, at each iteration, the most suitable instance is incorporated into the control points' set. We show that the proposed control points selection approach is able to improve the projection quality, using a reduced number of samples. Also, the proposed technique present a general mathematical formulation that does not assume the original data to be embedded in a Cartesian space, an interesting feature not present in recent MP techniques such as LAMP, PLP, LSP and PLMP.

Section 3 presents a detailed description of the proposed technique and the control points' selection approach is detailed in Section 4.

3 Multidimensional Projection with Radial Basis Function Interpolation

Radial basis function (RBF) is a popular technique to approximate multivariate functions [4]. In general terms, given data samples $x_i \in \mathbb{R}^m$ and function values $y_i = f(x_i) \in \mathbb{R}, i = 1 \ldots n$, an approximant $s : \mathbb{R}^m \to \mathbb{R}$ is sought, in such a way that s interpolates the function f between the data samples. The approximant $s(x)$ is formed by a linear combination of radial basis functions $\phi(x)$ with centers in x_i. Figure 1 illustrates a one dimensional RBF interpolation using five data samples and Gaussian radial basis function.

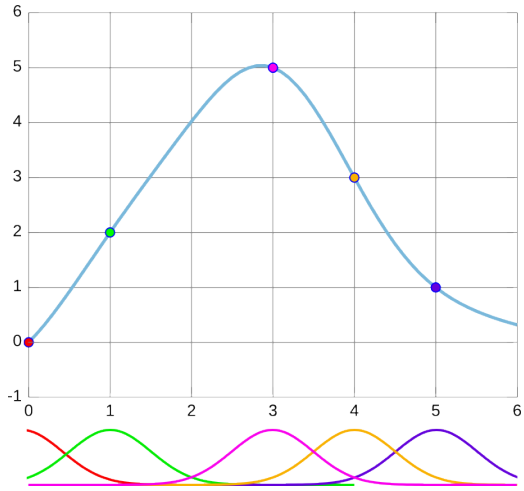

Figure 1: Radial Basis Function interpolation with five data samples $X = \{0,1,3,4,5\}$ and function values $Y = \{0,2,5,3,1\}$. The data samples are represented as colored points. The blue curve is the function obtained with RBF interpolating between the data samples. Below the graph, the Gaussian radial basis functions $\phi(r) = e^{-(\varepsilon r)^2}$ are represented, where r is the distance between a point and the data sample, with $\varepsilon^2 = 0.5$. The colors of the curves make the correspondence to the data samples.

The proposed MP method uses RBF to create a function s that maps high-dimensional data into a low-dimensional space. Given a subset of samples (control points) and their low-dimensional positions, RBF is used to create a function to map the remaining samples into the projection space. In the next Section we present a detailed description of the technique, and a discussion about different radial basis functions is found in Section 3.2.

3.1 Mathematical Formulation

Consider a data set $X \subset \mathbb{R}^m$ with n elements. Let $X_S = \{x_1, x_2, ..., x_k\} \subset X$, $k \ll n$, be a set of control points, for which the corresponding low-dimensional position $Y_S = \{y_1, y_2, \ldots, y_k\} \subset \mathbb{R}^d, d < m$ is calculated a priori using a force-based multidimensional projection technique [27] (from now on we will consider $d = 2$). The goal of the proposed RBF projection is to find a function $s : \mathbb{R}^m \to \mathbb{R}^2$ of the form

$$s(x) = \sum_{x_i \in X_S} \lambda_i \phi(\|x - x_i\|), \tag{1}$$

in such a way that s interpolates the position of each control point, i.e., $s(x_i) = y_i, i = 1, \ldots k$. Function $\phi : \mathbb{R}_+ \to \mathbb{R}$ is called *kernel* of the RBF, and its definition, together with the set of control points, dictate the final approximant s. There are numerous functions that can be used as a kernel, and we discuss this topic in Section 3.2.

The real-value coefficients λ_i need to be calculated as to satisfy the interpolation condition, giving rise to a linear system with k equations $s(x_i) = y_i, i = 1 \ldots k$. The system can be written in matrix form as

$$\Phi\lambda = y, \tag{2}$$

where Φ is the *interpolation matrix* with dimensions $k \times k$, with $\Phi_{ij} = \phi(\|x_i - x_j\|)$; the right-hand side of the system y and the unknowns λ are 2-columned vectors, each column accounting for one of the final dimension of the output. Let $\phi_{i,j} = \phi(\|x_i - x_j\|)$, $\lambda_i = (\lambda_i^1, \lambda_i^2)$, $y_i = (y_i^1, y_i^2)$ and Equation 2 can be written as

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1k} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{k1} & \phi_{k2} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}. \tag{3}$$

Once the coefficients λ's are calculated, the function s is fully determined and can be used to approximate the remaining instances of the data set. The proposed multidimensional projection technique is summarized in Figure 2.

