SepEx: Visual Analysis of Class Separation Measures
Supplemental Materials

Jürgen Bernard[1] , Marco Hutter[2] , Matthias Zeppelzauer[3] Michael Sedlmair[2], and Tamara Munzner[1]

[1]University of British Columbia, British Columbia, Canada
[2]University of Stuttgart
[3]St. Pölten University of Applied Sciences

Abstract
In the three sections of this document, we extend the degree to which details could be given in the manuscript. The first section describes all characteristics of the set of datasets that was used in the usage scenario. To do so, we also included additional figures showing sample datasets in detail. The second section shows screenshots of the entire system for every state and interface (cut) which made it into the paper. With this additional context (multiple linked views), we also add more findings we made during analyses. The third section provides details about the analysis of TSNE inconsistencies, including figures of selected datasets which have been dimensionality-reduced.

1. Sequence of Datasets

The rationale of our usage scenario was to prove that our tool was actually able to reveal a series of characteristics of measures for class separation, both expected and unexpected. As the output of measures is the product of datasets and measure characteristics, our approach was to control the dataset, for being able to actually assign the variations in the measured outputs to the measures, rather than dataset characteristics. As such, the usage scenario builds upon principles of sensitivity analysis approaches, where typically the output of a dependent variable is observed while an independent variable is varied. To come up with a data study that meets the requirements of our rationales, we created a set of datasets. For every dataset in this set, we kept the values of all synthesis parameter constant, except the on that we varied in a controlled way. This variable parameter was the distance between centroids (centers of gravity) of the two classes. In dataset "process000", we started with two completely overlapping classes, and successively increased the Euclidean distance between the classes centroids across datasets. We went for constant change of distances between any two datasets, resulting in a set of datasets with a linear increase of class distances. The distance between the classes in the final dataset was 10 times the size of diameter of the classes (which was identical for both classes). In contrast to the parameter that was varied, the characteristics in Table 1 were all kept constant across datasets. The dataset is publicly available for usage and replication purposes [†]

Dataset Characteristics	Value
Instances	500 per class
Dimensions	5
Classes	2
Class balance	1:1
Diameter	constant, equals in both classes

Table 1: *Parameters for the synthesis of datasets for the usage scenario. These parameters were kept constant for all 100 datasets.*

In the following, we show a scatterplot matrix for a representative subset of the set of datasets. In each of the following figures, the two classes are shown with red and blue color. The figures are ordered according to the sequence of datasets. The interested reader may notice that the classes seem to become smaller across the series: this is NOT the case. In contrast, the axis scale changes within the scatterplots due to the increasing data space (see axis labels). Table 2 provides an overview.

[†] http://juergen-bernard.de/paperPages/euroVA2020Bernard/paper.html
100 Synthetical Datasets for the Assessment of Class Separation Measures

Figure 1: *Dataset ID 000.*

Figure 2: *Dataset ID 010.*

Figure 3: *Dataset ID 020.*

Figure 4: *Dataset ID 030.*

Figure 5: *Dataset ID 040.*

Figure 6: *Dataset ID 050.*
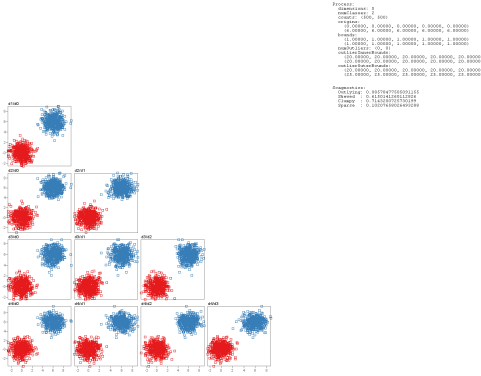
Figure 7: *Dataset ID 060.*

Dataset ID	Figure
"Process000"	Figure 1
"Process010"	Figure 2
"Process020"	Figure 3
"Process030"	Figure 4
"Process040"	Figure 5
"Process050"	Figure 6
"Process060"	Figure 7
"Process070"	Figure 8
"Process080"	Figure 9
"Process090"	Figure 10
"Process100"	Figure 11

