CheckViz : Sanity Check and Topological Clues
for Linear and Nonlinear Mappings

Sylvain Lespinats¹, Michaël Aupetit²

¹ CEA INES, Solar Systems Laboratory (L2S),
BP 332, 50 avenue du Lac Léman, F-73377 Le-Bourget-du-lac, France.
² CEA LIST, Information, Models and Machine Learning Laboratory (LIMA),
F-91191 Gif-sur-Yvette, France
sylvain.lespinats@cea.fr ; michael.aupetit@cea.fr

Abstract

Multidimensional scaling is a must-have tool for visual data miners, projecting multidimensional data onto a two-dimensional plane. However, what we see is not necessarily what we think about. In many cases, end-users do not take care of scaling the projection space with respect to the multidimensional space. Anyway, when using nonlinear mappings, scaling is not even possible. Yet, without scaling geometrical structures which might appear do not make more sense than considering a random map. Without scaling, we shall not make inference from the display back to the multidimensional space. No clusters, no trends, no outliers, there is nothing to infer without first quantifying the mapping quality. Several methods to qualify mappings have been devised. Here, we propose CheckViz, a new method belonging to the framework of Verity Visualization [WPL95]. We define a two-dimensional perceptually uniform colour coding which allows visualising tears and false neighbourhoods, the two elementary and complementary types of geometrical mapping distortions, straight onto the map at the location where they occur. As examples shall demonstrate, this visualisation method is essential to help users make sense out of the mappings and to prevent them from over interpretations. It could be applied to check other mappings as well.

Categories and subject descriptors: multidimensional data; nonlinear mapping; multidimensional scaling; evaluation; quality visualisation

1. Introduction

1.1. Multidimensional scaling

Mapping methods are generally designed to display data from a high-dimensional original space into a low-dimensional projection space. Such methods reduce the data dimensionality, and can be used to visualize the spatial organization of the dataset, as a preprocessing to escape from the curse of dimensionality phenomenon [Don00, AHK01] or to embed data from a metric space to a Euclidean vector space. Mapping methods can be used to unfold the dataset’s underlying manifold. In the sequel, ”items” denote original data and ”data points” denote their mapping into the projection space.

Since Torgerson’s ”embedding theorem” [Tor52], distance preservation is the objective of most of mapping methods. Indeed, the goal of Principal Component Analysis (PCA) [Jol02, Pea01] is equivalent to look for the linear projection that preserves Euclidean distances at best in terms of the mean-square-error criterion. PCA belongs to the family of linear projection techniques such as Projection Pursuit [FT74] and the Grand Tour [Asi85]. The Torgerson’s Classical MDS consists in mapping data points from a distance matrix. Both the PCA of a set of normalized vector items, and the Classical MDS of the Euclidean distance matrix of these same items, provide the same set of data points up to an isometry [Gow66].

In that framework, many methods have been designed to especially account for small distances, leading to nonlinear mappings. There are a large number of such methods known as nonlinear multidimensional scaling (NL-MDS).

In this work, we focus on some typical NL-MDS methods which are prone to exhibit specific types of mapping distortions, providing various enough examples to demonstrate our claims. Sammon’s Non Linear Mapping (NLM) [Sam69] and Demartines et al. Curvilinear Component Analysis (CCA) [DH97] minimize the difference between distances in the input and projection spaces, weighted by a decreasing function of the input or projection distances respectively. Lespinats et al. Data Driven High-Dimensional Scaling (DD-HDS) [LVG*07] makes a combination of both NLM and CCA weighting functions.
Tenenbaum et al. ISOMAP [TSL00] computes the geodesic distance matrix of the data estimated by the length of shortest paths [Dij59] in the K-nearest neighbour graph of the items, and uses Classical MDS to map this distance matrix. Notice that NLM, CCA and DD-HDS mapping results depend on the initial position of the data points often set randomly, while PCA and ISOMAP provide a unique mapping solution.

1.2. Mapping distortions

Several recent papers [Aup07, LVG*07, VK06] highlight two types of distortions which may occur when mapping items. A “false neighbourhood” occurs when a large distance in the original space becomes a small distance in the projection space (data points are neighbours whereas respective items are not). Reciprocally, a “tear” occurs when a small distance in the original space becomes a large distance in the projection space (true neighbours are mapped far apart). Tears and false neighbourhoods define two elementary and complementary types of distortions formally distinguished by the sign of the difference between the items’ pair-wise original distance and their respective data points’ projection distance. Both these types are sufficient to describe the full spectrum of possible distortions involved by the difference between two distances. However, a single data point having several neighbours defined up to some scaling parameter, may involve both types of distortions simultaneously (see Figure 1C).