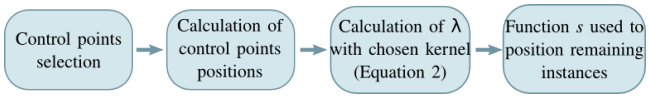

Figure 2: Summary of the flow of RBF projection technique. Control points selection and calculation of positions are the first steps, followed by determination of λ's and application of function $s(x)$ to position the remaining instances.

Note that function ϕ takes as argument the distance between a point in the domain and a control point. Even though the most commonly used distance metric is the Euclidean, in practice one could use different dissimilarity measurements not necessarily defined in a Cartesian space. This renders the proposed techniques more flexible than recent multidimensional projection methods, which require the original data set to be embedded in a Cartesian space. In the next Section we present more details about the radial basis functions ϕ.

3.2 RBF Kernels

We observed from the previous section that the solution of the radial basis function interpolation problem reduces to the solution of a linear system $\Phi\lambda = y$. Thus, it is important that matrix Φ is non-singular, in order to produce a uniquely determined system. Therefore, the kernel function ϕ needs to be chosen carefully, since it

determines the entries Φ_{ij} of the interpolation matrix Φ. There are kernels that guarantee that Φ will be invertible, with a minor assumption over the set of control points, which is that it can only contain unique samples.

In this work we experimented with three classic RBF kernels that guarantee matrix Φ to be non-singular: *Gaussian*, *Multiquadrics* and *Inverse Multiquadrics*. The definition of these kernels are given in Table 1, and Figure 3 presents the graph of these three functions in one-dimension with $\varepsilon = 1$.

Name	Definition of $\phi(r)$
Gaussian	$e^{-(\varepsilon r)^2}$
Multiquadrics	$\sqrt{c^2+(\varepsilon r)^2}$
Inverse Multiquadrics	$\frac{1}{\sqrt{c^2+(\varepsilon r)^2}}$

Table 1: Commonly used functions in the Radial Basis Function interpolation problem.

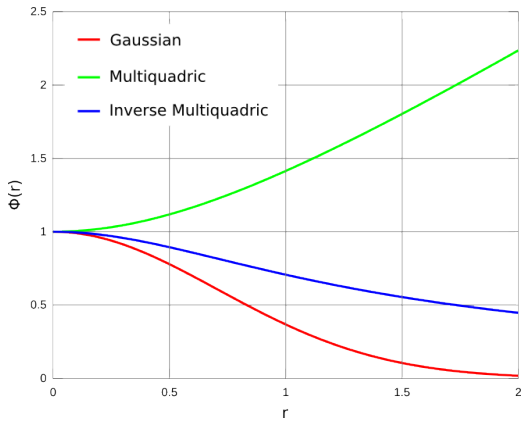

Figure 3: Gaussian, multiquadrics and inverse multiquadrics functions in one dimension. In this example all of the functions have $\varepsilon = 1$.

It is interesting to note that Pekalska's technique is a special case of RBF that uses multiquadrics kernel with $c = 0$ and $\varepsilon = 1$ (thus, $\phi(r) = r$), which we will call kernel *Norm*. However, generally this kernel does not give good projection approximations when used with a relatively small number of control points. This fact can be observed in Figure 4, where we present a comparison between the projection quality using the Norm, Multiquadrics with $c = 1$ and $\varepsilon = 1$ and Gaussian kernels. The projection quality is measured by a popular quality metric called *stress function*. The stress function indicates how well the original dissimilarities are preserved in the low dimensional space. It is given by

$$stress = \frac{\sum_{i=0}^{n}\sum_{j=i}^{n}(\delta_{ij} - d_{ij})^2}{\sum_{i=0}^{n}\sum_{j=i}^{n}\delta_{ij}^2}, \qquad (4)$$

where δ_{ij} is the original dissimilarity between instances i and j and d_{ij} are the distance of these same instances in the low dimensional representation. The smaller the stress, the better the projection's results.