Table 2: *Samples of datasets used of the set of datasets used for the usage scenario. Across datasets, the distance between classes increases in a linear way.*
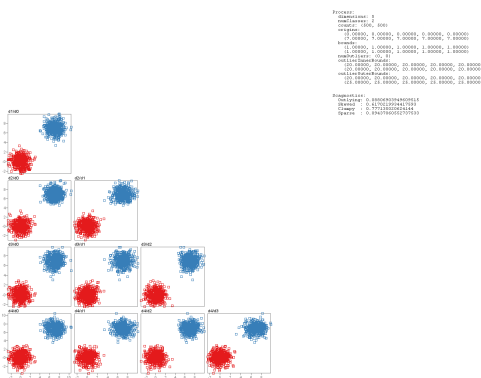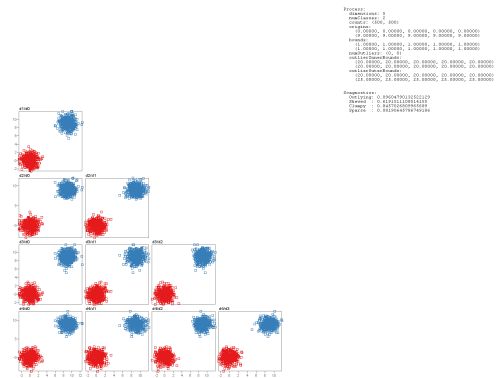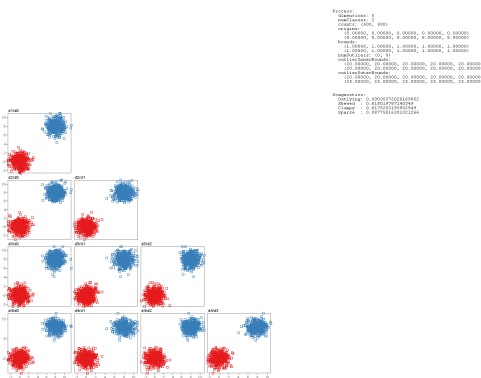

Figure 8: *Dataset ID 070.*

Figure 10: *Dataset ID 090.*

2. Depictions of the Overall System

Due to size limitations, we did not present a system screenshot of SepEx in the paper. Therefore, we presented individual views in the paper separately and at large, by the price that (brushing and) linking capability could not be depicted to its full extent. In the following, we show screenshots of the entire system, one for every figure which made it into the paper. In addition, we extend the analyses made in the usage scenario by incorporating all views. The system screenshots are presented in Figure 12, 14, and 16.
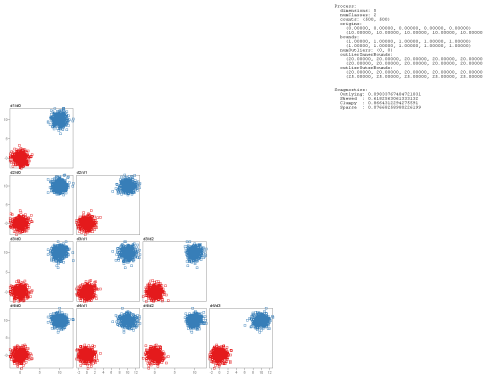

Figure 9: *Dataset ID 080.*

Figure 11: *Dataset ID 100.*

3. Usage Scenario: Detailed Analysis of TSNE Inconsistency

Figure 15 shows the system state at which we observed inconsistencies of TSNE (between measures applied to high-dimensional data and dimensionality-reduced data). Following finding 4), we select at a single dataset, and conduct a more detailed analysis to find out why in this case TSNE results seem to be less useful. We decide to use dataset "Process100", which is the dataset with the highest separation between the two classes among all datasets used (cf. Figure 11).

The nD-2D comparison of measures for the three DRs are shown in Figures 18, 19, and 20. We use the Dunn measure for every DR, and re-identify the inconsistencies of TSNE, as many errors exist when comparing ranks between nD and 2D measure results. Figures 21, 22, and 23 show details about the selected dataset ("Process100").

It can be seen that MDS separates the two classes very well, leaving a large space between the two class distributions. In contrast, TSNE aligns the two class distributions right next to each other. This may explain why class separation measures are rather inconsistent using TSNE for this particular setting.

