Biclustering Multivariate Data for Correlated Subspace Mining

Kazuho Watanabe*
Toyohashi University of Technology

Hsiang-Yun Wu
The University of Tokyo

Yusuke Niibe
Keio University

Shigeo Takahashi
The University of Tokyo

Issei Fujishiro
Keio University

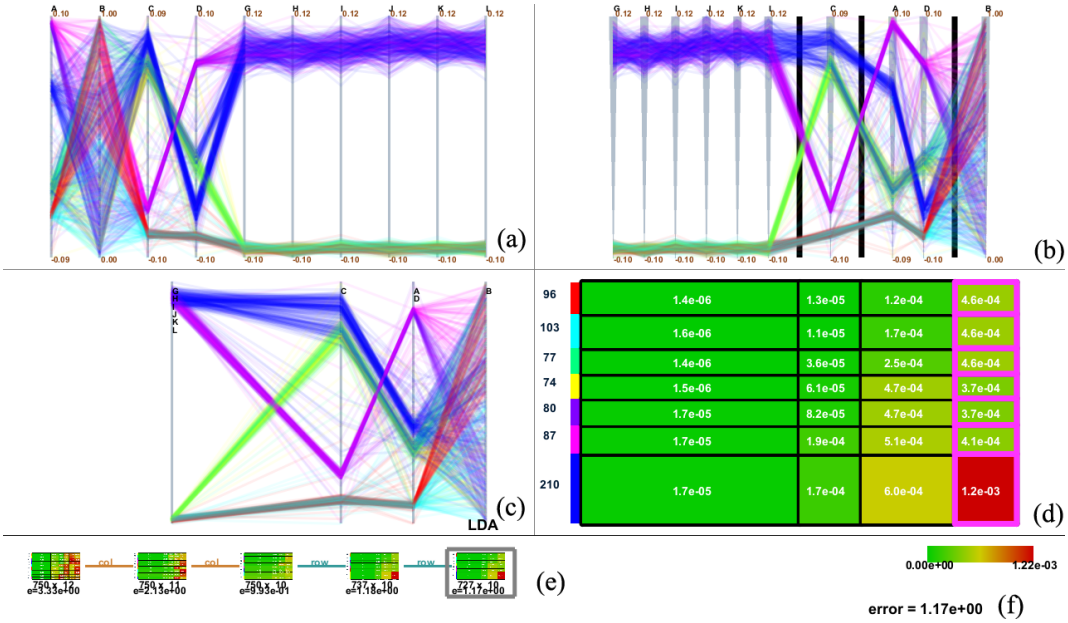


Figure 1: Screenshot of our system interface for finding correlated subspaces based on biclustering. (a) Classical parallel coordinate plot (PCP). (b) Clustered PCP. (c) Contracted PCP. (d) Block matrix diagram. (e) History tree. (f) Objective function value.

Abstract

Exploring feature subspaces is one of promising approaches to analyzing and understanding the important patterns in multivariate data. If relying too much on effective enhancements in manual interventions, the associated results depend heavily on the knowledge and skills of users performing the data analysis. This paper presents a novel approach to extracting feature subspaces from multivariate data by incorporating biclustering techniques. The approach has been maximally automated in the sense that highly-correlated dimensions are automatically grouped to form subspaces, which effectively supports further exploration of them. A key idea behind our approach lies in a new mathematical formulation of asymmetric biclustering, by combining spherical k-means clustering for grouping highly-correlated dimensions, together with ordinary k-means clustering for identifying subsets of data samples. Lower-dimensional representations of data in feature subspaces are successfully visualized by parallel coordinate plot, where we project the data samples of correlated dimensions to one composite axis through dimensionality reduction schemes. Several experimental results of our data analysis together with discussions will be provided to assess the capability of our approach.

Keywords: Multivariate data, subspaces, biclustering, correlation

Index Terms: I.3.8 [Computer Grahics]: Applications;

1 Introduction

With the rapid proliferation of high-performance computing and measurement facilities, the simulated/measured numerical datasets have been in common getting bigger and more complicated at an accelerated pace. A larger number of samples with higher precision and more associated attributes would deserve being analyzed, whereas it would become much harder for the data analysts to extract the important patterns due to their overwhelming structural complexity. Indeed, pursuit of effective mechanisms that allow the analysts to explore feature *subspaces* from the given dataset can be thought of as a central research topic in the current visualization and VAST communities [24, 31]. If relying too much on effective enhancements in manual interventions, as seen in several conventional approaches [24, 31], the associated results depend heavily on the knowledge and expertise of the data analysts.

This paper therefore builds upon *biclustering* techniques to come up with a novel approach to extracting feature subspaces from multivariate data [7, 17]. Unlike the conventional approaches, highly-correlated dimensions are automatically grouped to form subspaces, where we project the data samples of the correlated dimensions to one composite axis through dimensionality reduction schemes. To achieve *simultaneous* clustering of highly-correlated dimensions and data samples, we derive a novel *asymmetric* biclustering method that combines spherical k-means clustering for grouping highly-correlated dimensions [4, 3], together with ordinary k-means clustering for identifying subsets of data samples. A progressive style of visual exploration of subspaces, coupled with graceful elimination of uncorrelated blocks of subspace, can lead effectively to lower-dimensional representations of data in feature subspaces, which are successfully visualized by *parallel coordinate plot (PCP)* and its recent variants ameliorating intrinsic visual clutter artifacts.

Figure 1 shows a screenshot of our system interface of coordinated view, where a synthetic 12D dataset is being analyzed with the proposed approach. After two noisy dimensions and a small cluster of data samples have been identified and removed from the original dataset, the remaining 10D data are visualized with clas-

*e-mail: wkazuho@cs.tut.ac.jp

sical PCP (Figure 1(a)). We assume 4 column by 7 data sample block clusters, and delineate their dimensional correlations with the clustered PCP (Figure 1(b)) and their statistics with block matrix diagram (Figure 1(d)) where we can find another noisy dimension. Shown in Figure 1(c) is the resulting contracted PCP. In Figure 1(e), we are allowed to obtain an overview of the subspace search history.

The remainder of this paper is organized as follows. Section 2 provides a brief survey on multivariate data clustering and visualization. Section 3 gives a key idea to make use of biclustering for feature subspace exploration. Section 4 and Section 5 detail the underlying biclustering method and the visualization framework, respectively. Section 6 discusses effective usage guidelines of the proposed approach by using the synthetic dataset and reports empirical evaluation through the application to two practical datasets. Section 7 concludes the paper and refers to future extensions.

