DOI: 10.1111/j.1467-8659.2011.01961.x

Eurographics / IEEE Symposium on Visualization 2011 (EuroVis 2011)
H. Hauser, H. Pfister, and J. J. van Wijk
(Guest Editors)

Volume 30 (2011), Number 3

Visualizing High-Dimensional Structures by Dimension Ordering and Filtering using Subspace Analysis

Bilkis J. Ferdosi[†]

Jos B.T.M. Roerdink[‡]

Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen

Abstract

High-dimensional data visualization is receiving increasing interest because of the growing abundance of high-dimensional datasets. To understand such datasets, visualization of the structures present in the data, such as clusters, can be an invaluable tool. Structures may be present in the full high-dimensional space, as well as in its subspaces. Two widely used methods to visualize high-dimensional data are the scatter plot matrix (SPM) and the parallel coordinate plot (PCP). SPM allows a quick overview of the structures present in pairwise combinations of dimensions. On the other hand, PCP has the potential to visualize not only bi-dimensional structures but also higher dimensional ones. A problem with SPM is that it suffers from crowding and clutter which makes interpretation hard. Approaches to reduce clutter are available in the literature, based on changing the order of the dimensions. However, usually this reordering has a high computational complexity. For effective visualization of high-dimensional structures, also PCP requires a proper ordering of the dimensions.

In this paper, we propose methods for reordering dimensions in PCP in such a way that high-dimensional structures (if present) become easier to perceive. We also present a method for dimension reordering in SPM which yields results that are comparable to those of existing approaches, but at a much lower computational cost. Our approach is based on finding relevant subspaces for clustering using a quality criterion and cluster information. The quality computation and cluster detection are done in image space, using connected morphological operators. We demonstrate the potential of our approach for synthetic and astronomical datasets, and show that our method compares favorably with a number of existing approaches.

Categories and Subject Descriptors (according to ACM CCS): Information Search and Retrieval [H.3.3]: Clustering—; Computer Applications [J.2]: Physical Sciences and Engineering—Astronomy; Computer Graphics [I.3.6]: Methodology and Techniques—Interaction techniques

1. Introduction

High dimensionality is becoming a common feature of modern scientific datasets, such as astronomical data, gene expression data, etc. However, it is far from straightforward to visualize high-dimensional structures in a meaningful and user-interpretable way. Traditionally, low-dimensional representations of high-dimensional spaces, obtained by methods such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), etc., are used to perform visualization in a Cartesian coordinate system. Other methods to visualize high-dimensional data are the scatter plot matrix (SPM), the parallel coordinate plot (PCP) [Ins09], or tours [Asi85].

All of the above methods have shortcomings. The use of PCA, MDS etc., poses the problem of interpretation of the visualization, because of the transformation of the original feature space to a new coordinate system. Tours suffer from a similar problem. Pairwise relationships among dimensions can best be observed with SPM. However, if the number of dimensions is very high, it suffers from crowding and may become difficult to interpret. PCP does have the potential for visualization of high-dimensional structures in the original

[†] e-mail: b.j.ferdosi@rug.nl

[‡] e-mail: j.b.t.m.roerdink@rug.nl

feature space, as it does not have constraints on the number of dimensions that can be visualized at a time. However, to facilitate the visibility of high-dimensional structures in PCP, it is necessary to obtain a proper ordering of the coordinate axes. High data-dimensionality can make manual reordering unfeasible, hence an automatic method is required.

In the literature, there exist several approaches for ordering and filtering the dimensions of multi-dimensional datasets [AEL*10, ABK98, Guo03, JJ09, PWR04, TAE*09, YPWR03]. However, all of these approaches obtain the ordering of the dimensions by considering pairwise relationships between the dimensions only. In this paper, we propose methods that search for *higher-dimensional* structures to obtain the dimension ordering. In addition, the ordering of the dimensions obtained also indicates the importance of the features in terms of clustering.

Subspace ranking (of full feature spaces) is the process of identifying relevant subspaces for (later) clustering based on some quality criteria [KKZ09]. The proposed method in this paper builds on the method presented in our earlier work [FBT*10], which finds relevant subspaces for clustering according to a quality criterion obtained using connected morphological operators (see section 3). In addition, this method can give an indication of the number of clusters present in each subspace without doing the clustering itself.

In this paper, we use the quality criterion and the cluster indication capability of the method in [FBT*10] to present three algorithms: two for finding a suitable dimension ordering for PCP, and one for SPM using only quality criteria:

1. Structure-based full (SBF) ordering for PCP.
2. Structure-based partial (SBP) ordering for PCP.
3. Structure-based simple (SBS) ordering for SPM.

For the SBF and SBP ordering our contribution is a better visualization of high-dimensional structures; for SBS our main contribution is a significant reduction of computation time.

The SBF ordering tries to find the ordering of *all* of the dimensions present in the dataset so that high-dimensional structures become visible. The method starts by finding the highest ranked 1D subspace. It continues to find the next dimensions in the sequence of reordered dimensions, based on the number of clusters present in the corresponding subspace and its quality value, until it has found a complete sequence for all d dimensions. In SBP, which is supplementary to the SBF ordering, the process of finding a dimension order is repeated for every dimension present in the dataset. In contrast with SBF, it does not try to find the order of all the dimensions, but extracts an ordering of subspaces of the full feature space. For SPM, the most important goal is to reduce clutter in the plot to achieve better visualization of high-dimensional data. To improve readability we can identify the cluster and noise dimensions, and then either remove the noise dimensions from the plot or put them all together at one side of the plot. The SBS method uses the capability of the method of Ferdosi *et al.* [FBT*10] of identifying the cluster and noise dimensions even from the 1D density plot. Next we apply an automatic or user-defined threshold to remove some of the low-quality dimensions to make the SPM visualization better readable. In addition to the automatic ordering we provide the option of user interaction for manual adjustment of the ordering.