Finally, we explain the anomaly in the analysis of PCA-based 2D representations: there is a small gap between two datasets (cf. Figure 20). By looking at Figure 23, the scatterplot with the PCA result shows a data representation that only uses one dimension. This can be seen by the horizontal line the data points are aligned at. We infer that WEKA's PCA implementation seems to use a threshold of whether a second principal component is even needed at all. As the distance between the two classes rises, there seems to be some point at which the implementation neglects using a second axis.
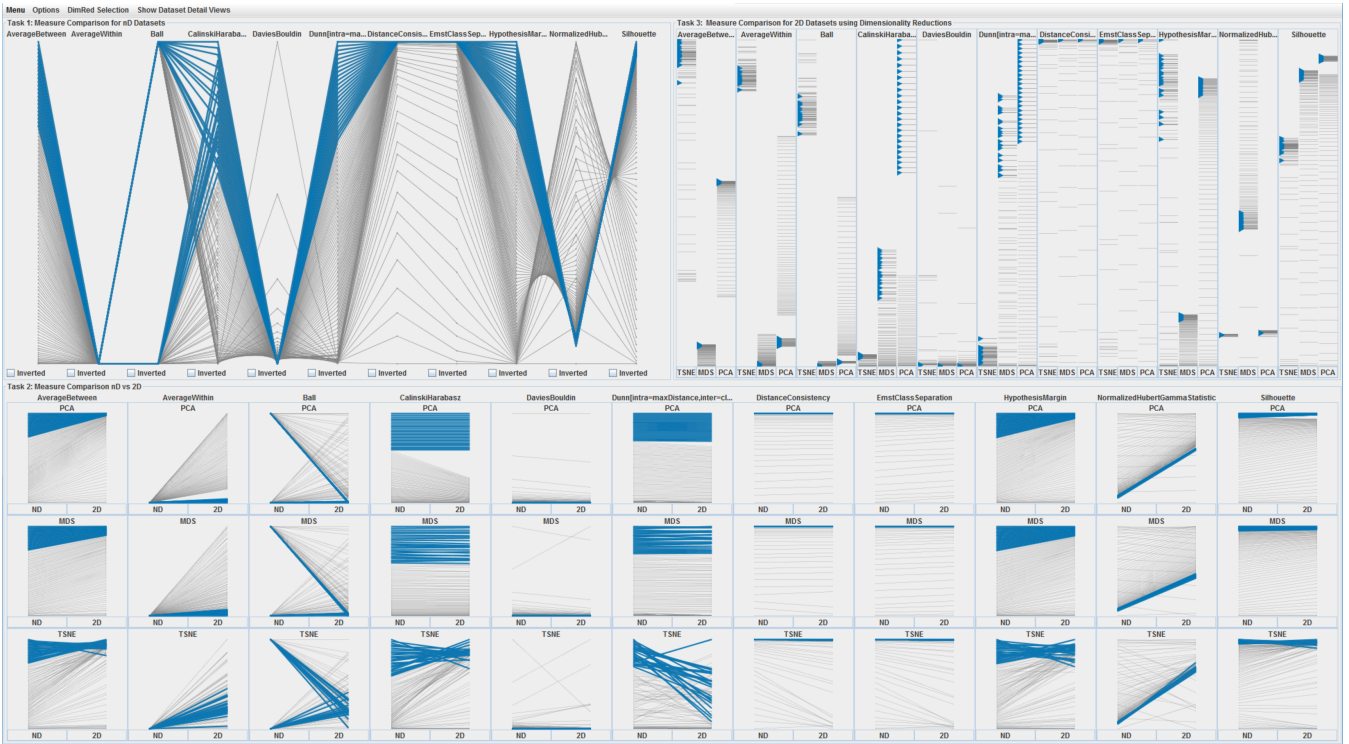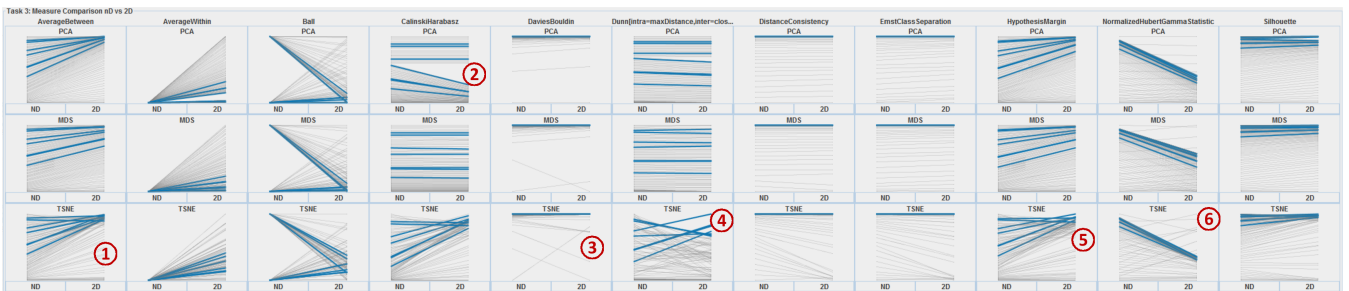
Figure 12: *System state at which the screenshot was taken to demonstrate the interface for T1 in the paper. To better comply with the overview-like nature of the system figure, we de-selected showing axis labels (in contrast to the figure in the paper). Brushing (blue) was used to select the datasets which are separated most. Across the three perspectives of SepEx, it can be seen how measures behave for this particular dataset subset.*
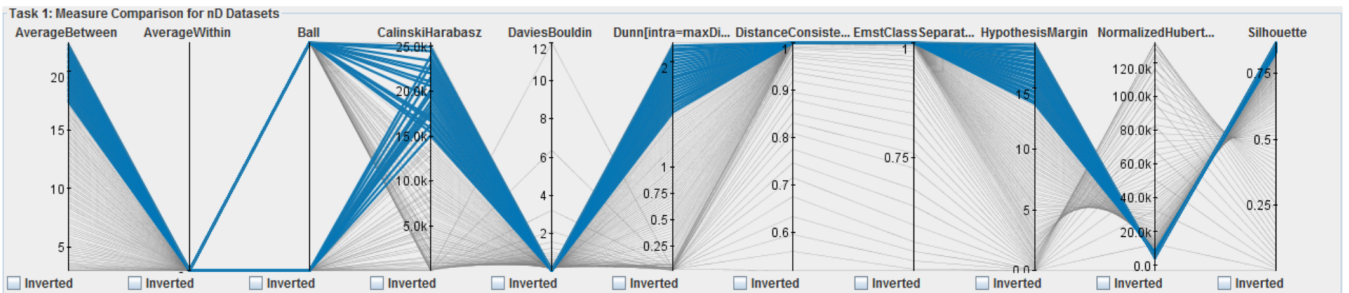