2 Related Work

This section provides a survey on relevant techniques for exploring feature subspaces from multivariate data, including data clustering and visual data mining.

2.1 Data Clustering

Classical clustering techniques such as k-means clustering provide us with a fundamental means of grouping data samples into a specific number of clusters, for better understanding of the underlying structure of the data [15]. While these techniques are effective for finding a set of data samples that are close to each other in the data domain, they do not offer any way to identify a set of highly-correlated dimensions at the same time. *Biclustering*, also known as co-clustering or two-mode clustering, is capable of solving this problem in the sense that it performs a simultaneous clustering of the rows (data samples) and columns (dimensions or axes) of a data matrix consisting of multivariate data samples [7]. A lot of methods have been proposed for biclustering and applied to data analyses in a variety of fields such as bioinformatics, sociometrics, and archaeology (see [17] and references therein). As one of the most popular methods of biclustering, we focus on partitioning methods, which have been extensively studied under the name of *block modeling*. The most fundamental approach can be considered as a two-way generalization of the classical k-means algorithm for both the rows and columns of a data matrix (Section 4.3), and hence is interpreted as a probabilistic modeling using the Gaussian distribution. In this view, an extension using the Bayesian inference method has been proposed for finite discrete variables and named as the stochastic block model [19]. Another extension using the exponential family distribution proposed in [23] can deal with diverse data types such as binary and non-negative integers more appropriately by choosing Bernoulli and Poisson distributions for instance. However, these approaches do not help us seek for highly-correlated subspaces and subsets of data samples from the given multivariate data since they just *symmetrically* treat the clusterings of rows and columns by sharing common mean values for each partitioned block. Thus, in this paper, we present an extension of the conventional block modeling that deals with the clusterings of rows and columns *asymmetrically*, by employing the correlation coefficient as a similarity measure between columns (i.e., dimensions). This is achieved by introducing the *spherical k-means clustering* for grouping of columns (dimensions), which maximizes the sum of the correlation coefficient of each dimension from the corresponding cluster centroid dimension [4, 3] (Section 4.2). This biclustering method is more suitable for common techniques of visualizing multivarite data such as PCPs and scatterplot matrices, because conventional biclustering methods based on the symmetric applications of k-means clustering do not identify highly-correlated dimensions in the given data. All the above mentioned methods including ours assume that the numbers of clusters are predefined both for rows and columns. However, several approaches have been proposed that can estimate the numbers of clusters from the given data, for example, by employing the infinite relational model based on the Bayesian nonparametrics [13]. It would be an interesting undertaking to generalize our method in this direction.

2.2 Visualizing High-Dimensional Data

Developing visual analytics models for exploring feature subspaces hidden behind the multivariate data has attracted more attention from the visualization community. Two significant research directions, *scalable data compression* for hierarchical representation of data samples and *dimension management* for rearranging dimensions for better visualization have been investigated so far. For the data compression, Fua et al. [6] proposed *hierarchical PCPs* together with interactive tools for controlling the level of details of the data, while Elmqvist and Fekete [5] conducted a survey on guidelines for hierarchical data aggregation and provided us with a useful insight into data abstraction. As for the dimension management, Yang et al. [29] filtered out insignificant dimensions according to their mutual similarities in the visualization of multivariate data, and Peng et al. [21] improved this technique for reordering dimensions by maximizing data correlation between each pair of the neighboring dimensions. The idea has been further extended in [10, 8, 32] by incorporating a *pairwise correlation graph* for visualization purposes. Contracting multiple dimensions into a single composite axis has been conducted in [18], where correlation was incorporated for visualizing the global trends inherent of the data.

Multivariate data exploration has recently been tackled since the aforementioned two research directions handle data samples and dimensions almost independently. Turkay et al. [25, 26] presented visual analysis models, which interactively project data samples and subspaces onto screen space through multivariate statistical analysis. Tatu et al. [24] introduced an algorithm for finding interesting subspaces in the multivariate data by referring to the similarity measure between a pair of subspaces. Furthermore, Yuan et al. [31] aggressively incorporated human intervention in the data exploration by visualizing distribution of data samples and correlation among dimensions. Another interesting approach has recently been developed by Yates et al. [30] where they employed glyph-based scatterplot matrices for finding feature subspaces. Although these approaches are effective, the results of the analysis heavily depend on knowledge and observation skill of the users and thus may fail to illuminate important feature subspaces. Our approach tries to maximally support such data exploration by suggesting possible subspace decomposition of the given data subsets together with correlation values of each extracted subspace, which allows us to visually select significant feature subspaces with less effort.

For better visualizing data correlation between dimensions in our approach, we employ PCP and its enhanced views as visual representations of the multivariate data [9]. However, PCP often suffers from visual clutter artifacts arising from overlaps among polyline plots especially with the increase in data complexity, which has been alleviated so far by improving its rendering styles. Zhou et al. [33] introduced edge bundling techniques to group highly-correlated polyline samples, while McDonnell and Mueller [16] applied translucent rendering styles on those bundled polylines, and Palmas et al. [20] employed density-based clustering for each dimension to provide an overview of bundled data. In this research, we will deploy such PCP representations of multivariate data to visually understand the subspace partitioning obtained by the proposed biclustering algorithm.

3 Overview of Visual Analytic Framework

This section shows the visualization framework of the present approach. Data analysts still find it difficult to analyze multivariate data since the characteristics of data are implicitly smoothed as

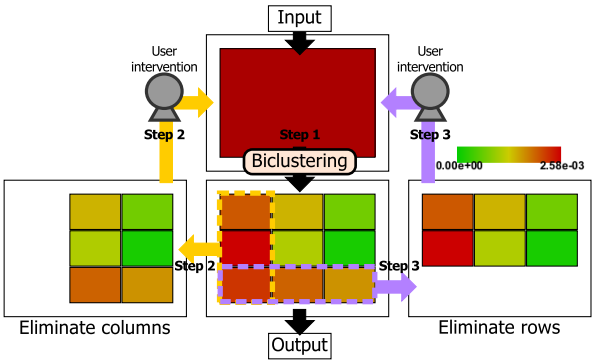
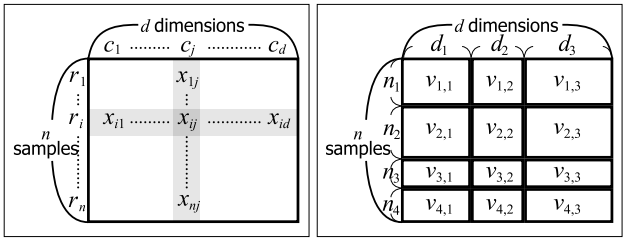


Figure 2: Visual analytic framework of correlated subspace mining.



