Eurographics Conference on Visualization (EuroVis) (2012)
M. Meyer and T. Weinkauf (Editors)

The Parallel Coordinates Matrix

J. Heinrich[1], J. Stasko[2], D. Weiskopf[1]

[1]Visualization Research Center (VISUS), University of Stuttgart
[2]School of Interactive Computing & GVU Center, Georgia Institute of Technology

Abstract

We introduce the parallel coordinates matrix (PCM) as the counterpart to the scatterplot matrix (SPLOM). Using a graph-theoretic approach, we determine a list of axis orderings such that all pairwise relations can be displayed without redundancy while each parallel-coordinates plot can be used independently to visualize all variables of the dataset. Therefore, existing axis-ordering algorithms, rendering techniques, and interaction methods can easily be applied to the individual parallel-coordinates plots. We demonstrate the value of the PCM in two case studies and show how it can serve as an overview visualization for parallel coordinates. Finally, we apply existing focus-and-context techniques in an interactive setup to support a detailed analysis of multivariate data.

Categories and Subject Descriptors (according to ACM CCS): Probability and Statistics [G.3]: Multivariate Statistics—, Computer Graphics [I.3.3]: Picture/Image Generation—Display algorithms

1. Introduction

The scatterplot is one of the most popular and widely applied visualizations of 2D data. While a single scatterplot represents two dimensions, the scatterplot matrix [Har75] (SPLOM) visualizes all 2D axis-aligned projections of a high-dimensional dataset. This is achieved by laying out 2D scatterplots in a matrix where every row and every column represents one dimension (Figure 1).

Multidimensional data can also be visualized using parallel coordinates [Ins85, Ins09]. Here, a set of parallel axes represent the dimensions while datapoints are rendered as polylines crossing all axes. Exploiting the point-line duality, parallel coordinates with two axes convey the same information as their dual scatterplots, although some training might be required to see the same patterns [LMvW08]. In addition, parallel coordinates allow to visually trace individual datapoints over all axes, providing a multidimensional "profile" of the datapoints. However, the parallel layout of axes also adds the constraint of a fixed ordering of dimensions, hindering the visualization of all pairwise relations in a single parallel-coordinates plot (PCP) without duplicating axes. As can be seen in Figure 1, laying out PCPs (with two dimensions each) in a scatterplot matrix breaks the traceability of lines over all axes and therefore one of the nice properties of parallel coordinates.

To combine the advantages of parallel coordinates and the scatterplot matrix, we introduce the parallel coordinates matrix (PCM) as the counterpart of the scatterplot matrix for parallel coordinates. The design goals of the PCM are to

1. visualize all pairwise correlations without redundancy using parallel coordinates while
2. all PCPs represent the same set of dimensions.

The first design goal is required to ensure that all pairwise correlations are presented to the user, while the second ensures comparability, consistency, and is required to obtain a matrix layout. As a result, the PCM is a list of high-dimensional PCPs, each with a different axis ordering. Since the PCM is composed of a set of PCPs, many existing ordering algorithms, interaction techniques, and visual representations can be used with the PCM.

2. Related Work

Hartigan [Har75] visualized pairs of variables placing two-dimensional scatterplots in a matrix. However, as the layout of 2D plots in the traditional SPLOM is symmetric, more than half of the scatterplots conveys redundant information. Giving order to dimensions using different measures was investigated extensively for the SPLOM [WAG05, SNLH09, ABK98, Hur04] as well as for parallel coordinates [WAG06, DK10, FR11].