Figure 13: *For the sake of completeness: the figure for T1 used in the paper. To be compared with Figure 12.*
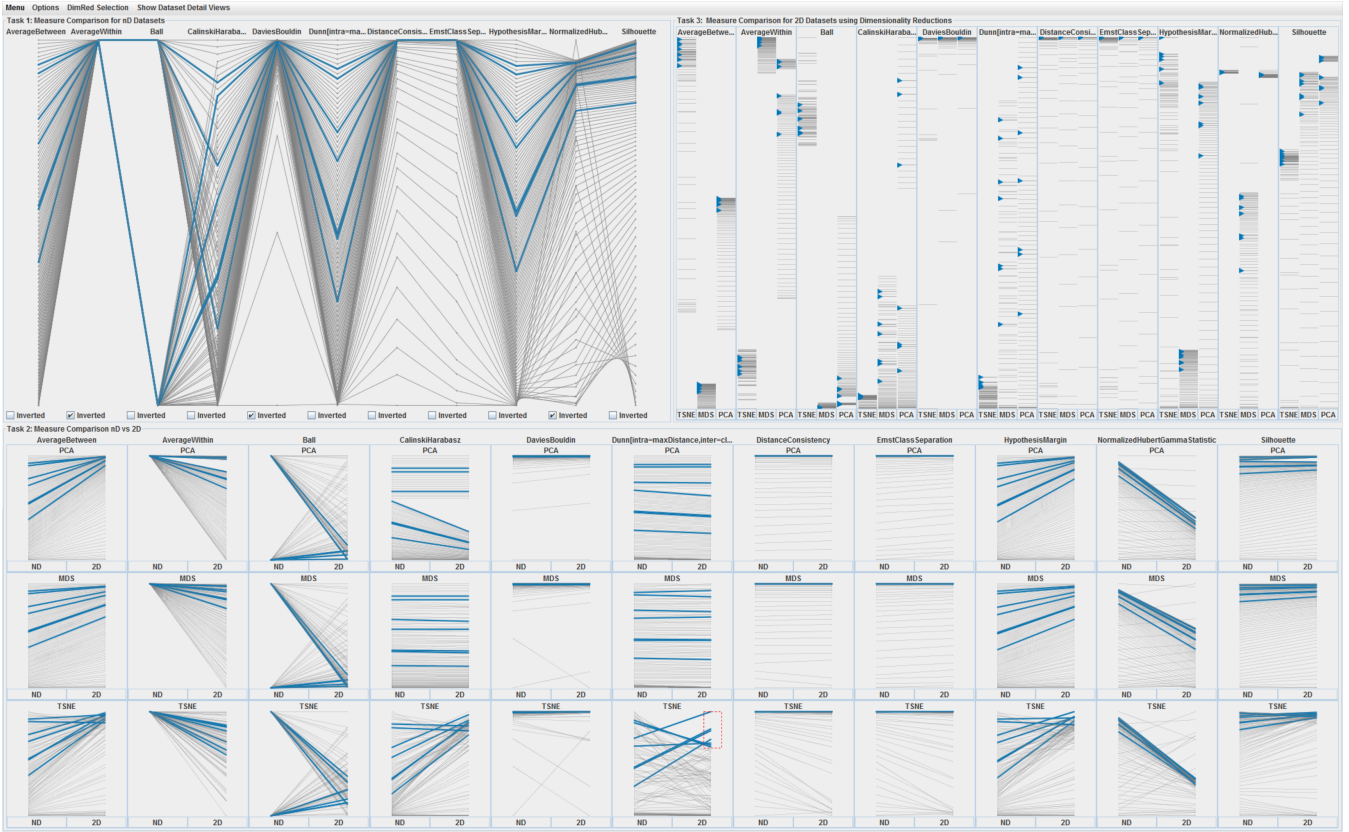
Figure 14: *System state at which the screenshot was taken to demonstrate the interface for T2 in the paper. The current selection was made in the lower left view (supporting T2) and includes the highest class separations measured with Dunn and projected with TSNE (red rectangle). In the nD analysis interface on the upper left (T[1]), it can be seen that this selection does not refer to the dataset with highest class separation. Thus, we identify the inconsistency of TSNE in combination with Dunn.*