1.3. Why visualizing distortions is necessary

In general, mapping distortions cannot be avoided because the topological structure of the items does not match necessarily the topology of the projection space, e.g. mapping a sphere onto a plane. Eventually, the mapping method itself is not a good one, or gets stuck into a local optimum of the stress function it attempts to optimise. So tears and false neighbourhoods are likely to be present without being noticed as shown in the figure 2. The main problem is that false neighbourhoods tend to collapse

Figure 1: Examples of the different types of distortions relative to the item marked as a star, whose original true neighbours are depicted as triangles. The large circle corresponds to the range of influence of the star item (see section 3.1 for more details). The original data (top left) mapped with a tear (original neighbours mapped far apart) (A), with a false neighbourhood (mapping neighbours originally far apart) (B) and with both types of distortions at the same place (C).

Figure 2: Several 2-dimensional mappings of the same originally 3-dimensional data not displayed on purpose. What really can we infer about the original structure of the items based solely upon these single mappings?
actually distinct structures while tears tend to separate actually connected structures, introducing mapping discontinuities. Therefore when such distortions occur, unveiling distorted areas should be a primary goal to avoid overinterpretation and wrong inference about the original items’ structure.

Indeed, a way to give a global “scale”, i.e. a meaningful link to known information, to the data points’ relative positions, is to project into the same space the unit vectors of the original space, but this is only possible for items given as elements of a vector space. Even in such a case, as far as the projection is nonlinear, the axes of the projection space have no meaning, no straightforward relationship with the original vector space because a unit item vector in the original space projects in a visually unpredictable manner onto the map depending on its original location.

However, as in the figure 2, many users do not take care about “scaling” the projection space even when they can. As a consequence, when clusters of data points eventually appear, their existence in the original space is taken for granted without measuring how much their cause is distributed throughout the original item structure which is intended to be discovered, and the mapping itself which is prone to distortions.

Linking the views is not a relevant way out because as far as no scale is provided, there is no more sense in a map obtained with any state-of-the-art nonlinear mapping method than in a map obtained using random scattering of the data points. One would not consider useful linking a random scatter plot to another one nor to any insightful scaled projection, in an attempt to extract new information. So is the case with nonlinear mappings. To make this issue even more conspicuous, we invite the reader to convince himself/herself that the naked scatter plot itself he/she is used to looking at, bears objectively no more information about the original items’ structures than a random projection despite whatever point patterns his/her eyes is prone to see.

Here we make this issue more formal. A mapping transforms the original (N,N) distance matrix D into an estimate D’ of it under topological constraints (connectedness and dimensionality) driven by the projection space. The projection space displays D’ without distortion (up to the rendering and perceptual distortions which may occur beyond that stage in the visualisation pipeline) as a scatter plot of N data points. Thus, visually inferring some information about D from this scatter plot showing D’ is idle if any minimal information about D, like the estimation error matrix (D-D’ ) is not provided too. But displaying the elements of this error matrix where they should be so, i.e. together with the elements of D’ onto the map itself, is still a big visualisation challenge. Indeed, as pointed out by one of the authors [Aup07], there is N(N-1)/2 error values to display in the vicinity of N data points only, a setting which has not been discussed by Wittenbrink et al. [WPL95] when they defined the Verity Visualisation framework. The visualisation methods which belong to this framework visualise together in the same space both the data and their uncertainty, providing a consistent and self-contained view of the data. In the present work the uncertainty does not deal with the absolute position of each single data point but with the relative distance between each pair of data points.

This work is an attempt to take up this challenge. We describe CheckViz: a new method to check visually the local quality of mappings and to characterize distortions when they exist. In this way, the mapping becomes usable as it makes inference possible from the projection space about the original structure in areas where the mapping is shown to be clear from any distortions.