2. Related work

Ankerst *et al.* proposed a method for arranging dimensions using pairwise similarity measures based on the Euclidean distance function [ABK98]. The arrangement of the dimensions is obtained using ant colony optimization [DG97], a global optimization method which only considers the pairwise relations of the dimensions. Thus the order in which dimensions with high-dimensional structure appear is a matter of chance. As will be shown in the results section, the pairwise method can reveal very simple structures, but fails when the structures are more complex. By contrast, our method considers local relationships of the dimensions. We not only use subspace quality but also clustering information when we add a new dimension to the sequence.

Guo proposed a human-centered exploration environment for high dimensional data [Guo03]. Both computational and visual measures are used to obtain the dimension selection and ordering. Maximum conditional entropy (MCE) is calculated in 2D data space as a measure of "goodness of clustering". The main difference with our method is that Guo considers only clustering in the 2D subspaces, whereas we take higher-dimensional clusters into account while obtaining the sequence of dimensions.

Peng *et al.* defined a clutter-based measure to rearrange the dimensions in such a way that a minimal number of outliers is present between two neighboring dimensions [PWR04]. For PCP, the proportion of outliers present in neighboring dimensions is used. The computational complexity of creating the outlier matrix of all the pairwise dimensions is $O(m^2n^2)$, where m is the number of data points and n is number of dimensions. The optimal dimension ordering is obtained using exhaustive search in $O(n \cdot n!)$ time. For SPM, two different measures are used. For the high-cardinality dimensions the Pearson correlation coefficient is used to obtain a clutter measure. The correlation matrix is obtained in $O(m * n^2)$ time. Then all the dimensions of similar correlation (above some threshold value) are searched in $O(n^3)$ time, and the optimal ordering is obtained in $O(n^2 \cdot n!)$ time by exhaustive search. The low-cardinality dimensions are sorted in descending order according to their cardinality value. In contrast to our approach this method is based on outliers instead of clusters. Therefore it can reduce the clutter but cannot ensure that the dimensions with d-dimensional clusters will be close to each other in the visualization.

To reduce the complexity of finding optimal arrangements

and improve interactivity, Yang *et al.* devised hierarchical dimension clustering using a similarity measure and a PCA-based importance measure [YPWR03]. Similar dimensions are joined together to form a dimension cluster. To handle large datasets they also extract data clusters using a bottom-up data clustering method. Only the data points in the clusters with extent much smaller than the minimum similarity value are used. However, the use of selected global clusters can restrict the finding of clusters which are hidden in subspaces. In contrast to this, our method searches for structures in subspaces, and dimensions are grouped together depending on their clustering structure and ordered according to their quality.

Tatu *et al.* [TAE*09] presented a method to rank scatter plots and parallel coordinate plots. For scatter plots they used rotating variance, class density, and histogram density measures. For parallel coordinates, Hough space, similarity, and overlap measures were used. To obtain the best PCP, the pairwise quality of the dimensions using Hough features is calculated first. Then, an algorithm to solve the Traveling Salesman Problem, such as the A^*-search algorithm, is used to obtain the optimal order. This method uses global optimization as in Ankerst's method, thus it neglects the local features. In addition, Hough-space quality computation may fail with very large datasets with a large amount of overlap.

Johansson and Johansson presented a dimension reduction system using a user defined quality matrix for correlation, outlier detection, and clustering [JJ09]. Pearson's correlation coefficient is used as a quality metric for correlation. For outlier detection a density and grid based approach is used. For computing quality, the Mafia clustering algorithm is applied [GC99]. In clustering-based dimension reduction interesting dimensions are selected based on cluster coverage. They also proposed dimension ordering of the reduced dimensions. The dimensions are ordered starting from the highest ranked cluster. Dimensions in that cluster are placed next to each other, removing any dimensions that do not belong to the reduced set of dimensions. This can result in a set of dimensions that contains some big clusters and can miss significant clusters with less coverage. Therefore, it would be more informative if the clusters with the associated dimensions were visualized with consecutive plots. By contrast, our method does not perform clustering. Instead, we use the quality of the subspaces and an indication of the number of clusters present in a subspace to obtain the ordering. In addition, our method targets subspaces with high-dimensional clustering instead of reducing the dimensionality of the dataset.

Albuquerque *et al.* proposed a parallel coordinate matrix (PCM) similar to a scatter plot matrix, a class-based scatter plot matrix and quality-aware dimension ordering for the proposed plots [AEL*09]. In PCM each row represents the relationship of a dimension j with all other dimensions, and each cell contains the relationship of j with two other dimensions. First the quality is computed for all 2D visualizations which are ranked according to quality value. Then two 2D visualizations are combined that share the main dimension of that row. Also the dimensions are ranked according to quality value. In quality-aware dimension ordering the quality of $(n-1)$ 2D visualizations of each dimension is used to compute the quality of every dimension. Then the dimensions are ordered according to quality values. Our proposed method bears similarity to Albuquerque's method in terms of using quality values for dimension ordering. However, our parallel coordinate plot can visualize relationships (in terms of clustering) among more than three dimensions, whereas the PCM in [AEL*09] can only show relationships among not more than three dimensions. Albuquerque's quality-aware dimension ordering has some similarity with our SBS ordering for scatter plot matrices. However, their method requires the computation of n^2 2D / 3D visualizations, whereas we obtain the SPM ordering using only n 1D density plots.

