UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

ON IMPROVED PROJECTION TECHNIQUES TO SUPPORT VISUAL EXPLORATION OF MULTI-DIMENSIONAL DATA SETS

EDUARDO TEJADA-GAMERO
ROSANE MINGHIM
LUIS GUSTAVO NONATO

Nº 207

RELATÓRIOS TÉCNICOS

São Carlos – SP
2003

On improved projection techniques to support visual exploration of multi-dimensional data sets

Eduardo Tejada-Gamero, Rosane Minghim and Luis Gustavo Nonato
{eduardo,rminghim,gnonato}@icmc.usp.br
High Performance Computing Laboratory
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, São Carlos, Brazil.

Abstract

Projection (or dimensionionality reduction) techniques have been used as a means to handling the growing dimensionality of data sets as well as providing a way to visualize information coded into point relationships. Their role is essential in data interpretation and simultaneous use of different projections and their visualizations improve data understanding and increase the level of confidence in the result. For that purpose projections should be fast to allow multiple views of the same data set. In this work we present a novel fast technique for projecting multi-dimensional data sets into bidimensional (2D) spaces that preserves neighborhood relationships. Additionally, a new technique for improving 2D projections from multi-dimensional data is presented, that helps reduce the inherent loss of information yielded by dimensionality reduction. The results are stimulating and are presented in the form of comparative visualizations against known and new 2D projection techniques. Based on the projection improvement approach presented here, a new metric for quality of projection is also given, that matches well the visual perception of quality. We discuss the implication of using improved projections in visual exploration of large data sets and the role of interaction in visualization of projected subspaces.

1 Introduction

Visual exploration of multi-dimensional data sets has become a common task in the last years and a necessary one to handle complexities of interpreting large multi-dimensional data sets. In order to make visual exploration feasible, different information visualization approaches have been developed for dealing with multi-dimensionality. These approaches are known as *multi-dimensional data visualization techniques*. Amongst these, we find 2D and 3D scatterplots of bidimensional and tridimensional projections of the data, scatterplot matrices, Sammon plots, heatmaps, heightmaps, table lens, survey plots, inconographic visualization, dimensional stacking, parallel coordinates, multi-linear graphs, pixel-oriented techniques, circle segments, polar charts, RadViz, PolyVis, grand tour, projection pursuit, and Kohonen self-organizing maps (SOM). A good survey of these techniques can be found in the work by Grinstein *et al.* [4].

That work also presents metrics for identifying how visualizations deal with n dimensions when displayed on screen, namely the Intrinsic Dimension (ID), the Intrinsic Record Ratio (IRR), and the Intrinsic Coordinate Dimension (ICD). Given an n-dimensional space, the ID of a visualization measures the maximum number of unit vectors that could be uniquely perceived, the IRR represents the percentage of records that can be distinguished, if one had reasonably distributed records, and the ICD is the maximum number of coordinates of any vector in that n-dimensional space that can be uniquely identified in the visualization.

These metrics are useful to measure the capacity of a visualization technique to deal with multi-dimensional and large data sets. However, since the screen is a bidimensional domain,

some of the most used visualization techniques need a pre-processing step for projecting multi-dimensional data into bidimensional spaces. Such projection must be done in a way that the loss of information is as low as possible. Different techniques have been proposed to accomplish this task such as SVD (Single Value Decomposition) [9], PCA (Principal Components Analysis) [5], Fastmap [3], and Fractal-based techniques [6, 7]. Visualizations of such projections can help identify structure in the original data. However, since a large number of projections can be resulted from any of those techniques, the choice of useful projections is an important step in data interpretation.

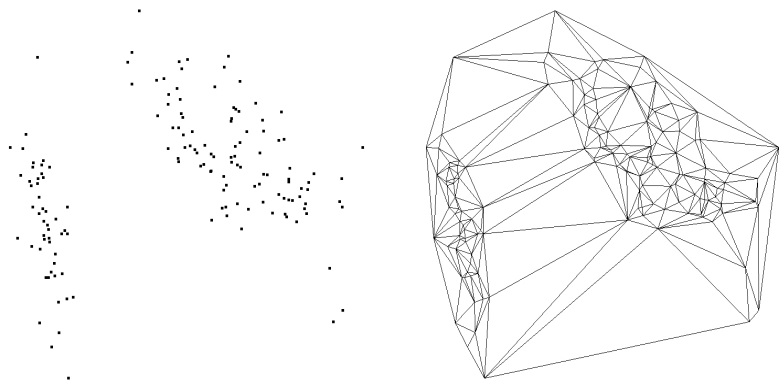
Thus, we are set to find bidimensional projections and their visualizations that offer alternative useful views of the data set. The problem at hand can be stated as follows:

Let X be a set of points in $\mathbb{R}^n$ and $d : \mathbb{R}^n \to \mathbb{R}$ be a criterion of proximity between points in $\mathbb{R}^n$.

We wish to identify a set of points P in $\mathbb{R}^2$ such that if $\alpha : X \to P$ is a bijective relation and $d_2 : \mathbb{R}^2 \to \mathbb{R}$ is a proximity criterion in $\mathbb{R}^2$, then $|d(x_i, x_j) - d_2(\alpha(x_i), \alpha(x_j)|$ is as close to zero as possible, where $x_i, x_j \in X$.

In this paper we call the set P a projection. Visually, they can be thought of as points on a plane, corresponding to the screen surface. To simplify expressing this representation along the text, we will also call the visual display of the projected points (that is, their scatter plot) a projection. We also want to have alternative projection schemes to reveal different structures and patterns in the data set. Those projection schemes should be fast so as to allow multiple evaluations. Under certain circumstances, we wish to explore individual points and their neighborhoods concerning the criterion d.

Over bidimensional projections we can apply a range of visualization and interaction techniques in order to allow a better understanding of the data by the user. For instance, figure 1 shows an scatterplot and a Delaunay triangulation [2] of the bidimensional projection of the Iris data set (see table 1 for data set specification) using Fastmap.



(a) (b)

Figure 1: Fastmap projection of the Iris data set (see table 1 for data set specification). (a) Scatterplot of the projection. (b) Triangulation of the projection.

Besides helping identify structure, performing a triangulation over point projections is useful to identify relationships between instances (data points) using interaction techniques based on polygonal meshes such as the spider-cursor, presented in section 5. It also allows

the mapping of attributes to visual properties (such as height or color), as in figure 2, or even aural properties (such as pitch), to help increase representational power.

