Indexed-Points Parallel Coordinates Visualization of Multivariate Correlations

Liang Zhou and Daniel Weiskopf

Abstract—We address the problem of visualizing multivariate correlations in parallel coordinates. We focus on multivariate correlation in the form of linear relationships between multiple variables. Traditional parallel coordinates are well prepared to show negative correlations between two attributes by distinct visual patterns. However, it is difficult to recognize positive correlations in parallel coordinates. Furthermore, there is no support to highlight multivariate correlations in parallel coordinates. In this paper, we exploit the indexed point representation of *p*-flats (planes in multidimensional data) to visualize local multivariate correlations in parallel coordinates. Our method yields clear visual signatures for negative and positive correlations alike, and it supports large datasets. All information is shown in a unified parallel coordinates framework, which leads to easy and familiar user interactions for analysts who have experience with traditional parallel coordinates. The usefulness of our method is demonstrated through examples of typical multidimensional datasets.

Index Terms—Multidimensional data visualization, multivariate correlations, parallel coordinates

1 Introduction

The visualization of multivariate data is a long-standing topic of visualization research [1]. In our era of big data, multivariate data has become ubiquitous and proper visualization is vital for gaining insight into such complex data. This paper addresses the general problem of visualizing data that depends on several variables (multivariate data) or dimensions (multidimensional data), i.e., we use the terms 'multivariate' and 'multidimensional' data synonymously. In particular, we want to provide specific support for identifying multivariate correlations in such data because they serve as an important means of understanding the structure of that data.

Our visualization approach is based on parallel coordinates [2], [3], a popular visual mapping method for multivariate data. Parallel coordinates show all attributes of multivariate data on parallel axes where a data point is mapped to a polygonal line (polyline) across all axes. In this way, we can see patterns of multivariate relationships. A great advantage of parallel coordinates is their good scalability with the number of data dimensions: With increasing number of data dimensions, we just need to add further axes (which works fine up to limitations imposed by the display space).

However, parallel coordinates lead to an asymmetric visual representation of positive and negative correlations. Negative correlations are easy to detect as lines form crossing patterns between axes. Visualizing positive correlations in parallel coordinates, however, is difficult. Positively correlated samples are shown as line segments intersecting outside their own pair of axes, or as almost parallel line segments if the slope of the linear regression line is close to 1. Such patterns are hard to observe as users have a bias toward negative correlations even if the true correlation is positive [4]. Furthermore, correlations (both negative and positive) are in general underestimated by people [4].

Parallel coordinates are also limited in the capability of showing multivariate correlations, e.g., correlations between three or more data attributes. Only certain cases for planes can be visualized using polyline-based parallel coordinates [3].

In this paper, we advocate the use of indexed points of p-flats [3] to highlight multivariate correlations. The term *p-flat* denotes a generalized flat surface of dimension p. For example, 0-, 1-, 2-flats are points, lines, and planes, respectively, and $p > 2$ indicates p-dimensional linear planars. An *indexed point* is a point representation of its corresponding p-flat within the space of the parallel coordinates. Inselberg [3] uses indexed points of p-flats to visualize and study geometry problems in great detail.

We exploit the geometry of p-flats and their highly compact representation by indexed points in order to support the visualization of multivariate correlation. In general, there is not just a single planar (flat) correlation in multivariate data; in fact, there might be non-linear relationships or correlations of varying direction (slope) at different parts of the dataset. Therefore, we first identify *local multivariate correlations* numerically by fitting p-flats of varying dimensionality p in the local neighborhoods of all data points. Second, these locally fitted p-flats are transformed into parallel coordinates as indexed points. Finally, the indexed points are plotted into the 2D image plane of parallel coordinates and blended together with the regular line plots of multivariate data.

The local fitting process is closely related to multivariate linear regression [5], which refers to the prediction of multiple dependent variables from multiple independent variables with a linear model. Our method does not distinguish

• *The authors are with the Visualization Research Center (VISUS), University of Stuttgart, Allmandring 19, 70567 Stuttgart, Germany. E-mail: {Liang.Zhou, Daniel.Weiskopf}@visus.uni-stuttgart.de.*

Manuscript received 16 Nov. 2016; revised 6 Mar. 2017; accepted 26 Mar. 2017. Date of publication 25 Apr. 2017; date of current version 27 Apr. 2018. Recommended for acceptance by V. Pascucci.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2017.2698041

between dependent and independent variables but allows for interpreting the regression information from indexed points by choosing dependent and independent variables and visualizing such information.

Our main contributions are:

- The representation of indexed points of p-flats to show local multivariate correlation in the same 2D domain of parallel coordinates.
- The combined visualization of indexed points and parallel coordinates.
- Adapted brushing-and-linking interaction including data point brushing and data dimension brushing.

Our method has several important benefits: It supports the explicit and direct visualization of multivariate local correlations for multidimensional datasets; the compact representation is suitable for visualizing large datasets; the use of indexed points allows for clear visual signatures for negative and positive correlations in parallel coordinates; the indexed points can help the user to visualize regression information, i.e., the type of relationship between data attributes; the visualization is embedded in traditional parallel coordinates in a single view that does not require context switching and transitions between visual mappings; the visual representation and the user interactions are easy to understand and use by those who are already familiar with traditional parallel coordinates.

2 Related Work

Multivariate or multidimensional visualization is an important topic of visualization research, with a large body of papers and much ongoing work. For surveys of the general topic, we refer to papers by Wong and Bergeron [1] and Liu et al. [6].

Our visualization technique is based on parallel coordinates, an important and popular approach to multidimensional data visualization. A survey of parallel coordinates can be found elsewhere [7]. The origin of parallel coordinates as a visualization method and an approach for multidimensional geometry goes back to early work by Inselberg [2] and Wegman [8]. Inselberg developed an extensive mathematical description of the geometry of parallel coordinates and multidimensional data [3]: basic geometric entities such as points, lines, curves, planes, and hyperplanes in multidimensional space can be represented in parallel coordinates. Constructions of p-flats of lines and planes serve as the mathematical foundation of our method and will be reviewed in more detail in Section 3. While the mathematical and geometric basis is covered in the above previous work, we are not aware of any prior work that would have used indexed points of p-flats for the visualization of local multivariate correlation.