3. Dimension reordering

3.1. Overview of the method

Let us denote by *DATA* a set of N data points (rows) with d dimensions (columns), i.e., $DATA \subseteq \mathbb{R}^d$. Let $A = \{A_1, ..., A_d\}$ be the set of all attributes A_i of *DATA*. There may exist a natural grouping among these attributes that contains high-dimensional structures such as clusters. The goal is to make groupings so that such clusters are visible in PCP and SPM visualizations.

We present three approaches for dimension reordering, two for PCP and one for SPM, using the concept of subspace clustering and ranking. A subspace of *DATA* is a set S with $S \subseteq A$. Following the approach of [FBT*10], we rank the subspaces according to certain quality criteria. The quality of a subspace depends on the structures present. Emphasis is given to multimodality of the density distribution of the subspaces, where each density mode is indicative of a cluster. In addition, significance and separability of each mode contribute to the quality value. The search for the density modes and determination of significance and separability is performed in grey-level image space. Therefore, a transformation of parametric space to image space is required. This transformation is obtained by grid-based density estimation. Thus, modes in the distribution are transformed into high-intensity peaks (local maxima) in the density image.

To search for modes (local maxima) in the density image we use connected morphological operators, implemented using the Max-tree data structure [SOG98]. Each node of the Max-tree with a certain grey level contains all the connected components at that level. The root of the tree contains the connected components with lowest intensity and the leaves contain those with highest intensity. Therefore, counting the number of leaves gives us the number of clusters. The significance and separability of modes is determined using the

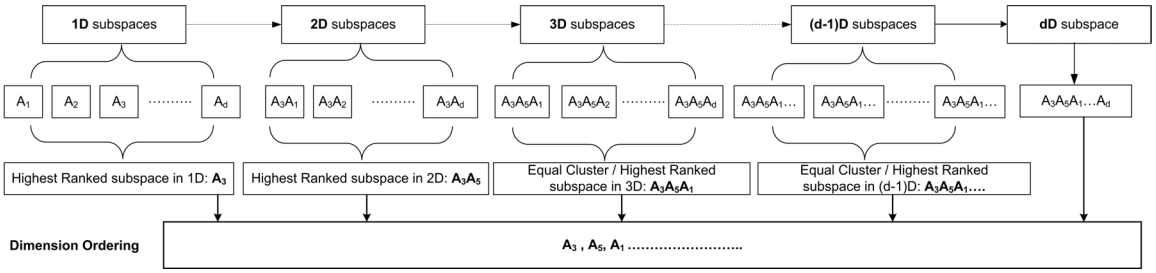

Figure 1: *Structure-based Ordering for PCP.*

concept of *relative dynamics*. In image analysis the concept of "dynamics" is used as a measure of contrast. It can be used to rank the local maxima of an image [Ber07]. Relative dynamics of a local maximum m is defined as

$$RelativeDynamics(m) = (H_1 - H_2)/H_1, \qquad (1)$$

where H_1 is the height of the maximum m and H_2 is the height of the deepest nearest minimum. Significant and well-separated modes will have a higher relative dynamics than overlapping clusters [FBT*10].

To derive a quality criterion for subspaces we use the number of modes (i.e., leaves in the Max-tree) and their relative dynamics as follows. The quality of a subspace S of the space of attributes, denoted by $Quality(S)$, is defined as:

$$Quality(S) = \begin{cases} N_L^{-1} \sum_{i=1}^{N_L} RD(i) & \text{if } N_L > 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where N_L is the number of leaves in the Max-tree and $RD(i)$ is the relative dynamics of local maximum i. The sum of the dynamics of all local maxima is normalized by the total number of local maxima; so the value of *Quality* ranges from 0 to 1. A subspace that contains modes/clusters with high dynamics will have a higher value of *Quality* than a subspace with clusters of lower dynamics.

Up to dimension three the creation of the density image, Max-tree construction and computation of the quality index is done in the original feature space. For subspaces of dimension higher than three we apply PCA and use the first three principal components for creating the density image and subspace ranking. The main reason is that for higher dimensions the current Max-tree implementation becomes prohibitive in terms of computing time and memory use.

In [FBT*10] we reported that the use of PCA has an effect on identifying the number of clusters, but not on identifying the important subspaces. For example, if a subspace has four clusters, say two very distinct and two overlapping clusters, then our method is able to find three of the four clusters in the space of the first three principal components of the original subspace. However, the use of PCA does not restrict us in finding subspaces with high-dimensional structures, so the dimension is not limited to 4/5 or 6 (see the results section). For example, consider a dataset with seven dimensions: a,b,c,d,e,f,g, where a,c, and f contain clusters and the others contain noise. The 4D subspace $acfb$ will always have a much higher quality than the 4D subspace $bdeg$, even if we only use the first three principal components, because, whatever method we use, noise as input will generate noise as output. However, if we compare subspace $aceg$ and $acfb$ the result will depend on many factors, such as the number and quality of the clusters, their separation etc.

3.2. Structure-based Full Ordering (SBF) for PCP

The process of subspace creation for reordering a d-dimensional dataset is depicted in Figure 1. In step 1 we compute the quality of all 1D subspaces and rank them according to quality value. The highest ranking subspace (A_3 in Figure 1) is chosen to appear first in the reordered sequence of dimensions. In step 2 we compute the quality of 2D subspaces, but only of those which include the highest ranking subspace from step 1 as one of the dimensions. The highest ranking 2D subspace thus found defines the second dimension in the reordered sequence. For example, in Figure 1 the highest ranking 2D subspace is A_3A_5, so the second dimension in the reordering will be A_5.