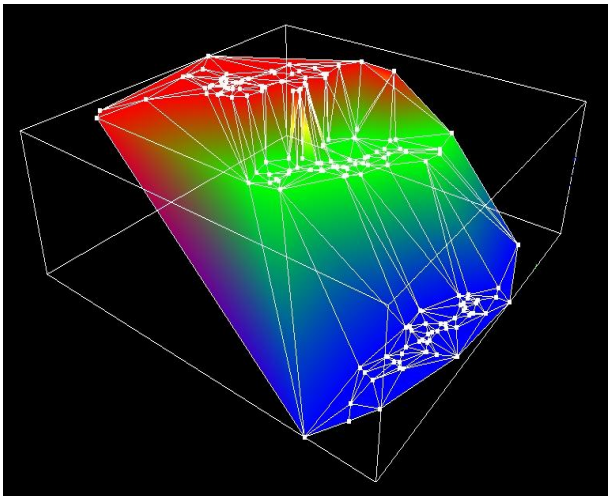


Figure 2: Delaunay triangulation of the Fastmap projection of the Iris data set (see table 1 for data set specification). The height and color map the class to which the instance belongs.

Delaunay triangulations present properties that make them preferable over other triangulations for generating meshes over which we will apply visualization and interaction techniques. Particularly for the case presented here, a mesh generated by the Delaunay triangulation shows good visual quality and, for visual and aural exploration, the fact that the star of each point is formed by the nearest neighbors allows a smooth transition when one walks through the mesh.

Having alternative projections is the first step to locate structures visually in the data, and in that respect, fast algorithms must be found to generate such projections. This paper proposes a novel technique for performing a fast mapping of multi-dimensional metric data sets into a bidimensional space without the need for vector information. For each new data point to be projected, it uses the two nearest neighbors amongst all data points already mapped to establish its position in the projected space. This will generate a projection that intends to preserve local neighborhoods as much as possible, offering information about mutual relationships between adjacent instances. This technique, called NNP (nearest-neighbor projection), is presented in section 2.

We exploit the characteristic of neighborhood preservation to propose an approach to generate improved mappings that recovers information lost during projection. This projection improvement approach, called the Force approach (section 3), can be used over projections generated with any dimensionality reduction technique. However, best results were obtained over NNP mappings since local relationships are already preserved in them (section 6).

Based on this Force approach, we define a metric for the relative error of projections obtained with different dimensionality reduction algorithms. Although information visualization techniques are useful to get an intuitive idea of the quality of a projection, this analytical measurement can confirm the evaluation made visually and also support applications where a projecting step is performed without user interference.

These analytical and algorithmic efforts can considerably reduce the loss of information when combined with multiple views and multiple visualization techniques for direct visual exploration of high-dimensional data sets. This combination improves the accuracy and increases

the confidence level of the knowledge gained by the user about the data and the domain it belongs to. Both the approach for projection improvement and the metric for projection quality are presented in section 3.

Also, user interaction over the various visualizations is a very important characteristic that must be added to applications for data exploration. It restores some of the information lost during projection and can reveal new local structures in data, otherwise hidden by the global aspect of data set mapping.

The interaction and visualization techniques discussed this far are suitable for moderate-sized data sets or subsets of very large data sets. However, also with large data sets it is important to provide tools for visual exploration. Contrary to visualization techniques that allow identification of each instance individually (which we call region-oriented), the visualization techniques for exploring large data sets must provide a general overview of the entire data set (global-oriented techniques). In global-oriented visualizations, exploration at an individual point level is impaired by resolution of devices and user perception. Examples of these global visualization techniques are the heightmap of a density estimation of a projection and a bidimensional self-organizing map of the data. As with region-oriented techniques, different projections determine added insights into the data. In section 4 we discuss the role of improved projections in visualization of large data sets.

Section 5 is devoted to argue for the importance of using interaction in the visualization of projected subspaces during data exploration. In section 6 we present some results of the above mentioned techniques using various data sets, as well as plots of projection errors. Finally, section 7 presents the conclusions and follow-up of this work. All software developed to implement the techniques presented here is available on the internet, as well as its source and examples.

2 Fast projection of multi-dimensional data

In this section, we present a novel algorithm for performing dimensionality reduction from multi-dimensional to bidimensional spaces, while attempting to preserve the relation between the local neighborhoods in the original and the projected spaces.

We use the positions in the bidimensional space of the two nearest neighbors from all points already mapped for determining the position of a new data point. Let $\widetilde{X} \subset X$ be the subset of points already projected. Suppose $x \in X$ is a new point to be projected and q and r two points in $\widetilde{X}$ such that $\forall x_k \in \widetilde{X},\ x_k \neq q, r,\ d(x, x_k) \geq d(x, r)$ and $d(x, x_k) \geq d(x, q)$, i. e., x is closer to r and q than to any other point in $\widetilde{X}$. The position of the new point $x' = \alpha(x)$ in the projected space is at an intersection between the two circles with centers in $q' = \alpha(q)$ and $r' = \alpha(r)$, and radii equal to $d(x, q)$ and $d(x, r)$. Figure 3a illustrates this scheme. When there is no intersections between the circles, an intermediary point between its centers is used. This intermediary point is determined by the radii of the two circles and the relative position between them. Figures 3b and 3c show the two main exception cases. In the case of tangent circles the only intersection point is used. We detail the approach in the algorithm presented below.

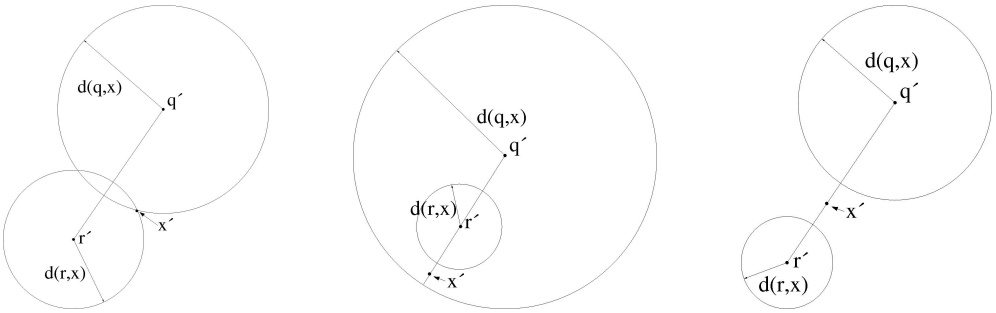
NNP algorithm

---

1. Project the first two data points such that the distance between them in the bidimensional subspace is equal to the distance in the original

multidimensional space.

