EuroVis Workshop on Visual Analytics (2022)
M. Angelini and J. Bernard (Editors)

# Reordering Sets of Parallel Coordinates Plots to Highlight Differences in Clusters

Elliot Koh[1], Michael Blumenschein[2], Lin Shao[1,3] and Tobias Schreck[1]

[1] Graz University of Technology, Austria
[2] University of Konstanz, Germany
[3] Fraunhofer Austria Center for Data Driven Design, Austria

**Abstract**

*Visualizing high-dimensional (HD) data is a key challenge for data scientists. The importance of this challenge is to properly map data properties, e.g., patterns, outliers, and correlations, from a HD data space onto a visualization. Parallel coordinate plots (PCPs) are a common way to do this. However, a PCP visualization can be arranged in several ways by reordering its axes, which may lead to different visual representations. Many methods have been developed with the aim of evaluating the quality of reorderings of given PCP view. A high-dimensional data set can be divided into multiple classes, and being able to identify differences between the classes is important. Then, besides overlaying the groups in a single PCP, we can show the different groups in individual PCPs in a small multiple fashion. This raises the problem of* jointly reordering sets of PCPs *to create meaningful reorderings of the set of plots. We propose a joint reordering strategy, based on maximizing the pairwise visual difference in PCPs, such as to support their contrastive comparison. We present an implementation and an evaluation of the reordering strategy to assess the effectiveness of the method. The approach shows feasible in bringing out pairwise difference in PCP plots and hence support comparison of grouped data.*

## 1. Introduction

Analyzing high-dimensional (HD) data is a challenge, and many visualization techniques for it have been proposed to date. For example, dimension reduction techniques [EMK*21] or visualization techniques including scatter plot matrices, pixel-based techniques, glyphs, and Parallel Coordinate Plots (PCP) [BBK*18] can be used. PCPs are useful in visualizing HD data due to their ability to reveal patterns spanning across multiple dimensions in the data and relationships between dimensions. A PCP consists of several vertical axes, which correspond to the dimensions in the data. Data points are then represented by polylines that are connected along the dimension axes and represent feature values in terms of intersections. However, the effectiveness of PCPs highly depend on the ordering of their axes, and different orderings result in different patterns.

Previous work provided different reordering methods for PCP. These methods typically look at properties of individual PCP orderings, such as visually emphasizing clusters [TAE*09] or outliers [PWR04]. Also, reduction of clutter and overlap [ED07] are typical criteria to guide the reordering. However, most approaches only consider the quality of *a single* PCP view. In this paper, we consider the simultaneous reordering of a *set of PCPs*, which show different groups in data in a small multiple fashion. We want to jointly reorder the plots such as to support the comparison of the groups, i.e., find the distribution of values among the axes that are in agreement and that differentiate between the data groups. This is a specialization of the reordering problem of a single PCP showing all groups in the data in one plot. We aim to reorder such that the aggregate pairwise visual difference between all plots in the set is maximized. We assume that as the visual difference gets larger, it becomes easier to compare and contrast the different groups (or clusters) in the data. We present and implementation of this idea, and results of applying it on data sets with success.

## 2. Related Work

PCPs are frequently used to for a wide range of analytical tasks and have been continuously expanded over recent years [YGX*09, CvW11]. Heinrich and Weiskopf [HW13] have presented a classification of common PCP tasks to the estabilshed KDD taxonomy by Fayyad et al. [FPSS96]. To accomplish these tasks, Behrisch et al. [BBK*18] introduced four visual patterns: grouping, correlation, outlier and trend. Furthermore, they named three major challenges for creating reasonable PCP views: (1) By increasing the number of data records, available patterns may vanish due to overplotted lines; (2) The perception of patterns in PCPs heavily depends on the ordering of the dimension axis; (3) A large number of dimensions limits the exploration and screen space between two axes, and may result in cluttered views. To tackles these issues, *quality metrics* may be defined to objectively measure qualities and characteristics of a certain view and determine how good, for example, a PCP is and in which areas it can be improved. In [BBK*18],



DOI: 10.2312/eurova.20221080

different quality metrics are indicated to tackle these issues. For example, Ellis and Dix [ED07] introduced a set of clutter reduction techniques that estimate the occlusion of lines in PCPs. The proposed methods consider (1) overplotting (percentage of pixels with more than one plotted point), (2) overcrowded display (percentage of pixels with more than one existing point), and (3) hidden patterns (percentage of plotted points hidden due to overplotting).

