Quality-Based Visualization Matrices

Georgia Albuquerque[1], Martin Eisemann[1], Dirk J. Lehmann[2], Holger Theisel[2] and Marcus Magnor[1]

[1]Computer Graphics Lab, TU Braunschweig, Germany
[2]Visual Computing, University of Magdeburg, Germany
Email: [1]{georgia, eisemann, magnor}@cg.tu-bs.de, [2]{dirk, theisel}@tu-bs.de

Abstract

Parallel coordinates and scatterplot matrices are widely used to visualize multi-dimensional data sets. But these visualization techniques are insufficient when the number of dimensions grows. To solve this problem, different approaches to preselect the best views or dimensions have been proposed in the last years. However, there are still several shortcomings to these methods. In this paper we present three new methods to explore multivariate data sets: a parallel coordinates matrix, in analogy to the well-known scatterplot matrix, a class-based scatterplot matrix that aims at finding good projections for each class pair, and an importance aware algorithm to sort the dimensions of scatterplot and parallel coordinates matrices.

1 Introduction

With the exponentially increasing amount of acquired multivariate data, several multi-dimensional visualization techniques have been proposed during the last decades [10]. Based on the fact that human perception cannot deal well with more than three continuous dimensions simultaneously, such techniques usually project the data in low-dimensional embeddings and combine these representations in a single plot or present them to the user in an interactive way. Some well-known examples of multi-dimensional visualization techniques are glyph techniques [17], parallel coordinates [9], scatterplot matrices [7] and pixel level visualizations [11]. But even these techniques do not scale well to high-dimensional data sets. In this work we focus on parallel coordinates plots (PCP) and scatterplot matrices (SPLOM), and propose extensions to these well-known visualization techniques.

Scatterplots are one of the oldest and widely used visualization methods.We can define them as graphs where the values of two variables for a sample in a data set are used to plot a point in 2-dimensional space, resulting in a scattering of points. Scatterplots are very useful for visually determining the correlation between two variables. A SPLOM is a symmetric matrix of adjacent scatterplots and allows the user to analyze the diverse dimensions at once. If there are n variables, the SPLOM has dimension $n \times n$ and the element at the i-th row and j-th column is a scatterplot of the i-th and j-th variable. Related to this kind of visualization, we propose two extensions: a class based scatterplot matrix and an importance oriented reordering of the dimensions of the matrix. The proposal of the *class based SPLOM* (C-SPLOM) is to support the visual analysis of labeled (classified) data sets. In such data sets, the analyst often searches for projections where distinct clusters can be observed. Previous approaches aim at finding good views of a data set considering all classes at once. The problem with such approaches is that this global optimization may ignore views that separate two classes well, because of the distribution of the remaining classes. To deal with this, our C-SPLOM presents the best projection for each class pair, based on a ranking index. This class based visualization method is useful to analyze labeled data sets with a large number of variables that cannot be well visualized using traditional SPLOMs.

Another popular visualization technique are parallel coordinates plots (PCPs) [9]. In such plots, each sample of an N-dimensional data set is represented by a polyline that intersects N vertical axes (dimensions). The intersection point represents its value in the respective dimension. Similar to the scatterplot matrices the parallel coordinates plots, do not scale well when the number of dimensions grows, as important dimensional relationships might not be visualized. Addressing this shortcoming, we propose an importance oriented *parallel co-*

ordinates matrix (PCM). Unlike the SPLOM the PCM is not symmetric, each row i of the matrix represents the relation of one dimension d to the others of the data set, ordered by the inherent information value. Additionally, we propose a quality aware dimension reordering framework for visualization matrices, like SPLOMs, C-SPLOMs and PCMs, to improve the visual analysis task of high-dimensional data sets.

2 Related Work

SPLOM and PCP are two of the most popular multi-dimensional visualization techniques and are implemented in diverse popular visualization tools as for example in the XmdvTool [17] and GGobi [14].

