Making Many-to-Many Parallel Coordinate Plots Scalable by Asymmetric Biclustering

Hsiang-Yun Wu*
Keio University, Japan

Yusuke Niibe
Keio University, Japan

Kazuho Watanabe
Toyohashi University of Technology, Japan

Shigeo Takahashi
University of Aizu, Japan

Makoto Uemura
Hiroshima University, Japan

Issei Fujishiro
Keio University, Japan

Abstract

Datasets obtained through recently advanced measurement techniques tend to possess a large number of dimensions. This leads to explosively increasing computation costs for analyzing such datasets, thus making formulation and verification of scientific hypotheses very difficult. Therefore, an efficient approach to identifying feature subspaces of target datasets, that is, the subspaces of dimension variables or subsets of the data samples, is required to describe the essence hidden in the original dataset. This paper proposes a visual data mining framework for supporting semi-automatic data analysis that builds upon asymmetric biclustering to explore highly correlated feature subspaces. For this purpose, a variant of parallel coordinate plots, many-to-many parallel coordinate plots, is extended to visually assist appropriate selections of feature subspaces as well as to avoid intrinsic visual clutter. In this framework, biclustering is applied to dimension variables and data samples of the dataset simultaneously and asymmetrically. A set of variable axes are projected to a single composite axis while data samples between two consecutive variable axes are bundled using polygonal strips. This makes the visualization method scalable and enables it to play a key role in the framework. The effectiveness of the proposed framework has been empirically proven, and it is remarkably useful for many-to-many parallel coordinate plots.

1 Introduction

Due to current advanced measurements, collected datasets tend to be large and multivariate. To visualize such complex data, parallel coordinate plots (PCP) has become a popular technique [5]. In PCP, all dimension variables are represented as parallel line axes and each data sample is represented as a polyline connecting the positions of the sample on the axes. Although PCP has been a prominent approach for observing correlation between adjacent axes, investigating all combinations of axis order and comparing two separate axes are still time-consuming tasks. One solution to this problem relies on an extension of PCP called many-to-many parallel coordinate plots (many-to-many PCP) [6], which is extended by placing the drawing of all pairs of axes so that users can compare every pair of dimension variables at a glance. One drawback of PCPs is that they cause visual clutter due to the increase in samples and dimensions because the drawing area for each pair of axes becomes limited and thus unexpected distortion occurs when the axis needs to be placed away from the diagram center.

Indeed, reducing visual complexity in analysis processes has been recently tackled. One typical solution is subspace exploration techniques [10, 12, 13, 15, 17], which present extracted feature subspaces for users to accelerate the analysis process; however, these incorporated with PCPs still suffer from the aforementioned axis ordering problem and cannot intuitively show axis correlation of all pairs. For this purpose, in this paper, we employ the asymmetric biclustering approach [15] to reduce the number of data samples and dimension variables without editing their original properties and then present the variable correlations using many-to-many PCPs. Our main contribution can be summarized as follows:

- Provide a framework with multiple views to visually conduct appropriate feature subspace extraction;
- restate the drawing algorithm of many-to-many PCP and extend the algorithm to any dimension; and
- perform several analyses and demonstrate the effectiveness of using many-to-many PCP.

The paper is organized as follows: Section 2 provides a survey and Section 3 summarizes our visualization framework. Section 4 details the employed asymmetric biclustering approach, followed by an explanation of the visualization in Section 5. Results will be presented in Section 6, and the paper will be concluded in Section 7.

2 Related Work

2.1 Subspace Exploration

Multivariate data exploration has become popular due to the time-consuming process of feature extraction. In visual analysis models proposed by Turkay et al. [12, 13], multivariate statistical analysis is employed to project a multivariate dataset onto screen space. Tatu et al. [10] introduced a similarity measure between a pair of subspaces for finding significant subspaces. Yuan et al. [17] presented data sample distribution and dimension correlation for users to accelerate interactive data exploration. Recently, Watanabe et al. [15] proposed a semi-automatic approach, which simultaneously clusters data samples and dimension variables to visualize highly correlated feature subspaces with enhanced PCPs.Nonetheless, aligning axes in parallel increases the visual difficulty of comparing two nonadjacent individual axes. In this paper, we revisit the approach of Watanabe et al. [15] for extracting highly correlated subsets and subspaces of data while incorporating many-to-many PCPs to overcome the weakness of classical PCPs.