Next we consider subspaces of dimension three and higher. Now an additional constraint is applied which takes precedence over the quality values, as follows. If the sequence obtained so far has a number p of k-dimensional clusters, then the highest ranking $(k+1)$-dimensional subspace with p clusters will contribute the next dimension to the ordering. If there is no subspace that contains p clusters, then the $(k+1)$-highest ranking subspace will contribute the next dimension. For example, in Figure 1 the chosen subspace A_3A_5 in 2D has 4 clusters, whereas the highest ranking 3D subspace $A_3A_5A_4$ has 2 clusters and there are other 3D subspaces with 4 clusters, of which $A_3A_5A_1$ is the highest ranking one; therefore $A_3A_5A_1$ will be chosen to contribute the next dimension A_1 to the order. In this way we obtain a dimension reordering which provides a good view of the dataset that emphasizes the high-dimensional structures, if present.

3.3. Structure-based Partial Ordering (SBP) for PCP

It may happen that one feature (dimension) contributes to different clusters involving different combinations of features. Therefore, it is possible to obtain multiple partial orderings of the dimensions which are basically the subspaces of the d-dimensional space. The process of finding a partial dimension reordering is similar to the full d-dimensional reordering, except that the sequence creation process stops when no subspaces contain $q = p$ clusters. Then it restarts the same process to find another partial ordering with the next dimension in the 1D ranking. It repeats the process until all of the 1D subspaces are used as seed to produce partial orderings. Partial ordering is also helpful for datasets with a very large number of dimensions, since visualizing all the dimensions simultaneously will make the screen crowded and unreadable.

In this paper we obtain the SBP ordering using the 1D ranking. It can find the ordering of n subspaces where n is the number of dimensions in the dataset. However, it is possible to extend the algorithm to find more interesting subspaces by starting the SBP from other rankings than the 1D ones. The starting dimension can be automated or the user can choose any dimension to start with.

3.4. Structure-based Simple Ordering (SBS) for SPM

SBS is very simple and it is based on the quality computation of the 1D density images. However, a useful property of the approach in [FBT*10] is its capability of identifying noise dimensions and cluster dimensions even in 1D. This property can be explained with the monotonicity lemma stated in [AGGR98]: "If a collection of points S is a cluster in a k-dimensional space, then S is also part of a cluster in any $(k-1)$-dimensional projections of this space." So, although SBS is very simple it can find the informative dimensions with very low computational cost compared to other methods. In addition to separating cluster dimensions from noise dimensions, dimension filtering can be applied in this ordering depending on the quality values in order to visualize only the most important relations. In our approach an automatic threshold for filtering is set to the average quality of the 1D subspaces. The user also has the freedom to change the threshold. This filtering is most helpful for SPM of very high dimensional datasets.

4. Visualization

We implemented both SPM and PCP on the GPU, reusing and adapting the implementation from Blaas *et al.* [BBP08] developed for PCP. In our implementation the plots can now be created with a varying number of dimensions, with different orderings, and with more extensive interaction. For PCP we use a histogram-based approach, as described in section 4.1. The GPU implementation of SPM is particularly helpful, considering the fact that SPM needs to visualize a lot more than PCP. Even for a large number of dimensions the SPM computation is now quite fast.

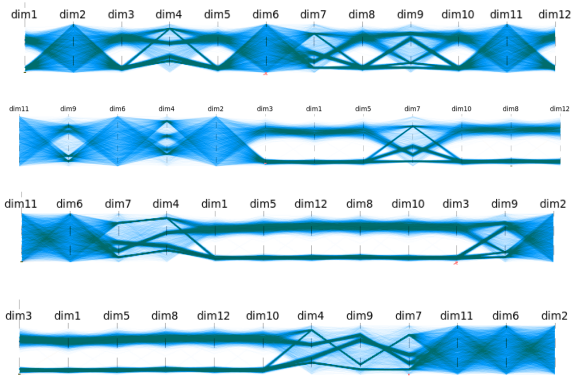

Figure 2: *Synthetic dataset 1 (see section 5.1.1). From top to bottom: (first) original ordering: rendered with the PCP version of Blaas* et al.*; structures are visible but high-dimensional structures are hard to identify; (second) the clutter based method of Peng* et al. *is unable to put the proper dimensions together to visualize the 6D subspace with two clusters and the 3D subspace with four clusters present in the dataset; (third) Ankerst's method: the two clusters in the 6D subspace are visualized properly but the more complex four clusters in the 3D subspace are missed; (fourth) ordering with our SBF method: dimensions are ordered in such a way that clusters in both the 6D and 3D subspaces are visible.*

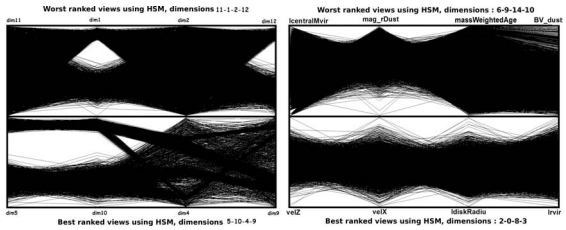

Figure 3: *The method of Tatu* et al. *(Left) Synthetic dataset 1. The first 4 dimensions, with the worst (top) and best (bottom) ranked visualization. In the best ranked visualization, the first two dimensions (dim5 and dim10) belong to the 6D subspace with two clusters and the others (dim4 and dim9) are two of the dimensions of the 3D subspace with four clusters. (Right) millMillennium dataset. The first four dimensions, with the worst (top) and best (bottom) ranked visualization. No structures can be observed in the best ranked view. (*Image courtesy: Tatu *et al.)*

4.1. Histogram-based Parallel Coordinate Plot

PCP in its original form prohibits the visualization of structures (such as clusters) if the number of data points in the

dataset is very large ($>$ 1000 data points). It may be possible to find structures with the help of brushing; however, discovering structures in such a way is very tedious and difficult.