Depending on the axis orderings, different patterns are revealed, and can be made more (or less) visually prominent. Quality measures also allow to compare different axis orderings and find an optimal ordering for displaying the data based on a certain task. In [JJ09] an interactive reordering approach is proposed that is based on user-defined and weighted quality metrics. The metrics consider user-defined weighting for multiple data spaces including dimension correlations, outlier, and cluster detection. Another reordering approach by Lu et al. [LHZ16] uses nonlinear correlation coefficient and singular value decomposition algorithms to measure the contribution of each dimension and reduces computational complexity. In [ARI20], a smart mutation operator for evolutionary algorithms is used to enhance the views of PCPs. The goal is to reduce the line intersections between neighboring coordinates by swapping the two axes that have maximum crossing line values.

Hu et al. [HCX*21] use reinforcement learning to tackle a similar problem than we. They search for optimal ordering of star glyphs to maximize visual difference. First, they extract features of the star glyph set using an encoder network, and then perform reinforcement learning on the features using a distance measure as a reward function to compare orderings. Then they evaluate the performance by conducting user studies, as well as test for generality of the solution by varying the properties different data sets. While this may not give the optimal solution, they show it to be successful in computing orderings that noticeably improve visual difference. In essence, this method tries to solve the same problem we set out to do: optimize for visual difference in visualization of high-dimensional data where dimension order is important. As a result, this method can be likely extended to our problem in the area of parallel coordinates, as well as to other high-dimensional visualizations where dimension order is important.

## 3. Our Approach

We describe how we compute aggregate visual dissimilarity among a set of PCPs, and use this to simultaneously reorder the set.

### 3.1. Representation with Feature Signatures and Total Distance Score

We presume the data is min-max normalized and class labels exist for grouping. First, we compute a feature signature representation ($FS$) for each group of data. This in turn serves to compute an aggregate visual difference measure between the PCPs of a set, called total distance score (TDS). We implemented the following three feature signatures, which have increasing descriptive information: $FS_1$: a vector of data means for each dimension, $FS_2$: the mean and respective standard deviation of each dimension, and $FS_3$: a histogram with 5 bins for each dimension. All these feature signatures are candidates to use (Fig. 1 illustrates).

As the feature signature represents the data, it is also a proxy of the visual appearance of the PCPs created from it. We require an appropriate distance function to compute the visual differences. Computing e.g., Minkowski distances like $L_2$ or $L_1$ between feature signatures is not suitable, as they are invariant regarding the ordering of the dimensions of the data (respectively the PCPs). We require a distance function that takes into account, for each dimension, also the neighboring dimensions, as these are responsible for the visual properties provided by the PCPs.

The *quadratic form distance* (QFD), which is typically used in shape or multimedia retrieval achieves this, as it can compare all dimensions from one vector with all dimensions of the other vector, according to a weight matrix. It is formulated in vector notation as:

$$QFD(FS^a, FS^b) = \sqrt{diff(FS^a, FS^b) \cdot \mathbf{W} \cdot diff(FS^a, FS^b)^T}$$

with $diff$ a vector of difference values between feature signatures $FS^a$ and $FS^b$ of two data groups $a$ and $b$, and $\mathbf{W}$ a weight matrix indicating the influence of each pair of dimensions from the feature signatures to the overall quadratic form distance.