2. For every data point x.
   2.1 Perform a kNN query in the subset of the projected points $\widetilde{X}$ returning the two nearest neighbors q and r.
   2.2 Find the intersection points of the two bidimensional circles with center in the projected data points q' and r' and radii equal to $d(x, q)$ and $d(x, r)$ respectively.
   2.3 If there is no intersection point
       If one circle contains the other
           Set the intersection point as in figure 3b
       Else
           Set the intersection point as in figure 3c
   2.4 If there is one intersection point
       2.4.1 Set the new projected data point as being the intersection point.
   2.5 If there are two intersection points
       2.5.1 Choose randomly one of the intersection points and set it as being the new projected data point x'.



(a) (b) (c)

Figure 3: Intersection Cases for NNP mapping (a) Circles intersect. (b)No intersections, one internal circle. (c) No intersection or inclusion.

For a data set with n instances, this technique has complexity $O(n^2)$ for sequential access data structures. This complexity is dramatically improved using data structures that support fast kNN (k-Nearest Neighbors) queries of the order of $O(log(n))$, over with the complexity of the algorithm is $O(log(n!))$.

Employing a Delaunay Triangulation for meshing the projected points, the projected data set supports similarity queries with complexity equal to $O(n_s)$, where $n_s = k$ for kNN queries and $n_s = l$ for range queries with radius r and center in q, where l is the number of instances in the r-neighborhood of q.

As stated before, this technique attempts to preserves local neighborhood information, that is given a projected point, it is likely that a neighboring point to it in the original space (according to the similarity criterion d) will be in a reasonable neighborhood in the projected space also. This characteristic is useful to apply interaction and visualization techniques in such a way that the transition between the instances when the user walks through the mesh

will be smooth.

This algorithm is suitable for metric data sets where the triangular inequality is valid and there is no vector data. Its advantage lies on the preservation of neighboring relationships between data points. However, it suffers from the same problem every of dimensionality reduction techniques presents: loss of information.

We can exploit the neighborhood preservation feature of NNPs to devise a mechanism that improves the projection by recovering similarity information lost during the projection process. A projection with such a particularity is a good initial state for the improvement scheme developed, which is presented in the next section.

3 Improving and measuring projections

This section presents a novel technique for improving projections, together with a metric to evaluate projection quality. The approach uses the fact that the relation between the distances in both the original and the projected spaces should be constant for every pair of data points (x'_i, x'_j). The idea is to separate instances projected too close, and to approach instances projected too far.

The basis of the projection improvement scheme is as follows: for an instance x'_i, we calculate the vector $\overrightarrow{v_{ij}} = (x'_j - x'_i)$, $\forall x'_j \neq x'_i$. Then, we apply a perturbation to x'_j in the direction of $\overrightarrow{v_{ij}}$. This perturbation depends on the actual and the ideal distances between the projected instances. We work with normalized distances in order to avoid inconsistencies derived from the difference between the ranges of the bidimensional and the original multi-dimensional domains. Since the normalization of the distances is a costly process, it cannot be performed in each iteration for the projected space. To improve performance we normalize the distances only once for the original space, and for each iteration we apply a normalization to the coordinates of the projected instances in the bidimensional space, instead of the projected distances. This improvement scheme has been named the Force approach (a reference to the impression that points are being 'attracted to' or 'repelled from' one another). The process for each iteration is presented in the algorithm below.

Force algorithm

---

1. For each projected data point x'.
   1.1 For each projected data point $q' \neq x'$.
      1.1.1 Calculate $\overrightarrow{v}$ as being the vector from x' to q'.
      1.1.2 Move q' in the direction of $\overrightarrow{v}$ a fraction of Δ.
2. Normalize the projection coordinates to the range [0,1] in both dimensions.

Δ in the algorithm is an approximation for the actual difference between the projected distance and the distance in the original space (that is, the error in relative positions of the two points x' and q').

This approximation is given by:

$$\Delta = \frac{d(x,q) - dmin}{dmax - dmin} - d_2(x', q') \tag{1}$$

where $dmax$ and $dmin$ are the maximum and minimum distances in the n-dimensional space respectively, and $x' = \alpha(x)$ and $q' = \alpha(q)$.

The above approach to improve bidimensional projections of multi-dimensional data is capable of handling all sorts of projections matching the original problem statement. Section 6 shows interesting results of its application to projections realized using NNP and Fastmap.

Another important feature of the Force scheme is the possibility of returning to pseudo-previous states inverting the direction of vector $\overrightarrow{v}$. Thus, the user (as well as an automatic projection generator) can use this mechanism in order to converge to a better projection for the specific application at hand.

Based on the Force approach, one can measure the relative error $Q(P)$ of the projection P obtained in each iteration through the expression:

$$Q(P) = \frac{1}{M} \sum_{k=1}^{M} |\Delta_k| \tag{2}$$

It is important to notice that $Q(P)$ is not an absolute measure for the error of a single projection since the distances in the projected subspace are not normalized when Δ is calculated. However, it allows to establish a means for comparing different projections of the same data set. Because of that, the optimal projection will have an error $Q(P) \geq 0$.

This metric is actually very useful not only to confirm visual interpretation of projected data, but also as an initial step towards optimization of projections done in an automatic or semi-automatic way. Plots of this projection quality metric have shown a good match to visual evaluation of the improved projections, as will be illustrated in section 6.

The projection improvement technique and the metric for projection quality presented in this section can be used over the results of any dimensionality reduction algorithm. In turn, the result of the improvement technique can be used both by region- and global-oriented visualization and interaction techniques. In the next section we focus in global-oriented techniques to support exploration of very large data sets. It presents some results of the work we performed with large data sets using well known approaches for visualizing their general properties, and the utility of improved projections for exploring such data sets.

4 Visual exploration using bidimensional projections of large data sets

It is not always possible to explore very large data sets visually using techniques like those presented in sections 1, 2, and 3, because one type of loss of information in this case is the individuality of the points. For very large data sets we have three choices: (a) sacrifice individuality using only global-oriented visualization techniques, that offers insight into the entire data set, (b) explore subsets using region-oriented techniques, or (c) combine both approaches using a global-oriented technique for identifying interest areas and then focusing on them using a region-oriented technique. The latter presents the ideal scenario for visual data exploration.

To employ a global-oriented approach for gaining insight into large data sets, one can use, for instance, heightmaps and colormaps of the density estimation of projected data sets instead of looking at the projected points themselves. Density estimation calculations can be looked up in the book by Jolliffe [5]. These visualization techniques, combined with a multiple-projection scheme, would be useful to identify clusters visually in the data set as shown in figure 4.