Blaas *et al.* [BBP08] proposed an extension of PCP for very large datasets, using quantization and compression. They rendered the PCP on the GPU using the joint histogram of each pair of dimensions. Instead of drawing a line for each data point, histogram bins are used to draw the primitives. Then, additive blending is applied to combine all primitives. A logarithmic intensity scale provides good contrast between low- and high-intensity (density) regions. This method produces a smooth and continuous PCP and thus structures become better visible, even if the dataset is very large. However, in this approach the original ordering of dimensions of the dataset is used. Without reordering, it is hard to perceive high-dimensional structures, even in this PCP.

User Interaction. Even though automated dimension reordering can assist the user to analyze high-dimensional data and identify structures and grouping, it is always helpful if the user can change the ordering. We provide the following interaction techniques in our implementation of the reordering methods. In SBF, the user can change the ordering by drag and drop of the dimensions. For SBP, the PCP with automatic ordering does not contain all the dimensions because SBP produces a partial ordering. The remaining dimensions are not part of the plot, but are visible to the user. If the user wants to swap dimensions within the group or add a new dimension to the group, this can be done by drag and drop.

5. Experimental results

We compare the performance of our SBF method with the similarity clustering method of Ankerst *et al.* [ABK98], the clutter-based method of Peng *et al.* [PWR04], the Hough space method by Tatu *et al.* [TAE*09], and the hierarchical dimension clustering method by Yang *et al.* [YPWR03]. The method of Ankerst *et al.* was implemented by us following the algorithm described in their paper. For the method of Peng *et al.* and Yang *et al.*, we used the implementation integrated in XmdvTool (`http://davis.wpi.edu/xmdv/downloadxmdv.html`). For the method of Tatu *et al.*, we supplied our datasets to the authors and they provided us with the results obtained with their system. The SBP performance was evaluated by studying its capability to find subspaces with high-dimensional clusters. The performance of the SBS ordering and SPM filtering methods are compared with the clutter-based method for SPM of Peng *et al.* [PWR04], as integrated in XmdvTool.

5.1. Datasets

5.1.1. Synthetic datasets

We created several synthetic datasets with varying numbers of clusters of varying dimensionality with different noise levels. Clusters were created as multimodal Gaussian distributions with different mean and variance. Depending on the value of the variance, the density of the clusters varies. Impulse noise was inserted uniformly, where the number of noise points varied between $0-10\%$ of the number of points in the clusters. Along with subspaces containing clusters, the datasets also contain dimensions with uniformly distributed random noise, to test if the methods can separate noise dimensions from cluster dimensions. The detailed description of the synthetic datasets used is as follows:

- *Synthetic Dataset 1* : 12D dataset. Six of the dimensions contain two clusters without any noise, three of the dimensions contain four clusters with some impulse noise, and the remaining dimensions contain uniform random noise.
- *Synthetic Dataset 2* : 15D dataset. Five of the dimensions contains four clusters with noise, four of the dimensions contains three clusters with noise, and the remaining dimensions contain uniform random noise.

5.1.2. Astronomical dataset: milliMillennium Galaxy Sample

The Millennium Simulation is one of the largest simulations ever made to study the development of the universe [SWJ*05], involving nearly 2×10^{10} particles. It was created to make predictions about the large-scale structure of the universe and compare these against observational data and astrophysical theories. We use the much smaller "milliMillennium" simulation, which sampled only $\sim 2 \times 10^7$ particles, and its associated L-Galaxies data [DB07, Ger05]. The actual dataset that we used is a subset of "milliMillennium" and contains 28,998 points and 15 attributes.

5.2. Performance of the Methods

5.2.1. SBF method

Synthetic Dataset 1. In the top of Figure 2, synthetic dataset 1 is visualized with the original ordering. Even though the presence of structures can be observed in this view, it is hard to understand the high-dimensional structures present. Second from the top is PCP with the clutter-based dimension ordering of Peng *et al.* In this view, the dimensions of the 6D subspace with two clusters appear in two groups (dim3, dim1, dim5) and (dim10, dim8, dim12), separated by one dimension (dim7) from the 3D subspace with four clusters present in the dataset. On the other hand, the other two dimensions of this 3D subspace are mixed up with noise dimensions. A possible explanation is the variation in noise level of different dimensions, since Peng's method needs to set a global parameter to define the outliers (points that are not in clusters). Third from the top the ordering by Ankerst's method is presented. It did put the dimensions of the 6D subspace with two clusters together, but the dimensions of the 3D subspace with four clusters got mixed with the noise dimensions. This method gives emphasis only to

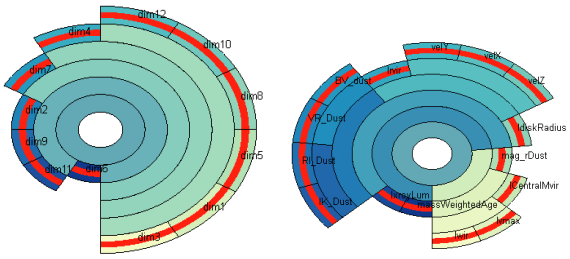

Figure 4: *Dimension hierarchy obtained with Yang's method and visualized with InterRing [YWR02]. Left: for synthetic dataset 1, the dimensions (dim 3-1-5-8-10-12) of the 6D subspace with two clusters without any noise are in one cluster; however, two of the dimensions (dim4 and dim7) of the 3D subspace with four clusters present in the dataset do not form any 'dimension cluster', and another dimension (dim9) of this 3D subspace forms a 'dimension cluster' with two noise dimensions. Right: millMillennium dataset; similarly to Ankerst's method, the x,y,z-components of velocity and colors form 'dimension clusters'.*