Figures 4(a) and 4(b) present the behavior of the kernels for three data sets (*Pima-indians*, *WDBC* and *Wine Quality*), with 10 and 100 randomly selected control points, respectively. In our experiments, 100 runs were executed, each with a different set of control points,

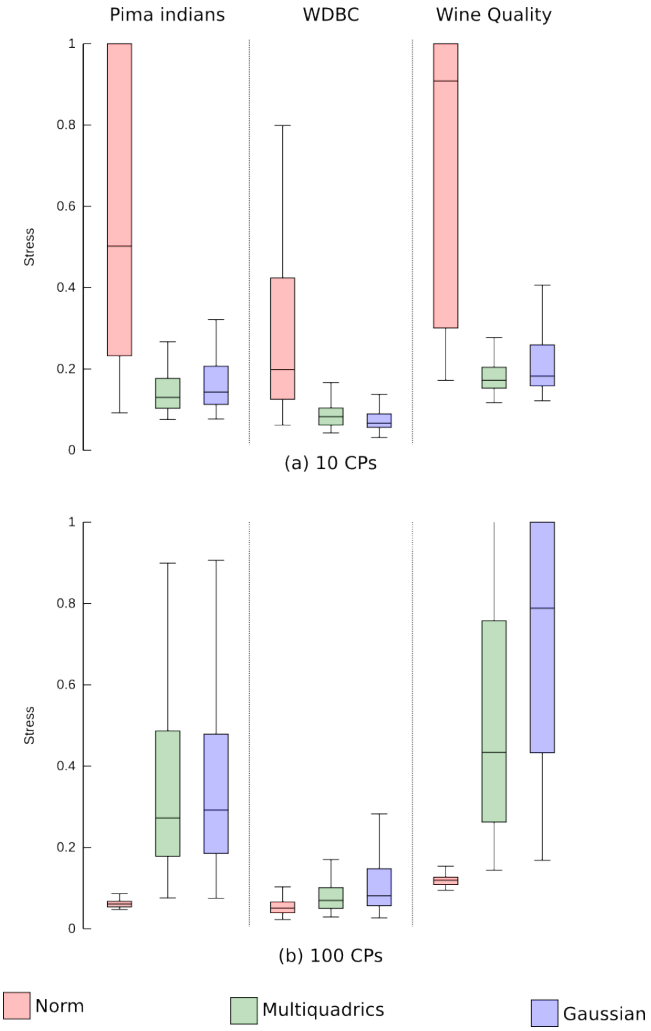

Figure 4: Boxplots with a comparison of projection quality (stress) using Norm (Pekalska), Multiquadrics and Gaussian kernels, for three data sets (Pima, WDBC and Wine Quality); (a) 10 and (b) 100 randomly selected control points. Each experiment was executed 100 times, with a different set of control points for each execution.

for each experiment. These examples indicate that the Norm kernel gives better results when a larger number of control points is used, while the Gaussian and Multiquadrics kernels work well with fewer control points, but give poor results when the number of control points is increased.

These experiments also suggest that, increasing the number of control points does not necessarily improves the projection quality, as one could expect. What happens is that, when a random control points' selection strategy is employed, chances are that similar, or even identical instances are used as control points, what causes matrix ill-conditioning or singularity. This is one of the motivations for performing a conscious selection of control points.

In the next Section we present an automatic technique for control points selection that can be easily built-in with our proposed multidimensional projection workflow and automatically avoid ill-conditioning. Besides, we also show that applying the proposed selection we are able to considerably improve the projection results by reducing the stress (Equation (4)) using just a few control points.

4 Control Points Selection Through Regularized Orthogonal Least Squares

As shown in the previous section, the set of control points is an essential part of the RBF technique and it sure can have a great impact on the quality of the final projection. Ideally, the set of control points should represent well the entire data set domain and still be reasonably sized, creating an RBF with low redundancy. In this work, we propose to employ a method based on Orthogonal Least Squares (OLS), introduced by Chen et al. in [6] for center selection in RBF. To understand how OLS works for control points selection, it is important to view RBF as a linear regression model. Assume we have N control points *candidates* $\{x_i, y_i\}_{i=1}^{N}$, where y_i is the output corresponding to control point x_i. If all x_i are used as control points, Equation (1) can be rewritten as:

$$s(x_t) = \sum_{i=1}^{N} \lambda_i \phi(\|x_t - x_i\|), 1 \leq t \leq N. \tag{5}$$

Let $\phi_i(t) = \phi(\|x_t - x_i\|)$, we can express the desired output y_t as

$$y_t = \sum_{i=1}^{N} \lambda_i \phi_i(t) + e_t, 1 \leq t \leq N, \tag{6}$$

where $e(t)$ is the error between the desired output y_t and the approximated output $s(x_t)$, i.e., $e(t) = y_t - s(x_t)$. (Of course $e(t)$ will be zero when all candidates are used as control points, but the goal of the method is to reduce this set). Finally, we can write Equation (6) in matrix form as

$$\mathbf{y} = \Phi\lambda + \mathbf{e}, \tag{7}$$

where $\mathbf{y} = [y_1 \dots y_N]^T$, $\Phi = [\phi_1 \dots \phi_N]$, $\phi_i = [\phi_i(1) \dots \phi_i(N)]^T$, $\lambda = [\lambda_1 \dots \lambda_N]$ and $\mathbf{e} = [e_1 \dots e_N]$.