2.1 Scatterplot Matrix

The SPLOM was first published by John Hartigan [7] and later explored and extended in diverse visual exploration tools. As aforementioned, SPLOMs lose their effectiveness when the number of variables is large; to deal with this problem different approaches have been proposed: The grand tour [3] is a dynamic tool that presents a continuous sequence of lower dimensional (e.g. 2-dimensional) point scatters. However, an exhaustive exploration of a high-dimensional data set requires prohibitive time. Projection pursuit [6, 8] was proposed as an alternative to an exhaustive visual search, a statistical technique to search for low dimensional (one or two-dimensional) projections that expose interesting structures of the high dimensional data set. Later on, different projection pursuit indices [5, 8] and a combination of the grand tour and projection pursuit [4] as a visual exploration system have been proposed. In a similar direction, the Scagnostics method [16, 18] was proposed.In this technique, different scagnostics indices (e.g. Convexity, Skinny, etc.) are computed and presented as a scatterplot matrix of the indices themselves (the scagnostics SPLOM). Such scagnostic indices can be used to reveal structures of the data set in the form of trends, hypersurfaces, clusters, or anomalies in the data set.

In our method we make use of projection pursuit like measures in a twofold way, to select information-bearing projections for the C-SPLOM and PCM, and to perform dimension reordering. Considering classified datasets, a class consistency visualization algorithm has been proposed by [12]. Similar to our class based matrix, the class consistency method proposes measures to rank lower dimension representations. The method proposed in [12] filters the best scatterplots based on their ranking values and present them in an ordinary scatterplot matrix. One problem of this method is that the SPLOM does not scale well for high-dimensional data sets and even if a zoom option is available, the overall visualization of the SPLOM is prejudiced. Another problem happens when all classes are analyzed together to rank the projections, in this case, projections that separate two classes very good might receive a bad ranking because of the distribution of the remaining classes. Our method reduces the matrix size to the number of classes of the data set and presents to the user the best projections for each class pair individually.

2.2 Parallel Coordinates

Another very popular multivariate visualization technique are parallel coordinates [9]. In a parallel coordinate plot each dimension appears just once and the relation with other dimensions may be difficult to pinpoint depending on the distance between them in the plot. Diverse linking and brushing algorithms [17, 14] together with transparency levels have been proposed to help visualizing these relations. However, they do not solve the problem when one dimension shares important correlations with more than its two neighboring dimensions in the visualization. Opposingly, we propose a parallel coordinate matrix, where there is the possibility to plot all possible 3-dimensional combinations for each dimension. In this matrix we have for each dimension d up to $(n-1)/2$ 3D parallel plots, where d is the central dimension, theoretically revealing all important relations for this dimension. An important issue for parallel coordinates is how to order the dimensions in the plot. Different proposals to solve this problem focus on ordering the dimensions by similarity [2, 19], and a recent work [15] proposes a sorting of the dimensions based on the quality of the plots. In this second case, a rank function evaluates each 2D dimensional parallel plot and the result is used to determine the order of the dimensions in the final plot. Our PCM capitalizes on this second approach to order the 3D individual plots. For each dimension we sort its respective

3D plots using a ranking function; the plots with a higher ranking are presented first and the ones with a lesser amount of useful information are presented last.

3 Visualization Matrices

In the following subsections we describe our information bearing visualization matrices in more detail and define the measures we use to rank the low-dimensional projections. We then discuss reordering of scatterplot matrices using such quality measures and how it can help to visualize high-dimensional data sets.

3.1 Parallel Coordinates Matrix

Parallel coordinates plots [9] are one of the techniques which allow to visualize an arbitrary number of dimensions of a data set within the same plot. This makes them very attractive for high-dimensional data sets but comes at a cost. The amount of information bearing content is very sensitive to the ordering of dimensions [2]. In addition, every dimension can be paired with only two other dimensions. Therefore important relations to a third or fourth dimension might be missed.

Our approach aims at presenting all those relations in a single matrix, where each entry shows the relationship between only two dimensions. This way all n^2 possible combinations of dimensions are represented and no information is lost. The problem with such an approach is the overstraining of the user as he would have to check every single visualization for possible information content. It is therefore important to sort the visualizations inter- as well as intradimensionally, so that important visualizations are spatially close together and at known positions in the matrix. We found that such a matrix is most legible if three constraints are fulfilled:

1. Every row should contain one main dimension, which appears in every visualization in this row. A label is assigned at the left of the row for faster indexing.
2. The visualizations in each row should be sorted in descending order according to their inherent information value. The best should be positioned on the left, the worst on the right.
3. The dimensions itself, i.e. the rows of the matrix, should be rearranged so that the most valuable rows are on top, while dimensions with lesser information value are closer to the bottom of the matrix.