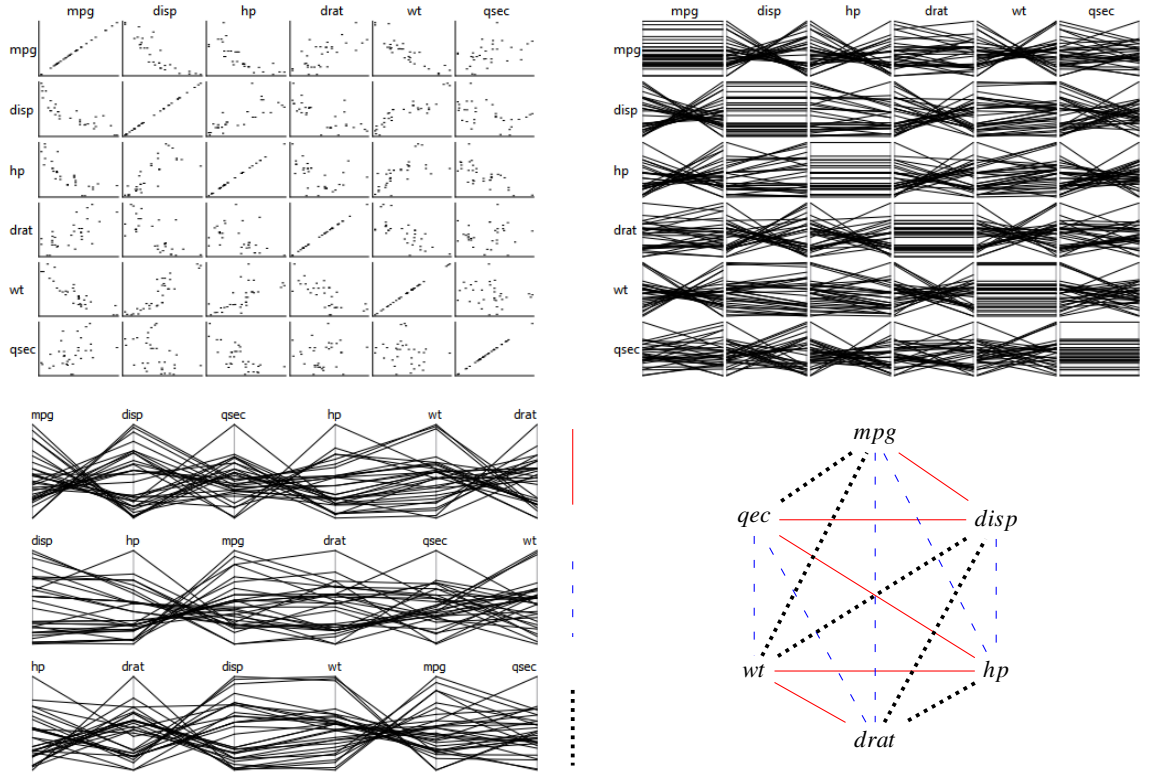

Figure 1: *Replacing scatterplots in the SPLOM (top, left) with 2D parallel-coordinates plots (top, right) conveys the same information, but breaks the continuity of lines. Both visualizations are symmetric such that the whole information is represented by* $\frac{n(n-1)}{2}$ *2D plots. The corresponding parallel coordinates matrix (bottom, left) comprises* $\lfloor \frac{n}{2} \rfloor$ *parallel-coordinates plots, each representing n dimensions, while all pairwise correlations occur exactly once. The nodes of the complete graph* K_6 *(bottom, right) denote the dimensions of the dataset, while edges represent pairwise relations. To construct the parallel coordinates matrix, the graph is decomposed into three Hamiltonian paths (red, blue, and black) describing the order of axes of the three parallel-coordinates plots in the matrix. Together, they form the complete graph such that all pairwise relations are covered. While all correlations can be seen in all three visualizations, the parallel coordinates matrix further shows lines expressing a similar pattern over a subset of variables. This is probably most striking in the third row, where a small set of lines with high values for "disp" move to the top of "wt" before dropping to low values for "mpg".*

Other layouts for 2D PCPs were proposed to visualize single-to-many [JCJ05] and many-to-many [LJC09] relations. For the latter, line continuity is not achieved while the first does not represent all pairwise relations. The P-SPLOM [VMCJ10] comprises the same number of plots as the SPLOM and thus contains the same redundancy. Albuquerque et al. [AEL*09] order PCPs with 3 axes in a matrix of $(n-1)/2$ columns and n rows, rendering a total of n^2-1 pairwise relations.

In the general framework for the layout of 2D plots presented by Claessen and van Wijk [CvW11], axes can be placed freely in Cartesian space such that both a SPLOM and a PCM could be generated. However, doing so still requires a significant amount of manual labor, even for low-dimensional datasets.

3. The Parallel Coordinates Matrix

We describe how the parallel coordinates matrix is obtained from an n-dimensional dataset based on the work by Wegman [Weg90] and Hurley and Oldford [HO10]. Wegman describes how to compute all orderings of PCPs required to see all pairwise relations using a graph-theoretic approach. Hurley and Oldford use a slightly modified algorithm to create a single PCP with all possible pairwise permutations.