Equation (7) has the form of a linear regression model and the vectors ϕ_i can be referred to as regressors. Thus, the question of how to select control points can be translated into the problem of selecting significant regressors. The adopted technique is based on a "forward selection", i.e., the process starts with an empty set of regressors and one regressor from the set of candidates is selected at a time. Each selection is made in such a way to maximally decrease the squared error $\mathbf{e}^T\mathbf{e}$.

Since the regressors are generally correlated, it is not clear how to measure their individual contributions to the error decrement. Applying the concept of the OLS method, which transforms the set of ϕ_i into a set of orthogonal basis vectors, it is possible to "isolate" the regressors and calculated their individual contributions. The regression matrix Φ can be decomposed as

$$\Phi = WA, \tag{8}$$

where A is an upper-triangular matrix with diagonal 1 and $W = [w_1 \dots w_N]$ with orthogonal columns that satisfy $w_i^T w_j = 0$, if $i \neq j$. The model (7) can be rewritten as

$$\mathbf{y} = Wg + \mathbf{e} \tag{9}$$

with $A\lambda = g$.

However, as discussed in [5], the minimization of only the squared error $\mathbf{e}^T\mathbf{e}$ is prone to overfitting, i.e., even though the produced approximant may interpolate the control points candidates, it may not be good to describe the overall behavior of the target function. To prevent this problem, a regularization term penalizing large λ values is added to the error. Observe that, since $A\lambda = g$, penalizing λ is equivalent to penalizing g; thus, the final formulation for the error we aim to minimize is:

$$\mathbf{e}^T\mathbf{e} + \beta g^T g, \tag{10}$$

where $\beta \geq 0$ is the regularization parameter. This error formulation renames the technique to *regularized* Orthogonal Least Squares (ROLS). Equation (10) can be rewritten as

$$\mathbf{e}^T\mathbf{e} + \beta g^T g = \mathbf{y}^T\mathbf{y} - \sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2. \tag{11}$$

Dividing (11) by $\mathbf{y}^T\mathbf{y}$ we have

$$\frac{(\mathbf{e}^T\mathbf{e} + \beta g^T g)}{\mathbf{y}^T\mathbf{y}} = 1 - \frac{\sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2}{\mathbf{y}^T\mathbf{y}}, \tag{12}$$

and the regularized error reduction ration due to w_i is defined as

$$[rerr]_i = \frac{\sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2}{\mathbf{y}^T\mathbf{y}}. \tag{13}$$

At each step of the selection, the control point x_i associated with vector w_i and maximum *rerr* is included in the control points' set. We also calculate the stress, given by Equation (4), of the remainding candidates. It is clear that using all candidates as control points reduces the error *rerr* (13) at most, but not necessarily the stress. The goal is to select a limited amount of points that better explains the data set and potentially reduces the stress. This is achieved by introducing stop criterion in the selection process: (1) Akaike-type criteria reaches a minimum, as suggested in [6] and (2) a maximum number of control points is reached. At the end of the selection process, the iteration with minimum stress is identified. As a final step, we seek a trade off between the minimum stress and the reduced number of control points, by finding an iteration with less control points than the one with minimum stress, but with stress slightly higher (in our experiments, between 0 and 5% higher proved to be a good range).

To prevent ill-conditioning, a simple check can be built into the procedure. The relation $w_i^T w_i = 0$ implies that ϕ_i is a linear combination of the previously selected control points. Thus, if $w_i^T w_i$ is less than a preset threshold γ, ϕ_i will not be selected as a control point.

The ROLS technique for control points selection is summarized in Algorithm 1. Note that the orthogonalization of matrix Φ is done step by step until a stop criterion is met and the selection process is terminated. The orthogonalization process is acquired through the *Modified Gram-Schmidt* algorithm and a detailed description of the algorithm can be found in [5].