Parallel coordinates are typically used to show polylines that correspond to multidimensional data points, ignoring any higher-order geometric entities such as general p-flats. Already this polyline visualization is effective for many examples of visual data analysis. In particular, negative correlation leads to distinct visual patterns in the form of high-density intersections of line segments. Unfortunately, the visual patterns are not symmetric under changes in sign and positive correlation is much underestimated by users [4]. On top of this asymmetry, parallel coordinates are less effective for visual correlation analysis than scatterplots, as shown in a controlled user experiment [4].

To improve the capacity for correlation analysis, there are hybrid methods that combine parallel coordinates with other visual mappings such as scatterplots, scatterplot matrices (SPLOMs), or variants thereof. For example, small scatterplots can be placed on top of parallel coordinates by rotating the scatterplots by 45 degrees [9]. Alternatively, scatterplots can be embedded between adjacent axes in parallel coordinates [10]. When two adjacent axes are selected, one of the axes is rotated by 90 degrees to obtain a Cartesian coordinate system of the two attributes. In case of more than two axes, multidimensional scaling is utilized to form a single scatterplot. The P-SPLOM technique [11] unifies SPLOM, 2D parallel coordinates, and 3D parallel coordinates with smooth transitions between them. Finally, there are flexible user-customized visualizations in the style of parallel coordinates or SPLOMs [12]. However, none of the above works directly supports the visualization of multivariate correlations.

Multivariate relationships involve more than one pair of data axes of parallel coordinates. Curved lines (instead of piece-wise straight lines) can improve the traceability of those lines and, thus, the perception of structures in more than two data attributes [13]. Unfortunately, some important geometric properties of parallel coordinates are destroyed by replacing straight lines with curves. For example, the distinct high-density intersection points are lost as an indicator of negative correlation—even though some correlation perception is still supported for curved parallel coordinates [14]. In short, curved parallel coordinates do not explicitly highlight multivariate relationships.

A general issue of parallel coordinates is scalability with the number of data points: Since data points are represented by lines, they cover a large number of pixels (as opposed to scatterplots, which come with a one-to-one mapping between data points and points on the display). An approach to data scalability replaces line plots by density representations, e.g., by different kinds of binning methods [15], [16], [17] or continuous modeling [18]. Unfortunately, density representations tend to blur visual (line) structures, making correlation recognition harder than for traditional parallel coordinates plots. Alternative approaches employ clustering [19] or splatting-based clutter reduction [20], but again at the cost of removing visual indicators of correlation. The aforementioned hybrids of parallel coordinates and scatterplots [9], [10], [11], [12] provide a point-to-point mapping from data to visualization in their scatterplots. However, this part of the visualization is identical to traditional scatterplots; in particular, there is no direct support for showing multivariate correlation between more than two data dimensions. Our technique, in contrast, maps p-flats to individual indexed points, i.e., we provide a very condensed visual representation of quite complex p-dimensional structures such as lines and planes, which is not possible with the above visualization techniques.

In general, brushing-and-linking interaction is vital for visual analysis of multivariate data, including parallel coordinates. Combining multiple 1D range selections enables higher-dimensional range brushes [21], [22] with simple logical operators. Angular brushing is based on the slope of line segments between adjacent axes [23], defining a range

of angles in parallel coordinates that selects line segments within the slope range between two axes. The angular brush allows for easy selection of positive or negative correlations by the user. We adopt the brushing concepts for our visualization technique to make interaction easy to users who have experience with traditional parallel coordinates and scatterplots. We include range brushes working on axes, and rectangular and lasso brushes on the 2D parallel coordinates domain. Our extended brushing includes selection of indexed points; in this way, we are able to apply brushing-and-linking not only to data points but also to p-flats in data sets. Furthermore, we introduce dimensional brushing to facilitate local correlation visualization in subspaces.

Our approach is based on numerical estimation of local multivariate correlation in the input data. There is previous work with similar local fitting to multivariate data. For example, the shape of locally fitted data points can be used to illuminate 3D scatterplots for better shape perception [24]. An eigen-analysis on the covariance matrix for a neighborhood of each data point is performed to classify the neighborhood into linear, planar, or volumetric structures. Then, different illumination models are applied based on the classification result. Nevertheless, the method does not scale to more attributes, whereas our method supports higher-dimensional datasets. Another technique uses dot-line representation with streamlines to visualize trend lines in 2D scatterplots for sensitivity studies, building on local correlation computation [25], [26]. Again, this method does not support the visualization of multivariate correlations of attributes. A recent technique [27] estimates local linear relationships to visualize correlations in parallel coordinates using the point-line duality (Section 3.1). We adopt their approach for the visualization of negative correlations between two data attributes but use a different visual representation of positive correlations. Furthemore, we extend the approach to correlations between more than two attributes.

Correlation coordinate plots provide another focused view on correlation in multidimensional data [28]. Here, correlation strength and a mapping to a linear global model are used to show the extent of correlation as well as its shape. The technique is primarily designed for (many) pairwise correlations of two attributes but has some support to combine several of such pairs. The visualization is based on scatterplots and star plots of the transformed data points. In contrast, our technique is based on parallel coordinates and supports full multivariate correlation.

3 Theoretical Background

This section summarizes the relevant theoretical background of parallel coordinates as covered by Inselberg [3]. Specifically, representations of p-flats, i.e., points ($p = 0$), lines ($p = 1$), planes ($p = 2$), and higher dimensional flats ($p > 2$) in parallel coordinates will be explained. To compactly represent p-flats in parallel coordinates, we use indexed pooints, which are a point representation whose location in parallel coordinates encodes the parameters of p-flats. The construction of geometries in parallel coordinates is based on the m-dimensional domain of data values that we refer to as *data domain*.

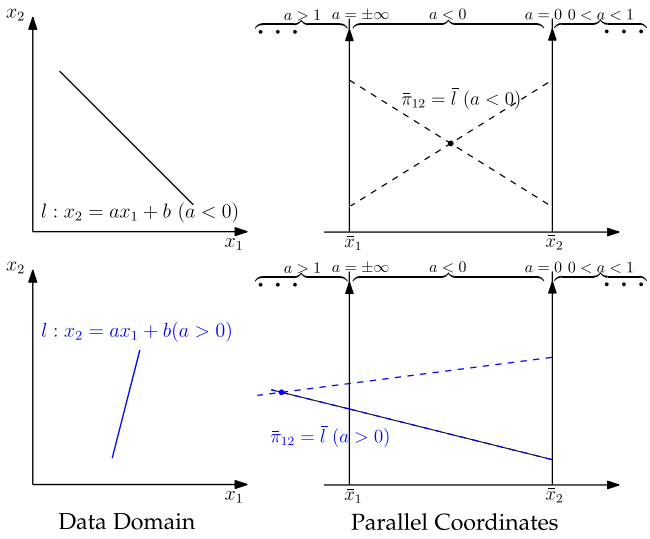


Fig. 1. Point-line duality. A point in the data domain is a line in parallel coordinates, whereas a line in the data domain is a point in parallel coordinates, which is found by intersecting polylines for points on the line in the data domain.