2.2 Variants of Parallel Coordinate Plots

A pioneering work on PCP in visualization can be traced back to Inselberg [5], which has been commonly used for visualizing multivariate datasets. PCP aligns dimension variables using parallel axes and uses polylines to connect the value of a data sample along each axis, and thus is effective to present correlations between adjacent axes. To overcome the weakness of PCP, many-to-many PCP was proposed by Lind et al. [6]. It aligns all pairs of axes in a star shape, so that the correlations of all pairs of dimension variables appear exactly once in the diagram. Nonetheless, many-to-many PCPs causes visual clutter due to the amount of information being visualized, thus making it difficult to directly apply to multivariate datasets. Several techniques have also been investigated to reduce the visual complexity of classical PCP. Nohno et al. [8] employed Pearson's correlation coefficient to contract dimensions into

*e-mail: hsiang.yun.wu@acm.org

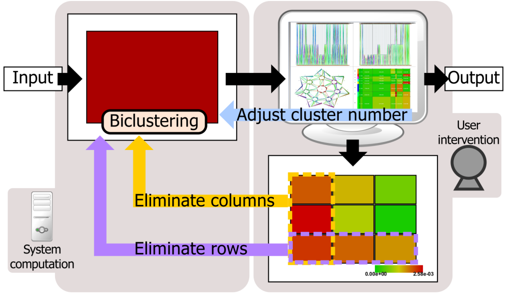
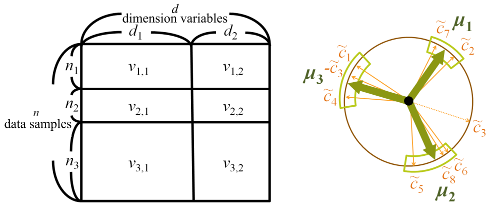


Figure 1: Overview of our visualization framework.



(a) (b)

Figure 2: (a) Schematic data matrix representation of the block model for $K = 3$ and $L = 2$. (b) A concept of a spherical k-means algorithm.

a single composite axis to show trends inherent in multivariate data. Besides, edge bundling was introduced to group highly correlated data samples [7, 18]. Palmas et al. [9] introduced strip rendering styles to enhance the readability of grouped data samples. In our approach, we also bundle multiple data samples [9] to eliminate the visual complexity inherited from the many-to-many PCP.

3 Overview of the Visualization Framework

This section summarizes the visualization framework of the proposed approach. The present approach consists of two main functions, which include (1) biclustering highly correlated data samples and dimension variables and (2) providing a coordinated view for visual data analysis. Specifically, our approach begins with automatically extracting highly correlated data samples and dimension variables using the asymmetric biclustering technique [15], thus providing multiple PCP views together with a newly introduced many-to-many PCP view for data exploration to give full correlation information after reducing the data. Figure 1 depicts the analysis scenario using our prototype system. First, users can input their initial guess on the number of clusters of data samples and dimension variables and then visually confirm the clustering results by refining the initial choice of cluster number. In other words, if the correlation between a pair of clusters is still high, users can try to reduce the dimension cluster number before deciding to remove them from the target dataset. Alternatively, they can further investigate what relationships these clustered dimensions implicitly have, since many-to-many PCP shows all combinations of correlation between each pair of dimension variables. With this scheme, users can thus justify the removal of poorly-correlated dimension variables first or merge separated clusters into one more effectively.