3.1. Pairwise Correlation Graph

The first design goal of the PCM is to visualize all (unordered) pairwise relations of an n-dimensional dataset. Finding these relations can be translated to visiting all edges in the undirected complete graph $K_n = (V, E)$, where the set

of vertices $V = \{1,...,n\}$ represent dimensions and the set of edges $E = \{e_{ij} | i,j \in V, i \neq j \text{ with } e_{ij} = e_{ji}\}$ represent 2D relations between dimensions i and j. The number of pairwise relations (edges) is $|E| = \frac{n(n-1)}{2}$.

Using independent 2D data representations (such as scatterplots), all plots may be distributed arbitrarily over the available space in any table layout. For parallel coordinates, however, it might be beneficial to exploit the multidimensional nature of the plot, such that polylines representing individual data points can be traced across more than two axes. Hence, it is desirable to see every dimension at least once in every PCP. Using a graph description, this translates to a path in the corresponding complete graph that visits all vertices at least once.

3.2. Eulerian Trails and Hamiltonian Decomposition

A *Hamiltonian decomposition* is an edge decomposition of a graph into *Hamiltonian paths* or *Hamiltonian cycles*. An *Eulerian trail* is a trail in a graph that visits every edge exactly once and an *Eulerian cycle* is an Eulerian trail that ends in the starting vertex. A Hamiltonian path is a path in a graph that visits every vertex exactly once and a Hamiltonian cycle is a Hamiltonian path ending in the starting vertex.

There are $(n-1)!$ Hamiltonian cycles for the complete graph K_n. We employ the *Lucas-Walecki Hamiltonian decomposition* to obtain $m = \frac{n}{2}$ Hamiltonian paths for even n and $m = \frac{n-1}{2}$ Hamiltonian cycles for odd n. In the following, we use the construction algorithms described by Hurley and Oldford [HO10]. For $n = 2m$, we construct the $m \times n$ layout-matrix H^n by defining

$$\begin{aligned} H^n[1,1] &= 0 \\ H^n[1,j] &= (H^n[1,j-1] + (-1)^j (j-1)) (\text{mod } n) \\ H^n[k,j] &= (H^n[k-1,j] + 1) (\text{mod } n) \end{aligned}$$

where $j = 2,...,n$ and $k = 2,...,m$. Adding one to every value results in a matrix of indexes to dimensions that we use to layout axes on the available canvas. The rows of H^n are Hamiltonian paths in K_n. For $n = 6$ the layout matrix is:

$$H^6 = \begin{matrix} 1 & 2 & 6 & 3 & 5 & 4 \\ 2 & 3 & 1 & 4 & 6 & 5 \\ 3 & 4 & 2 & 5 & 1 & 6 \end{matrix}$$

Hurley and Oldford [HO10] concatenate the rows to form an Eulerian trail T that is used to render one "long" PCP. This has the disadvantage of introducing duplicate edges between vertices $H^n[i,n]$ and $H^n[i+1,1]$. Instead, we use the rows of H^n as axis order for $\frac{n}{2}$ independent PCPs.

For $n = 2m+1$, H^n is constructed by adding n at the beginning and the end of each row of H^{n-1}. This results in m Hamiltonian cycles in K_n. Concatenating the cycles and duplicating the common vertices at $H^n[i,n]$ and $H^n[i+1,1]$ results in an Eulerian cycle of K_n. Using this algorithm, H^7 reads:

$$H^7 = \begin{matrix} 7 & 1 & 2 & 6 & 3 & 5 & 4 & 7 \\ 7 & 2 & 3 & 1 & 4 & 6 & 5 & 7 \\ 7 & 3 & 4 & 2 & 5 & 1 & 6 & 7 \end{matrix}$$

4. Results

Figures 2 and 3 show the PCM for two datasets [Ins09, EDF08] of different dimensionality. Figure 3 exemplifies linking & brushing as well as the use of a focus-and-context technique similar to the Table Lens [RC94] with the PCM. The accompanying video [Hei12] further demonstrates the same analyses in an interactive setup. Note that the analysis conducted here was driven by looking for patterns first, followed by investigating which dimensions contribute to these patterns. This complies with the visual information-seeking mantra [Shn96], as no particular question about the data has been raised prior to the analysis.

5. Discussion and Conclusion

To the best of our knowledge, the PCM is the first visualization presenting all pairwise correlations using parallel coordinates without redundancy for any number of dimensions. Using a simple layout algorithm, the PCM serves as a promising overview for PCPs making it a valuable tool to get an idea of a dataset and then focus on individual relations or plots. Due to the fact that the rows of a PCM are composed of independent high-dimensional PCPs, different rendering or interaction techniques can easily be incorporated, as we have shown in a small example using linking and brushing as well as a focus-and-context technique. Highlighting axes representing the same data dimension is another possible interactive addition.