The first step in the proposed process is to randomly select N candidates for control points and use a force-based technique to calculate their low-dimensional positioning. Of course this creates an extra overhead in computational time, however, as soon as the number of control points is reduced, the RBF becomes extremely simple. Also, the projection quality measured by the stress (Equation (4)) presents significant improvements using this careful selection process. In fact, Figure 5 experiments with three data sets, Page Blocks, Ionosphere and Yeast, with RBF using a Multiquadrics kernel ($c = 1$ and $\varepsilon = 30$). The green boxplots present the stress for random selection of control points, while the pink boxplots present the results for ROLS selection. Each experiment was executed 100 times.

The ROLS parameters used are: 30 maximum number of final control points and $\gamma = 1.e-5$, selected heuristically. The average number of control points selected is shown in the row *#FCP*. Observe that the stress achieved with ROLS selection with a few control points (less than 30) is lower than the ones achieved with a larger number of randomly selected control points. Another interesting fact to observe is that, increasing the number of candidates N, the chances of selecting more meaningful control points are increased and the stress is improved, but there is a tradeoff between

Algorithm 1 Control points selection with ROLS

Given $X_c = x_1, \ldots x_N$ (set of control points candidates);
Given $Y_c = y_1, \ldots y_N$ (candidates' position in the projection space);
Construct matrix Φ
$W = \Phi$
$X_0 = \varnothing$
$it = 1$
while Stop-criteria not met **do**
  for each candidate i where $w_i^T w_i > \gamma$ **do**
    Calculate $[rerr]_i(w_i)$ (Equation (13))
  end for
  $w_k = \arg\max([rerr]_i)$
  $X_{it} = X_{it-1} \cup \{x_k\}$ (Select k as control point)
  Remove w_k from W
  Calculate stress for instances in W
  for each $w_i \in W$ **do**
    $w_i =$ Orthogonalization of w_i with respect to w_k
  end for
  $it = it + 1$
end while
Find iteration it with minimum stress (s_{min})
$X = X_i$, with i the iteration with less CPs and $stress < 1.05 * s_{min}$
Return X

projection quality and computational time. In the next Section we present an evaluation of the proposed RBF projection and comparison with other techniques.

5 Evaluation of Technique

The experiments presented in this section were executed in a 2.80 GHz Intel Core i7 CPU 860 with 12 GB of RAM. We compare our technique to four other methods, namely LAMP [18], PLMP [22], Pekalska [24] and Fastmap [14]. The first three techniques are control-points-based, LAMP being the only one able to handle a reduced amount of control points. All of these techniques propose a random strategy for control points selection. The Fastmap technique, in turn, is known for its computational efficiency.

For the RBF setup in these experiments we adopted the Multiquadrics kernel, with $c = \varepsilon = 1$. For the control points selection process we used the following parameters: number of candidates $N = 150$, maximum number of control points 30 and $\gamma = 1.e-5$ to avoid matrix ill-conditioning. The data sets used in the experiments are described in Table 2.

Data set	# Instances	# Dimensions
Ionosphere	350	35
Pima indians diabetes	768	8
Yeast	1152	8
WDBC	569	30
Segmentation	2085	16
Wine Quality	3961	11
Page blocks	5405	10
Letter Recognition	18667	16
Shuttle	42365	8

Table 2: Data sets used in the experiments, downloaded from the UCI Machine Learning Repository [2].

For each data set 100 experiments were executed, to account for the variance introduced in the results due to the random factors of the techniques (except for Fastmap, which is a deterministic approach). We compared the methods in terms of mapping quality (stress - Equation (4)) and execution time. The results are shown in Figure 6 and we discuss the experiments in Section 5.1.

5.1 Discussion

Figure 6(a) presents the comparison of projection quality between the RBF approach and the other four techniques. The y-axis was