(a) (b)

Figure 3: (a) Data matrix, and (b) schematic representation of the block model for $K = 4$ and $L = 3$.

the dimension increased, and visual clutter occurred when projecting the data to the limited screen space. Thus, they aim to reduce the numbers of dimensions and data samples in order to obtain the global knowledge through exploring significant correlation in the multivariate data. To achieve this, we develop a block modeling framework where each block represents a subspace in the dataset, so that we can systematically reduce poorly-correlated dimensions and data samples and provide an efficient guide to the data analysts.

Figure 2 shows the overall framework of the present approach. We provide two main functions, including *simultaneous clustering of highly-correlated dimensions and data samples* and *data exploration* for sophisticated data analysis. Our approach begins with an automatic clustering step, where the highly-correlated dimensions and data samples are simultaneously grouped through the proposed biclustering algorithm (step 1). By referring to the objective function values, analysts can justify the goodness of the current clustered results with colored blocks as shown in Figure 2. The HSV color model was employed in our prototype system, where the red color indicates poorly-correlated data samples and green color represents highly-correlated ones. Thanks to this color assignment scheme, analysts can selectively delete poorly-correlated dimensions (step 2), which is followed by eliminating a limited number of outlier data samples (step 3). Note that user intervention is involved here in the second and third steps, because analysts can intentionally generate a highly-correlated subspace and remove outlier data samples for further exploration iteratively until reaching a satisfactory result. The present feature subspace extraction is accomplished by incorporating the biclustering algorithm (Section 4), while the data exploration is composed of several view designs together with the history tree recording (Section 5).

4 Biclustering Method

In this section, we describe the biclustering method, which clusters dimensions and data samples simultaneously. As the distance measure for the clustering of dimensions, we focus on the correlation coefficient, which is suitable for PCP, and is in fact applied to the ordering of dimensions [28, 21]. More specifically, we use the clustering method, called spherical k-means, which has the sum of the correlations from the cluster center vectors as the objective function, and conducts k-means clustering on a high-dimensional unit sphere [4, 3]. After describing the objective function of the classical k-means algorithm (Section 4.1), we introduce an extension of the spherical k-means that takes into account negative correlation as well as positive correlation (Section 4.2). Then, we describe a fundamental biclustering algorithm based on the block model (Section 4.3). Finally, incorporating the same constraint as the spherical k-means, we extend the biclustering algorithm so as to deal with the correlation coefficient as the objective function (Section 4.4). We develop an optimization method for the block model which maintains the spherical constraint during the optimization to guarantee the monotonic improvements of the objective function.

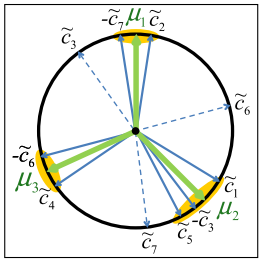


Figure 4: A concept illustration of spherical k-means.

4.1 K-Means Algorithm

Given n samples in d-dimensional space, $\{r_1,\cdots,r_n\}$, $r_i = (x_{i1},\cdots,x_{id}) \in \mathbf{R}^d$, the k-means clustering algorithm with K clusters iteratively updates the cluster mean vectors $\{\theta_k = (\theta_{k1},\cdots,\theta_{kd}) \in \mathbf{R}^d\}_{k=1}^K$ and cluster labels $\{\kappa(i) \in \{1,2,\cdots,K\}\}_{i=1}^n$, which is guaranteed to converge to a local minimum of the following objective function [15],

$$\sum_{i=1}^{n} ||r_i - \theta_{\kappa(i)}||^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} (x_{ij} - \theta_{\kappa(i)j})^2. \quad (1)$$

4.2 Spherical K-Means Algorithm

Let us consider clustering of dimensions $c_1,\cdots,c_d$ where $c_j = (x_{1j},\cdots,x_{nj}) \in \mathbf{R}^n$. Normalizing each dimension, we define $\tilde{c}_j = (\tilde{x}_{1j},\cdots,\tilde{x}_{nj}) \in \mathbf{S}^{n-1}$ (i.e. $\tilde{c}_j \in \mathbf{R}^n$ and $||\tilde{c}_j|| = 1$), where

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}},$$

and $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the average of the jth dimension for $j = 1,\cdots,d$. Figure 3(a) summarizes the notation of the data matrix.

The spherical k-means clustering uses the following objective function,

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

where $\cdot$ denotes the inner product, $\mu_l \in \mathbf{S}^{n-1}$ is the lth mean vector of dimensions, and $\lambda(j) \in \{1,2,\cdots,L\}$ is the cluster label representing which cluster the jth dimension belongs to. Here we have also introduced $s(j) \in \{-1,+1\}$ which shows that $\mu_{\lambda(j)}$ is positively (negatively) correlated with $\tilde{c}_j$ if $s(j) = +1$ ($s(j) = -1$) while the original spherical k-means deals only with the positive correlation, i.e., $s(j) = +1$ for all j [4, 3].

The objective function (2) is equivalent to

$$\sum_{j=1}^{d} \left|\tilde{c}_j \cdot \mu_{\lambda(j)}\right|,$$

when the signs $\{s(j)\}$ are optimized. Since the mean vector μ_l satisfies $\mu_l \cdot \mathbf{1} = 0$, where $\mathbf{1}$ is the vector of all 1, during the optimization, the objective function is the sum of the absolute values of the correlation coefficients between the dimensions $\{c_j\}$ and the assigned mean vectors $\{\mu_{\lambda(j)}\}$.

The spherical k-means can be viewed as the k-means algorithm on the unit hypersphere, as illustrated schematically in Figure 4 for $n = 2$ and $d = 7$. Specifically, the spherical k-means algorithm, initializing the cluster labels $\{\lambda(j)\}$, iterates the following two steps, which monotonically increase the objective function (2), and hence is guaranteed to converge to a local maximum:

Update step: For $l = 1, \cdots, L$,

$$\mu_l = \frac{\sum_{j:\lambda(j)=l} s(j)\tilde{c}_j}{\left\| \sum_{j:\lambda(j)=l} s(j)\tilde{c}_j \right\|},$$

where the numerator sums the (signed) dimensions that are assigned to the lth cluster, and the denominator normalizes μ_l.