3.1 Point and Line Representations in Parallel Coordinates

The point-line duality is the foundation for parallel coordinates: Points in m-dimensional data space are represented as polylines crossing m axes in the 2D domain (the image plane) of parallel coordinates. Parallel coordinates allow the construction of the final plot by splitting m-dimensional data into $m - 1$ independent 2D data. The overall plot is then generated by placing the parallel axes consecutively in the 2D image plane. A simple case of a point P in the 2D space spanned by two attributes x_1 and x_2 is a line segment $\bar{P}$ in parallel coordinates as illustrated in Fig. 1.

The duality states that a line in a 2D subspace of the data space is mapped to a point in the parallel coordinates. A line l in the Cartesian coordinates can be described as

$$l : x_2 = ax_1 + b.$$

To describe l in parallel coordinates, we transform two points on l into line segments between $\bar{x}_1$ and $\bar{x}_2$ in parallel coordinates using the point-line duality, and then calculate the intersection point of these two line segments. Therefore, line l is represented as point $\bar{l}$ in parallel coordinates

$$\bar{l} = \left(\frac{1}{1-a}, \frac{b}{1-a}\right), a \neq 1. \quad (1)$$

In fact, the point representation $\bar{l}$ is the indexed point of 1-flat l, and as it resides in the 2D space x_1x_2, we denote $\bar{l}$ as indexed point $\bar{\pi}_{12}$

$$\bar{\pi}_{12} = \bar{l}\,.$$

The horizontal location of $\bar{\pi}_{12}$ depends only on the slope a of line l. The relationship between a and the horizontal location of $\bar{\pi}_{12}$ is indicated in Fig. 1 (Parallel Coordinates). For visualization purposes, it is important to note that if $\bar{\pi}_{12}$ is inside axes $\bar{x}_1$ and $\bar{x}_2$ it indicates that l has a negative slope ($a < 0$); when $\bar{\pi}_{12}$ is outside of $\bar{x}_1$ and $\bar{x}_2$, a positive slope ($a > 0$) is indicated. If $a = 1$, $\bar{\pi}_{12}$ is at positive/negative

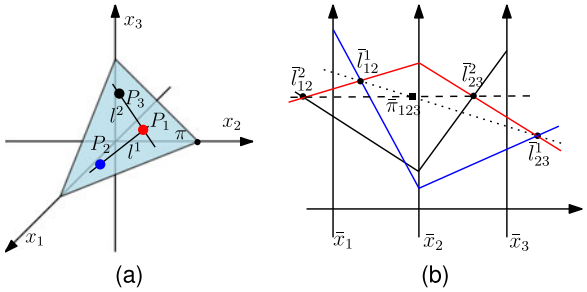


Fig. 2. A 2-flat (plane) π in Cartesian coordinates (a) can be represented by indexed points (for simplicity, here we show only the first one $\bar{\pi}_{123}$) in parallel coordinates (b). The indexed points of 2-flats can be constructed from indexed points of 1-flats as shown in (b).

infinity. Note that l and its indexed point $\bar{l}$ can also represent the local linear regression of attributes.

3.2 Indexed Point Representation of 2-Flats in Parallel Coordinates

The indexed point representation of 2-flats or planes in parallel coordinates can be constructed using indexed points of 1-flats. Assume a plane π spanned by attributes x_1, x_2, and x_3 in the data domain as shown in Fig. 2a is written as implicit surface

$$\pi : c_1x_1 + c_2x_2 + c_3x_3 = c_0, \tag{2}$$

where (c_1, c_2, c_3) is the normal vector of the 2-flat and c_0 is the minimum distance from the origin to the 2-flat. The 2-flat π can be described by a 2-flat indexed point $\bar{\pi}_{123}$ which can be recursively constructed from indexed points of lower dimensions as shown in Fig. 2b. We take two intersecting lines on 2-flat π: l^1 and l^2 that pass through points P_1, P_2 and P_2, P_3 respectively. Indexed points of 1-flats are found by intersecting the polylines of these points in each subspace, e.g., line l^1 is described by $\bar{l}^1_{12}$ for subspace $\bar{x}_1\bar{x}_2$ together with $\bar{l}^1_{23}$ for subspace $\bar{x}_2\bar{x}_3$. Next, we connect the 1-flat indexed points of l^1 and l^2 respectively, i.e., connect $\bar{l}^1_{12}$, $\bar{l}^1_{23}$ (the dotted line in Fig. 2b), and connect $\bar{l}^2_{12}$, $\bar{l}^2_{23}$ (the dashed line in Fig. 2b). The validity of this operation is based on the collinearity theorem ([3], pp. 127, Theorem 5.2.1). Finally, the 2-flat indexed point π_{123} of π is the intersection of these two newly constructed lines in parallel coordinates.

The construction of a 2-flat indexed point can also be achieved analytically

$$\bar{\pi}_{123} = \left(\frac{c_2 + 2c_3}{c_1 + c_2 + c_3}, \frac{c_0}{c_1 + c_2 + c_3}\right). \tag{3}$$

Here, the vertical coordinate of the indexed point is determined by the distance of the 2-flat from the origin. In fact, to fully define the 2-flat π, four indexed points are required. Each indexed point is shifted horizontally in proportion to a component in the normal vector of the 2-flat from the previous indexed point. In our method, only the first indexed point is used for reasons discussed in detail in Section 4.4.

3.3 Indexed Points for Higher-Dimensional Flats

The collinearity theorem can be generalized to higher dimensions, which allows the recursive construction of p-flats for $2 \le p \le m-1$. A p-flat can be written as a linear combination of $p+1$ attributes

$$\pi : c_1x_1 + c_2x_2 + \cdots + c_{p+1}x_{p+1} = c_0. \tag{4}$$

To construct a p-flat, we need two $(p-1)$-flats and use the collinear property to find the intersection of the two lines joining the two indexed points of each $(p-1)$-flat. For example, to construct a 3-flat π^3, we need two 2-flats, namely π^{2_1} and π^{2_2}. First, we connect indexed points $\pi^{2_1}_{123}$, $\pi^{2_1}_{234}$ of π^{2_1}, and $\pi^{2_2}_{123}$, $\pi^{2_2}_{234}$ of π^{2_2} respectively to form two lines. The indexed point π^3_{1234} is then found by the intersection of these two lines.