4 Asymmetric Biclustering Algorithm

As described previously, the reduced data presented on the screen space is extracted by the approach of Watanabe et al. [15] and thus the clustered content is expected to be highly correlated. In this section, we briefly introduce how this biclustering approach works in our visualization system; for more details, refer to [15]. Practically, the asymmetric biclustering approach allows us to cluster data samples and dimension variables simultaneously and thus effectively extract the highly correlated portion of the data. This asymmetric biclustering approach is extended from block modeling techniques [4], which decompose a data matrix into several subblocks and find the optimal partition under a predefined objective function.

Figure 2 shows an example of the data structure in this approach, which includes n samples in d-dimensional space that are divided into $K \times L$ submatrices. The data matrix with its rows and columns sorted is therefore approximated as:

$$\begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} \simeq \begin{pmatrix} v_{1,1}E_{n_1 d_1} & \cdots & v_{1,L}E_{n_1 d_L} \\ \vdots & \ddots & \vdots \\ v_{K,1}E_{n_K d_1} & \cdots & v_{K,L}E_{n_K d_L} \end{pmatrix}, \quad (1)$$

where E_{kl} is the $k \times l$ matrix with all ones. Based on the decomposed blocks, the conventional biclustering approach defines the following squared error as the objective function:

$$\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij} - v_{\kappa(i),\lambda(j)})^2, \quad (2)$$

where $v_{k,l} \in \mathbf{R}$ indicates the mean value of each block ($k = 1,2,\cdots,K$ and $l = 1,2,\cdots,L$). Here, $\kappa(i) \in \{1,\cdots,K\}$ ($i = 1,\cdots,n$) is the data sample cluster assignment and $\lambda(j) \in \{1,\cdots,L\}$ ($j = 1,\cdots,d$) is the dimension variable cluster assignment.

Once sample and dimension cluster labels have been initialized, the algorithm is accomplished by iteratively updating Eq. (3) and assigning Eq. (4) and Eq. (5), respectively.

$$v_{k,l} = \frac{1}{n_k d_l}\sum_{i:\kappa(i)=k}^{n}\sum_{j:\lambda(j)=l}^{d} x_{ij} \quad (3)$$

$$\kappa(i) = \underset{k\in 1,2,\ldots,K}{argmin}\sum_{j=1}^{d}(x_{ij} - v_{k,\lambda(j)})^2 \qquad (i = 1,2,\ldots,n) \quad (4)$$

$$\lambda(j) = \underset{l\in 1,2,\ldots,L}{argmin}\sum_{i=1}^{n}(x_{ij} - v_{\kappa(i),l})^2 \qquad (j = 1,2,\ldots,d) \quad (5)$$

Nonetheless, this classical biclustering algorithm uses k-means clustering for clustering both data samples and dimension variables and the correlations between dimensions of each cluster could be missed. To address this, Watanabe et al. introduced spherical k-means-based clustering techniques for clustering of dimensions. Generally, the spherical k-means can be considered as the k-means algorithm on the unit hypersphere, as illustrated schematically in Figure 2(b). In the spherical k-means algorithm, the mean vector is normalized so that the center of each cluster (i.e., $\tilde{c}_j \in \mathbf{R}^n$ and $||\tilde{c}_j|| = 1$) has a minimal angle between components. For this reason, the block mean value is redefined as

$$v_{k,l} = \frac{\frac{1}{n_k}\bar{v}_{k,l}}{\sqrt{\sum_{k=1}^{K}\frac{(\bar{v}_{k,l})^2}{n_k}}}, \quad (6)$$

and the objective function is then updated as

$$Errs = \sum_{j=1}^{d}||\tilde{c}_j - s(j)\mu_{\lambda(j)}||^2 = 2d - 2\sum_{j=1}^{d}s(j)\tilde{c}_j\cdot\mu_{\lambda(j)}, \quad (7)$$

where $\mu_{\lambda(j)}$ is the lth mean vector of dimensions and $s(j) \in \{-1,+1\}$ are indicators showing that $\mu_{\lambda(j)}$ is positively or negatively correlated with $\tilde{c}_j$ to accomplish asymmetric property between data samples and dimension variables. The algorithm initializes the cluster labels using the k-means++ method and iteratively computes the solution until convergence.