1.4. Paper outline and contribution

In section 2, we review the state of the art methods to visualize mapping distortions. In section 3, we present the CheckViz method. We compare the CCA and NLM methods, both of which map data points onto a plane while attempting to preserve their relative distances. We derive two criteria to measure the tears and false neighbourhoods of local mapping distortions from their respective stress functions. Then we propose CheckViz as a visualisation technique to display these criteria straight onto the map, exactly where they must be shown to make the map usable and useful for making inference about the original items’ structures. We also provide two additional inference rules to help the user to read this representation. At last, in section 4, we apply this method on synthetic and real datasets to demonstrate our claims.

Our visualisation method is the combination of four components:

- we propose to show the local mapping distortion onto the map itself;
- we propose to use the pair of complementary indicators based on the NLM and CCA stress functions presented in [LA09], to characterize for each data point local false neighbourhoods and tears respectively;
- we propose to use a single 2-dimensional colour map sliced in the perceptually uniform CIELab 3-dimensional colour space to show the pair of distortion values simultaneously. The colour of some region around each data point location is given by the position in this colour map of a point having as coordinates the local false neighbourhoods and tears criteria of this data point;
- and following many authors [Aup07, BD05, NSB04, RMC04, PFM*06], we use the Voronoi cell of each data point as the region to colour to make visible the values of the local distortions.
This whole combination is our original contribution both in terms of the informational and graphical parts. On the informational side, this work is the first one to propose to display simultaneously the local pair of complementary distortion indicators at each data point onto the map itself. On the graphical side, both indicators of a data point are mapped to a bicolour table and the resulting colour is set to its Voronoi cell. As a by-product, this work makes the MDS maps fitting within the framework of Verity Visualisation.

2. Related work

2.1. A practical point of view

In practice, as data miners using nonlinear multidimensional scaling methods and aware of their potential mapping distortions, we must answer two major questions:

- Which mapping method displays the best map of our data?
- In which part and how much can we trust the resulting map?

The former question deals with global mappings quality evaluation, the latter with local evaluation. In this section, we display in the figure 3 some of the state of the art evaluation methods based on the CCA and NLM mappings of the same data as in the figure 2.

2.2. Global quality assessment

Many methods like in [Sam69, DH97, LGV*07, VK06] attempt to find the position of data points in the mapping in order to optimise some stress function. The optimal stress value reached can be used to evaluate the global mapping quality and to compare several mappings obtained optimising the same stress function. However it is unfair to compare two different methods based on the stress function of one of them, so more general criteria have been devised to tackle this problem. For instance, Venna and Kaski [VK01] propose the trustworthiness and continuity rank-based criteria to measure global mapping distortions, but many other criteria have been devised too (See [Ven07] for a survey). However, the stress value or such general criteria are single or pairs of numbers which summarize very crudely all the mapping distortions contained in the (D-D^*) matrix, and do not answer the where and how much questions. In the sequel, we focus on evaluation methods based on visualisation.

The Shepard diagram [Kru64, DH97] is such a model-free summary which allows visual comparison of mappings. Each pair of items is displayed as a single point in a 2-dimensional scatter plot, where the x-axis shows the pairwise distance in the original space and the y-axis the pairwise distance in the projection space. All the points lying on the diagonal show a perfect mapping, while points below and above the diagonal show false-neighbourhoods and tears respectively. The number of points to display being proportional to the square of the number of items, Lespinats et al. [LVG*07] propose to use a density plot as shown in the figure 3, when this number is too large.

The Shepard diagram displays all the possible pairs of items so it shows all the possible mapping distortions (D-D^*) at once. However it shows the distortions in a separate view which makes it difficult to link these distortions with the evaluated mapping, because it displays N(N-1)/2 points, one for each pair of items, while the mapping displays N points, one for each single item. Thus it cannot tell

![Figure 3: Four quality assessment methods tested on projections by CCA [DH97] and Sammon’s mapping [Sam69]). The lower the distortions, the less the stress values (E_{CCA} and E_{NLM}), the closer to the diagonal of the Shepard diagram [Kru64, DH97], the brighter the CCA and NLM pressures [LA09], the more disentangle the geodesic links, and the smoother the proximity mosaic [Aup07].](image-url)
where and which mapping distortions occur straight onto the map, making the map still difficult to interpret.

2.3. Local quality assessment

In practice, the local quality of nonlinear mappings can be very heterogeneous. However, few methods have been proposed to visualize this local quality exactly where it makes sense, i.e. straight onto the map.