Like the case of 2-flats, p-flat indexed points can be calculated analytically using coefficients $c_0, c_1, \ldots, c_{p+1}$, and the first indexed point is

$$\bar{\pi}_{12\ldots p+1} = \left(\frac{\sum_{k=1}^{p+1} d_k c_k}{\sum_{k=1}^{p+1} c_k}, \frac{c_0}{\sum_{k=1}^{p+1} c_k}\right), \tag{5}$$

where d_k is the distance of the kth attribute's axis to the first axis in parallel coordinates. A total number of $p+2$ indexed points are required to fully define the p-flat, and they can be derived by shifting the first indexed point horizontally similar to the case of 2-flats. For visualization, however, the interpretation of higher dimensional flats ($p > 2$) is hard and therefore we use only 1- and 2-flats in our method, although the computation of higher-dimensional p-flat indexed points is straightforward once the implicit representation of the p-flat is given (Equation (4)).

4 Local Multivariate Correlations with Indexed Points

We propose using indexed points in parallel coordinates to represent local multivariate correlation between two or three attributes in large high-dimensional datasets. Given an m-dimensional data, the goal is to numerically fit p-flats of the local neighborhood for all data points and calculate indexed points for these p-flats. Specifically, we use principal component analysis to find eigenvectors of the neighborhood of a sample in data domain. Then, we calculate 1-flat indexed points from the major eigenvector, and use the major and the second major eigenvectors to compute 2-flat indexed points using Equation 3. The work flow of our method is illustrated in Fig. 3.

4.1 Finding Nearest Neighbors in Data Domain

Local fitting is computed for every data point in its neighborhood for the m-dimensional data in data domain. Since we handle both unstructured high-dimensional point cloud datasets and high-dimensional scientific datasets on regular grids (e.g., 2D image or 3D volume datasets, where data is defined on a continuous spatial domain), the first step is to construct the data domain. For point cloud datasets, data samples naturally reside in the data domain and no processing is needed, whereas for regular grid datasets, where data points are given on regular grids with assumed piecewise trilinear interpolation, we construct the data domain using Monte-Carlo sampling. If only samples on the original grid points were used, it might lead to artifacts [18], [29]; therefore, we adopt the better solution in the form of Monte-Carlo sampling [18].

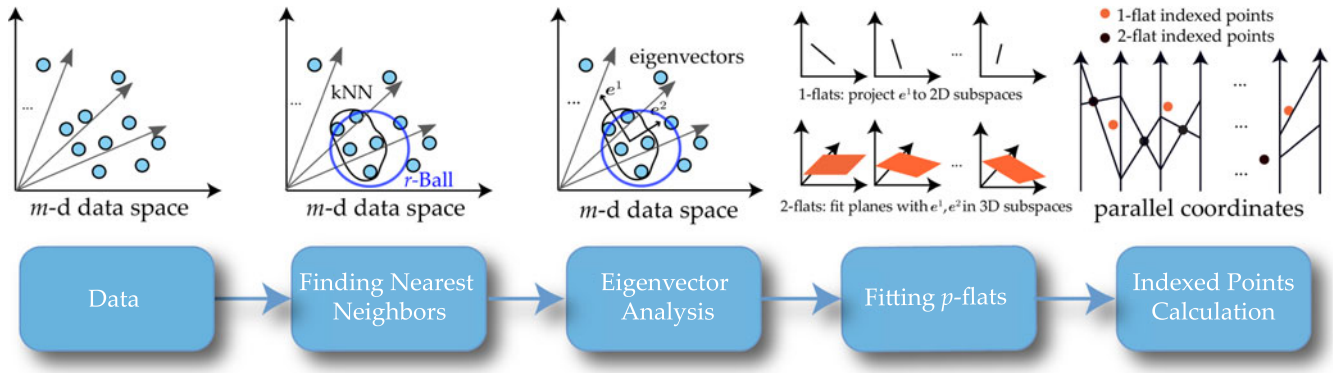


Fig. 3. The work flow of our method.

To efficiently find nearest neighbors in the high-dimensional data domain, a kd-tree is constructed. A kd-tree [30] is a k-dimensional binary tree structure where each non-leaf node partitions one data dimension into half in terms of the number of samples. Since high-dimensional space is sparse, we gain from a kd-tree only when the number of samples N is much larger than 2^k ($N \gg 2^k$) [31], otherwise, it reduces to a brute force search. Using a kd-tree is appropriate for our case as the number of data samples is typically greater than 5,000, whereas the dimensionality m is less than 10. We use the kd-tree to partition the m-dimensional data domain, and then search nearest neighbors for each sample ξ in the data domain.

The set of the n nearest neighbors, X, is used for local fitting. We use two alternatives to specify the size of the neighborhood: The k-nearest neighbors (kNN) method (with a fixed number of samples n, now identical with "k" in kNN) or a region of a ball with fixed radius (with variable n). The choice of an appropriate neighborhood size affects subsequent computations, and there is not single best selection for all datasets. For the kNN method, we start with an n by the empirical rule-of-thumb recommended by Duda et al. [32] and check the visual feedback of the impact of neighborhood size to find an n that gives stable visual patterns. For the fixed radius ball method, an empirical radius of 10 percent of the size of the data domain is chosen. Both methods yield similar results for datasets having rather uniform densities in the data domain.

4.2 Eigenvector Analysis

Principal component analysis finds an orthogonal set of largest variances [33] of the point set of the nearest neighbors, X. We use the principal components to fit local planes formed by adjacent attributes in parallel coordinates (Section 4.3) and determine the correlation strength (Section 4.4). Our visualization technique is independent of the regression technique and, therefore, other linear regression estimation methods [34] could be used as well.