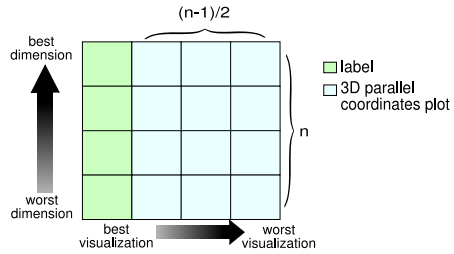

Figure 1: Structural overview of the PCM. Each row has one main dimension appearing in the middle of each 3D plot and as a label on the left for a better overview. The rows are ordered according to the overall importance of the main dimension in ascending order. The visualizations in each row are again odered according to their relative importance.

This way, only looking at the $n \times p$ submatrix, starting at index $(0, 0)$ reveals the most valuable relationships, i.e. visualizations to the user. An example of this concept is given in Figure 1.

In a first step, all n^2 2D visualizations are created. A quality measurement is applied to characterize the possible information value of each visualization. Most approaches in the literature aim at finding the best ordering of all n dimensions globally, or choosing a subset of them.

Only recently an approach has been presented, which rates every PCP consisting of only two dimensions and combines them in a second step to the complete visualization [15].

We exemplarily make use of their *overlap measure* for our test data, which measures the similarity between the different classes of the data set in Hough space. Visualizations with distinctive classes therefore receive a high quality value and visualizations with very similar classes receive a low value. Other measurements, class and non-class based, would be possible as well and can be easily included in our framework, like [2].

We initialize the matrix so that each row of the matrix has one main dimension, e.g. each visualization in row 1 contains the dimension 1, each in row 2 the dimension 2 and so on. We then sort the visualizations intradimensionally, i.e. per row. As each visualization is associated with a quality value,

we can easily apply a simple standard sorting algorithm. We always combine two 2D visualizations to a 3D visualization, as both share the same main dimension, which is then positioned in the middle.

In a last step we reorder the dimensions itself, i.e. the rows of the matrix. We tested different criteria, like summation of all quality values in each row or linear and Gaussian falloffs, increasing the importance of the first visualizations in each row, while decreasing the importance of the lesser valued ones, and found that the linear falloff gives good results for the PCM. More details and a more general description for dimension reordering visualization matrices is given in Section 3.3. Therefore the quality value of the j-th dimension is computed by

$$D_j = \sum_{i}^{n} \frac{(n-i)}{n} Q(p_{(j,i)}) \quad , \qquad (1)$$

where n is the number of dimensions and $Q(p_{(j,i)})$ is the quality value for the i-th visualization in the j-th row of the matrix.

3.1.1 Evaluation and Results

We used the *Wisconsin Diagnostic Breast Cancer* (WDBC) as well as others from the UCI data base to test the usefulness of our PCM. The WDBC data set consists of 569 samples with 32 real-valued dimensions each [13]. The task is to find the best dimensions separating the malign and benign cells in the data set. We created our PCM for this data set using the *overlap measure* from [15]. Other measurements could be used as well, depending on the task. Figure 2 shows the complete PCM with the best and worst ranked visualizations enlarged. Visualizations with higher information value are found in the top left of the matrix, as desired, while the visualizations on the bottom right are hardly of any use. Looking only at the best parallel coordinates plots, like other approaches did [19, 15], one might miss important information. E.g. dimension 22 (radius (worst)) in combination with dimension 9 (concave points (mean)), 29 (concave points(worst)), 25 (area (worst)) and 5 (area (mean)) all separate the malign and benign cells comparingly well, but in usual parallel coordinate visualizations only two combinations could be displayed in one visualization.

One could argue that SPLOMs fulfill a similar task as PCMs, but there are major differences between these two approaches. First, SPLOMS are not sorted. This limits their usefulness for data sets of up to a dozen dimensions only, otherwise exhaustively investigating each plot is overstraining for a user. Even when sorting the dimensions beforehand, as proposed in 3.3, additional information as color encoding or ranking values are needed to guide the visual search. Using PCMs, looking at the $n \times p$ submatrix, starting at index $(0, 0)$ always reveals the most valuable relationships, i.e. visualizations to the user, no matter how high the dimensionality of the data set is. Of course the choice of Parallel Coordinates could also be exchanged with Scatterplots, which one is more beneficial depends on the preference of the user.