5 Coordinated Views for Visual Analytics

Once the highly correlated data samples and dimension variables have been extracted, we display the data using our coordinated

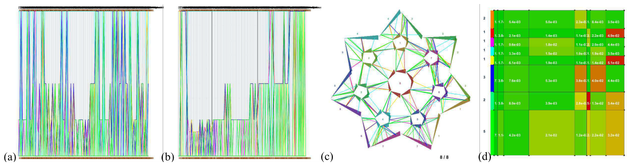


Figure 3: Coordinated view of the system, including (a) a classical PCP, (b) a clustered PCP, (c) a many (one)-to-many PCP, and (d) a block matrix diagram.

view. In our visualization framework, the coordinated view includes the (a) classical PCP, (b) clustered PCP, (c) many (one)-to-many PCP, and (d) block matrix diagram, to illustrate how the multivariate data is reduced and depict the correlation between the clustered dimensions (Figure 3). Because the other three views were introduced in [15], this section concentrates on the restatement of many-to-many PCP and its variant one-to-many PCP.

5.1 Many-to-Many PCP

As mentioned in Section 2, Lind et al. proposed the concept of many-to-many PCP and evaluated the effectiveness of the techniques [6]. Nonetheless, the idea of how they replicate dimension variables can be traced back to Theisel's work [11], and the position of each axis in the drawing was not fully described. The solution for small-dimension variables (< 3) is trivial, and Lind et al. solved the problem for drawing seven-dimension variables. Claessen and van Wijk then extended Lind's work and resolved the problem until eight-dimension variables [3]. In our work, we revisit this problem by referring to the related work from the study of Claessen and van Wijk [3] to extend the problem and summarize the solutions for variables with any integer number of dimensions for general usage.

Algorithm 1 presents our approach for drawing many-to-many PCP. Note that for the drawing algorithm for less than eight dimension variables, we integrate the solution from Claessen and van Wijk [3]. The proposed drawing algorithm consists of three portions, including drawing axes in the outer, middle, and inner scopes from the diagram center (see Figure 4(d) and (h)). For each scope, we iteratively draw the corresponding number of axes in the local area from the outer to the inner scope. Figure 4 provides variations of many-to-many PCP with different numbers n of dimension variables for $3 \le n \le 10$. For $3 \le n \le 6$, we draw the many-to-many PCPs as shown in Figure 4(a)-(d). When $n = 2k+1 \ge 7 (k \in \mathbb{N})$, axes are repeatedly and convexly drawn toward a predefined circular boundary (outer scope) and are aligned on the boundary of a $k+2$ polygon (middle scope), respectively. In addition, if $n = 2k+2 \ge 8 (k \in \mathbb{N})$, an $n-1$ polygon is added at the center accompanied by an $n-1$ many-to-many PCP.

5.2 One-to-Many PCP

To more deeply investigate a specific dimension variable, we present another view called one-to-many PCP, which is partially extracted from the aforementioned many-to-many PCP. This view is presented because comparing a specific dimension variable against the remaining dimension variables is a common procedure during the analysis process. Thus, showing the correlation between a combined dimension variable again, the other clustered dimension variables can be considered, thus promoting further user analysis. By removing the data not related to the selected dimension, more screen space can be reused for representing the information.

6 Results and Discussion

This section presents two experimental results together with discussions on the present limitations of the approach. Our system is implemented on a desktop PC with Intel Core-i7 CPU (3.4GHz) and 4GB RAM. The source code was written in C++ using GSL, OpenGL, Mesa, and GLUI libraries for advanced computation.