The largest eigenvector $\boldsymbol{e}^1$ and the second largest eigenvector $\boldsymbol{e}^2$ of the m-d space are recorded and normalized, and then projected to 2D or 3D subspaces to fit 1- and 2-flats respectively. For cases of $p > 2$, the p largest eigenvectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^p$ are normalized and projected to $(p+1)$-dimensional subspace to fit p-flat. We use the full m-d space for eigenvector computation and then perform projection to preserve as much information of the high-dimensional space as possible, since a study [26] has shown that the interpretation of trends in high-dimensional datasets requires the information of all dimensions rather than just the projected dimensions. Using the complete set of dimensions may miss some well correlated samples in subspaces. However, for the datasets of this paper, we find no significant differences if neighboring subspaces from parallel coordinates are used. Nevertheless, for other datasets, one could use subspace data mining methods [35] or visual exploration methods [36], [37] to analyze subspaces in detail.

4.3 Fitting p-Flats

Fig. 4 illustrates the geometric meaning of the numerical fitting of p-flats for the case of 3D data space. To fit 1-flats, we project the major eigenvector $\boldsymbol{e}^1$ located at data point ξ into all adjacent 2D subspaces, i.e., $x_1x_2, x_2x_3, \ldots, x_{m-1}x_m$, which yields the corresponding projected line segments $l_{12}, l_{23}, \ldots, l_{m-1,m}$. Line segment $l_{i,i+1}$ is the projection in 2D subspace x_ix_{i+1} of the m-d line segment l with end points $P^1 : \xi$ and $P^2 : \xi + \boldsymbol{e}^1$. Therefore, $l_{i,i+1}$ can be described by the two-point form line equation

$$l_{i,i+1} : y - P^1_{i+1} = \frac{P^2_{i+1} - P^1_{i+1}}{P^2_i - P^1_i}(x - P^1_i), \quad \text{where } i = 1, 2, \ldots m-1. \tag{6}$$

Fig. 4 shows line segment l_{12} as the projection in subspace x_1x_2 of the line segment from the 3D space with end points $P^1 : \xi_{123}$ and $P^2 : \xi_{123} + \boldsymbol{e^1}_{123}$.

For 2-flats, eigenvectors $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ are projected to adjacent 3D subspaces, i.e., $x_1x_2x_3, x_2x_3x_4, \ldots, x_{m-2}x_{m-1}x_m$. In a 3D

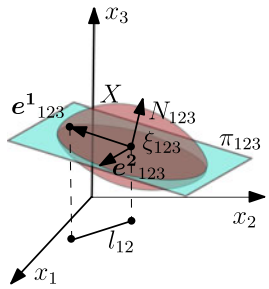


Fig. 4. Local p-flat fitting for the neighborhood X of data point ξ_{123}. The major eigenvector $\boldsymbol{e^1}_{123}$ is the linear trend of X, and is projected to 2D subspaces ($\boldsymbol{e^1}_{12}$ here for example) for 1-flat indexed point computation. The fitted 2-flat π_{123} is described by point ξ_{123} and normal vector N_{123}, which is calculated as the cross product of $\boldsymbol{e^1}_{123}$ and the second major eigenvector $\boldsymbol{e^2}_{123}$.

subspace spanned by x_i, x_{i+1}, and x_{i+2}, a local plane $\pi_{i,i+1,i+2}$ is fitted. The normal vector of $\pi_{i,i+1,i+2}$ is given by the cross product of the two projected eigenvectors ${e^1}_{i,i+1,i+2}$ and ${e^2}_{i,i+1,i+2}$, and the projected data point $\xi_{i,i+1,i+2}$ is used as a point on the plane. The plane $\pi_{i,i+1,i+2}$ is therefore written as

$$\begin{aligned} \pi_{i,i+1,i+2} : c_1 x_i + c_2 x_{i+1} + c_3 x_{i+2} = c_0, \\ \text{where } c_1 = N_i, c_2 = N_{i+1}, c_3 = N_{i+2}, \\ c_0 = N_i \xi_i + N_{i+1} \xi_{i+1} + N_{i+2} \xi_{i+2}, \end{aligned} \tag{7}$$

with normal vector N. Fig. 4 illustrates a case in 3D with plane π_{123} in subspace $x_1x_2x_3$.

For higher-dimensional p-flats ($p > 2$), the normal vector of the p-flat is calculated from the projected eigenvectors $\boldsymbol{e^1}, \boldsymbol{e^2}, \ldots, \boldsymbol{e^p}$. The computation can be done in analogy to the cross product in 3D. A $(p+1) \times (p+1)$ square matrix is constructed by setting $x_i, x_{i+1}, \ldots, x_{i+p+1}$ as the first row, and filling rows 2 through $(p+1)$ with eigenvectors $\boldsymbol{e^1}, \boldsymbol{e^2}, \ldots, \boldsymbol{e^p}$. Then we calculate the determinant of this square matrix. Coefficients $c_1, c_2, \ldots, c_{p+1}$ are then given by coefficients associated with $x_i, x_{i+1}, \ldots, x_{i+p+1}$ respectively in the determinant, and c_0 can be found by plugging data point ξ into Equation (4).

4.4 Indexed Points Calculation

The computation of indexed points is straightforward given information of fitted p-flats.

For indexed points of 1-flats, the slope and intercept of a projected line segment $l_{i,i+1}$ can be derived from Equation (6)

$$\begin{aligned} a_{i,i+1} &= \frac{P^2_{i+1} - P^1_{i+1}}{P^2_i - P^1_i}, \\ b_{i,i+1} &= P^1_{i+1} - P^1_i a_{i,i+1}. \end{aligned} \tag{8}$$

The indexed point $\bar{\pi}_{i,i+1}$ is then computed by plugging $a_{i,i+1}, b_{i,i+1}$ into Equation (1). The process is repeated for all $i = 1, 2, \ldots, m-1$ to get all indexed points of projections of $\boldsymbol{e^1}$.

Recall that four indexed points are needed to fully define a 2-flat. However, for visualization, this setting is redundant. The reason is threefold: First, it is unlikely in real-world multivariate data to have different planes sharing the first indexed point, i.e., coefficients of different planes yield same values in Equation (3); second, the associated 2-flat of the indexed point can be visually acquired through brushing and linking, while knowing the exact 2-flat equation is hard to interpret by the user and unnecessary; finally, it is difficult to associate the four indexed points of a 2-flat in the visualization. We therefore decide to use only the first indexed point $\bar{\pi}_{i,i+1,i+2}$ for a 2-flat located in subspace $x_ix_{i+1}x_{i+2}$. Such an indexed point is calculated by plugging Equation (7) into Equation (3).