When using projection methods which build a graph to estimate geodesic distances, e.g. ISOMAP [TSL00] or Curvilinear Distance Analysis [LLV04], it is usual to draw in the projection space the links which connect neighbouring items in the original space. It is applied for Self-Organizing Maps as well by Tasdemir and Merényi in [TM09]. The links connecting data points far apart show false neighbourhoods. However, these methods are prone to the “hairy ball” problem frequently occurring in dense graph drawing: too many “tears” links hiding areas with no distortions, and too many “false neighbours” links being difficult to disentangle to make clear at first glance which data are really neighbours in the original space and which are not.

Another method called proximity measure is proposed by Aupetit in [Aup07]. It consists in showing the original neighbourhood of an item whose data point is selected by the user. The Voronoi cell of each data point is coloured with respect to the original distance between its respective item and the selected one. This method shows the original neighbourhood of an item as an image made of a grey-scale mosaic of Voronoi cells, rather than as a set of links. It conveys a continuous distance-based measure of neighbourliness instead of the above discrete graph-based one, and makes clear at first glance all tears and false neighbours involving the selected data point. However, it focuses on a single data point at a time, so the user has to explore the neighbourhood of each of the N items by selecting each data point one by one. Colouring Voronoi cells is proposed by many authors [Aup07, BD05, NSB04, RMC04, PFM*06] and alternative representations like heat maps are discussed in [SLB10] and [SSK10]. Aupetit also suggests in [Aup07] a global view to help focusing on the data points gathering most of the distortions, by colouring Voronoi cells with a grey level proportional to the accumulated tears and false neighbourhoods of each data point. Hermann et al. [HGK09] propose to display distortions measured through the Kruskal stress with a colour coding of the data points. The Kruskal stress essentially emphasizes false neighbourhoods. Similar proposals are studied in [SLB10] and [SSK10] with other stress functions. Anyway, these four approaches which display a single value for each data point cannot show when both tears and false neighbourhoods occur simultaneously at the same place. Finally, Lespinats and Aupetit propose in [LA09] two separate views of the mapping distortions, one for the tears and another one for the false neighbourhoods measured with criteria based on CCA and NLM stress functions respectively. However, these two separate views do not allow seeing at a single glance all the distortions. Some work exists to visualize SOM distortions, which might be applied to continuous nonlinear mappings as well. Kaski et al. [KVK00] propose a similarity colouring to display with perceptually similar colours neurons of the map which are neighbours in the original space. This is intended to provide a global view of the tears and false neighbourhoods, but the colouring is obtained by unfolding the SOM in an auxiliary perceptually uniform 2 dimensional colour space, making the colouring itself prone to mapping distortions. Davidson et al. [DWB01] propose to analyse the stability of clusters obtained through mappings initialized at random. The only way to make sense out of this analysis is to link the views to check whether a cluster in one view exists or not in the others. Thus a single view is not self-content, as it must be linked to others to make eventually the truth coming out. This process is misleading as not a single view makes sense for itself, and as stated above, linking views without scale is unlikely to produce meaningful conclusions. Correa et al. [CCM09] propose to visualize uncertainties using sensitivity analysis, but problems come with the amount of perturbation to use and the difficulty to apply this method to nonlinear mappings such as ISOMAP, NLM, CCA or DD-HDS which do not provide an explicit continuous function from the original space to the projection space as PCA does. Warning et al. [WGD*00] compare the original and projection areas of triangles resulting from a Delaunay triangulation of the data points. This distortion measure is difficult to handle because triangles with very different shapes may still have the same area. Many works have been done in the cartographic visualisation community (see references in Brainerd et al. [BP98]) such as the one of Brainerd et al. themselves who propose to visualise mapping distortions by displaying how a circle lying in the sphere is distorted while being mapped onto the plane. The projected circle can be moved across the map to show how the distortions evolve from place to place. However, these approaches are specific to the projection of a sphere onto a plane, so the extension to original spaces with intrinsic dimension higher than 2 and no analytical expression of their topological structure is not straightforward. They are also difficult to apply when projection is defined as the solution of an optimization problem where no continuous mapping is defined between original and projection spaces.

In this work, we use the idea of colouring Voronoi cells and we put in a single view both tears and false neighbourhoods’ views proposed in [LA09], by designing a 2-dimensional perceptually uniform colour map.