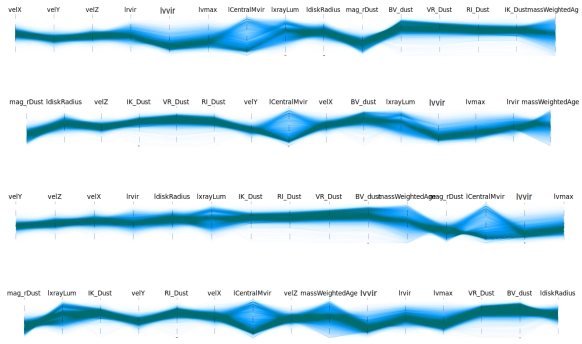

Figure 5: *millMillennium dataset. Top: original ordering. Second from top: ordering by Peng's clutter-based method. Third from top: reordering by Ankerst's method. Bottom: SBF reordering. Bimodality of the galaxies is better visible in the SBF ordering.*

the (distance-based) similarity of the dimensions. Therefore, it was able to find the dimensions of the 6D subspace which are very similar, but not the dimensions of the 3D subspace with four clusters that contain more complex structures and are not very similar in terms of distance. It is also possible to obtain multiple reordered sequences from the ant colony optimization method. We looked into many such sequences, but none of these put the dimensions of the 3D subspace with four clusters together.

In the bottom of Figure 2, the reordering obtained from our SBF algorithm is presented. We see that the method did find the dimensions of the 6D subspace with two clusters, and the clusters are well separated from each other. Next in the sequence the method put the dimensions of the 3D subspace with four clusters with some impulse noise. Finally, all the dimensions with uniform random noise were put together. However, it can also happen that the noise dimensions end up between the other two groups, as we use the first three principal components for dimensions higher than three. Added noise dimensions will not change the quality of a subspace because of the noise reduction capability of PCA. Anyhow, the method can separate the dimensions with clusters from those with noise.

In Figure 3 (left), the first four dimensions of the worst (top) and best (bottom) ranked visualization obtained by Tatu *et al.*, can be seen for synthetic dataset 1. In the best ranked visualization, the first two dimensions (dim5 and dim10) belong to the 6D subspace with two clusters, and the others (dim4 and dim9) belong to the 3D subspace with four clusters. Therefore, it is possible to derive from this view that the two clusters of the 6D subspace present in the dataset will not be visible in their entirety in the best view with this method.

Dimension hierarchy ordering by Yang's method produced similar results as Ankerst's method. In the left of Figure 4 the dimension clustering by Yang's method can be seen for the synthetic dataset 1. The method did put the dimensions of the 6D subspace, which has two very clear clusters, in one 'dimension cluster'. However, it failed to group the dimensions of the 3D subspace with four clusters in one 'dimension cluster'.

millMillennium Dataset. Bimodality of galaxies is a very well-known phenomenon in astronomy [BWM*04], corresponding to the separation of galaxies in red and blue groups. Red galaxies are elliptical and compact galaxies with mostly old stars, and blue galaxies are spiral and extended galaxies with mostly young stars.

Figure 5 shows that the bimodality of galaxies can best be observed in the ordering obtained by the SBF method (bottom of Figure 5), especially from the dimensions "CentralMvir" to "BV_Dust". The dense cluster basically represents the red galaxies that can be identified from their high values in color dimensions (such as VR_dust). This phenomenon can also be observed in the ordering by Peng's method (second from top in Figure 5) but not as prominent as in the SBF ordering. Ankerst's method put all similar dimensions such as velocities ('velx', 'vely','velz'), or colors ('BV_dust', 'VR_dust', 'RI_dust', 'IK_dust') together (third from top in Figure 5). However, most of the time domain experts know beforehand that x, y, and z-components of velocity, or colors in different bands, will be similar. A relation like galaxy bimodality is more interesting to them than the obvious relations. Results by the method of Tatu *et al.* for synthetic data are shown in Figure 3 (right): the first 4 dimensions, with (top) the worst [lcentralMvir (6), mag_rDust (9), massWeightedAge (14), BV_dust(10)] and (bottom) the

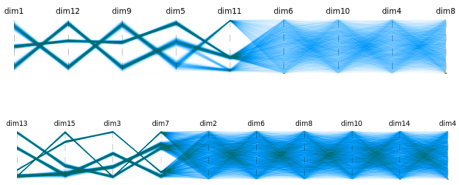

Figure 6: *Synthetic dataset 2 (see section 5.1.1). Two of the subspaces revealed by the SBP reordering method. Top: three clusters in a 5D subspace can be observed. Bottom: four clusters in a 4D subspace are visible.*

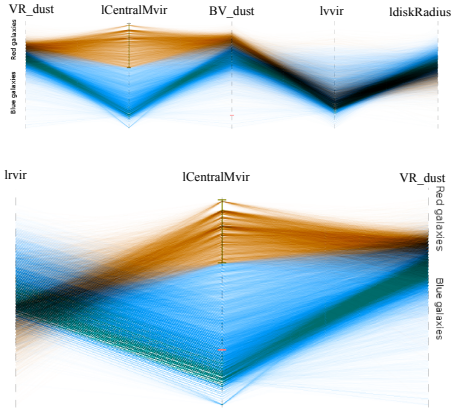

Figure 7: *millMillennium dataset. Two of the subspaces revealed by SBP reordering. Top: two clusters in a 5D subspace (VR_dust, lCentralMvir, BV_dust, lvvir, ldiskRadius) can be observed. Bottom: two clusters in a 3D subspace (lrvir, lCentralMvir, VR_dust) are visible. For better visualization, clusters of red galaxies are colored in orange by manual selection.*

best [velZ (2),velX (0), ldiskRadius(8), lrvir(3)] ranked visualization. Surprisingly, two of the well-known attributes, i.e., magnitude (mag_rDust) and color (BV_dust), that can show bimodality of galaxies, are in the worst ranked view. On the other hand, in the best ranked view no visible structures can be observed.