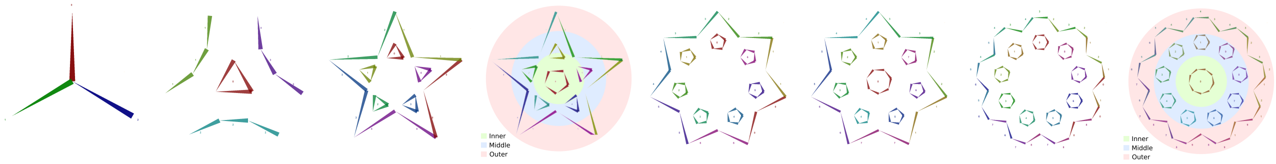
Algorithm 1 Constructing Many-to-Many PCP

```
Suppose the number of dimension variables is equal to $n$.
if $n < 3$ then
    Draw as a classical parallel coordinate plot
else
    if $n < 5$ then
        Draw dimension variables as shown in Figure 4(a) and (b);
    else
        if $5 \le n < 7$ then
            for $i = 1$ to 5 do
                Rotate $2\pi/(2 \times \lfloor \frac{n+1}{2} \rfloor - 1)$ degrees and draw 2 convex edges as shown in Figure 4(e) (outer scope);
                Rotate $2\pi/(2 \times \lfloor \frac{n+1}{2} \rfloor - 1)$ degrees and draw a triangle as shown in Figure 4(e) (middle scope);
            end for
        else
            Set $k = \lceil \frac{n-2}{2} \rceil$;
            for ($i = 1$ to $2 \times k + 1$) do
                Rotate $2\pi/(2 \times \lfloor \frac{n+1}{2} \rfloor - 1)$ degrees and draw a $k-1$ convex polygon as shown in Figure 4(h) (outer scope);
                Rotate $2\pi/(2 \times \lfloor \frac{n+1}{2} \rfloor - 1)$ degrees and draw a $k+2$ polygon as shown in Figure 4(h) (middle scope);
            end for
        end if
        if $(n \bmod 2) = 0$ then
            Draw $n-1$ polygon at the center and align axes on the boundary of the polygon (see Figure 4(h) (inner scope));
        end if
    end if
end if
```

6.1 USDA National Nutrient Data

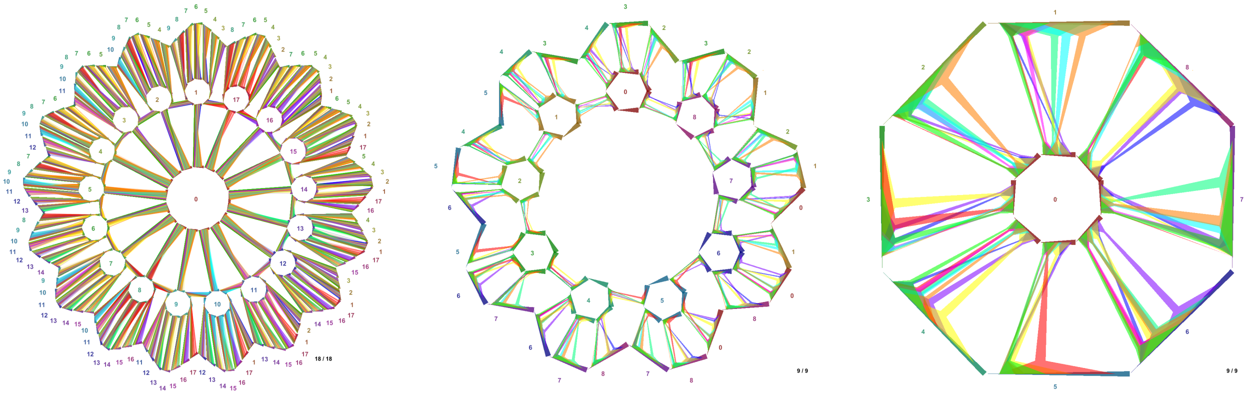
In our first experiment, we employed the USDA food composition dataset [1], which has been used in several studies [10, 15, 17]. Each data sample here corresponds to a specific food and the dimensions represent different nutrients in this food. To compare with the aforementioned results [15], we set $K = 9$ and $L = 9$ in this experiment and followed their analysis process. Figure 5 presents the visualization result for the USDA dataset, where image (a) shows the result before and (b) and (c) show the result after applying the biclustering approach. The distortion of axes in many-to-many PCPs is improved and display area is saved, and resulting in a scalable result for using many-to-many PCPs.