3. A colour table to exhibit mapping distortions
3.1. A pair of criteria based on NLM and CCA

The purpose of NLM is to find the positions for data points in the projection space, which minimize the following stress function:

\[ E_{NLM} = \sum_{i,j} \left| \left| d_{ij} - d_{ij}^* \right| \right| \times F\left( d_{ij} \right) \]

where \( d_{ij} \) and \( d_{ij}^* \) represent the original distance (matrix D) between items \( i \) and \( j \) and the Euclidean distance (matrix \( D' \)) between data points \( i \) and \( j \) in the projection space respectively, and \( F \) is a monotonically decreasing weighting function designed to emphasize small original distances. While NLM fairly penalizes tears, it lightly penalizes false neighbours. Indeed, let us suppose that \( d_{ij} \) is large but \( d_{ij}^* \) is small, so \( i \) and \( j \) are false neighbours, then \( F(d_{ij}) \) is low and the difference between \( d_{ij} \) and \( d_{ij}^* \) does not give much weight to \( E_{NLM} \). Thus, as noticed by Sammon [Sam69], NLM is prone to false neighbours.

The CCA [DH97] stress function is close to the NLM one but the weighting function \( F \) relies on the distance in the projection space rather than in the original space:

\[ E_{CCA} = \sum_{i,j} \left( |d_{ij} - d_{ij}^*|^2 \right) \times F\left( d_{ij}^* \right) \]

Now false neighbours are fairly penalized, but tears can easily occur: if \( d_{ij} \) is low and \( d_{ij}^* \) is high, \( F_{CCA}(d_{ij}^*) \) is low. As noticed in [DH97], CCA is prone to tears.

We proposed in [LA07] to define the pressure at data point \( i \) as the part of the stress function which involves only this data point:

\[ P_{NLM}(i) = \sum_j \left( |d_{ij} - d_{ij}^*|^2 \right) \times F(d_{ij}) \]

\[ P_{CCA}(i) = \sum_j \left( |d_{ij} - d_{ij}^*|^2 \right) \times F(d_{ij}^*) \]

The stress is the sum of the local pressures. The more there are false neighbours or tears, the larger the value of \( P_{CCA} \) or \( P_{NLM} \) respectively. So we propose to use the pair of indices \( (P_{CCA}, P_{NLM}) \) to quantify locally both mapping distortions.

We propose to use for \( F \) a Heaviside step function with the parameter \( \sigma \) : \( F_{\sigma}(x) = 1 \) if \( x < \sigma \) and \( F_{\sigma}(x) = 0 \) otherwise. Notice that \( P_{NLM} \) is similar to \( P_{CCA} \), up to an interchange between original and projection distances as input to the \( F \) function. If original and output distances are similar, which the mapping enforces, the two indices have a low value and the same order of magnitude, and otherwise at least one of these two indices has a high value, which is a clue that the mapping failed.

3.2. A two-dimensional perceptually uniform colour table

Colouring data points so as to display local information on the mapping is not new. See for example [PFM*06]. This is easy to set up when dealing with a 1-dimensional index using for example a grey-scale. However, in our setting, we consider a pair of indices \( (P_{CCA}, P_{NLM}) \): the first one accounts for false neighbourhood and the second one for tears. We want to display both indices simultaneously to render the fact that both distortions may occur at the same place as shown in the figure 1, so we project the data points onto a two dimensional colour map using this pair of indices.

Some rules to select good colour maps are given in [R99]. It appears that perceptually uniform two-dimensional colour map is suitable to recover the pair of parameters from the colour they encode. A perceptually uniform colour space as defined in [CIE86] is such that if the pair of criteria \( P_{NLM}(P_{CCA}(i), P_{NLM}(i)) \) is twice as close to \( P_j \) as to \( P_k \), then the data point’s colour \( c_i \) is perceived twice as much closer to \( c_j \) as to \( c_k \). The data provided by G. Hoffmann [Hof09] shows perceptually uniform colour maps. We slice a square coloured with white and black at two opposite corners, and being oriented in the 3 dimensional CIELab colour space to get a set of opposite hues on both the other corners. We use the transformations between RGB and CIELab colour spaces based on [ITU90] using the D65 white point reference which was implemented as a matlab code by Ruzon [R09]. The square in the 3 dimensional CIELab colour space has corners with the following coordinates: (100, 0, 0) for “white”; (65, -30, 20) for “green”; (30, 0, 0) for “dark grey” and (65, 30, -20) for “purple”. A slice with yellow-blue opposite hues instead of green-purple can be used as well to comply with colour blindness.