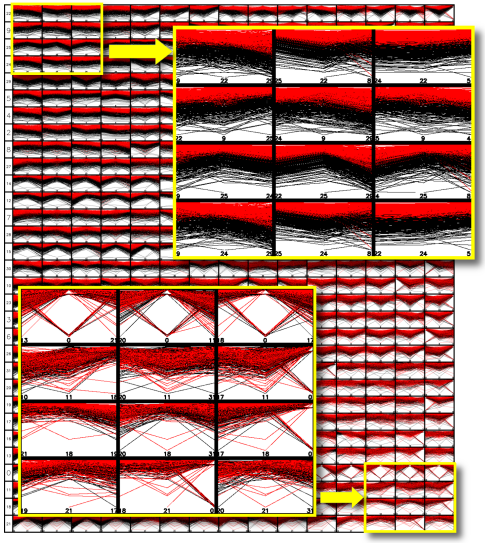

Figure 2: Results of the PCM for the WDBC data set. Malign nuclei are colored black while healthy nuclei are red. Visualizations with only few overlap are preferred, so the difference between malign and benign cells becomes more clear, and can be found in the top left of the matrix. The worse visualizations in the bottom right hardly convey any useful information.

3.2 Class-Based Scatterplot Matrix

A common task in visual analytics is to search for projections of high-dimensional data that shows well defined clusters. The same occurs when class information is available; finding the projec-

tions or dimensions that can well separate the distinct classes is a desired outcome. To serve this purpose, we introduce a new visualization matrix called *class-based scatterplot matrix* (C-SPLOM). We assume that each point in the high-dimensional space has a class label c. Diverse data sets have a clear definition of classes, but this assumption does not limit the use of this technique to these data sets, as class labels can be assigned through an automatic clustering algorithm.

Similar to the well known SPLOM, the class-based version is also a matrix of pairwise scatterplots $s(a, b)$, with data dimensions a and b. The difference is that the classes are listed on the rows and columns instead of the original dimensions. If there are m classes in a data set, the C-SPLOM has dimensions $m \times m$ and the element at the i-th row and j-th column is the scatterplot of the k-th and h-th variable. The projection axes k and h are chosen in a way to maximize the information content for the pairwise relation of the i-th and j-th classes.

An important issue of the C-SPLOM is to choose an appropriated analysis algorithm to compute the quality index $Q(s(a, b))$ of the scatterplots. Different algorithms can be used to this end [12, 15] as long as they consider the pairwise relationships between classes. The problem in considering all classes at once, as proposed in [12, 15], is that the global optimization may ignore views that separate two classes well, because of the distribution of the remaining classes.The Figure 3 shows examples of scatterplots generated from the *Olives* data set. (Section3.2.1). The first scatterplot $s(4, 5)$ has the highest rank $Q(s(4, 5)) = 1$ considering all classes. However the scatterplot $s(2, 8)$ with rank $Q(s(2, 8)) = 0.44$ presents a better separation of 3-th and 4-th classes presented in the data set (the *South-Apulia* region in red and *Sicily* region in green, respectively), as can be seen in the third plot. This outcome is only possible if the adopted measure analyzes the pairwise relationships between classes instead of a global measure. The resulting quality index Q is then used to rank the scatterplots, and the best scatterplot is selected for the respective class pair.

We tested our C-SPLOM with two similar algorithms to measure the quality of scatterplots with class information. The first algorithm is the *class density* method proposed in [15]. It assigns high values to plots with few overlap between the classes

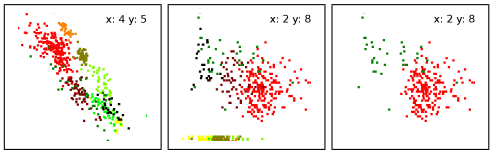

Figure 3: The first scatterplot is the one with the highest rank $Q = 1$, when considering all classes, however the second one with rank $Q = 0.44$ presents a better separation of the 3-th and 4-th classes (red and green), as can be seen in the third plot.

and dense clusters. We adopted their algorithm with the difference that only one class pair is considered per time. To rank projections considering a specific class pair, the algorithm is applied only to the data of the respective classes and the best ranked scatterplot will represent this class pair in the C-SPLOM. The second measure, as the first one, presents high values for plots with well separate clusters and instead of dense clusters, this measure has a bias towards larger distances between the clusters. The distance at a pixel p is defined as r, where r is the radius of the enclosing sphere of the $k-$nearest neighbors of p:

$$r = max_{i \in N_p} ||\mathbf{x} - \mathbf{x}^i||, \tag{2}$$

as defined in [15]. Both measures then compute the sum of the mutual differences of these images. To decide which algorithm is the best one depends strongly on the user task. Figure 4 shows an example of the differences between the C-SPLOMs for the *Wine* data set (Section3.2.1), using these two approaches.