In contrast to the SPLOM, the PCM makes more efficient use of the available screen real-estate, as pairwise relations appear only once. However, the layout of the SPLOM facilitates labeling and navigation to particular scatterplots. We hypothesize that the SPLOM performs better at finding the relation of a particular pair of dimensions, which however needs yet to be confirmed by a user study. Conversely, if the task is exploratory such that recognition of patterns is more important than finding a specific pair or dimensions, we argue that the analyst might benefit from the space gained using a PCM instead of a SPLOM. In any case, it is important to note that the PCM is not intended to replace the SPLOM, but to be its natural counterpart for parallel coordinates.

Acknowledgments

In part, this work was supported by the German Research Foundation (DFG) within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

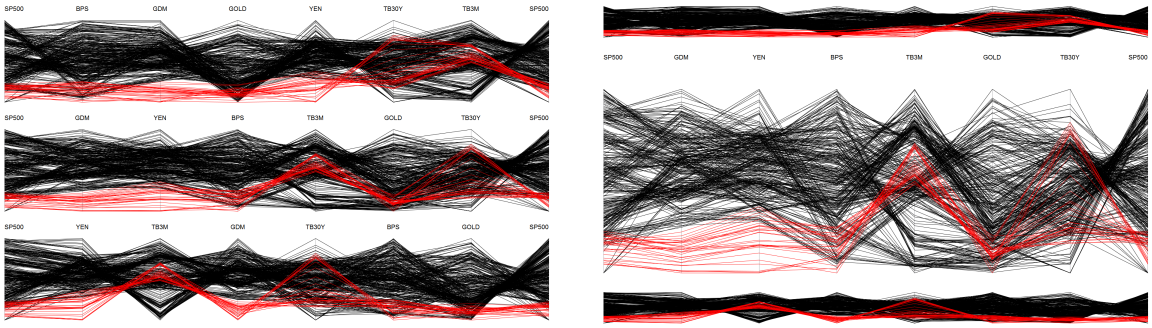

Figure 2: *PCM of a 7-dimensional financial dataset [Ins09]. In this dataset, every line represents weekly stock-market quotes over a period of several years. Starting in the bottom row of the left PCM, we note a small cluster at the bottom between the "SP500" index and "GOLD" prices, indicating a positive correlation. The lower-left part of the middle row shows another positive correlation between "SP500", "GDM" (German Dmark), and "YEN". Being interested in this pattern, we brush it and see the corresponding lines in the other plots. For a detailed view, the middle PCP has been focused (right). Now we see that low "SP500", "GDM", "YEN", and the British Pound Sterling "BPS" go with a negative correlation between "BPS" and "TB3M" (interest rates in percent for the first three months). As expected, "GOLD" prices are low, while "TB30Y" (interest rates in percent for 30-year bonds) varies in the mid-price section.*