In this example, the density $\hat{f}_D$ is estimated using a gaussian kernel estimator given by:

$$\hat{f}_D = \frac{1}{nh^d} \sum_{i}^{n} \frac{1}{\sqrt{2\pi}h} e^{-\frac{(x-x_i)^2}{2h^2}} \tag{3}$$

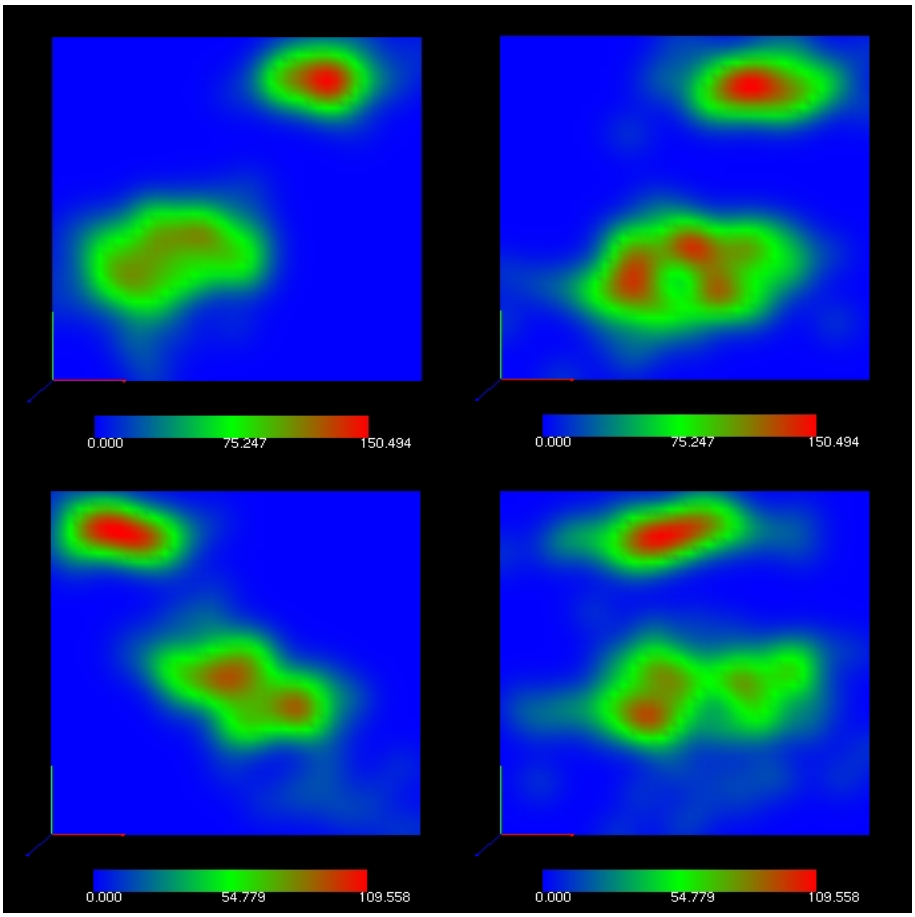


Figure 4: Multiple views of the density estimation of various projections of the Iris data set (see table 1 for data specification). We can see from the projections that there are three clusters, two of them considerably close to each other and undistinguishable in some of the projections, while revealed in others.

where x_i are the data points, h is the smoothing factor, n the number of instances in the data set, and x represents a point in the bidimensional cartesian grid where the density is stored.

Visually plotting the density map can help determine interest areas to be explored using region-oriented techniques. This 'preview' of the entire data set is affected by the quality of the projection over which it is applied. The amount of information preserved determines how accurate the insight gained is. Thus, approaches such as NNP and the Force scheme (sections 2 and 3) are of key importance for these kinds of tasks.

Whatever the means used for mapping data, information will always be lost with dimension reduction. Here again it is necessary to use a number of visualizations of the same data set in order to minimize such loss. Another useful example of global-oriented technique for large data set exploration is self-organizing maps (SOM).

Figure 5 shows a SOM of the IDH data set (see table 1 for data specification), where one can see features, such as class distribution, that can not be perceived with region-oriented techniques or other global-oriented techniques.

As stated before, these approaches, as well as most visualization techniques, can be combined with interaction mechanisms in order to provide meaningful insights into the data. We

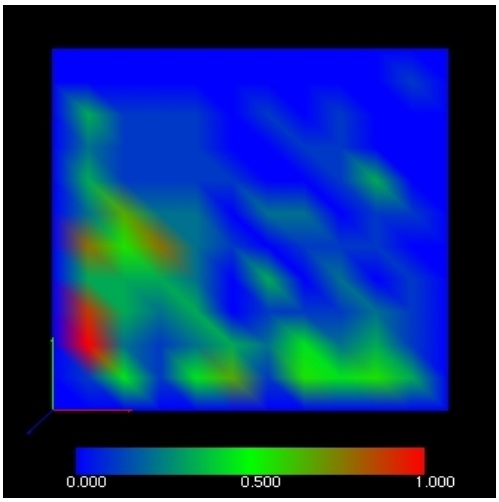


Figure 5: Self-organizing map of the IDH data set (see table 1 for data set specification).

dedicate the next section to discuss the role of interaction within the context of projections of small and large data sets.

5 The role of interaction in data exploration using projections

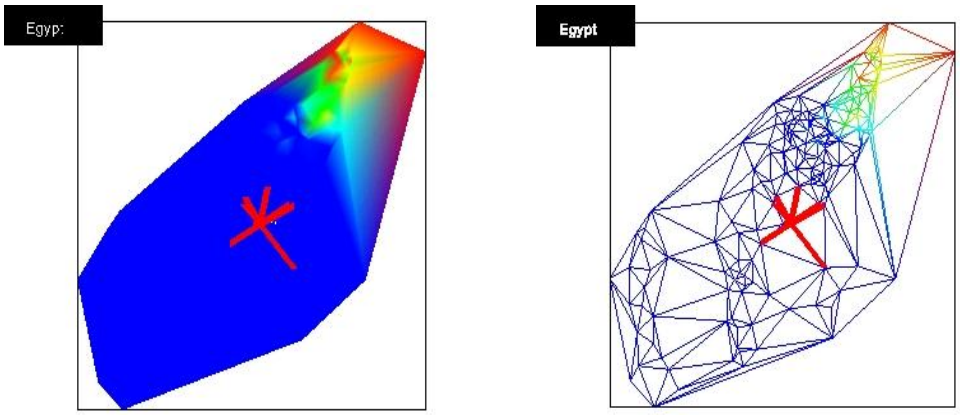
Through interaction techniques, the user is capable of finding structures that otherwise would remain hidden. Interaction is particularly useful in lossy techniques, such as bidimensional projections of data sets. It can also be applied to gather information about a particular instance or group of instances, to select sub-domains, to choose attributes, and so on. The range of applications for interaction is as wide as the needs of the user.