Note that for the 1-st and 2-nd class (in black and red respectively) the class density presents a scatterplot with more dense clusters as best result, while the class distance measure presents a scatterplot where the distance between the center of the clusters is larger. The same happens for the 1-st and 3-rd class (in black and green respectively), and for the 2-nd and 3-rd the same scatterplot is chosen.

3.2.1 Evaluation and Results

To evaluate our C-SPLOM, we tested it on diverse real data sets from the UCI repository [1] with labeled information. The first presented data set is

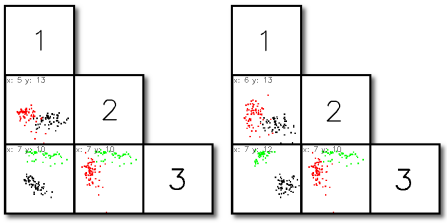

Figure 4: The resulting C-SPLOM for the class density (left) and class distance quality measures (right).

the *Wine* data set, a classified data set with 178 instances and 13 attributes describing chemical properties of Italian wines derived from three different cultivars. The user task here is to find the projections (dimensions) that separate these classes well. Figure 5 shows the comparison of the C-SPLOM (upper-right) and its counterpart SPLOM (bottom-left). The C-SPLOM was computed by means of the *distance measure* described previously. Another data set we used to evaluate the C-SPLOM is the *Olives* [20] data set. With 572 olive oil samples from nine different regions of Italy; for each sample the normalized concentrations of eight fatty acids are given. Figure 3 show two scatterplots of this data set, the first one with the 4-th and 5-th dimensions (concentrations of the oleic and linoleic acids), and the second with the 2-nd and 8-th dimensions (concentrations of the palmitoleic and eicosenoic acids).

3.3 Dimension Reordering

Often, n-dimensional datasets are represent as a series of 2D scatterplots. Such scatterplots are commonly arranged in a SPLOM and usually, the dimensions are arranged as provided by the data set. Dimension reordering methods for SPLOM based on the similarity between the projections have been proposed [18]. But no *quality-aware sorting* methods have been presented.. This motivated us to adopt a *quality-aware sorting* concept and to start investigating the advantages of such an approach. Note that the concept of dimension reordering can be applied to any matrix-based visualization, e.g. in Section 3.1 we also apply a dimension reordering for our PCM, but for the ease of understanding we will use SPLOMs in this chapter.

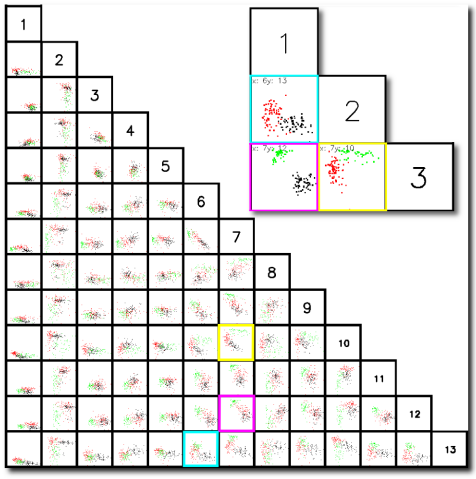

Figure 5: Results of the C-SPLOM for the Wine data set. Visualizations with only few overlap are preferred, so the difference between the wine cultivars becomes clearer. The best visualizations for each class pair are shown in the C-SPLOM.

The pipeline of our framework for quality-aware reordered SPLOMs (D-SPLOM) is shown in Figure 6 and explained in the following. As a pre-process for the reordering, we initially apply a quality-measure $Q(s_{(a,b)})$ to each scatterplot $s_{(a,b)}$. This quality-measure ought to be a scalar one, so that it rates the scatterplot unambiguously with a single number. Apart from that, it could be any useful measure [18, 12, 15]. Furthermore, we need this quality-measure to estimate the quality of each dimension itself. Once we have $n-1$ scatterplots for each dimension in a n-dimensional dataset, we consider $n-1$ quality measures (one per plot) to compute the dimension overall quality-measure.