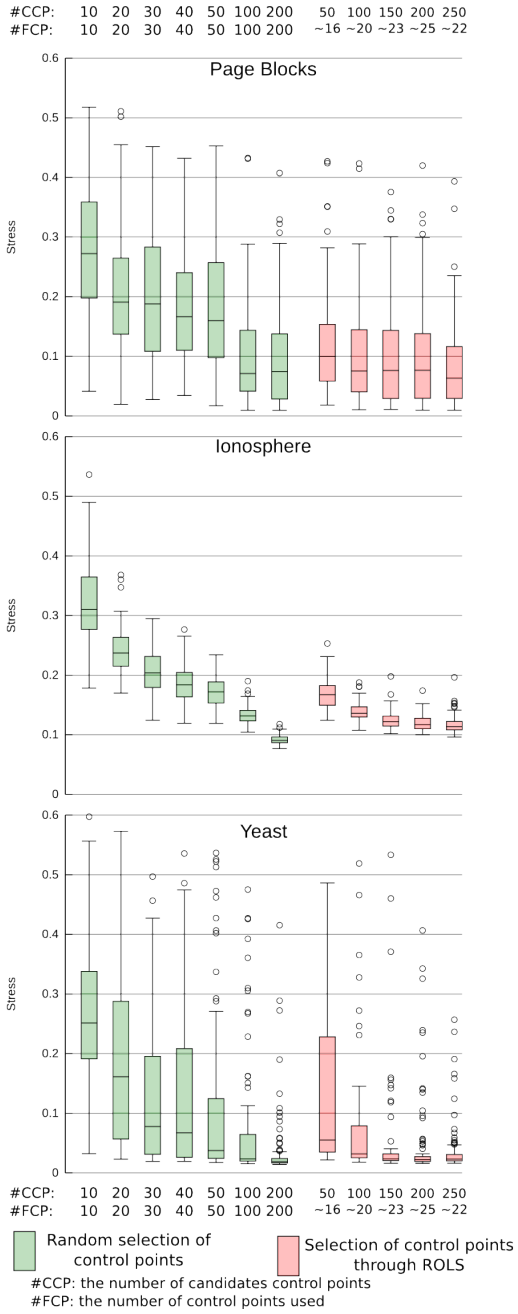

Figure 5: Impact of control points' selection with ROLS in projection stress, for three data sets. In ROLS results, *#FCP* indicates the average number of control points selected in the experiment.

truncated for $stress = 0.6$. We can see that the stress achieved by our method is very low and outperforms all the other methods, with the exception of Pekalska, which present stress slightly smaller. Pekalska, however, needs to use a large number of control points to produce good-quality results, while our method used a maximum of 30 control points in every test case.

Figure 6(b), in turn, presents the boxplots with the time variance for each technique. The results indicate that the overhead created by the procedure for control points' selection gives a small disadvantage to RBF compared with LAMP, PLMP and Fastmap. However, we note that the execution time was around 1 second, which is small in practical terms. Moreover, the created overhead is only caused by the control points' selection process, which depends on

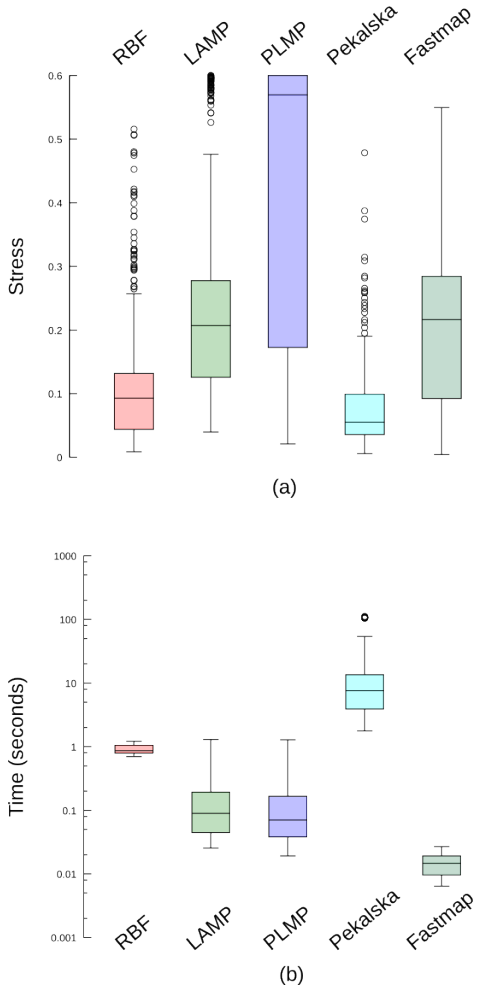

Figure 6: Stress and time comparison for RBF, LAMP, PLMP, Pekalska and Fastmap.

the number of candidates N chosen by the user. Our experiments indicate that a fixed number of 150, independent of the size of the data set, gives satisfactory results. Accordingly, this only creates a lower bound in terms of computational time, since the final RBF process with only a few control points is extremely fast.

Regarding ROLS for control points selection, we want to point out that the results achieved with this technique were very satisfactory. We applied this process to all 9 data sets presented in Table 2 (a reduced number of results was presented in Figure 5) and the stress produced by less than 30 control points was always equal or better than the ones produced by 50 randomly selected control points. Figure 7 presents an illustration of how the ROLS-based control points selection work in the *Mammals* dataset, which contains data that characterizes dogs, cats, horses and giraffes, forming four well-separated clusters [1]. Figures 7(a) and 7(b) present 150 randomly selected and 15 ROLS-selected control points, respectively. We can observe that the method automatically selects representatives of each cluster, maintaining the general behavior of the projection (results in the top row).