Figure 15: *For the sake of completeness: the figure for T2 used in the paper. To be compared with Figure 14. The dataset selection was made very close to the mark of finding (4).*
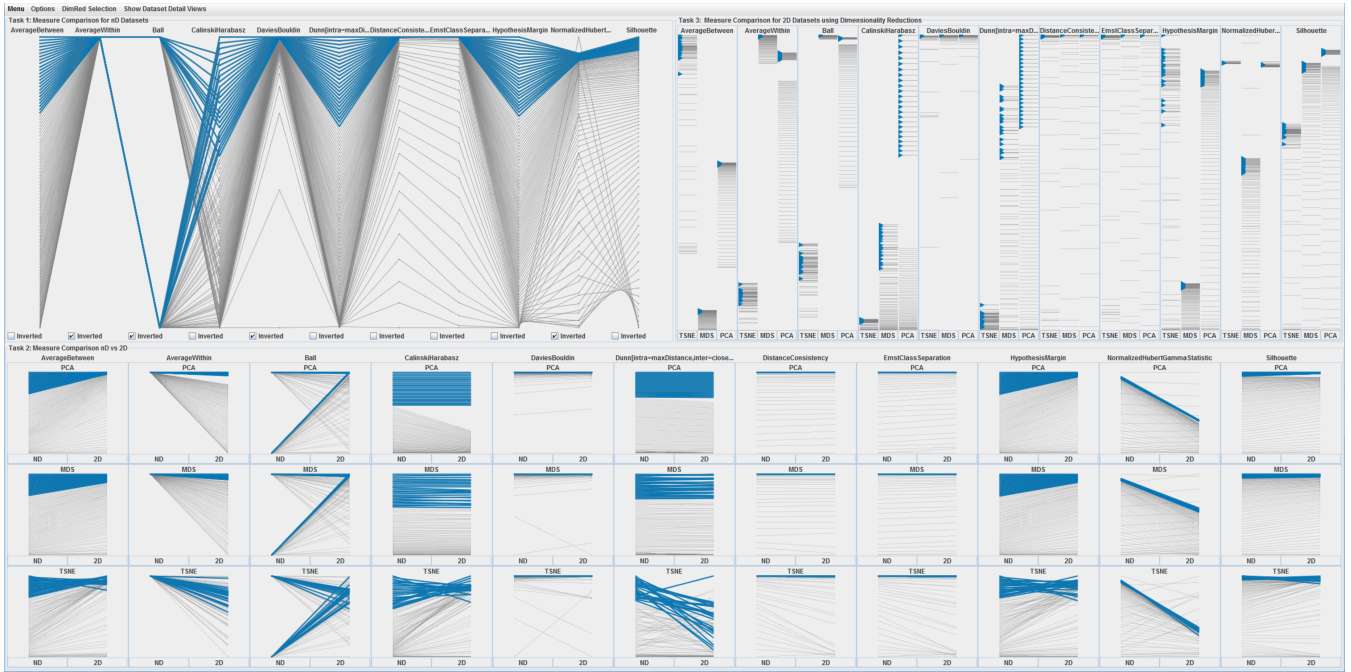
Figure 16: *System state at which the screenshot was taken to demonstrate the interface for T3 in the paper. No selection was made as different findings referring to different subsets have been discussed. In the usage scenario where Figure 17 was used in the paper, we identified patterns of measure behavior (groups with similar measure outputs for DR-reduced 2D data) T3. We now analyze these four patterns (1), (2), (3), and (4) in the other views of SepEx, aiming at generalizing these findings from 2D to nD. Interestingly, the measures of group (1), (3), and (4) also show commonalities for nD data. The situation is different for group (2) where the similar behavior in 2D does not seem to exist for nD data.*
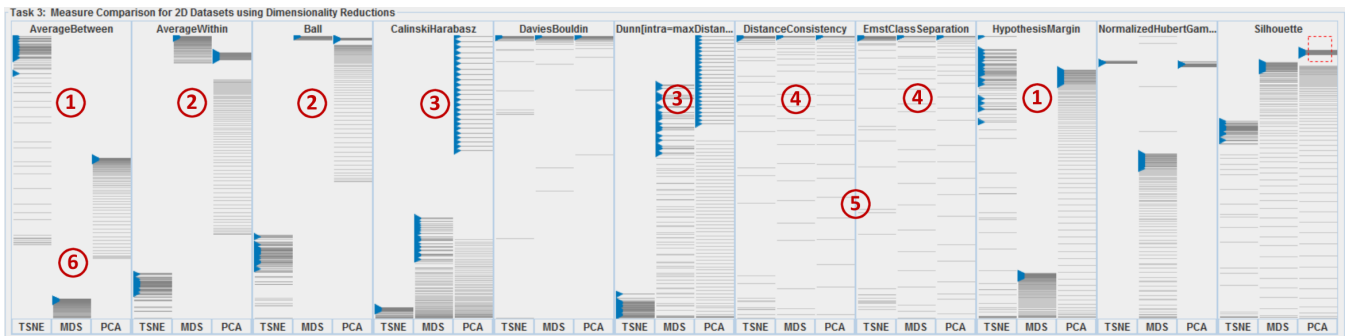