We define $diff$ depending on the type of feature signature used. For $FS_1$ (vector of data means) absolute differences between the dimensions can be used. $FS_2$ (mean and standard deviation) and $FS_3$ (histogram) are multidimensional vectors, where each dimension consists of the mean and variance measure, or bins of the histogram. For these, we compute a difference value for each dimension as the Earth Mover Distance [RTG00].

We define the weight matrix $W$ heuristically as

$$\mathbf{W_{ij}} = 1 - (d_{ij}/(D-1)) \quad \forall i, j \quad \text{where} \quad d_{ij} = |i - j|$$

indicating that the influence of a pair of dimensions gets lower, as the dimension indices in their feature signatures are farther apart.

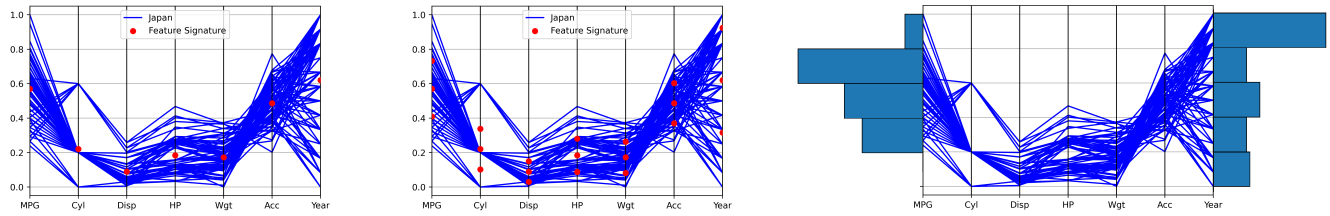
Finally, we compute the Total Distance Score (TDS) for a set of PCPs as the weighted average of all pairwise differences measured by QFD. The weights are set proportional to the size of the respective data group.

### 3.2. Reordering the PCPs

To reorder a set of PCPs, we create all possible reorderings, then compute their TDS, and then select the one with the highest TDS. We note this is an exhaustive search approach with high computational complexity, and hence works for data sets with limited numbers of dimensions and groups. We apply it here for feasibility and evaluation purposes. For larger data sets, speed-up techniques will be required (cf. Section 5).

## 4. Experimental Results

We present qualitative and quantitative experimental results to characterize and compare the effects of our set-based reordering for the three defined feature signatures.




**(a)** $FS_1$*: Mean* **(b)** $FS_2$*: Mean and Standard deviation* **(c)** $FS_3$*: Histogram*

**Figure 1:** *Feature signatures are the basis for computing visual dissimilarity between PCPs for groups of data. We implemented feature signatures based on mean, mean and standard deviation, and histogram.*

### 4.1. Analysis Goals

First, we want to visually inspect the reorderings of highest and lowest TDS scores, to validate the general functionality. Then, we want to analyze the relationship between the TDS scores of reorderings and two established PCP quality measures: polyline length and average correlation of adjacent dimensions in the PCP. Polyline length is a proxy for the complexity of the polyline shape, with lower complexity supposed to be easier to visually grasp. Correlation between dimensions is a key data analysis goal, hence it is also proxy for PCP quality.

### 4.2. Data Sets

We experimented with several data sets (#records, #classes and #dimensions in brackets): Iris [Fis88] (150, 3, 4), Cars [Qui93] (398, 3, 7), Cars_year [Qui93] (398, 6, 6), Wine [For91] (178, 3, 8), Seeds [CNK*12] (210, 3, 7), Glass [Ger87] (205, 5, 7), Ecoli [Nak96] (336, 8, 5), and 2C_6_mod [BZP*20] (245, 5, 5). For the wine and 2C_6_mod data sets, only a subset of the original number of dimensions was used to keep computation times reasonable. The cars_year data set is a modification of the original cars data set (the classes of the cars_year data set were changed to be the year the car was released.) Due to size limitations we only present exemplary results in detail, and give a generalization afterwards.