For each dimension d, we compute a dimension-measure as the base for reordering. A dimension-measure Q_d is a scalar function $Q_d : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ over all quality-measures $Q(s_{(a,b)})$ of a dimension d: $D_d = Q(s_{(d,i)})$, where $i \neq d$ and $i \in [1, \ldots, n]$. It appraises the quality-aware impact over all scatterplots, which contains the dimension d. There are exactly n dimension-measures for a n-dimensional dataset. Different functions could be used for computing D_d, as long as it is guaranteed that the measure-values are comparable to each other, as the *mean*, a *PCA* or the *variance* over

pre-computed quality-measures. We decided to use the sum over all quality-measures as dimension-measure for this paper, as a proof of concept:

$$D_d = \sum_{i=1,\ i \neq d}^{n} Q_{(s_{(d,i)})}. \qquad (3)$$

This measure produces the same partial order between the dimension-measures as the mean, with the advantage that it is easier and faster to compute. In this last step, we make use of the computed information to reorder a $n \times n$ SPLOM. Because such a SPLOM is symmetric, we use the upper triangular matrix for display. First, we allocate to each dimension its quality-measure value D_d. We sort all quality-measure/dimension pairs (D_d, d) by means of a simple partial order $(\geq)$ with respect to the quality-measure D_d, which gives us a dimension-ranking $r = (sort\{(D_d, d)\}; \geq)$. The dimension $r[0].d$ from ranking r describes the *best* dimension and $r[n].d$ the *worst* one.

We map a dimension d to its position in the ranking r and, depending on that mapping, we reorder the scatterplots in the SPLOM and get the dimension-based reordered SPLOM (D-SPLOM), as can be seen in Figure 6.

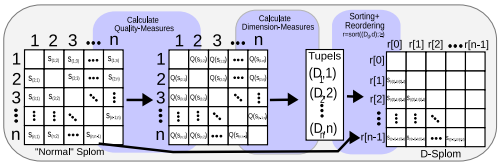

Figure 6: Overview of the dimension reordering Process

3.3.1 Evaluation and Results

To evaluate our concept, we tested it on real class-based and non-class-based multi-dimensional datasets. For classified data, we applied the *Class Density Measure* (CDM) [15] as a quality-measure to the *Olives* dataset [20], see Section 3.2.1 for a description. The CDM assigns higher values to scatterplots with a better separation between the classes. The result of the reordering is shown in Figure 7. For non-classified data we applied the *Rotating Variance Measure* (RVM) [15] as quality-measure to the Parkinson-dataset. This set has 13 dimensions, no classes and 197 items. The RVM rates the linear and non-linear correlation within the scatterplots with respect to its two dimensions. The result is shown in Figure 8.

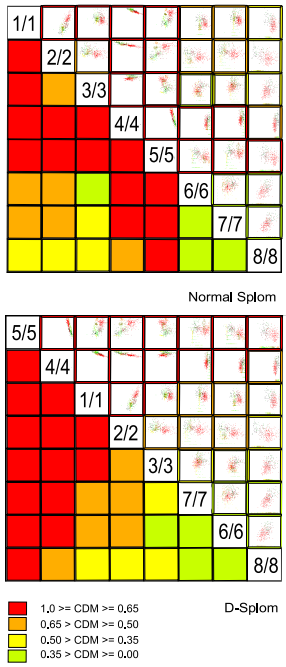

Figure 7: Evaluation on class-based *Oliveoil* data set using the CDM: The ordinary SPLOM (top), the resulting D-SPLOM (bottom)

In Figure 7 and 8, relevant scatterplots are colored more red than non-relevant. It is easy to see that both types of scatterplots are distributed over the whole SPLOM before the reordering. After the reordering, relevant and non-relevant scatterplots in the D-SPLOM are mostly separated from each other. Therefore, we can observe that the quality-aware reordering reduces the region of interest, speeding up the visual search. I.e., a quality-aware reordering has practical advantages and enhances the visual quality of SPLOMs. Depending on the data set, some dimensions might contain outliers. This may happen, when the used quality-measure assigns a low value to most visualizations of one dimension, but a high value to only a few, as the dimension 4 in the SPLOM, shown in Figure 8. Our applied color coding allows for easy recognition of such plots. In the future, we should inves-