This 2 dimensional colour map is shown in the figure 4. In this colour coding, pure false neighbourhood areas appear as purple and pure tear areas as green. A white area means an area “clear” from any distortion, and a black area means a “dark” area where no structure is faithful.
3.3. Colouring Voronoi cells

As proposed in several works [Aup07, BD05, NSB04, RMC04, PFM*06], we choose to colour Voronoi cells rather than data points themselves for the following reasons:

- The background is visually expected to bear contextual information about the foreground. So Voronoi cells cover the plane without overlapping and fill the whole background of the data points with colour encoding the faithfulness of the foreground scatter plot.

- According to the Gestalt theory [War04], colouring the data points’ markers would induce similarity relation between points with similar colour while these points are not necessarily close with respect to their original distance. Using different graphical supports (Markers’ relative position and Voronoi cells’ colour) for different variables (estimated position and distortion respectively) is likely to avoid this mismatch. Moreover, it allows releasing the colour variable of the data points’ markers which then can be used to encode a more familiar variable like the class label of the data point.

- Each data point is the perceptually nearest one among all the data points, to any point of its Voronoi cell. So anywhere we look at, the colour we see is the one of the closest data point, i.e. the data point to which we would have “naturally” assigned the colour of the point of focus.

- Voronoi cells adapt their size to the local data points’ density, avoiding the clutter problem of fixed-size markers.

However, Voronoi cells could also lead to visualisation bias. The size of cells depending on the density of data points, the attention of the analyst might be attracted towards data points with larger cells. But the data points being displayed at the foreground also create patterns which seem to catch the visual attention at first. In our previous work [LA09] we only coloured the data points themselves and felt the result less effective at catching the amount of local distortions at a glance (see figure 3).

Another issue concerns the nearest-neighbour interpolation visually induced by the use of Voronoi cells. Here the vacuum between the data points is more than two-dimensional, i.e. the value which is displayed at any point in the projection space is potentially the one to give to the whole multidimensional subspace of the original space which would be mapped at that point. So we must emphasize that the Voronoi cells are not meant to infer the distortion level in vacuum areas through interpolation but as a mean to make visible this distortion level by spreading its respective colour “perceptually at best” around each data point.

At last, there is still a possibility for “light contrast illusion” [CSS89] because no reference colour is provided between the cells, but it is probably not an issue here because the colouring is used to guide the analyst during his exploration rather than to provide very precise distortion level at each location. The overall visualization process is given in the figure 5.

3.4. Inferring from coloured maps

The colouring makes the map distortions visible. It also allows inferring structures of original data from the map. Indeed, a bright background in some area reveals that the observed structures in this area can be trusted. A green

![Figure 4: The two-dimensional colour-scale. Each data is related to a colour according to \( P_{CCA} \) and \( P_{NLM} \).](image)

![Figure 5: The distortions visualized straight onto the map. Data are mapped into a 2-dimensional space using ISOMAP (A). Voronoi cells are subsequently coloured (B) according to a colour-code displaying \( P_{CCA} \) and \( P_{NLM} \) (C).](image)
background reveals that the structure has been torn and a purple background reveals that some distinct structures have been overlapped at the same place. Both types of distortion happen simultaneously in dark background areas.

Despite that structures observed in dark areas are not faithful; some conclusions can be drawn from green and purple areas:

- **True Separation Rule**: If several purple structures are disjoint on the map, then up to the scale $\sigma$, the gap between them is true in the original distance matrix $D$ and so is the observed partition (see left insert in the figure 6). Otherwise, it would correspond to a tear and associated cells would be green or dark. Nevertheless, each of the observed structures within purple areas artificially connects originally distinct ones or is denser than the original ones.

- **True Overlap Rule**: If several green structures are placed side by side or overlap, then up to the scale $\sigma$, the side by side location or the overlap is real in the original distance matrix (see right insert in the figure 6). Nevertheless, the observed structures within green areas are part of an original one artificially disconnected or are less dense than the original ones.