As an example, figure 6 shows the use of an interaction technique we call *spider cursor*. This technique can be applied over any meshing of a bidimensional projection in order to find relationships between instances. It allows identification of neighboring instances to the one selected with the cursor. It is also useful to identify specific instances as we can see in figure 6, which is a projection of multi-dimensional information on quality of life for countries.

The spider cursor can also be accompanied by aural attribute mapping of the instance pointed by the cursor. This can guide exploration of the instances, allowing the user to perceive characteristics that are visually indistinguishable. For instance, in figure 6 ranking of countries was mapped to both color and pitch. While various countries are coded with the same color, pitch for them was different, allowing distinction of their rank.

Interaction techniques suitable for region-oriented visualization techniques are different from those suitable for global-oriented techniques. An appropriate interaction tool for the type of global-oriented techniques presented in section 4 would be a *cutting plane*. It can, for instance, assist in the task of defining a noise level for eliminating outliers in further analysis steps [1] (see figure 7).

There is a wide range of interaction approaches that can be applied over different visualization techniques, and for those, the way the projection is generated determines the insight gained. For instance, in the case of the spider cursor, the fact that the local neighborhoods are preserved by NNP mapping and by the Force scheme techniques, allows an smooth and consistent walk through the graphical representation. For instance, it is possible for the user



(a) (b)

Figure 6: Spider cursor interaction tool over the IDH data set (see table 1 for data set specification). (a) Surface color map view. (b) Delaunay triangulation of a bidimensional projection.

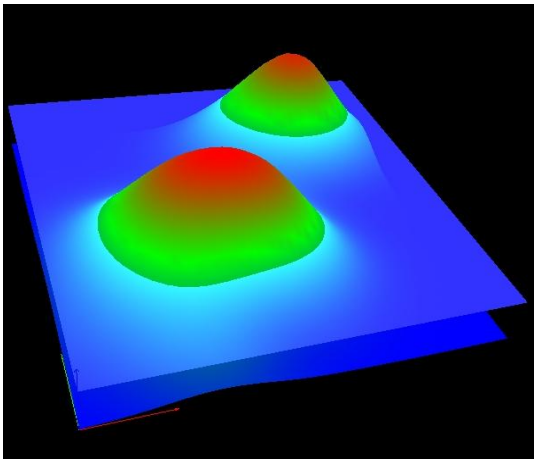


Figure 7: Cutting plane for specifying a noise level λ over a heightmap representing the density estimation of a bidimensional projection of the Iris data set (see table 1 for data set specification). Data below the cutting plane are considered outliers for further processing.

to find a specific instance in the visualization following a path starting in a point with a color mapped to a value close to the value of the queried instance.

The next section presents comparative results of the techniques developed.

6 Results

In this section we show results of applying the techniques developed in comparison with related techniques to various domains. Table 1 summarizes the main characteristics of the data sets used for the tests[1]. Similarity criterion in all cases was defined by the Euclidean distance

[1]All data sets, with the exception of Synt5d and Synt10d, were taken from the UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository).

between points.

	Features					
Data set	Instances	Num. of inst.	Dim.	Clusters	Synt./Real	Class attrib.
IDH	countries	173	9	-	real	quality of life
Iris	plants	150	4	3	real	type
Wine	Italian wines	178	13	3	real	type
Housing	houses in Boston	506	13	-	real	value in US$
Quadrupeds	quadrapeds	96487	72	4	synthetic	type
Synt5d	-	200	5	4	synthetic	cluster
Synt10d	-	1000	10	4	synthetic	cluster

Table 1: Data sets used for the experiments performed.

In section 3 we proposed an improvement scheme for dimensionality reduction techniques. In figure 8 we show that it improves projections generated by the Fastmap [3] algorithm applied to different data sets. Measuring the error as defined in section 3, we could notice that numerically the improved projection converges to an acceptable state as shown in figure 9. The Force scheme improved projection in all cases. That can be recognized in the color mappings (points of similar color are neighbors according to the similarity criterion d) and in te aspect of the triangulation (points are grouped together better). The housing data set is not very suitable to this kind of projection due to the 'unnatural' meaning of the Euclidean distance for the data.

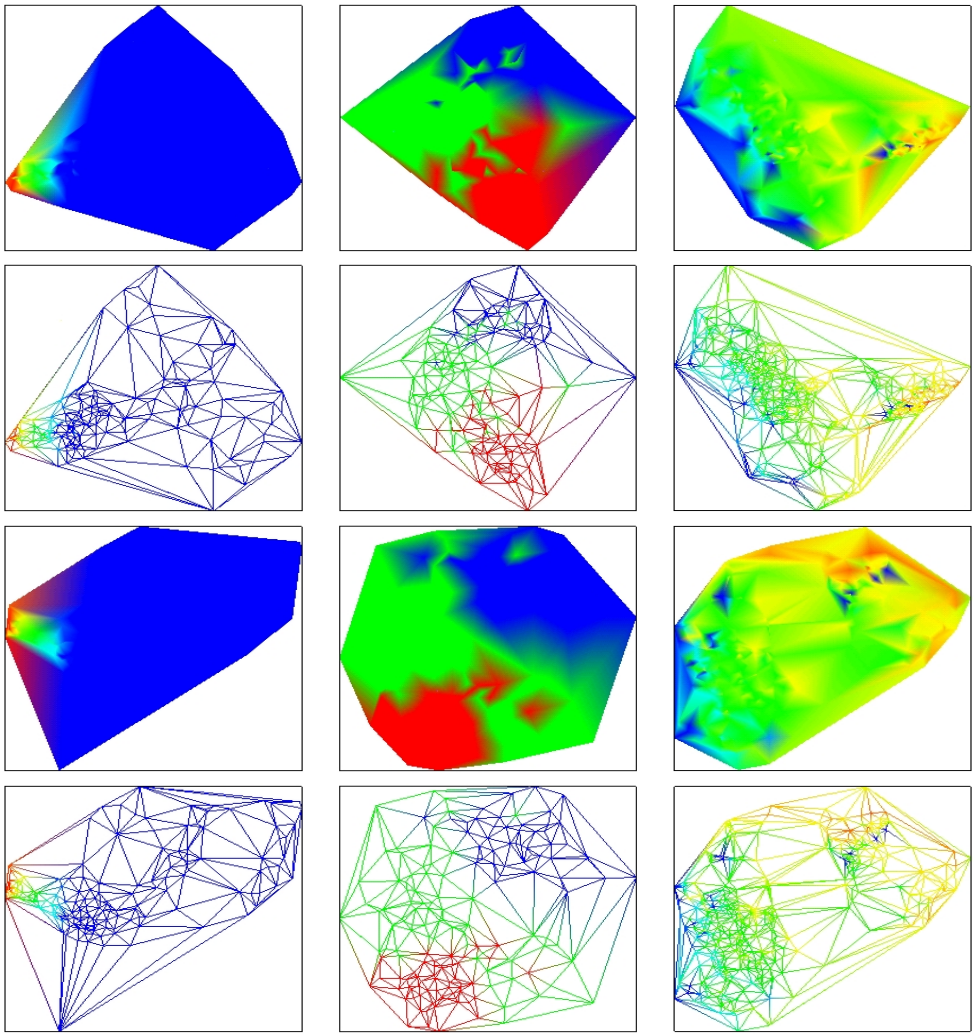
Figure 10 illustrates the use of the Force scheme over the projection of the same data sets generated using the nearest-neighbors (NNP) approach proposed in section 2, as well as the original results of our projection algorithm. As with Fastmap, improvemente in this case was achieved in all cases.