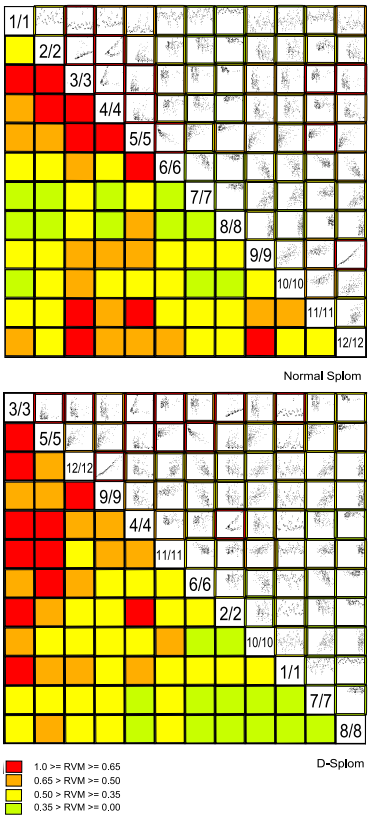

Figure 8: Evaluation on non-class-based *Parkinson* data set: The ordinary SPLOM (top), the resulting D-SPLOM (bottom).

tigate how far the quality-aware framework is appropriate and stable to fading out non-relevant scatterplot from SPLOMs, speeding up even more the visual search task.

4 Conclusion

In this paper we, presented two new visualization matrices to support the visual analysis of high dimensional data sets: A class based scatterplot matrix for data sets with label information that supports the analysis of pairwise relationship between classes, and a parallel coordinates matrix that allows examining correlations between all possible dimensions using parallel coordinates plots. Additionally, we proposed an information-bearing reordering framework that can improve the visual analysis task for any matrix-based visualization method. We have shown that our quality-based visualization matrices together with the presented reordering framework successfully reduces the region of interest of the visualization matrices. In the future, we intent to evaluate our methods more thoroughly with an adequate user study and to allow user interaction while forming and reordering the matrices. Furthermore, we would like to test more sophisticated ranking functions for dimension reordering.

Acknowledgements

The authors gratefully acknowledge funding by the German Science Foundation from project DFG MA2555/6-1 and DFG TH692/6-1.

References

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

[2] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *Information Visualization, IEEE Symposium on*, 0, 1998.

[3] D. Asimov. The grand tour: a tool for viewing multidimensional data. *Jurnal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[4] D. Cook, A. Buja, J. Cabreta, and C. Hurley. Grand tour and projection pursuit. *Jurnal of Computational and Statistical Computing*, 4(3):155–172, 1995.

[5] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. *Prim-9: An interactive multidimensional data display and analysis system*, volume In W. S. Sleveland, editor. Chapman and Hall, 1987.

[6] J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on*, C-23(9):881–890, Sept. 1974.

[7] J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–273, 1975.

[8] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[9] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, December 1985.

[10] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[11] D. A. Keim, M. Ankerst, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, pages 279–286, Washington, DC, USA, 1995. IEEE Computer Society.

[12] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis 2009)*, 28(3):831–838, 2009.

[13] W. Street, W. Wolberg, and O. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *IS&T / SPIE International Symposium on Electronic Imaging: Science and Technology*, 1905:861–870, 1993.

[14] D. F. Swayne, D. Temple Lang, A. Buja, and D. Cook. GGobi: evolving from XGobi into an extensible frameworkfor interactive data visualization. *Computational Statistics & Data Analysis*, 43:423–444, 2003.

[15] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high dimensional data. *IEEE Symposium on Visual Analytics Science and Technology*, page to appear, 2009.

[16] J. Tukey and P. Tukey. Computing graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics 85. Nat. Computer Graphics Assoc.*, 1985.

[17] M. O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 326–333, 1994.

[18] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 157–164, 2005.

[19] J. Yang, M. Ward, E. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets, 2003.

[20] J. Zupan, M. Novic, X. Li, and J. Gasteiger. Classification of multicomponent analytical data of olive oils using different neural networks. In *Analytica Chimica Acta*, volume 292, pages 219–234, 1994.