It must be emphasized that the term “structure” refers to the value of the $\sigma$ parameter which defines the radius of the ball centred at an item and its respective data point, in which both original and projected neighbours respectively are considered.

3.5. Setting the parameters

The parameter $\sigma$ is the most important to be set. It provides the scale at which the data points and items are considered to form a structure. Parameters with the same meaning are used in many nonlinear MDS methods (see for examples [DH97], [VK06]) so it is quite natural to tune $\sigma$ to the same value as the one used within the mapping method itself. Otherwise, the value of $\sigma$ can be based on some criterion derived from the distribution of all the distances in the original space or can be set interactively by the analyst. Normalisation of the distances in original and projection spaces could also be applied enforcing the sigma scale to be similar in both spaces thus masking any global overall distance stretching or compression. Here we use a “rule of thumb” setting the $\sigma$ value to the average distance between each item and its $5^{th}$ nearest neighbour in the original space. Experiments about how this parameter affects the results are provided in the section 4.2.

The maximum value for the colour scale can also be set to a limit to prevent some outliers from masking less important but still existing mapping distortions in other areas. But in any case, a similar maximum colour key value must be chosen for $P_{CCA}$ and $P_{NLM}$.

4. Experiments

4.1. Datasets and experimental setup

The proposed mapping quality visualization is tested on three datasets. The first one is our guiding example: a unit empty cubic box with one face missing sampled with a regular grid of 193 items in the 3-dimension real space. The second one is the “Oil Flow” synthetic dataset [Sve99] obtained from a physical simulation of oil flow in a pipeline. It is a set of 1000 items in the 12 dimensional real space, partitioned into 3 classes (in the present case, we use a subset of 500 randomly drawn data). The third one is a real dataset called “Optical Recognition of Handwritten Digits” obtained from the UCI Machine Learning Repository database [AN07], containing 5264 items in the 64 dimensional pixel space of the 0 to 9 digits (in the present case, we use a subset of 200 data equally distributed between the classes).

The local quality is observed through the pair $(P_{CCA}, P_{NLM})$. We use the PCA, NLM, CCA, ISOMAP and DD-HDS mapping methods with defaults parameters provided in the original publications. For ISOMAP, we build the 5 nearest neighbours graph. For CCA and NLM stresses and pressures, the function $F$ is given at the end of section 3.1 and distances are Euclidean. The default value for $\sigma$ is set to $2.5$ ; 0.5 and 35 for the guiding example, the Oil Flow and the Digit data respectively, it is the same value as the one which was used to optimise the stress function of NLM, CCA and DD-HDS in order to get the maps shown in this work.

The first experiment based on the guiding example dataset, is intended to show the influence of the parameter $\sigma$ and the basic properties of the colour coding we propose. The second and third experiments present our method facing multidimensional synthetic and real data. Both of them are intended to demonstrate that our method is really necessary to make sense out of the mapping.

We emphasize that we do not intend to show whether one mapping method is worse or better than another one,
but rather to show that without displaying distortions, the mappings cannot be trusted, they are not usable for inference, whereas using the CheckViz method they become usable and so may be useful.

4.2. The cubic open box guiding example

Items lie onto the faces of a three-dimensional open box (figure 9). Despite the fact that these data lie onto a 2-dimension manifold, there is no way to project them onto a plane preserving all the distances. This example allows us to show the main distortions of mapping methods. As expected, CCA generate only tears, PCA, ISOMAP and NLM generate only false neighbourhoods, and DD-HDS provides the most faithful map of these data.

The DD-HDS mapping provides a faithful map of the original structure because every Voronoi cells are almost white. Data points within the shape shown by CCA mapping are brighter while its borders are clearly torn: two lateral faces of the cubic box have been ripped out of each other through the mapping. Two opposite faces of the box overlap in the NLM map while PCA and ISOMAP show the cubic box projected along an axis orthogonal to its bottom face.

The values for $P_{NLM}$ and $P_{CCA}$ depend on the value of the parameter $\sigma$: the lower the value, the smaller the considered neighbourhood. If $\sigma$ is lower than the smallest distance, each pressure is null and no distortion can be detected. Conversely if $\sigma$ is greater than the largest distance, every pressure is maximal. In the figure 7, we set $\sigma = 2.5$; in the figure 8, two other values are tested. When $\sigma = 1.5$ (lower than the default value), colours are brighter and some distortions are understated, such as at the border of torn corners of the bottom side in the CCA map. When $\sigma = 3.5$ (higher than the default value), colours are darker and distortions are reported even in fair areas close to distorted ones: false neighbourhoods are detected in the centre of the bottom face in the PCA map and green colour floods the central area in the CCA map.