The performance of Yang's method, shown in Figure 4 (right), is also similar for this dataset. This method also put similar dimensions such as velocities or colors in 'dimension clusters'. As remarked above, such straightforward groupings are less useful than the interesting relations these dimensions might have with other dimensions.

5.3. SBP method

Synthetic Dataset 2. The SBP method found a 9D subspace with three clusters (top of Figure 6). Five of the dimensions are actual cluster dimensions and the remaining ones are noise dimensions. Also a 10D subspace was found with four clusters (bottom of Figure 6). Four of the dimensions are actual cluster dimensions and the remaining ones are noise. This addition of noise dimensions to the sequence is due to the use of PCA (see section 3).

Astronomical Dataset. In Figure 7 two of the subspaces of the millMillennium dataset can be observed. The first subspace (top of Figure 7) is 5D and at least two clusters can be identified visually. If we observe the axis "VR_dust", which is the first dimension of this ordering, galaxies with high value represent the red galaxy group. A similar bimodality can be observed even more clearly in the 3D subspace (bottom of Figure 7).

5.4. SBS method

In Figure 8, we show SPM visualizations of synthetic dataset 2. In the original ordering the noise dimensions are so dominant that it is very hard to perceive any clustering. The ordering obtained by the SBS method is comparable to that by the method of Peng *et al.* [PWR04]. However, Peng's method ordered the dimensions in such a way that cluster dimensions appear in certain groups among the noise dimensions. On the other hand, our method separated cluster dimensions and noise dimensions in two distinct groups, making it possible to filter out the noise dimensions using some threshold value.

Another difference of our approach with Peng's method is that the latter obtains the reordering at a very high computational cost (see section 2). On the other hand, our method computes the density image in $O(N)$ time where N is the number of data points. The computational cost of the Max-tree creation is linear in both the number of pixels and in the connectivity. For the SBS method we only compute the Max-tree for 1D images, therefore the complexity of Max-tree computation is $O(I)$ where I is the number of pixels in the 1D image (we chose $I = 512$). For example, to find the reordering for SPM of a dataset with 744 data points with 11 dimensions Peng's method took 3 : 13 min with exhaustive search and 7 sec with random swapping. In comparison, our method took 0.23 sec to perform the ordering and filtering for the Millennium galaxy sample dataset with 28998 data points and 15 dimensions.

5.5. Comparison between PCP and SPM

An interesting observation is that pairwise structural relationships are better visible in scatter plot matrices than in parallel coordinate plots. An example is shown in Figure 9, where three dense clusters are clearly visible in SPM, whereas in PCP these clusters are not so clear.

On the other hand, high-dimensional structures seem to

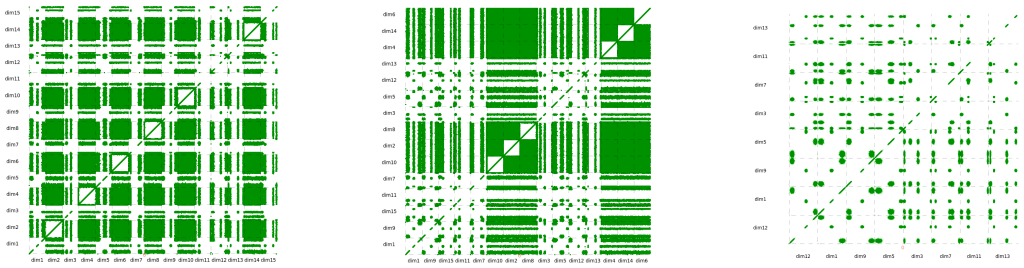

Figure 8: *SPM visualization of synthetic dataset 2. Left: original ordering. Middle: ordering by the clutter-based method of Peng* et al.*; some grouping of cluster and noise dimensions can be observed. Right: ordering by the SBS method after filtering; only cluster dimensions are visualized since all the noise dimensions are filtered out by the method.*

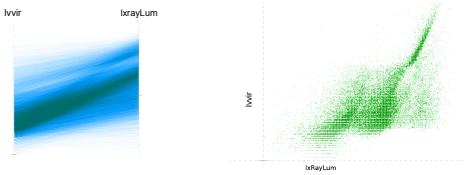

Figure 9: *milliMillennium dataset: dimensions lxrayLum vs. lvvir. Left: visualized with PCP. Right: visualized with SPM. The presence of three dense clusters is apparent in SPM but less obvious in PCP.*

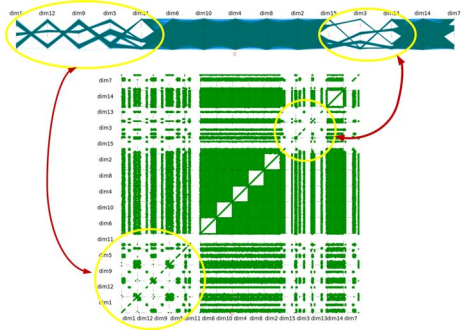

Figure 10: *SBF reordering of synthetic dataset 2. Top: for PCP. Bottom: for SPM. Two groups of dimensions, one with three clusters in a 5D subspace and one with four clusters in a 3D subspace, are better visible in PCP than in SPM.*

be more intuitively visible in PCP than in SPM. We compare the reordering by the SBF method with SPM vs. PCP in Figure 10. Here two groups of dimensions, one with three clusters in a 5D subspace and one with four clusters in a 3D subspace, are almost immediately noticeable with PCP, whereas with SPM it may require some in-depth analysis.