For general p-flats ($p > 2$), the indexed point is computed by plugging coefficients $c_0, c_1, \ldots, c_{p+1}$ from Section 4.3 into Equation (5).

To describe the importance of an indexed point $\bar{\pi}$, a weight $w_{\bar{\pi}}$ is set. The weight denotes how well the full m-d space is preserved after the projection to the associated subspace. For a 1-flat indexed point $\bar{\pi}_{i,i+1}$, its weight $w_{\bar{\pi}_{i,i+1}}$ is formulated as

$$w_{\bar{\pi}_{i,i+1}} = ||{\boldsymbol{e^1}}_{i,i+1}||, \tag{9}$$

where $||\cdot||$ is the L2-norm. Since all eigenvectors have been normalized before projection, the weight measures the relative length of the projected eigenvector $\boldsymbol{e^1}$ to that of the original version $\boldsymbol{e^1}$. Imagine that if the 2D subspace is orthogonal to $\boldsymbol{e^1}$, the projected eigenvector becomes a point of length 0 and therefore the associated indexed point is not reliable and is set to a weight of 0. In the case of 2-flat indexed point $\bar{\pi}_{i,i+1,i+2}$, we formulate the weight to measure how well the two tangent vectors ${\boldsymbol{e^1}}_{i,i+1,i+2}$, ${\boldsymbol{e^2}}_{i,i+1,i+2}$ are projected from the m-d space to the 3D subspace $x_ix_{i+1}x_{i+2}$. We define the weight $w_{\bar{\pi}_{i,i+1,i+2}}$ as the product of lengths of the two vectors

$$w_{\bar{\pi}_{i,i+1,i+2}} = ||{\boldsymbol{e^1}}_{i,i+1,i+2}|| \cdot ||{\boldsymbol{e^2}}_{i,i+1,i+2}||. \tag{10}$$

The weight of a p-flat ($p > 2$) indexed point can be derived from Equation (10) by replacing the product of lengths of the two tangent vectors with the product of lengths of p projected eigenvectors in $(p+1)$-dimensional subspace

$$w_{\bar{\pi}_{i,i+1\ldots i+p}} = \prod_{j=1}^{p} ||{\boldsymbol{e^j}}_{i,i+1\ldots i+p}||. \tag{11}$$

A p-flat indexed point encodes the linear regression information and the correlation strength of the attributes at the same time: Its location represents the linear regression parameters; its weight indicates the correlation strength. Our choice of an indexed point's weight is simple and scalable to any number of dimensions. Other definitions of weight could also work, e.g., using the correlation coefficient as the weight of 1-flat indexed point; we have experienced no substantial difference between this definition and our choice of weights for the examples in this paper.

5 Visualization and User Interaction

In this section, we elaborate on how to visualize local multivariate correlations with indexed points, discuss user interactions, and briefly explain how we implemented our method. Fig. 5 shows an overview of the visualization system with an example of a white wine quality dataset [38]. Explanations of this visualization and further visual analysis of the data will be covered in detail in Section 5.8.

5.1 Plotting Indexed Points

The visualization of our method consists of three layers of plots: The original polyline-based parallel coordinates, the 1-flat indexed points, and the 2-flat indexed points. To support dense datasets, we use three layers of accumulation buffers for visualization. The first layer accumulates the density of original polyline-based parallel coordinates. The second and third layers accumulate weights of indexed points of 1- and 2-flats respectively. Specifically, the weight $w_{\bar{\pi}}$ of an indexed point $\bar{\pi}$ is accumulated into the buffer at its location in the 2D image plane of parallel coordinates.

Density-based rendering is achieved for each layer by taking the logarithm of the accumulated value at each pixel normalized to $\log(N)$, and then this normalized value is color mapped. The polyline layer is mapped to grayscale where higher density has lower luminance. Indexed points are colored with a discrete qualitative color map [39] based on the first axis of the subspace they belong to, and 1- and 2-flats with the same first axis share the same color. A colorbar is shown below parallel coordinates to indicate the association between subspaces and colors of p-flats. The contrast of

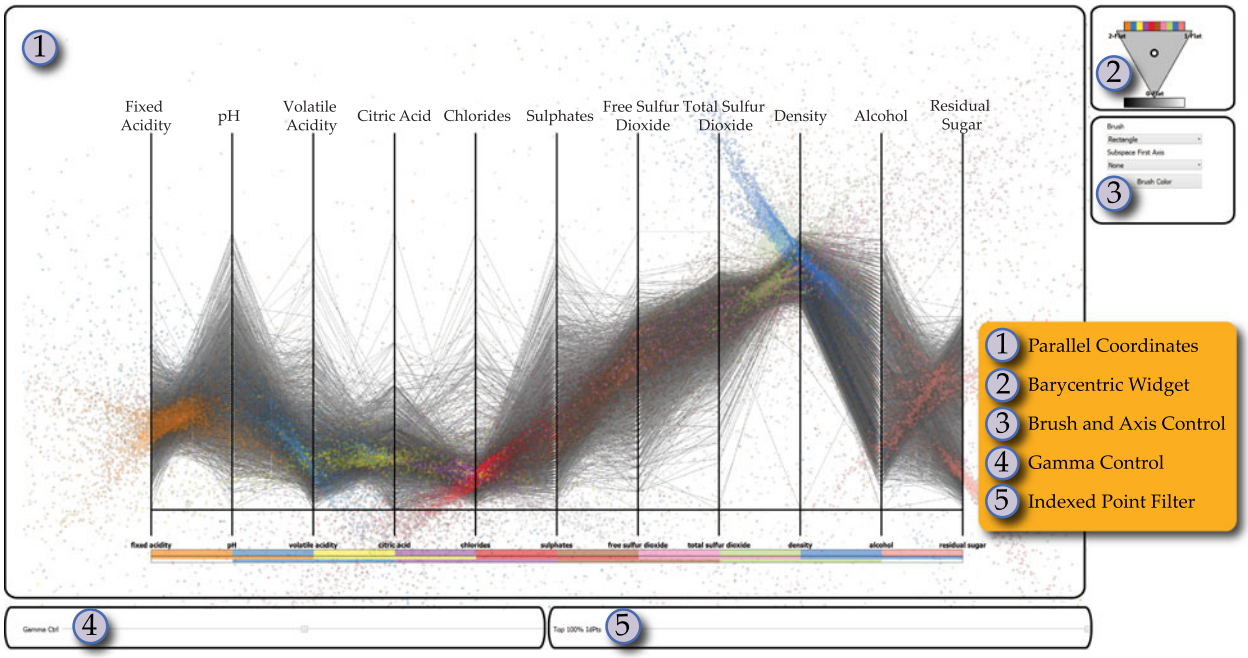


Fig. 5. The visualization system of our method for visualizing correlations in a white wine quality dataset.

the visualization can be adjusted via gamma mapping by a slider shown as (4) in Fig. 5.