### 4.3. Cars Data Set with Mean Feature Signature ($FS_1$)

Fig. 2 (top row) shows the reorderings of the Cars data set with lowest (left) and highest (right) TDS scores, respectively, using feature signature $FS_1$. We observe the PCPs with lowest TDS are visually similar, especially the zig-zag shapes of the blue and green plots. The PCPs with hightest TDS, on the other hand, show more dissimilarity between the blue and green plots. They also show more complex PCP patterns among the first three dimensions. In addition, the red plot overall shows more discernible patterns (fan, correlation). In that, the TDS-based reordering may improve the visual analysis potential when using this reordering.

We also observe the TDS-based reorderings do well on PCP quality measures. Fig. 2 (bottom row) shows the correlation between TDS scores, and the per-axis correlation of the PCPs (left) and the average polyline length (right). It indicates that higher-ranked reorderings according to our TDS score provide also higher quality in terms of PCP axis-correlation (as a key analysis target), and shorter average polyline lengths (as lower plot complexity). The correlation strength is moderate with $R^2$ of 0.23 and 0.36.

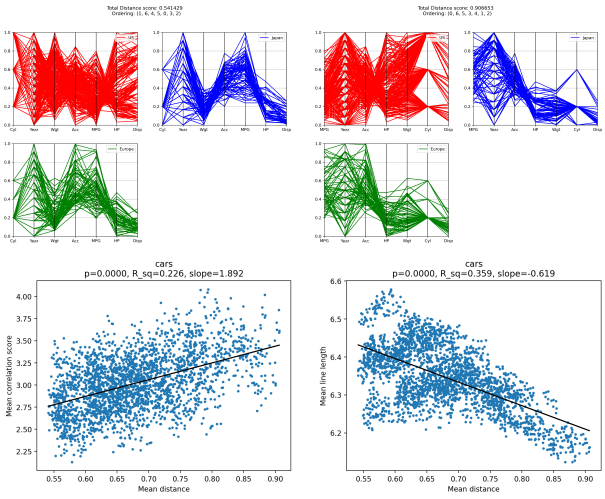



**Figure 2:** *Cars data set,* $FS_1$

### 4.4. Glass Data Set with Mean and Standard Deviation Feature Signature ($FS_2$)

Fig. 3 (top row) shows the reorderings of the Glass data set with lowest (left) and highest (right) TDS scores, respectively, using feature signature $FS_2$. Both sets of reorderings appear similar among and between them. The numeric differenc between the TDS scores is low (0.72 and 0.83, respectively). However, the groups in the reorderings with hightest TDS (top right) appear more crisp. Specifically, the red, blue and green plots (right) show a tighter bundling of the PCP polylines.

For this data set, the reorderings show only weak relationships to the two observed quality measures, correlation score (left) and polyline length (right). The trend seems at first glance to contradict the relationship of the preceding example, but $R^2$ seems too low for validation (0.05 and 0.13, respectively). It indicates that our reordering strategy may not relate to these quality measures for certain data sets like this.

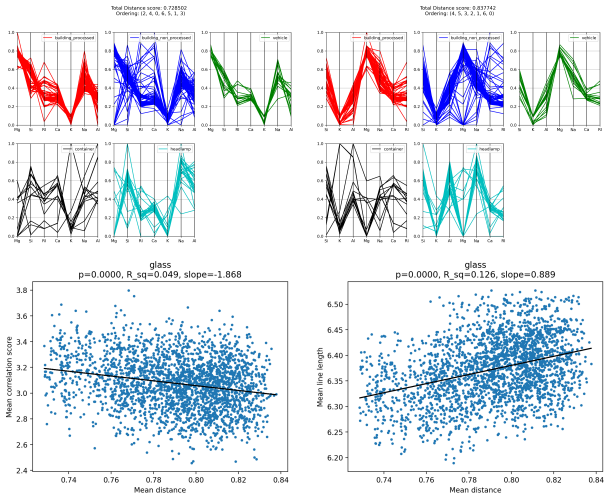



**Figure 3:** *Glass data set, $FS_2$*

### 4.5. Seeds Data Set with Histogram Feature Signature ($FS_3$)

Fig. 4 shows the results for the Seeds data set, using the feature signature $FS_3$. We observe the reordering of highest TDS score (top row, right) produces a flattening effect, amounting to low average polyline length, leading to lower visual complexity than the longer polylines in the reordering with lowest TDS (left). The dependency between TDS scores and PCP axis correlation (left) and polyline length (right) are positive and negative, respectively, like in the Cars data set example (Sect. 4.3) albeit at lower $R^2$.