5.6. Limitations

The use of PCA for dimensions higher than three is currently one of the limitations of our method. As already discussed in section 3, PCA does not restrict us in finding subspaces with high-dimensional structures (> 6), but it imposes limitations for finding the proper number of subspace clusters present. Another limitation concerns the SBP method: with the current implementation we can only obtain n subspaces where n is the number of dimensions.

6. Summary and future work

We have presented methods for dimension ordering and filtering for the parallel coordinate plot (PCP) and the scatter plot matrix (SPM), based on structures present in subspaces of the full feature space. For PCP we obtained two orderings: one providing a reordering for all the dimensions; the second one producing groups of dimensions which are subsets of the full feature space. We also presented a simple ordering and filtering scheme for dimension ordering in SPM.

Evaluation on synthetic and astronomical datasets confirmed that the methods are able to find a proper order of dimensions that facilitates the perception of high-dimensional structures. We observed that high-dimensional structures can be more easily perceived in PCP, whereas pairwise structures are better visible in SPM. We showed that our method compares favorably with a number of existing approaches. Future work will include other high-dimensional data visualization techniques in our structure-based subspace analysis approach.

Acknowledgments

We thank Jorik Blaas of Delft University of Technology for providing his PCP code, and Hugo Buddelmeijer of the Kapteyn Astronomical Institute for providing the millMillennium galaxy sample. We also thank Andrada Tatu of the University of Konstanz and Georgia Albuquerque of the Technical University Braunschweig for providing us with the results of their method on our datasets. This research is funded by the Dutch National Science Foundation (NWO), "STARE" program, project no. 643.200.501.

References

[ABK98] ANKERST M., BERCHTOLD S., KEIM D. A.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1998), IEEE Computer Society, pp. 52–60. 2, 6

[AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D., THEISEL H., MAGNOR M.: Quality-based visualization matrices. In *Proc. Vision, Modeling, and Visualization (VMV'09)* (Braunschweig, Germany, Nov. 2009), pp. 341–349. 3

[AEL*10] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M.: Improving the visual analysis of high-dimensional datasets using quality measures. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)* (Oct. 2010). 2

[AGGR98] AGRAWAL R., GEHRKE J., GUNOPULOS D., RAGHAVAN P.: Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD 27* (1998), 94–105. 5

[Asi85] ASIMOV A.: The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Statist. Comp. 6*, 1 (1985), 128–143. 1

[BBP08] BLAAS J., BOTHA C., POST F.: Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1436–1451. 5, 6

[Ber07] BERTRAND G.: On the dynamics. *Image Vision Comput. 25*, 4 (2007), 447–454. 4

[BWM*04] BELL E. F., WOLF C., MEISENHEIMER K., RIX H., BORCH A., DYE S., KLEINHEINRICH M., WISOTZKI L., MCINTOSH D. H.: Nearly 5000 distant early-type galaxies in COMBO-17: A red sequence and its evolution since z˜1. *Astrophysical Journal 608* (June 2004), 752–767. 7

[DB07] DE LUCIA G., BLAIZOT J.: The hierarchical formation of the brightest cluster galaxies. *Monthly Notices of the Royal Astronomical Society 375*, 1 (2007), 2–14. 6

[DG97] DORIGO M., GAMBARDELLA L. M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Trans. Evolutionary Computation 1*, 1 (1997), 53–66. 2

[FBT*10] FERDOSI B. J., BUDDELMEIJER H., TRAGER S., WILKINSON M. H. F., ROERDINK J. B. T. M.: Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In *Proc. IEEE Conference on Visual Analytics Science and Technology (IEEE VAST), October 2010* (2010), pp. 35–42. 2, 3, 4, 5

[GC99] GOIL S., CHOUDHARY H. N. A.: *MAFIA: Efficient and scalable subspace clustering for very large data sets*. Tech. rep., Northwestern University, Evanston IL, USA, 1999. 3

[Ger05] GERMAN ASTROPHYSICAL VIRTUAL OBSERVATORY: Virgo - Millennium Database. `http://www.g-vo.org/Millennium`, 2005. 6

[Guo03] GUO D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2* (December 2003), 232–246. 2

[Ins09] INSELBERG A.: *Parallel Coordinates : VISUAL Multidimensional Geometry and its Applicationss*. Springer, New York, 2009. 1

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 993–1000. 2, 3

[KKZ09] KRIEGEL H.-P., KRÖGER P., ZIMEK A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data 3*, 1 (2009), 1–58. 2

[PWR04] PENG W., WARD M. O., RUNDENSTEINER E. A.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *In INFOVIS'04: Proceedings of the IEEE Symposium on Information Visualization* (2004), IEEE Computer Society, pp. 89–96. 2, 6, 8

[SOG98] SALEMBIER P., OLIVERAS A., GARRIDO L.: Anti-extensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing 7* (1998), 555–570. 3

[SWJ*05] SPRINGEL V., WHITE S. D. M., JENKINS A., FRENK C. S., YOSHIDA N., GAO L., NAVARRO J., THACKER R., CROTON D., HELLY J., PEACOCK J. A., COLE S., THOMAS P., COUCHMAN H., EVRARD A., COLBERG J., PEARCE F.: Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature 435* (2005), 629–636. 6

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST), Atlantic City, New Jersey, USA* (2009). 2, 3, 6

[YPWR03] YANG J., PENG W., WARD M. O., RUNDENSTEINER E. A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization* (2003), pp. 105–112. 2, 3, 6

[YWR02] YANG J., WARD M. O., RUNDENSTEINER E. A.: InterRing: An interactive tool for visually navigating and manipulating hierarchical structures. *Information Visualization, IEEE Symposium on 0* (2002), 77. 7