5.2 Layer Blending and Selection

Once the data is loaded, a full plot with original parallel coordinates, 1-flats, and 2-flats is presented to the user. The three layers are alpha blended in back-to-front order with original parallel coordinates in the back and 2-flats in the front, and by default with an equal alpha value ($\frac{1}{3}$) assigned to each layer. The user can start the visual analysis by zooming in to any combination of layers by setting alpha values for layers. Barycentric coordinates allow easy and flexible control for blending between three variables [40]. We adopt barycentric coordinates for alpha blending as in (2) in Fig. 5. A control point $\boldsymbol{p}$ is a point inside or on a triangle with vertices $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$ representing three individual layers, and $\alpha_1, \alpha_2, \alpha_3$ are associated alpha values

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \alpha_3 \boldsymbol{v_3},$$
$$\text{where } \alpha_1 + \alpha_2 + \alpha_3 = 1.$$

In this way, setting the alpha value is intuitive: If $\boldsymbol{p}$ is closer to an vertex, higher alpha will be assigned to the associated layer.

5.3 Visual Pattern Interpretation

With indexed points of 1-flats, our method allows clear representations for both negative and positive correlations. Fig. 6 shows visual signatures of 1-flats for negative correlation (Fig. 6 left) and positive correlation (Fig. 6 right). The negative correlation case has indexed points mostly inside the two axes of focus, whereas indexed points are outside the selected axes for the positive correlation case. Note that positive correlations cannot be directly visualized in traditional parallel coordinates.

The effect of 2-flat indexed points is illustrated in Fig. 9. The underlying dataset contains three attributes in which all data points reside on a single plane in data domain. Such relationship could be seen in a 3D scatterplot with camera interaction. However, as the screenshot in Fig. 9a shows, it is hard to see this relationship in a static image. Similarly, it is impossible to identify such relationship in a SPLOM (Fig. 9b). Our method, in contrast, clearly shows the coplanar behavior of these attributes. Indexed points of 2-flats shown as orange points in Fig. 9c converge in nearly a single location, indicating that all local planes are almost on a single plane.

Identifying relevant p-flat patterns is important to make the visualization useful. Some interesting patterns of 1-flat indexed points are shown in Fig. 7. The underlying dataset, which is generated with random samples spanning the whole 3D data domain, contains a strong positively correlated pattern in the subspace of attribute 0 and attribute 1, and a strong negative correlation pattern in the subspace of attribute 1 and attribute 2. Both patterns are visible in the SPLOM as seen in Fig. 7d. With original parallel coordinates (Fig. 7a), it is hard to identify either correlation pattern. Fig. 7b shows 1-flat indexed points over parallel coordinates. Noticeable are two compact high-density patterns that are then brushed in Fig. 7c. As seen in Figs. 7e and f, the two brushes (yellow for

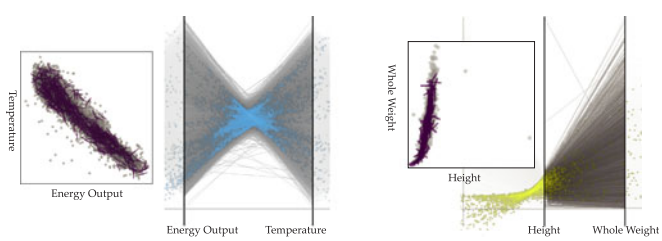


Fig. 6. Indexed points of fitted lines enable clear visual signatures of both negative and positive correlations. A negative correlation is shown between the "energy output" and "temperature" of a combined cycle power plant data (left). A positive correlation is seen to the right between the "height" and "whole weight" measurements of an abalone data (right). The direction of local correlation is shown by short purple lines in the scatterplots, adopting the dot-line representation [25], [26], i.e., drawing the local linear regression line at the corresponding data point.

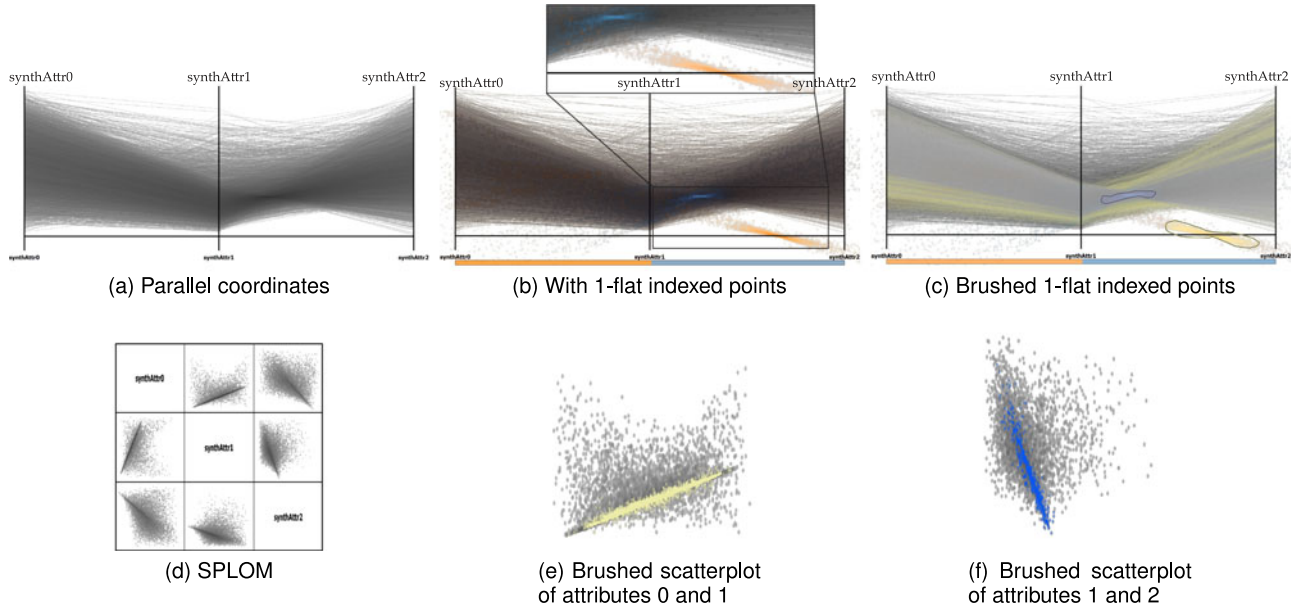


Fig. 7. A synthetic data to illustrate patterns of interest of 1-flat indexed points.