Assignment step: For $j = 1, \cdots, d$,

$$\lambda(j) = \operatorname*{argmax}_{1 \le l \le L} \left| \tilde{c}_j \cdot \mu_l \right|, \tag{3}$$

and set $s(j)$ to the sign of $\tilde{c}_j \cdot \mu_{\lambda(j)}$.

Note that the maximization of the above objective function is equivalent to the minimization of

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

which is the objective function of the original k-means algorithm in (1) applied to the columns of the data matrix instead of the rows when $s(j) = 1$ for all j and $||\mu_l|| = 1$ for $l = 1, \cdots, L$. In the update step, the mean vector is normalized, so as to have the norm 1, which is the main difference of the spherical k-means from the original k-means algorithm.

4.3 Biclustering

We turn to the biclustering algorithm which simultaneously clusters dimensions and data samples. Basic biclustering methods are based on the block model [7, 17], where the data matrix of the size $n \times d$ is divided into $K \times L$ submatrices (blocks), each of which has the size $n_k \times d_l$ ($k = 1, \cdots, K$ and $l = 1, \cdots, L$). Here n_k is the number of data samples assigned to the kth cluster of samples, and d_l is the number of dimensions assigned to the lth cluster of dimensions. Hence, $\sum_{k=1}^{K} n_k = n$ and $\sum_{l=1}^{L} d_l = d$ hold. The biclustering algorithm defines as the objective function, the following squared error,

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

which is minimized with respect to the mean value of each block $\nu_{k,l} \in \mathbf{R}$ ($k = 1, \cdots, K$, $l = 1, \cdots, L$), and sample and dimension cluster assignments $\kappa(i) \in \{1, \cdots, K\}$ ($i = 1, \cdots, n$) and $\lambda(j) \in \{1, \cdots, L\}$ ($j = 1, \cdots, d$). Figure 3(b) shows a schematic representation of the block model where the rows and columns of the data matrix are permuted according to the row and column cluster assignments.

Initializing the sample and dimension cluster labels, $\{\kappa(i)\}$ and $\{\lambda(j)\}$, the biclustering algorithm iterates the following steps:

Update step: For $k = 1, \cdots, K$ and $l = 1, \cdots, L$,

$$\nu_{k,l} = \frac{1}{n_k d_l} \sum_{i:\kappa(i)=k} \sum_{j:\lambda(j)=l} x_{ij},$$

which is the mean of the elements of the data matrix assigned to the klth block.

Assignment step: For $i = 1, \cdots, n$,

$$\kappa(i) = \operatorname*{argmin}_{1 \le k \le K} \sum_{j=1}^{d} (x_{ij} - \nu_{k,\lambda(j)})^2,$$

and for $j = 1, \cdots, d$,

$$\lambda(j) = \operatorname*{argmin}_{1 \le l \le L} \sum_{i=1}^{n} (x_{ij} - \nu_{\kappa(i),l})^2.$$

This biclustering algorithm can be considered as applying k-means algorithms to samples and dimensions, where for samples, the mean vector of the kth cluster is $\theta_k = (\nu_{k,\lambda(1)}, \cdots, \nu_{k,\lambda(d)}) \in \mathbf{R}^d$, and for dimensions, the mean vector of the lth cluster is $\mu_l = (\nu_{\kappa(1),l}, \cdots, \nu_{\kappa(n),l}) \in \mathbf{R}^n$.

4.4 Spherical K-Means Based Biclustering

4.4.1 Spherical Constraints

The biclustering algorithm in Section 4.3 uses k-means clustering both for clustering of samples and clustering of dimensions. Hence, it does not take into account the correlations between dimensions of each cluster, which are directly dealt with by the spherical k-means presented in Section 4.2. To incorporate the correlation between dimensions, we propose to introduce the constraint on the block mean values, for $l = 1, \cdots, L$ and $k = 1, \cdots, K$,

$$\nu_{k,l} = \frac{\frac{1}{n_k}\bar{\nu}_{k,l}}{\sqrt{\sum_{k=1}^{K} \frac{(\bar{\nu}_{k,l})^2}{n_k}}}, \tag{4}$$

for $\bar{\nu}_{k,l} \in \mathbf{R}$ satisfying $\sum_{k=1}^{K} \bar{\nu}_{k,l} = 0$. Hereafter, we consider $\bar{\nu}_{k,l}$ as a parameter instead of the block mean value $\nu_{k,l}$. Its update rule will be given in (7) of Section 4.4.2. We see that under the constraint (4), the block mean values satisfy the constraints for the mean vectors of the spherical k-means, i.e.,

$$\|\mu_l\|^2 = \sum_{k=1}^{K} n_k \nu_{k,l}^2 = 1,$$

for $l = 1, \cdots, L$, where $\mu_l = (\nu_{\kappa(1),l}, \cdots, \nu_{\kappa(n),l})$ is the lth mean vector of dimensions. Furthermore, $\mu_l \cdot \mathbf{1} = \sum_{k=1}^{K} n_k \nu_k = 0$ holds for $l = 1, \cdots, L$.

Thus, we propose the biclustering algorithm that minimizes the following objective function subject to the constraint in (4),

$$D = \sum_{j=1}^{d} \|\tilde{c}_j - s(j)\mu_{\lambda(j)}\|^2 = 2d - 2\sum_{j=1}^{d} s(j)\tilde{c}_j \cdot \mu_{\lambda(j)}, \tag{5}$$

Note that in (4), the block mean value $\nu_{k,l}$ depends on the sample cluster label $\kappa(i)$ ($i = 1, \cdots, n$) through n_k.

The normalized error for the klth block is then defined by

$$E_{k,l} = \frac{1}{n_k d_l} \sum_{i:\kappa(i)=k} \sum_{j:\lambda(j)=l} \{\tilde{x}_{ij} - s(j)\nu_{k,l}\}^2, \tag{6}$$

which is used for the block matrix diagram, as will be described in Section 5.2. The total error D in (5) and the normalized block errors are related by $D = \sum_{k=1}^{K}\sum_{l=1}^{L} n_k d_l E_{k,l}$.