Figure 11 shows the convergence when the improvement scheme is applied to the results of NNP. We can see that the final errors for the Wine data set and the Housing data set are lower than the errors obtained when the original projection is generated using Fastmap. These results, that confirm the quality verified visually, is obtained due to one of the features of the NNP projection, namely that local neighborhoods are preserved when the data are mapped into a bidimensional space. This property imposes a characteristic to the projections that makes them a good initial state for the Force projection improvement scheme presented here.

For the IDH dataset, the error (as well as visual confirmation) indicate a better final state using Fastmap (less exception points), supporting the position for the use of more than one projection under the same type of visualization.

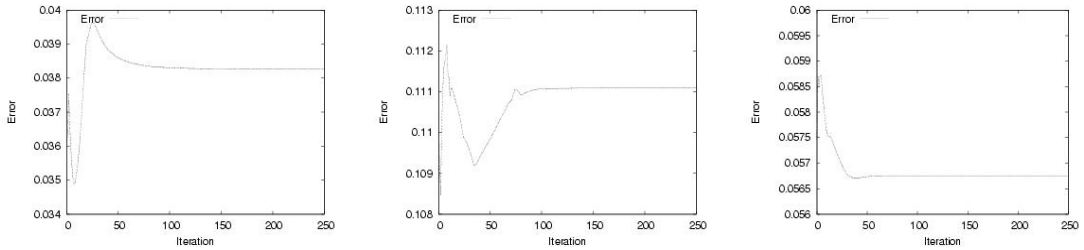
From the previous figures, it is noticeable that after the improvement, point position is better. From the error plots, it can be seen that the number of iterations necessary to estabilize the projection improvement is lower when the initial state is the result of a NNP mapping. It means that the (usually costly) projection improvement algorithm is faster when its initial state is generate by NNPs.

In figure 12, we can see the advantages obtained from the Force scheme, when used together with a global-oriented visualization technique. It shows the density estimation and the Delaunay triangulation of a 5-dimensional data set containing four groups projected into a bidimensional space (data set Synt5d in table 1). In the original projection (figure 12a), it can be seen that, the triangulation over projected points does not help the perception of the groups in the middle of the domain. For that projection, the density estimation suggests the presence of all the groups. Applying the improvement scheme to this projection, both of the visualiza-



(a) (b) (c)

Figure 8: Projections of data sets using the Fastmap technique and the Force scheme on the Fastmap projection for three data sets. (a) The IDH data set. (b) The Wine data set. (c) The Housing data set. In all cases, color maps a specific characteristic for each data point (column Class attrib. in table 1).



(a) (b) (c)

Figure 9: Error measured for iterations 0-250 of the Force scheme over Fastmap projections. (a) The IDH data set. (b) The Wine data set. (c) The Housing data set.