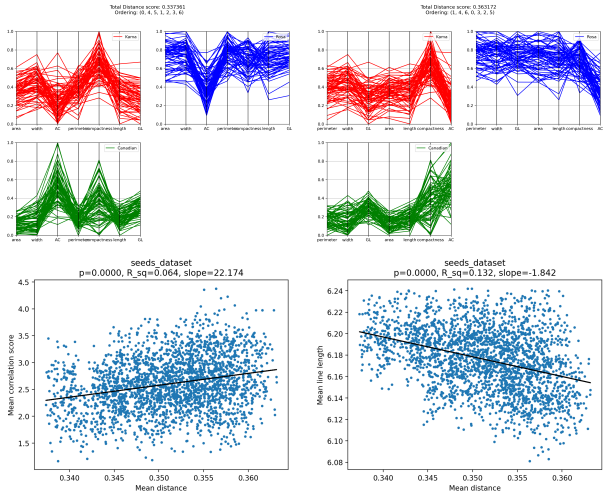



**Figure 4:** *Seeds data set, $FS_3$*

### 4.6. General Observations

Considering all our results, we observe that orderings with higher TDS score tend to have smoother curves compared to the more jagged shape of orderings with smallest distance score (flattening effect). While this appears to occur most often when the $FS_1$ mean feature signature type is used, the effect also occurs for $FS_2$ and $FS_3$. A negative correlation of TDS with polyline length indicates the visually observed flattening effect: As curves are smoothed, the average line length decreases. Not all data sets have PCP orderings that flatten it or that are significantly visually different. When the approach works well in terms of lower polyline length and visual differences, it works comparably so for all $FS$ types. Overall, the TDS score is moderately to weakly related to the PCP axis correlation quality metric. We observed cases of rather low numeric differences between smallest and largest TDS scores, which we found also to correspond to small visual differences in the PCP sets. It may indicate some data sets are hard to reorder for strongly noticeable differences.

## 5. Discussion

More analysis should be done to compare our reorderings with other established PCP quality measures. We considered a selection of data sets and two selected quality measures. From the correlation scatter plots, we see that reorderings of similar TDS values may have a larger spread in terms of axis-wise correlation and polyline length. This indicates we might select not just the reordering of the highest TDS, but a comparably high TDS value which may improve a quality measure. To this end, an interactive exploration of the distribution of TDS scores might be interesting to select the a good reordering for contrastive comparisons.A computational limitation of this method is that it does not scale to data sets with a large number of dimensions, due to the brute force search of reorderings. Improvements could be done by sampling reorderings, or applying optimization e.g., by genetic algorithms. Also, partitioning of the dimensions is possible, where either the user or a subspace search method selects groups of dimensions to sort separately, and then concatenate. In general, visual difference is subjective and not easy to quantify. User preferences could be obtained to guide the reordering strategies. Also, the user might want to fix certain dimensions to keep in place for the mental map. Eventually, a user study could compare our approach for time and error in solving PCP-based analysis tasks.

## 6. Conclusions

We provide a novel reordering approach for sets of PCP plots. We reorder such as to maximize the visual difference between PCPs of a data set. PCPs are represented by feature signatures and a quadratic form distance between all pairs of PCPs is computed. This resulted often in orderings where the PCPs were smoothed which allows for differences between them to be more quickly identified. Also, in some cases more subtle PCP patterns could be brought out, and the PCP plots each showed more bundled. The approach has extension possibilities, e.g., improving scalability and adapting to user preferences.