The spherical k-means described in Section 4.2 has $L(n-2)$ degrees of freedom. This flexibility may cause severe overfitting to the data matrix. The block model with the spherical constraint, reducing the degrees of freedom to $L(K-2)$, can control the flexibility of the method by choosing K.

4.4.2 Derivation of the Algorithm

We alternately optimize the objective function (5) with respect to one of the variables while other variables are fixed. This is iterated until convergence to obtain a local minimum of the objective function. Note that the convergence value of the objective function (5) is meaningful only when compared among multiple runs with different initializations.

For fixed $\{\kappa(i)\}$ and $\{\bar{v}_{k,l}\}$, the objective function (5) is expressed as a function of $\lambda(j)$ and $s(j)$ by $-2s(j)\tilde{c}_j \cdot \mu_{\lambda(j)}$ up to constant terms independent of $\lambda(j)$ and $s(j)$. Hence, the assignment step (3) of the spherical k-means can be directly used to optimize $\lambda(j)$ and $s(j)$.

For fixed $\{s(j)\}$, $\{\lambda(j)\}$, and $\{\bar{v}_{k,l}\}$, all the block mean values depend on the sample cluster label $\kappa(i)$ through n_k and the normalizing constant in (4). Hence, to optimize the ith sample label $\kappa(i)$ while other sample labels are fixed, we recompute all the block mean values by (4) for $\kappa(i) = 1, \cdots, K$, compare their objective values by (5), and choose the label that attains the minimum.

For fixed $\{\kappa(i)\}$, $\{\lambda(j)\}$, and $\{s(j)\}$, the Cauchy-Schwarz inequality yields that the optimal block mean value that minimizes the objective function (5) is given by

$$\bar{v}_{k,l} = \sum_{i:\kappa(i)=k} \sum_{j:\lambda(j)=l} s(j)\tilde{x}_{ij}. \tag{7}$$

4.4.3 Algorithm

The proposed algorithm is summarized in Algorithm 1. As for the initialization, we can optionally use the k-means++ method [2] to generate initial sample labels $\{\kappa(i)\}$ and dimension labels $\{\lambda(j)\}$ since the objective function (5) is the squared distance both for each sample and dimension.

Algorithm 1 Spherically constrained biclustering

Input: Data samples $\{x_i\}_{i=1}^n$. Number of sample clusters K. Number of dimension clusters L.
Output: Sample cluster labels $\{\kappa(i) \in \{1,2,\cdots,K\}\}_{i=1}^n$. Dimension cluster labels $\{\lambda(j) \in \{1,2,\cdots,L\}\}_{j=1}^d$.

Initialize $\kappa(i)$ for $i = 1,2,\cdots,n$ and $\lambda(j)$ for $j = 1,2,\cdots,d$.
repeat
  Update $\bar{v}_{k,l}$ for $k = 1,\cdots,K$ and $l = 1,\cdots,L$ by (7).
  For $j = 1,\cdots,d$, assign the cluster label $\lambda(j)$ by (3), and set $s(j)$ to the sign of the correlation of the maximum.
  For $i = 1,\cdots,n$, recompute block mean values $\{v_{k,l}\}$ for $\kappa(i) = 1,\cdots,K$ by (4), and set $\kappa(i)$ to the sample label minimizing the objective function (5).
until Convergence

5 Visualizing Highly-Correlated Subspaces

Once the biclustered dimensions and data samples have been generated, we are ready to describe the present visualization framework. The coordinated view in Figure 1 consists of the classical PCP, *clustered PCP*, *contracted PCP*, and *block matrix diagram*, to effectively illustrate how the multivariate data is decomposed into subspaces through biclustering techniques. In addition, it offers a history tree that effectively restores any previous transactions.

5.1 Enhanced Parallel Coordinate Plots

Classical PCP effectively presents data correlation between two adjacent axes, as shown in Figure 1(a). To further enhance the readability of the clustered dimensions in our system, we devised the *clustered PCP* in a way that we can group parallel axes according to the clustering result while inserting additional space to place thick separators between the clusters by employing Gestalt principles, as shown in Figure 1(b). Note that each parallel axis in this clustered PCP is represented by the orientation of a thin triangle which allows us to discriminate between the normal axes (i.e., those with $s(j) = +1$) and the inverted ones ($s(j) = -1$) (see Figure 4 also). For increasing the readability of the global trends inherent in the data, we also introduced the *contracted PCP*, where we project multiple axes in each axis cluster onto one composite axis, as shown in Figure 1(c). Several projection techniques including spherical k-means [3], PCA [12], and LDA [14] have been implemented in our system, so that we can effectively infer relationships between the clusters of axes. We also employed edge-bundled cluster rendering and strip rendering styles in our system in order to improve the visual readability of the data [16, 20, 33].

5.2 Block Matrix Diagram

Figure 1(d) shows our block matrix diagram for the biclustered data, which also serves as an interface for exploring meaningful subspaces in the data. Vertical color bars (in red, for example) attached to the left side of the diagram show the correspondence with the colors assigned to the data clusters in PCP representations, and the integer value next to each bar shows the number of data samples in the corresponding cluster. The value in each small block shows the normalized error of the corresponding subsets of data samples in the feature subspace, which is obtained by (6). We incorporated a color legend of popular heat map representation that ranges in hue from red to green according to the degree of data correlation. In our implementation, the blocks with low correlation are automatically sorted to the right bottom corner in the diagram. Users are also allowed to drag a set of blocks to reorganize the block matrix diagram itself, and delete a row or column of feature subspaces to further investigate the correlation in the remaining data. Note that the sequence of columns (i.e., subset of dimensions) in the block matrix diagram is matched with that of composite axes in the aforementioned contracted PCP view.