In the sequel we set sigma to a value which makes visible both types of distortions.
4.3. The Oil Flow data

Three classes correspond to three different relative positions of oil, water and gas in pipes: green squares for “stratified”, blue circles for “annular” and red triangles for “homogeneous”.

Looking at the maps without colouring (top of the figure 10), the red class seems to be disconnected but it is not clear in how many parts, and green and blue classes seem to overlap (PCA and NLM) or to be disconnected (DD-HDS and CCA) except onto the ISOMAP projection, so where is the truth?

When adding the colouring (bottom of the figure 10), the PCA insure that some points of the red class are really separated from the remaining points (True Separation Rule). Conversely, ISOMAP shows an interesting clue despite the large amount of distortions: blue and green data are separated by a bright area. Moreover, blue and green data are not involved in tears because they lie in a purple area. Therefore blue and green data are separated in the original space (True Overlap Rule). CCA and DD-HDS allow observing that red data are faithfully spread over nine clusters around blue and green data because they belong to white areas in these maps. Moreover, the contact between some green and blue points is faithful (True Overlap Rule) in CCA.

We obtain similar conclusions about the data structure as the one provided in [GAG08] using a topological analysis based on a statistical generative model.

4.4. The handwritten digits data

While the maps without colouring (top of the figure 11) lead us to think that each set of digit forms a distinct cluster structure in the original space, the decorated maps enjoin us not to be so affirmative.

Indeed, few can be said here from ISOMAP and NLM. However, the PCA map tells that the separation between digits 0, 2 and 4 is faithful (True Separation Rule), and several bright areas in CCA and DD-HDS maps show that the digits 0, 2, 6 and 7 form distinct clusters in the original space. The CCA also tells us (True Overlap Rule) that the contact between digits 3 and 5, and 5, 8 and 9 are
faithful. But it is still difficult to conclude about the original structure of the other classes.

5. Conclusion

Maps are used to show multidimensional patterns or to drill down into multidimensional data sets. However, there is no visible difference between a totally random map and one which minimizes some distortion measure. Moreover, non-specialists tend to believe whatever pattern they see with the only principle in mind that MDS or others maps are faithful black-boxes provided by their Computer Science department. Here we claim that the quality of the map must be checked thoroughly before it can be used safely or not used at all.

We proposed the CheckViz method to evaluate the mapping quality at one single glance. The CheckViz method acts as a sanity check to display as bright areas the location where inference can be made safely, and as a tool to infer some topological properties of the original structures using the True Separation and True Overlap rules. We showed on toys and real data that without this kind of quality visualization, maps cannot be trust; they are not usable for inferring the original items’ structures. As a by-product, the CheckViz method makes the MDS maps fitting within the framework of Verity Visualisation [WPL95].

We think it could be useful to draw in the two-dimensional colour map the points used to give a colour to the data points, showing at a glance the overall distribution of the mapping distortions (as presented in figure 12 with the open box dataset). We also think about applying CheckViz to other information mappings, such as tree-maps [JS91], and in fact CheckViz applies whenever original multidimensional data are displayed as point clouds as far as some similarity matrix can be computed in both original and projection spaces. It applies to Self-Organizing Maps [Koh88] as well considering nodes as data points.

CheckViz paves the way for a renewed viewpoint on exploratory data analysis based on Multi-Dimensional Scaling techniques and other projection-based approaches widely used in the Information Visualisation domain.

Acknowledgments

We are very grateful to the anonymous reviewers for their truly helpful comments which allowed us to improve the quality of this work.

References


Figure 12: Distribution of the items in the 2-dimensional colour table for the five open box mappings presented in the figures 2 and 7 (σ is set to 2.5).


[Pea01] Pearson K., On lines and planes of closest fit to systems of points in space, Philosophical Magazine n°2, pp. 559-572, 1901.


[TM09] Tasdemir K., Merényi E., Exploiting data topology in visualization and clustering of Self-