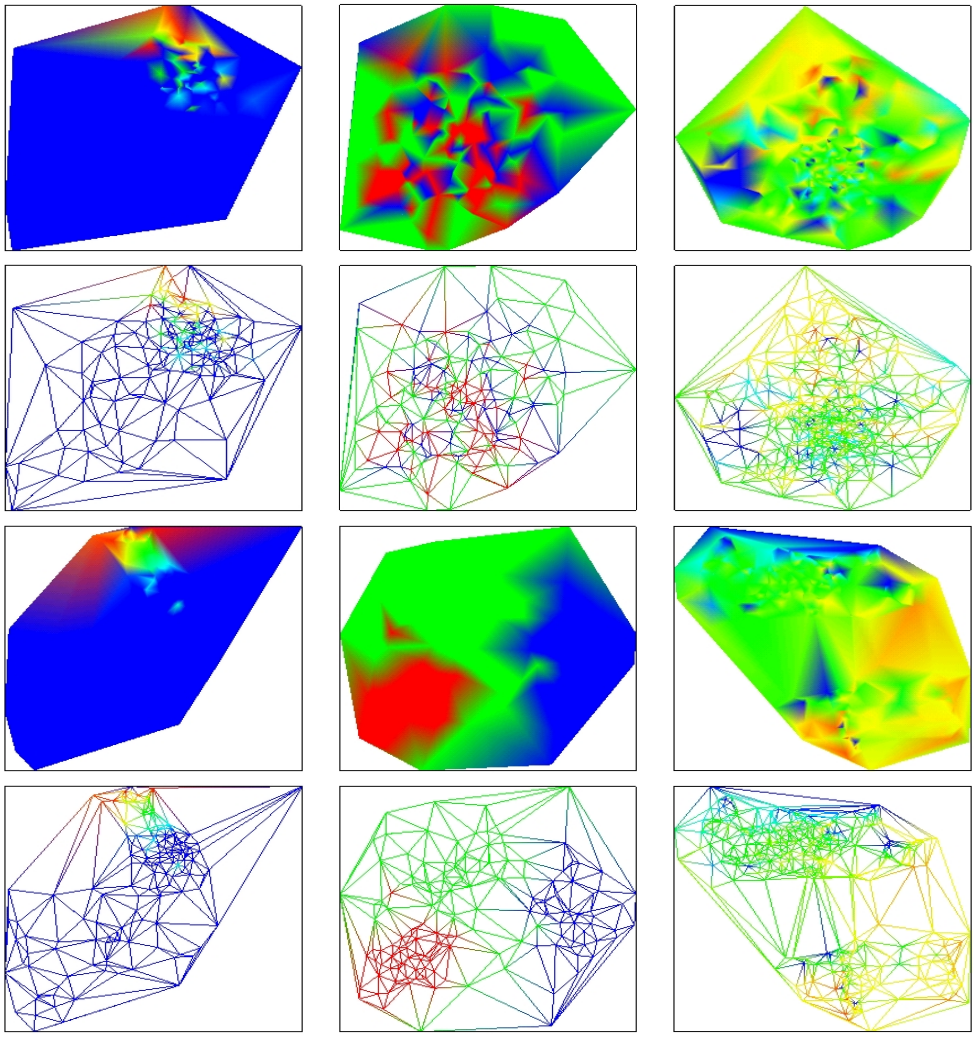


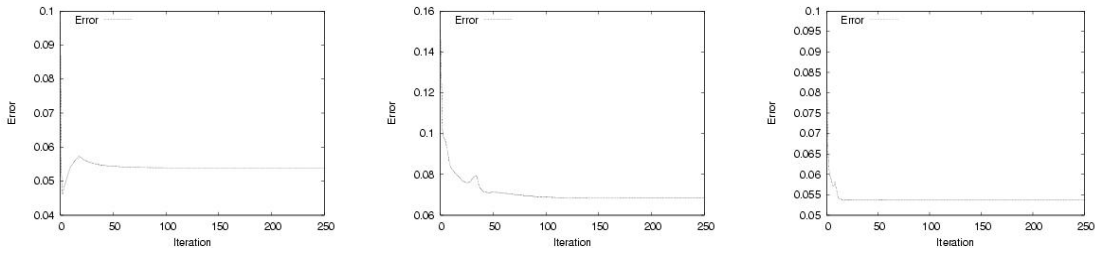
Figure 10: Projections of data sets using the NNP technique and the Force scheme on the NNP for three data sets. (a) The IDH data set. (b) The Wine data set. (c) The Housing data set. In all cases, color maps a specific characteristic for each data point (column Class attrib. in 1).

tion techniques show the four groups. Therefore the first visualization was an indicator that the projection improvement scheme should be used for this case.

This leads us back to the issue of multiple visualizations. Some characteristics of the data could be perceived with some visualization techniques whilst other characteristics will be perceived with other visualization techniques. On top of that, interaction improved local distinction of hidden features.

As an additional example, figure 13 illustrates a situation where using multiple visualization techniques helps find characteristics that would otherwise remain hidden. Figure 13a shows a test case where it is not possible to perceive the four existing groups using a scatterplot. However, in the density heightmap they are identifiable. The situation is reversed for the data set in figure 13b.

The data set in figure 13a has a considerable amount of instances. We can see that the



(a) (b) (c)

Figure 11: Error measured for iterations 0-250 of the Force scheme over NNPs. (a) The IDH data set. (b) The Wine data set. (c) The Housing data set.

density estimation gives us a better insight about the data. This global-oriented technique is useful for a first analysis of the data, which could allow to focus in interest regions and use region-oriented techniques to gain further knowledge about the data, as stated in section 4.

Conclusions and ideas for future works drawn for the work presented here are the matter of the next section.

7 Conclusion

We have presented a novel technique for performing a fast projection of multi-dimensional data sets into bidimensional spaces (NNP mapping) and a novel approach for improving projections resulting of dimensionality reduction techniques (the Force scheme) . Additionally we offered a metric to evaluate the quality of two different projections of the same data set. We presented the techniques inside a framework that foments discussions on the relation of improved projections to large data sets exploration and the role of interaction in data exploration of projected subspaces.

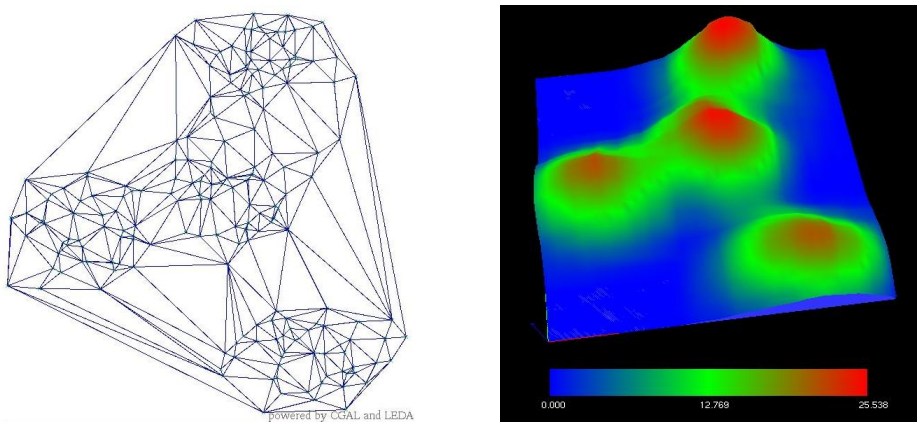
The NNP dimensionality reduction technique proposed generates projections that attempt to preserve local neighborhoods. This has the advantage that local exploration is supported by the similarity between neighbors. The NNP technique represents a very interesting alternative to existing dimensionality reduction techniques, and can be used in combination with them to highlight patterns and structures in data. Additionally, NNps do not use vector information during projection, keeping its computational cost low.

Using NNPs as an initial state for projection improvement schemes (such as the Force scheme presented here) benefits from the neighborhood preservation feature. The Force scheme is more effective (and faster) when run over NNPs.

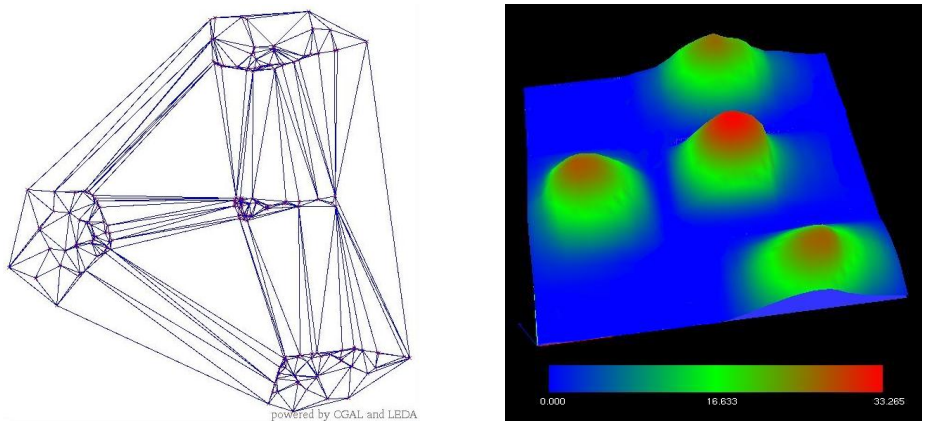
The Force scheme itself is a time consuming procedure, although worth employing for most data sets. Also, as illustrated in the results, using fast projections initially (such as Fastmap and the new NNPs) may help determine whether employing improvement procedures is worthwhile. The same occurs for global-oriented techniques, such as density plots or SOMs.