5.3 History Tree Visualization

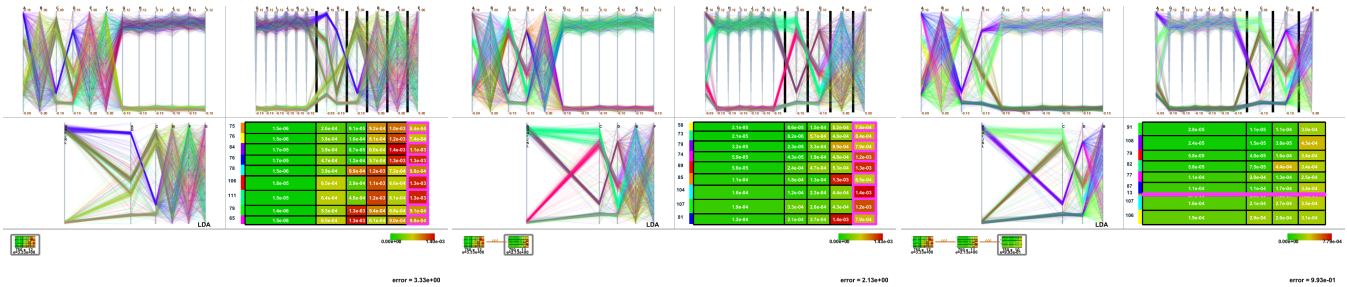
To allow users to arbitrarily retrieve their previous transactions, we introduced a history tree representation to record the exploration history, as shown in Figure 1(e). Here, each tree node is labeled with a thumbnail image of the corresponding block matrix diagram together with their matrix size and objective function value for better understanding the history. Users are allowed to estimate the goodness of biclustering results by referring to the objective function value in (5), so that they can navigate the history tree in a trial and error manner.

6 Results and Discussion

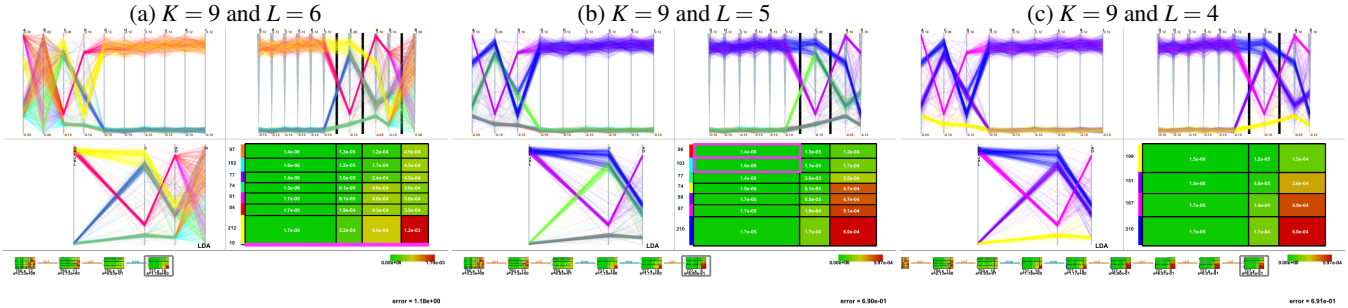
In this section, we present experimental results of the proposed approach on synthetic and real-world datasets, together with discussions of the present approach. Our prototype system has been implemented on a desktop PC with Quad-Core Intel Xeon CPUs (3.7GHz, 10MB cache) and 12GB RAM, and the source code was written in C++ using GSL for numerical computation, OpenGL for graphics, and GLUI library for the user interface.

6.1 Synthetic Data

We first demonstrate our experimental study on a 750 records of 12D synthetic data employed in [24]. This multivariate data includes four 3D clusters with 10% noise and two 6D clusters without noise, while the data samples are uniformly distributed in other dimensions. Figure 5(a) shows an initial subspace decomposition of the data obtained through the biclustering process. Note that our prototype system automatically conducts three trials of the biclustering process and takes the best one as the result. Here, we set the initial numbers of data and dimension clusters to $L = 6$ and $K = 9$,