Based on the previous studies, we know that Energy, Lipid, and Water are highly correlated [17], and cluster 0 (Energy and Water) and cluster 6 (Lipid and VitaminE) are also highly correlated [15]. However, with our many-to-many PCP ($n = 9$), we found that the correlation between cluster 0 (combined Energy and Water) and cluster 8 (combined Calcium, Carbohydrate and Sodium) is relatively and positively high, as indicated by the lower amount of crossing between these two combined dimension variables. Moreover, after confirming the correlation using one-to-many PCP ($n = 9$), we found that the correlation between cluster 0 and cluster 8 is higher than the one between cluster 0 and cluster 6 (combined Protein and Vitamin B6), as stated in Watanabe's work (see Figure 5(c)). Based on this result, we further investigate the mutual correlations between the elements in clusters 0 and 8. We discovered that if either Calcium, Carbohydrate or Sodium is removed from this combined cluster, we cannot find similar correlation result as described previously. In other words, the high correlation between cluster 0 and cluster 8 is newly found once the



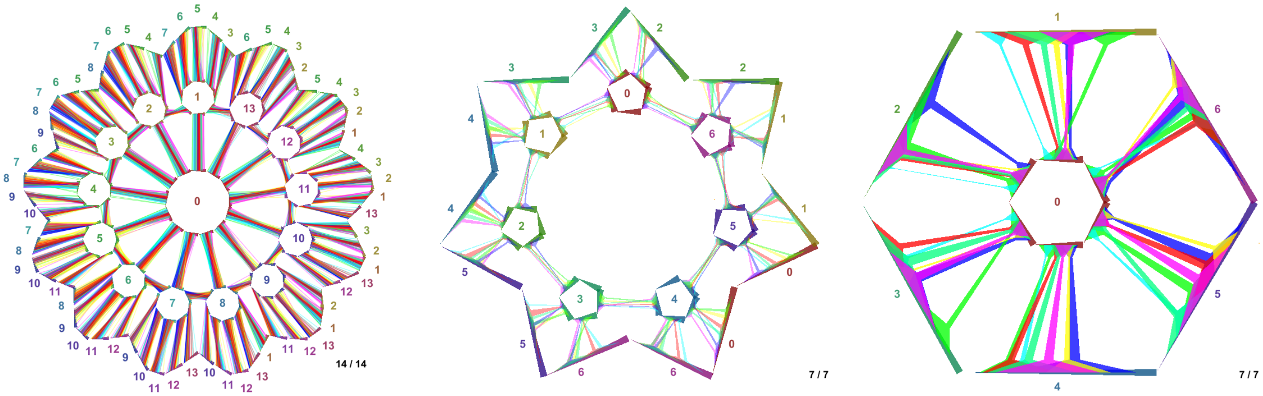
(a) $n = 3$ (b) $n = 4$ (c) $n = 5$ (d) $n = 6$ (e) $n = 7$ (f) $n = 8$ (g) $n = 9$ (h) $n = 10$

Figure 4: Many-to-many PCPs for the case of $3 \leq n \leq 10$.



(a) Initial many-to-many PCP (b) Many-to-many PCP (c) One-to-many PCP

Figure 5: Visualization results on USDA food nutrient data with 722 records and 18 dimensions [1]. Numbers here represent cluster IDs.



(a) Initial many-to-many PCP (b) Many-to-many PCP (c) One-to-many PCP

Figure 6: Visualization results on supernovae dataset with 132 data samples and 14 dimensions [2]. Numbers here represent cluster IDs.

corresponding dimension variables are merged. Previous work also shows that correlation of Energy and Water, and correlation of Calcium, Carbohydrate and Sodium are high [15]; we can therefore conclude that cluster 8 may be a latent variable behind the dataset. Using the many-to-many PCP allows users to review all correlated pairs and thus it reduces the probability of missing information during the analysis process.