Combined, NNP followed by Force projection improvement yield projections with reduced error and faster computations when compared with the results of applying the Force scheme over Fastmap projections.

On the subject of error estimation, a metric was defined, based on the principles of the Force scheme, that is useful to calculate the difference in quality of two distinct projections



(a)



(b)

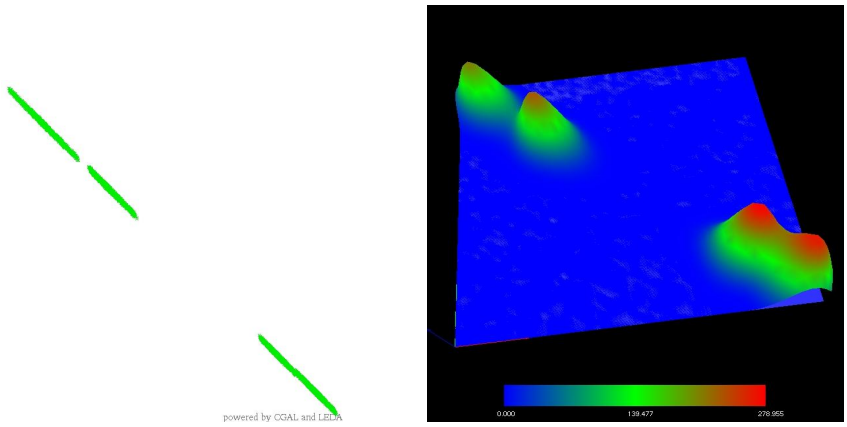
Figure 12: Visualizations of a 5-dimensional data set (see table 1 for data set specification). (a) Original projection and corresponding density plot. (b) Projection improved through the Force scheme, with corresponding density plot.

of the same data set. This metric was confirmed visually in all cases, and it represents a contribution to automatic detection and calculation of good projections with and without user interference, a necessary procedure for many data analysis procedures. Also, it is an indicator of when to stop an improvement scheme, even for an approach other than the Force scheme.

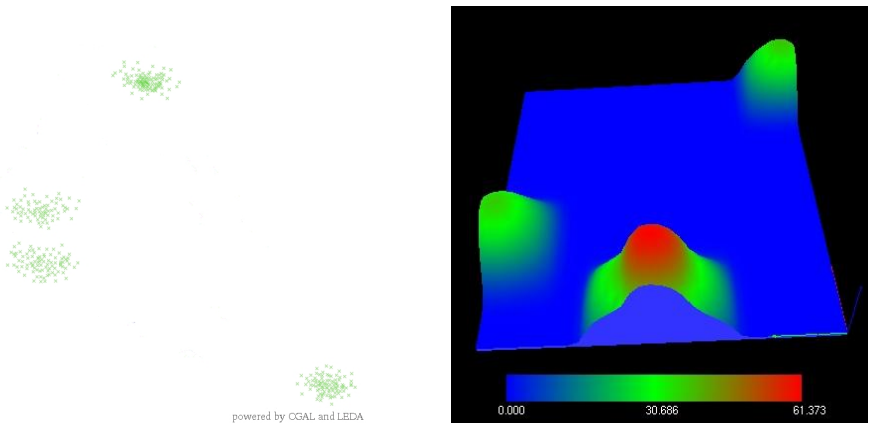
We have shown that the way a projection is generated determines the insight that will be gained using both region- and global-oriented visualization techniques. In the latter ones, an improved projection would allow to gain an accurate first notion of the data set that supports focusing on regions of interest.

During the whole process, interaction plays an important role. It allows the user to modify the visualizations to gain insight and improve the accuracy and confidence level of the knowledge gathered from the visualization.

We have also shown that multiple visualizations are necessary to reduce the loss of information and help increase accuracy and confidence level.



(a)



(b)

Figure 13: Gains using multiple visualization techniques for exploring data sets. (a) Four groups are perceived in the density heightmap, whilst the scatterplot shows only three (Quadrupeds data set in table 1). (b) The Delaunay triangulation over the scatterplot shows four groups in another data set, whilst the density estimation shows only three (Synt10d data set in table 1).

It is expected that the techniques presented here will contribute for a framework of multiple views to support data interpretation, in a way that the user is involved and that offer a measure of projections of multi-dimensional information.

All software developed, as well as documentation, source code and data files, are available to general users on the internet at http://www.lcad.icmc.usp.br/˜powervis/visual/. The spider cursor is a part of a library for interactive data visualization and sonification, available at http://www.lcad.icmc.usp.br/˜powervis/DSVol/. Software is written in C++. Many of the visualizations presented here were generated using the VTK software [8].

8 Acknowledgements

The authors want to acknowledge the finance of CNPq and FAPESP Brazilian financial agencies. Also, we acknowledge the ideas and the development of some pieces of code by our colleagues and students.

References

[1] C. Aggarwal. A human-computer cooperative system for effective high dimensional clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 221–226, CA, USA, 2001.

[2] Herbert Edelsbrunner. *Geometry and Topology for Mesh Generation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge, 2001.

[3] C. Faloutsos and K. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia databases. In Michael J. Carey and Donovan A. Schneider, editors, *ACM SIGMOD'95 International Conference on Management of Data*, pages 163–174, San Jose, California, 1995. ACM.

[4] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualizations. In *7th Data Mining Conference KDD Workshop 2001*, pages 7–19, San Francisco, CA, 2001.

[5] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[6] C. Traina Jr., A.J.M. Traina, Leejay Wu, and C. Faloutsos. Fast feature selection using fractal dimension. In *Simpósio Brasileiro de Banco de Dados - SBBD'00*, João Pessoa-PA, Brazil, 2000.

[7] B.-U Pagel, F. Korn, and C. Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. In *International Conference on Data Engineering - ICDE'00*, pages 589–598, San Diego-CA, USA, 2000.

[8] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit - An Object-Oriented Approach to 3D Graphics*. Kitware, 3rd edition, 2002.

[9] G. Strang. *Linear Algebra and its Applications*. Academic Press, 1980.