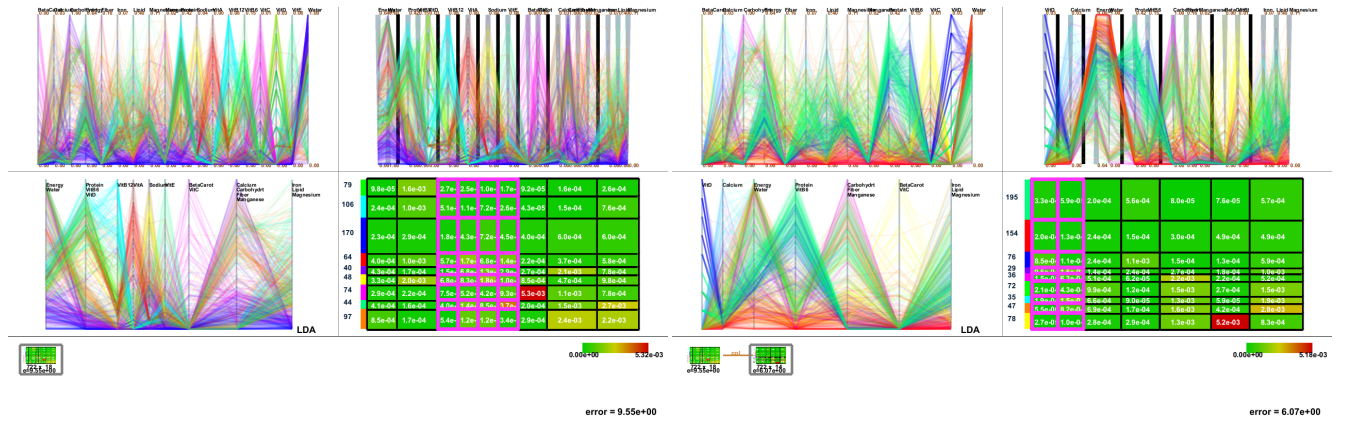


(a) $K = 9$ and $L = 6$ (b) $K = 9$ and $L = 5$ (c) $K = 9$ and $L = 4$

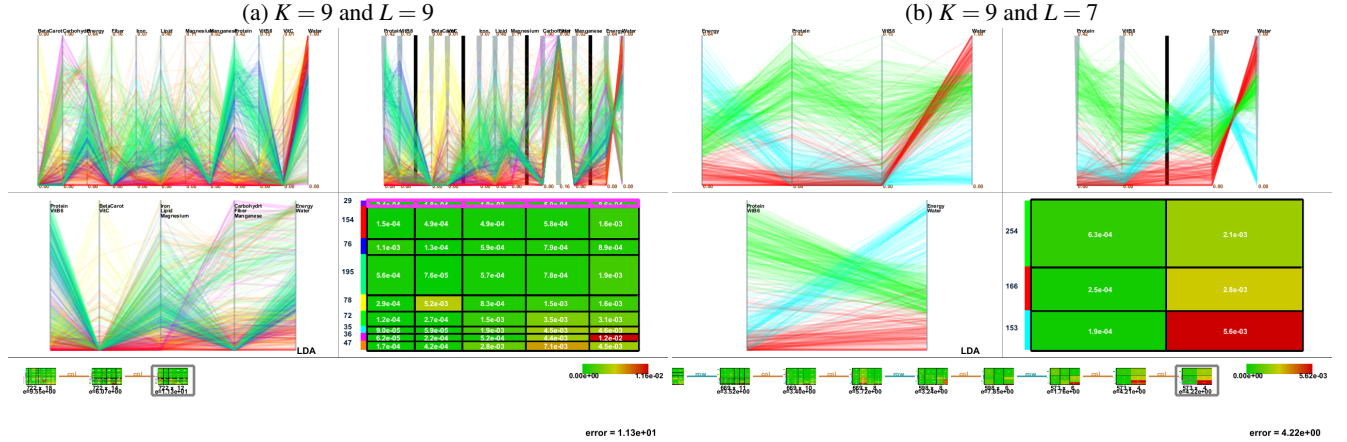


(d) $K = 8$ and $L = 4$ (e) $K = 7$ and $L = 3$ (f) $K = 4$ and $L = 3$

Figure 5: System snapshots for exploring subspaces in a 12-dimensional synthetic data.



(a) $K = 9$ and $L = 9$ (b) $K = 9$ and $L = 7$



(c) $K = 9$ and $L = 5$ (d) $K = 3$ and $L = 2$

Figure 6: Experiments on the USDA national nutrient data.

which are simply determined by half number of the original dimensions and the logarithm of data samples, respectively. Thus, the isolated dimensions generated by the algorithm can be considered as poorly-correlated, so that analysts can delete them to maintain the correlation of the reduced dataset. Our guideline here is to eliminate subspaces having uncorrelated and often isolated dimensions first, and then data samples by cutting off rows and columns of the block matrix diagram, until we can fully identify correlation among data samples and dimensions in the remaining subspaces. Figures 5(a) and (b) show that the selected columns (i.e., clusters of isolated dimensions) were interactively eliminated. Then, we were about to arbitrarily remove all the small uncorrelated data samples (23 only as shown in Figures 5(c) and (d)). Figure 1 presents an intermediate screenshot of our prototype interface after the above processes, where we finally obtained most correlated subspaces through merging similar clusters into ones (Figures 5(e) and (f)).

6.2 USDA National Nutrient Data

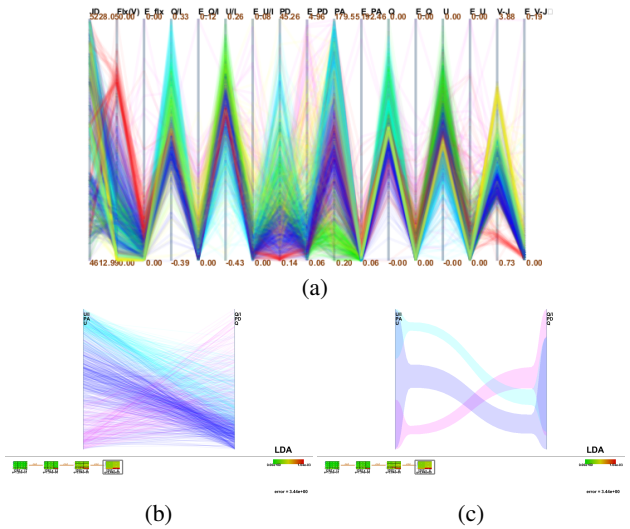
We employed the USDA food composition dataset [1] as the first real-world example, which has been also employed in the experiments conducted by Tatu et al. [24] and Yuan et al. [31]. In this dataset, each data sample corresponds to a specific food while each dimension represents a kind of nutrient composed in each food. After having deleted data samples having missing values and selected interpretable dimensions as a preprocessing by following [24, 31], we finally employed 722 records and 18 dimensions as the input.

We conducted biclustering analysis with aforementioned initial setting ($K = 9$ and $L = 9$) in our experiment. Dimensions labelled as Vitamin B12, Vitamin A, Sodium, and Vitamin E, which are not highly correlated with the remaining dimensions, were first deleted from the block matrix diagram to preserve the correlation in the data (Figure 6(a)). Following the guideline we discussed in Section 6.1, we deleted isolated dimensions step by step to extract a new set of feature subspaces as shown in Figures 6(b) and (c). Removing additional small data outliers (Figure 6(b)) finally resulted in the small set of correlated feature subspaces (Figure 6(d)). Throughout our experiments with 573 left records, we can observe that a pair of Energy and Water is the most strongly correlated while a pair of Protein and Vitamin B6 is also highly correlated.

Comparison with the case study by Tatu et al. [24] shows that our results also support their claim that Protein is dominant within a set of dimensions and strongly influences on the results of subspace clustering. Yuan et al. [31] also stated Energy, Lipid, and Water are highly correlated after removing unnecessary data samples and dimensions manually in their system, which was also supported by our results where we could clearly find three data clusters as shown in Figure 6(d). Meanwhile, we could further extract the second highly-correlated subspace spanned by Protein and Vitamin B6, and identify relationships between the two subspaces through our contracted PCP representation.

6.3 Blazar Data

Our second practical case study targets at diurnally-measured light of *blazars* being emanated from active galactic nuclei (supermassive black holes) [27]. The dataset of interest contains 12 blazars whose properties are commonly characterized with 8 parameters, that is, total intensity (I); two variables for polarization (Q and U); corrected variables for polarization (Q/I and U/I); degree of polarization ($PD = \sqrt{(Q/I)^2 + (U/I)^2}$); angle of polarization ($PA = 0.5 \arctan(U/Q)$); and color index ($V - J$). Note that each of the parameters is accompanied with its own measurement deviation, and the total dataset contains 1,285 samples in 17D, indexed with observational day (JD) [11]. The challenging task here is to visually explore correlations among these parameters, aiming at blazar classification and behavior abnormity isolation based primarily on such polarized light observations.



(a) (b) (c)

Figure 7: Analyzing blazar dataset of 1,285 samples in 17D. (a) Plotting original data. (b) Plotting final data clustered with 3 sample-by-2 axis clusters. (c) Plotting final data (strip-rendered view).