6.2 UC Berkeley Supernovae Dataset

In the second experiment, we employed the Berkeley Supernova Database (SNDB). Each data sample represents an observed supernova and each dimension indicates the observed parameters [2, 14]. In this experiment, we biclustered the data matrix using $K = 7$ and $L = 7$ and visualized the results with a many-to-many PCP (see Figure 6). Here, we see that cluster 0 is highly related to some combined dimension variables. We thus switch to one-to-many PCP ($n = 7$) and analyze the correlation between cluster 0 and other combined dimension variables (Figure 6(c)). Here, correlations between cluster 0 and clusters 4 and 6 are high. In astronomy, the magnitude is determined by the brightness parameter and our result clearly supports this statement. Weakly correlated Si5 parameters are also well known in the astronomy community, and these correlations can be discovered by even novice users of our system.

7 Conclusion and Future Work

In this paper, we presented an approach to make many-to-many PCPs scalable by introducing the asymmetric biclustering approach so that the visual complexity complexity of the result is effectively reduced. Potential extensions include comparison of our visualization framework to the probability-based asymmetric biclustering approach [16]. Effectively showing the correlation transitions of clusters as they vary with time is also an essential topic.

Acknowledgements

This work has been supported by MEXT KAKENHI under Grant-in-Aid for Scientific Research on Innovative Areas No. 25120014.

References

[1] Nutient data : USDA national nutrient database for standard reference. http://www.ars.usda.gov/Services/docs.htm?docid=8964.

[2] University of California, Berkeley. The Filippenko group's supernova database (SNDB), 2015. http://heracles.astro.berkeley.edu/sndb/.

[3] J. H. T. Claessen and J. J. van Wijk. Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2310–2316, 2011.

[4] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[5] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[6] M. Lind, J. Johansson, and M. Cooper. Many-to-many relational parallel coordinates displays. In *Proceedings of the 13th International Conference on Information Visualisation (iV2009)*, pages 25–31, 2009.

[7] M. T. McDonnell and K. Mueller. Illustrative parallel coordinates. *Computer Graphics Forum*, 27(3):1031–1038, 2008.

[8] K. Nohno, H.-Y. Wu, K. Watanabe, S. Takahashi, and I. Fujishiro. Spectral-based contractible parallel coordinates. In *Proceedings of the 18th International Conference on Information Visualisation (iV2014)*, pages 7–12, 2014.

[9] G. Palmas, M. Bachynskyi, A. Oulasvirta, H. Seidel, and T. Weinkauf. An edge-bundling layout for interactive parallel coordinates. In *Proceedings of the IEEE Pacific Visualization Symposium 2014 (PacificVis 2014)*, pages 57–64, 2014.

[10] A. Tatu, F. Maaß, I. Farber, E. Bertini, T. Schreck, T. Seidl, and D. Keim. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceeding of tje IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 63–72, 2012.

[11] H. Theisel. Higher order parallel coordinates. In *Proceedings of the 5th Fall Workshop on Vision*, pages 415–420, 2000.

[12] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions - A dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011.

[13] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, 2012.

[14] M. Uemura, K. S. Kawabata, S. Ikeda, K. Maeda, H.-Y. Wu, K. Watanabe, S. Takahashi, and I. Fujishiro. Data-driven approach to type Ia supernovae: Variable selection on the peak luminosity and clustering in visual analytics. *Journal of Physics: Conference Series*, 699(1):012009, 2016.

[15] K. Watanabe, H.-Y. Wu, Y. Niibe, S. Takahashi, and I. Fujishiro. Biclustering multivariate data for correlated subspace mining. In *Proceedings of the IEEE Pacific Visualization Symposium 2015 (PacificVis 2015)*, pages 287–294, 2015.

[16] K. Watanabe, H.-Y. Wu, S. Takahashi, and I. Fujishiro. Asymmetric biclustering with constrained von Mises-Fisher models. *Journal of Physics: Conference Series*, 699(1):012018, 2016.

[17] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, 2013.

[18] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. *Computer Graphics Forum*, 27(3):1047–1054, 2008.