Figure 17: *For the sake of completeness: the figure for T3 (comparison of measure outputs for 3 DRs across 11 measures) used in the paper. To be compared with Figure 16.*
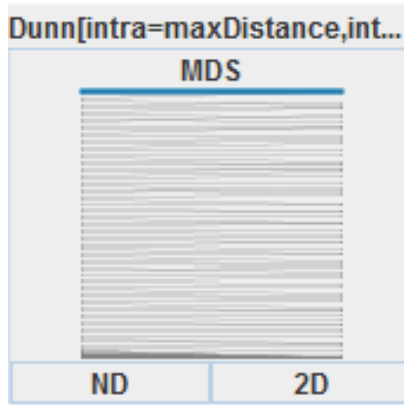
Figure 18: *Rank preservation of the 100 datasets between the high-dimensional datasets and the 2D representations using the Dunn measure and MDS: rank preservation is almost perfect.*
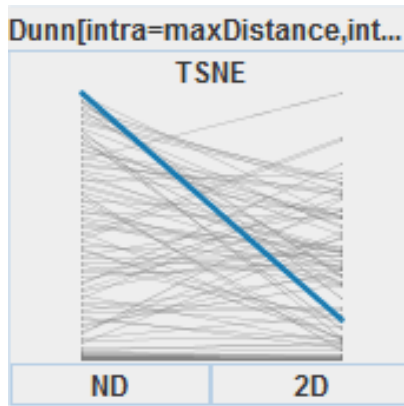
Figure 19: *Rank preservation of the 100 datasets between the high-dimensional datasets and the 2D representations using the Dunn measure and TSNE: we identify some inconsistencies regarding rank preservation.*
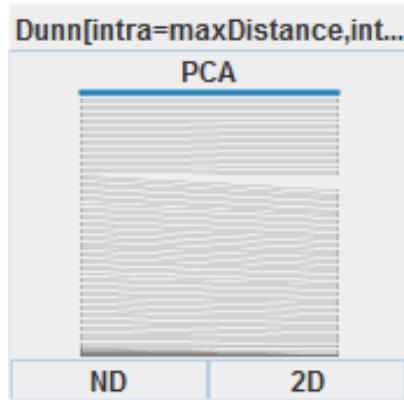
Figure 20: *Rank preservation of the 100 datasets between the high-dimensional datasets and the 2D representations using the Dunn measure and PCA: rank preservation is almost perfect. However, there seems to be a small gap in the 2D-PCA data (right axis), which at first we could not explain.*
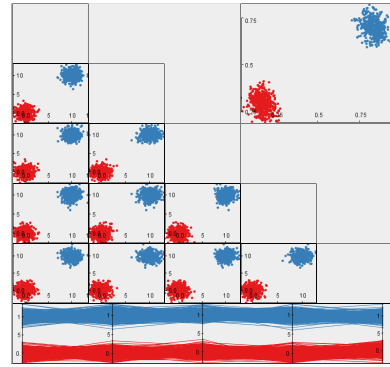
Figure 21: *Detailed analysis of dataset "Process100". A scatterplot matrix (main diagonal left away) and a parallel coordinates plot are showing the original 5D dataset. At the upper right a MDS-based 2D representation of the dataset is shown using a scatterplot. It can be seen that MDS separates the two classes very well.*
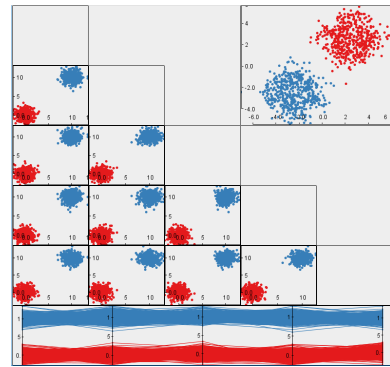
Figure 22: *Detailed analysis of dataset "Process100". A scatterplot matrix (main diagonal left away) and a parallel coordinates plot are showing the original 5D dataset. At the upper right a TSNE-based 2D representation is shown using a scatterplot. It can be seen that TSNE arranges the two classes right next to each other.*
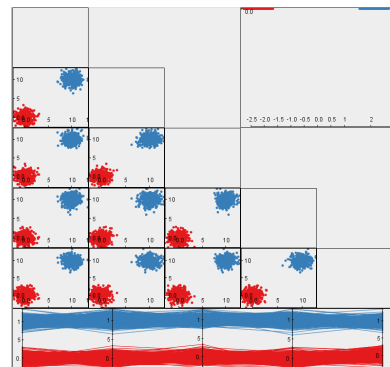
Figure 23: *Detailed analysis of dataset "Process100". A scatterplot matrix (main diagonal left away) and a parallel coordinates plot are showing the original 5D dataset. At the upper right a PCA-based 2D representation of the dataset is shown using a scatterplot. It can be seen that PCA separates the two classes very well, using (needing) only one principal component.*