Figure 7(a) plots the original dataset. Starting with the initial setting of $K = 6$ and $L = 6$, iterative axis contractions and elimination of outlier data samples (finally reduced to 1,093) made the dataset manifest its own correlation structure in a 3 sample-by-2 axis block cluster, as shown in Figure 7(b). Besides, the clustered structure of the final data plot can be enhanced with a strip-rendered view in Figure 7(c) [20]. It is not surprising that those deviations and temporal index have been eliminated as uncorrelated properties. Also, it can be observed that polarization variables (Q/I and U/I) corrected with total intensity (I) have strong correlations with their originals (Q and U), respectively. This suggests that the temporal behavior of blazars is governed not by the variations in unpolarized emission, but that by polarized emission. In addition, it was contrary to the expectation from astrophysics that color index ($V - J$) did not remain as a key blazar discriminator.

6.4 Discussion

In the previous subsections, we demonstrated several experimental results of our prototype system. We successfully retrieved highly-correlated subspaces of several experimental data by following the proposed applicable guideline, and the results can be reproduced through different trials. Moreover, in a reduced subspace, data samples are more understandable because of the lack of the visual complexity. Thus, our approach not only allows us to investigate data samples individually in different clustered dimensions, but also explore the behaviors between clusters simultaneously. The aforementioned benefit provides the users with a different impact when analyzing meaningful data samples, where they can compose the overall knowledge of the dataset from possible *local* clusters within a subspace as well as *global* clusters between the subspaces. A limitation of the present approach is that we cannot apply it for analyzing uncorrelated dataset, such as datasets for classification purposes. Moreover, although we have employed k-means++ method for initializing sample labels, we cannot always reach the best objective value without a few trial and error analyses. For the conventional block model, extensions based on Bayesian estimation have been proposed, which offer a way to select the numbers of clusters, K and L [23, 22]. It is an important undertaking to derive such an extension of the spherically constrained block model.

7 Conclusion and Future Work

We presented a novel data analysis approach based on a reformulated biclustering method that employs correlation coefficients as a similarity measure of dimensions. Simultaneously-clustered data samples and dimensions are visualized by enhanced PCPs together with a block matrix diagram and its history record. Through experiments on a synthetic and two real datasets, we demonstrated that our approach enables efficient interactive handling of the extracted subspaces.

Our future directions include the estimation of the numbers of row and column clusters only from the given data and the generalization of the block structure to the nested clusterings (i.e., partitionings), which were implemented for the ordinary block model in [13, 22]. Although our current system can already extract nested clusters by preserving the clusterings of the removed blocks of rows and columns, this sequential partitionings would be improved by the direct modeling of nested partitionings. Another issue to be addressed is more appropriate treatment of time-series data.

Acknowledgements

This work has been partially supported by MEXT KAKENHI under Grants-in-Aid for Scientific Research on Innovative Areas No. 25120014. The Food data and blazer data are courtesy of Dr. Andrada Tatu, and Professor Makoto Uemura and Hiroshima Astrophysical Science Center, Hiroshima University, respectively.

References

[1] USDA national nutrient database for standard reference. United States Department of Agriculture (USDA). http://www.ars.usda.gov/Services/docs.htm?docid=8964.

[2] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[4] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001.

[5] N. Elmqvist and J. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, 2010.

[6] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the Conference on Visualization '99: Celebrating Ten Years*, pages 43–50, 1999.

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

[8] J. Heinrich, J. Stasko, and D. Weiskopf. The parallel coordinates matrix. In *Proceedings of EuroVis 2012 Short Papers*, pages 37–41, 2012.

[9] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. In *Eurographics 2013 - State of the Art Reports*, pages 95–116, 2013.

[10] C. B. Hurley and R. W. Oldford. Pairwise display of high-dimensional information via eulerian tours and hamiltonian decompositions. *Journal of Computational and Graphical Statistics*, 19(4):861–886, 2010.

[11] Y. Ikejiri, M. Uemura, M. Sasada, R. Ito, M. Yamanaka, K. Sakimoto, A. Arai, Y. Fukazawa, T. Ohsugi, K. S. Kawabata, M. Yoshida, S. Sato, and M. Kino. Photopolarimetric monitoring of blazars in the optical and near-infrared bands with the kanata telescope. I. correlations between flux, color, and polarization. *Publications of the Astronomical Society of Japan*, 63:143–175, 2011.

[12] I. T. Jolliffe. *Principal component analysis*. Springer, 2002.

[13] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, pages 381–388, 2006.

[14] J. H. Lee, K. T. McDonnell, A. Zelenyuk, D. Imre, and K. Mueller. A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):351–364, 2014.

[15] J. B. MacQueen. Some Methods for classification and analysis of multivariate observations. In *Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297, 1967.

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

[17] I. V. Mechelen, H. H. Bock, and P. D. Boeck. Two-mode clustering methods: a structured overview. *Statistical methods in medical research*, 13(5):363–394, 2004.

[18] 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.

[19] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[20] G. Palmas, M. Bachynskyi, A. Oulasvirta, H. P. Seidel, and T. Weinkauf. An edge-bundling layout for interactive parallel coordinates. In *Proceedings of the 7th Pacific Visualization Symposium (PacificVis 2014)*, pages 57–64, 2014.

[21] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the 10th IEEE Conference on Information Visualization (InfoVis 2004)*, pages 89–96, 2004.

[22] D. M. Roy and Y. W. Teh. The Mondrian process. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[23] H. Shan and A. Banerjee. Bayesian co-clustering. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 530–539, 2008.

[24] 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.

[25] 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.

[26] 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.

[27] C. M. Urry and P. Padovani. Unified schemes for radio-loud active galactic nuclei. *Publications of the Astronomical Society of the Pacific*, 107(715):803–845, 1995.

[28] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the 9th Annual IEEE Conference on Information Visualization (InfoVis 2003)*, pages 105–112, 2003.

[29] J. Yang, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. *Computers & Graphics*, 27(2):265 – 283, 2003.

[30] A. Yates, A. Webb, M. Sharpnack, H. Chamberlin, K. Huang, and R. Machiraju. Visualizing multidimensional data with glyph sploms. *Computer Graphics Forum*, 33(3):301–310, 2014.

[31] 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.

[32] Z. Zhang, K. T. McDonnell, and K. Mueller. A network-based interface for the exploration of high-dimensional data spaces. In *Proceedings of the 5th Pacific Visualization Symposium (PacificVis 2012)*, pages 17–24, 2012.

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