We intend to fully explore the usability of a reduced number of control points in an interactive application. As a preliminary result, we created a simple application where the user manipulates control points in order to better visualize samples that are similar to a selected *pivot* sample. Figure 8 and 9 illustrate such an application. Figure 8-(a) presents the initial projection configuration, achieved automatically. Figure 8-(b) presents the final configuration after

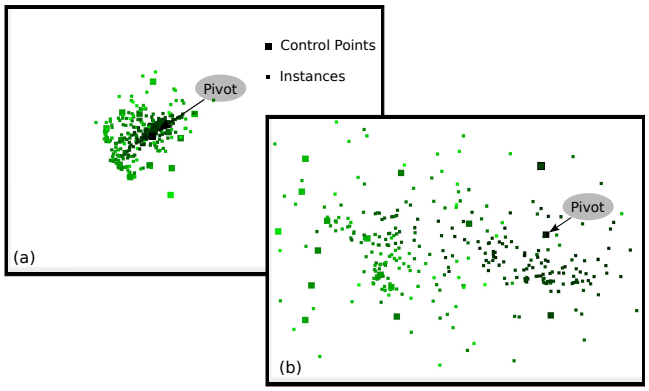

Figure 8: Example of control points manipulation with the Ionosphere data set, with the goal of unveiling samples close to a pivot. Through the manipulation of control points the user is able to examine a subset of the data (b) observed to be cluttered in the original projection (a). The color represents the distance value of the sample to a pivot, black being the most similar and light green the most dissimilar.

user intervention through the manipulation of control points. The black-to-green scale indicates how similar the sample is to a pivot, black being the most similar and light green the most dissimilar, and is used as a visual cue to aid the user to perform the separation task. We observe that after control point manipulation the data becomes more accessible to the user and the dark samples turn out to be easier to visualize. Figure 9 also contains an example of control points manipulation with the Segmentation data set divided into 7 clusters (represented by different colors). Starting from an automatic configuration of control points, a projection layout is generated (Fig. 9-(a)). The user is then able to interfere in the projection layout by changing the control points positions (Fig. 9-(b)). We can observe that the resulted projection allows the user to better explore instances that were originally hidden, for example the ones that belong to the red cluster.

6 Conclusion and Future Work

In this work we proposed a novel multidimensional projection technique, based on radial basis functions. The main advantage of the presented method is its ability to perform control points selection, a step generally not present in most projection techniques. Results indicate that the presented ROLS-selection procedure presents a good trade-off between time and stress, while reducing the number of control points, an important result if user interaction through control points is desired.

There are a few things we consider interesting to explore in this method, which are: how to create local projections, using the shape parameter ε of the Gaussian kernel to determine the radial of influence of each control point; understand how the projection behaves when different techniques, such as Sammon's mapping, PCA and MDS, are used to position control points; investigate how to automatically select a good RBF kernel and automatically tune their shape parameters for different data sets. Besides, we intend fully explore the usability of a reduced number of control points.

Finally, we intend to experiment the ROLS-based control points's selection procedure in other projection techniques. In fact, some tests were already performed using this approach and the results are encouraging.

Acknowledgements

We would like to thank the anonymous reviewers for their careful and valuable comments and suggestions. This research was supported in part by the NSERC / Alberta Innovates Academy (AITF) / Foundation CMG Industrial Research Chair program in Scalable

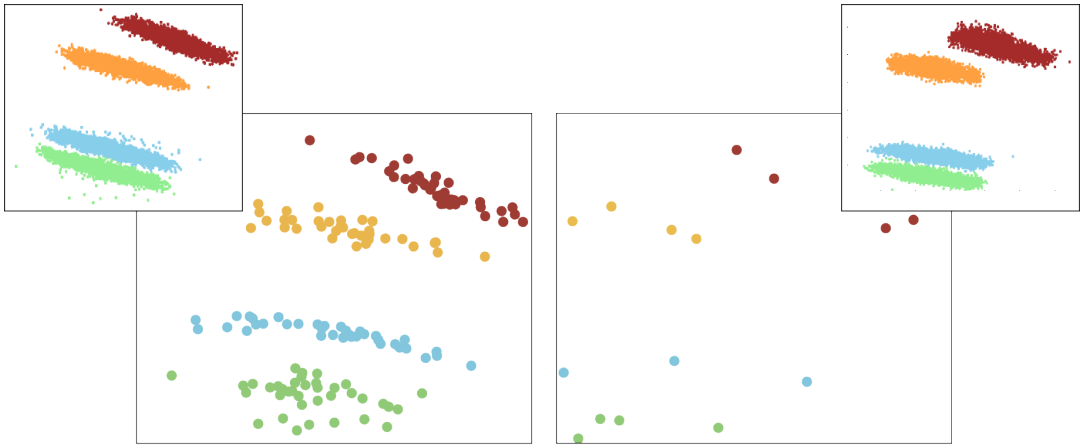

Figure 7: Example of control points selection through ROLS in the Mammals data set (20,000 instances and 47 dimensions). This data set contains four well-defined clusters, indicated by the different colors in the projection. Figures on the bottom present only the control points used to produce the projection, depicted on the top.