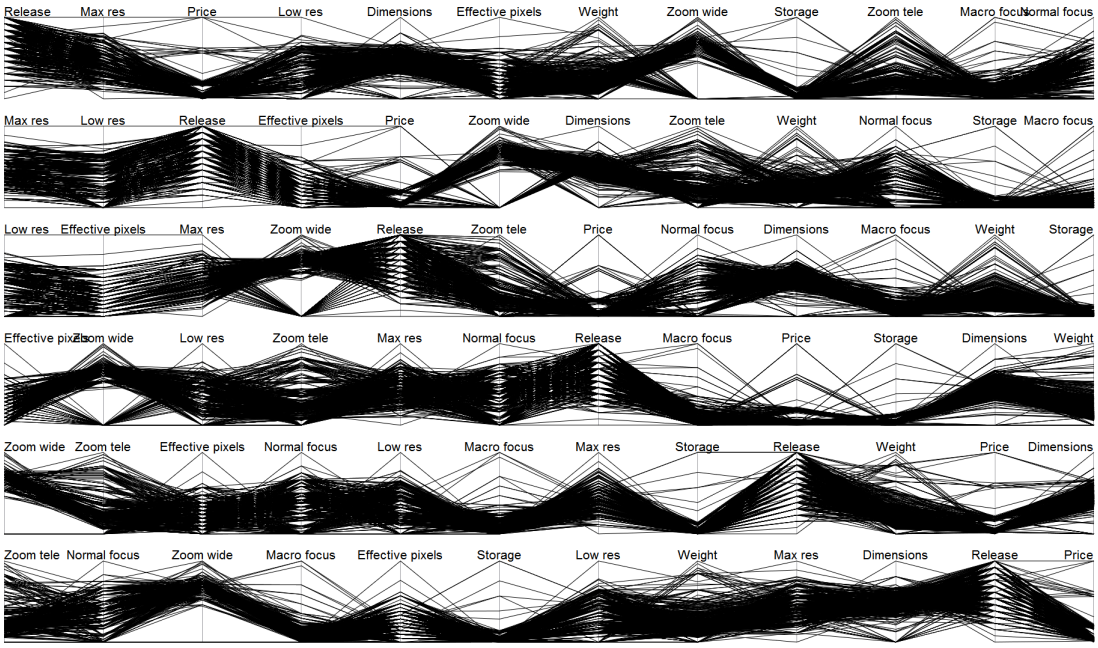

Figure 3: *PCM of the 12-dimensional cameras dataset [EDF08]. The "Price" and neighboring "Max res[olution]" and "Low res[olution]" in the first row show us that (1) there are three dense price-segments: two low-cost segments, a small set of mid-price models, and only three expensive cameras. From the direction of lines leaving the "Price" axis for the mid-priced models, we can tell that the distributions of "Max" and "Low" resolutions is similar and there are no outliers. This is more difficult to say for low-cost cameras, as their resolutions seem to have a "wider" distribution over the neighboring axes. The "Price" in row four suggests that the price for a camera does not necessarily predict the storage included. The most expensive models come without storage. Regarding the "Zoom wide" dimension, if an analyst only had the bottom-most PCP for analysis, he might think at first glance that there is single outlier with no zoom at all, as we see a perfectly horizontal line to the neighboring axes. Comparing this with row number four, it becomes evident that there are many of such models.*

References

[ABK98] ANKERST M., BERCHTOLD S., KEIM D.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings of the IEEE Symposium on Information Visualization* (1998), pp. 52–60. 1

[AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D., THEISEL H., MAGNOR M.: Quality-based visualization matrices. In *Vision, Modeling, and Visualization (VMV)* (2009), pp. 341–349. 2

[CvW11] CLAESSEN J. H. T., VAN WIJK J. J.: Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2310–2316. 2

[DK10] DASGUPTA A., KOSARA R.: Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1017–1026. 1

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J. D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1539–1148. 3, 4

[FR11] FERDOSI B. J., ROERDINK J. B. T.: Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Computer Graphics Forum 30*, 3 (2011), 1121–1130. 1

[Har75] HARTIGAN J.: Printer graphics for clustering. *Journal of Statistical Computation and Simulation 4*, 3 (1975), 187–213. 1

[Hei12] HEINRICH J.: The parallel coordinates matrix supplement page. http://www.vis.uni-stuttgart.de/pcm, 2012. 3

[HO10] HURLEY C. B., OLDFORD R. W.: Pairwise display of high-dimensional information via Eulerian tours and Hamiltonian decompositions. *Journal of Computational and Graphical Statistics 19* (2010), 861–886. 2, 3

[Hur04] HURLEY C. B.: Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics 13*, 4 (2004), 788–806. 1

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 4 (1985), 69–91. 1

[Ins09] INSELBERG A.: *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer, 2009. 1, 3, 4

[JCJ05] JOHANSSON J., COOPER M., JERN M.: 3-dimensional display for clustered multi-relational parallel coordinates. In *Proceedings of the 9th International Conference on Information Visualization* (2005), pp. 188–193. 2

[LJC09] LIND M., JOHANSSON J., COOPER M.: Many-to-many relational parallel coordinates displays. *Proceedings of the 13th International Conference Information Visualisation* (2009), 25–31. 2

[LMvW08] LI J., MARTENS J., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2008), 13–30. 1

[RC94] RAO R., CARD S. K.: The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1994), pp. 318–322. 3

[Shn96] SHNEIDERMAN B.: The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (1996), pp. 336–343. 3

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum 28*, 3 (2009), 831–838. 1

[VMCJ10] VIAU C., MCGUFFIN M. J., CHIRICOTA Y., JURISICA I.: The FlowVizMenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1100–1108. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization* (2005), pp. 157– 164. 1

[WAG06] WILKINSON L., ANAND A., GROSSMAN R.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1363–1372. 1

[Weg90] WEGMAN E. J.: Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association 85*, 411 (1990), 664–675. 2