## Acknowledgment

This work has been supported by the Austrian FFG-COMET-K1 Center Pro$^2$Future (Products and Production Systems of the Future), Contract No. 881844.

## References


[ARI20] ALDWIB K., RAHNAMAYAN S., IBRAHIM A.: Enhancing parallel coordinates visualization using genetic algorithm with smart mutation. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2020), pp. 3746–3752. doi:10.1109/SMC42975.2020.9282852. 2

[BBK*18] BEHRISCH M., BLUMENSCHEIN M., KIM N. W., SHAO L., EL-ASSADY M., FUCHS J., SEEBACHER D., DIEHL A., BRANDES U., PFISTER H., SCHRECK T., WEISKOPF D., KEIM D. A.: Quality metrics for information visualization. *Computer Graphics Forum 37*, 3 (2018), 625–662. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13446, doi:https://doi.org/10.1111/cgf.13446. 1

[BZP*20] BLUMENSCHEIN M., ZHANG X., POMERENKE D., KEIM D. A., FUCHS J.: Evaluating reordering strategies for cluster identification in parallel coordinates. *Computer Graphics Forum 39*, 3 (2020), 537–549. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14000, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14000, doi:https://doi.org/10.1111/cgf.14000. 3

[CNK*12] CHARYTANOWICZ M., NIEWCZAS J., KULCZYCKI P., KOWALSKI P. A., ŁUKASIK S., ŻAK S.: Seeds data set, 2012. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/seeds. 3

[CvW11] CLAESSEN J. H., VAN WIJK J. J.: Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2310–2316. doi:10.1109/TVCG.2011.201. 1

[ED07] ELLIS G., DIX A.: A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1216–1223. doi:10.1109/TVCG.2007.70535. 1, 2

[EMK*21] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics 27*, 3 (2021), 2153–2173. doi:10.1109/TVCG.2019.2944182. 1

[Fis88] FISHER R.: Iris data set, 1988. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/iris. 3

[For91] FORINA M. E. A.: Wine data set, 1991. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/wine. 3

[FPSS96] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI magazine 17*, 3 (1996), 37–37. 1

[Ger87] GERMAN B.: Glass data set, 1987. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/glass+identification. 3

[HCX*21] HU R., CHEN B., XU J., VAN KAICK O., DEUSSEN O., HUANG H.: Shape-driven coordinate ordering for star glyph sets via reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics 27*, 6 (Jun 2021), 3034–3047. URL: http://dx.doi.org/10.1109/TVCG.2021.3052167, doi:10.1109/tvcg.2021.3052167. 2

[HW13] HEINRICH J., WEISKOPF D.: State of the art of parallel coordinates. In *Eurographics* (2013). 1

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 993–1000. doi:10.1109/TVCG.2009.153. 2

[LHZ16] LU L. F., HUANG M. L., ZHANG J.: Two axes reordering methods in parallel coordinates plots. *Journal of Visual Languages & Computing 33* (2016), 3–12. SI:IVIMLA. URL: https://www.sciencedirect.com/science/article/pii/S1045926X15300379, doi:https://doi.org/10.1016/j.jvlc.2015.12.001. 2

[Nak96] NAKAI K.: Ecoli data set, 1996. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/ecoli. 3

[PWR04] PENG W., WARD M., RUNDENSTEINER E.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *IEEE Symposium on Information Visualization* (2004), pp. 89–96. doi:10.1109/INFVIS.2004.15. 1

[Qui93] QUINLAN R.: Cars data set, 1993. Data retrieved from UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/auto+mpg. 3

[RTG00] RUBNER Y., TOMASI C., GUIBAS L.: The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision 40* (01 2000), 99–121. 2

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNORK M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), pp. 59–66. doi:10.1109/VAST.2009.5332628. 1

[YGX*09] YUAN X., GUO P., XIAO H., ZHOU H., QU H.: Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1001–1008. doi:10.1109/TVCG.2009.179. 1