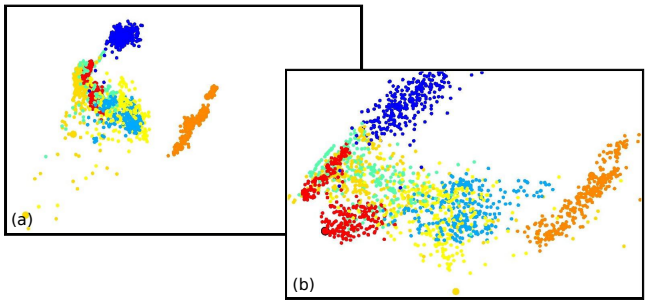

Figure 9: Example of control points manipulation with the segmentation data set. (a) Control points automatically positioned and (b) final projection after user manipulation.

Reservoir Visualization. We also acknowledge the Brazilian funding agencies Fapesp and CNPq, and GRAND NCE of Canada.

REFERENCES

[1] D. N. A. Asuncion. UCI machine learning repository, 2007.

[2] K. Bache and M. Lichman. UCI machine learning repository, http://archive.ics.uci.edu/ml, 2013.

[3] J. Bennett and W. Hays. Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 25(1):27–43, 1960.

[4] M. D. Buhmann. *Radial Basis Functions*. Cambridge University Press, New York, NY, USA, 2003.

[5] S. Chen, E. S. Chng, and K. Alkadhimi. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International Journal of Control*, 64(5):829–837, 1996.

[6] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, Mar. 1991.

[7] Y. Chen, L. Wang, M. Dong, and J. Hua. Exemplar-based visualization of large document corpus. *IEEE TVCG*, 15:1161–1168, 2009.

[8] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, 2nd edition edition, 2000.

[9] J. Daniels II, E. Anderson, L. Nonato, and C. Silva. Interactive vector field feature identification. *IEEE TVCG*, 16:1560–1568, 2010.

[10] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems 15.*, pages 705–712, 2002.

[11] V. de Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford, 2004.

[12] G. Deboeck and T. Kohonen. *Visual Explorations in Finance: With Self-Organizing Maps*. Springer Finance. Springer, 2010.

[13] P. Demartines and J. Herault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. on Neural Networks*, 8(1):148–154, 1997.

[14] C. Faloutsos and K.-I. Lin. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec.*, 24(2):163–174, 1995.

[15] M. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE TVCG*, 9(3):378 – 394, 2003.

[16] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel mds on the gpu. *IEEE TVCG*, 15:249–261, 2009.

[17] H. Janicke, M. Bottinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Trans. on Vis. Comput. Graph.*, 14(6):1459–1466, 2008.

[18] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE TVCG*, 17(12):2563 –2571, 2011.

[19] I. Jolliffee. *Principal Component Analysis*. Springer-Verlag, 3rd edition edition, 2002.

[20] F. Paulovich and R. Minghim. Text map explorer: A tool to create and explore document maps. *Info. Vis.*, pages 245–251, 2006.

[21] F. Paulovich, L. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE TVCG*, 14(3):564–575, 2008.

[22] F. Paulovich, C. Silva, and L. Nonato. Two-phase mapping for projecting massive data sets. *IEEE TVCG*, 16(6):1281–1290, 2010.

[23] F. V. Paulovich, D. M. Eler, J. Poco, C. P. Botha, R. Minghim, and L. G. Nonato. Piecewise Laplacian-based Projection for Interactive Data Exploration and Organization. *Computer Graphics Forum*, 30(3):1091–1100, 2011.

[24] E. Pekalska, D. de Ridder, R. P. Duin, and M. A. Kraaijveld. A new method of generalizing Sammon mapping with application to algorithm speed-up. In *5th Annual Conf. of the Advan. School for Comput. and Imag.*, 1999.

[25] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[26] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5):401–409, May 1969.

[27] E. Tejada, R. Minghim, and L. G. Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, 2(4):218–231, Dec. 2003.

[28] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.