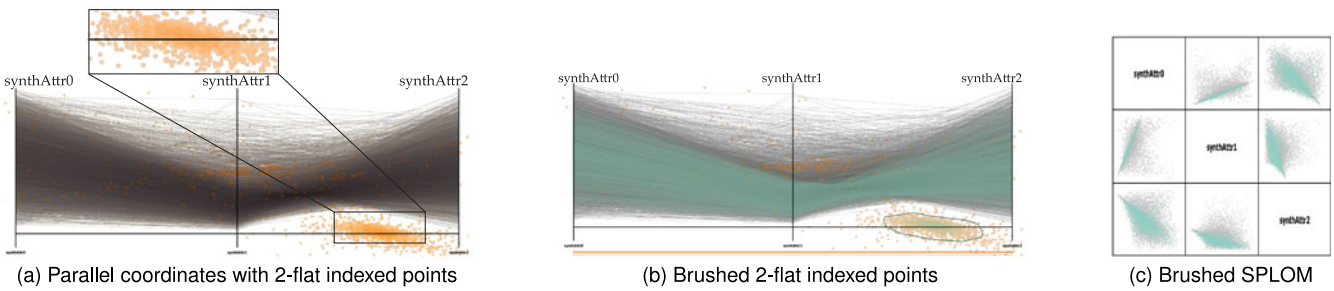


Fig. 8. Relevant patterns of 2-flats in the synthetic dataset.

the subspace of attributes 0 and 1, and light purple for the subspace of attributes 1 and 2) highlight the patterns with strong correlations. In general, to find meaningful 1-flat patterns, one should look for those where indexed points concentrate inside tight boundaries or have structures of high-density regions in original parallel coordinates. The same strategy should be taken to find 2-flat patterns, since 2-flat indexed points can be constructed from 1-flats as shown in Section 3. The same synthetic example is illustrated in Fig. 8, where the high-density 2-flat pattern in orange corresponds to a plane (Fig. 8c) as found through brushing (Fig. 8b).

Indexed points enable the visualization of multivariate nonlinear regression based on local linear correlations (Equations (1) and (3)). The case of 1-flats is of particular interest as the interpretation of indexed points is relatively easy. Fig. 10 shows visualizations of 1-flat indexed points in the subspace of attributes "pressure" and "temperature" of the hurricane Isabel simulation.[1] Nonlinear correlation is visible in the form of smeared out, often curved, high-density patterns of indexed points, as illustrated in Fig. 10a. After brushing with a lasso (Fig. 10b), a typical pattern of nonlinear regression with spread is seen in the linked 2D scatterplot (Fig. 10c). This example demonstrates that nonlinear patterns of 1-flat indexed points are associated with structures with nonlinear regression in data domain, and the variability of such structures in data domain is reflected by the variability of 1-flat indexed point patterns. The exploration of the hurricane Isabel data is discussed in more detail in Section 6.3.

5.4 Percentile Filtering

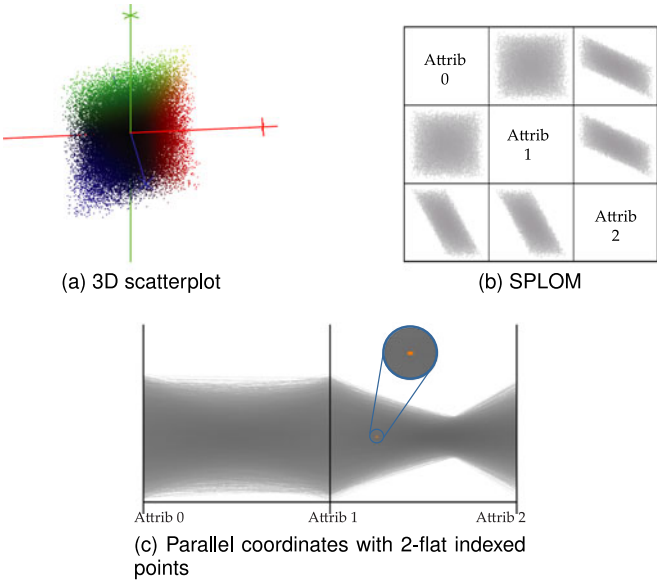
Based on weights of p-flat indexed points, a percentile filtering mechanism is adopted. For each subspace (2D for 1-flats, 3D for 2-flats), we compute the cumulative distribution function for the weight of indexed points and record the percentile of each indexed point. A slider shown as Equation (5) in Fig. 5 enables the user to perform percentile filtering.

Percentile filtering allows the user to remove indexed points with low weights while preserving those with high weights. Fig. 11 shows a comparison of before and after percentile filtering of 1-flat indexed points of the white wine data. The filtering enables the user to flexibly remove samples of less importance and focus on patterns with better projection quality.

5.5 Dimension Brushing

Dimension brushing, i.e., selection of a subset of attributes, has been used in parallel coordinates for interactive dimensional reduction [41]. In our work, dimension brushing is useful for closer analysis of local correlations between attributes as indexed points usually suffer from occlusions when all attributes are used. Patterns inside axes of attributes "chlorides", "sulphates" and "free sulfur dioxide" can be

1. http://vis.computer.org/vis2004contest/data.html



(a) 3D scatterplot (b) SPLOM

(c) Parallel coordinates with 2-flat indexed points

Fig. 9. A synthetic data with three attributes forming a plane in the data domain. (a) The relationship is hard to identify with a static image of a 3D scatterplot (even with sample location-based color coding for clarity). (b) Similarly, the relationship cannot be identified with a SPLOM. (c) Using 2-flat indexed points, it is clear that the samples (in orange, also in the zoomed-in inset) are on a plane.

observed in Fig. 12b in the white wine quality data. However, due to occlusions from indexed points of other attributes, one is not certain about the finding. Applying dimension brushing to highlight exclusively the subspace of "chlorides", "sulphates" and "free sulfur dioxide" allows a better visualization as shown in the top left inset in Fig. 12b.

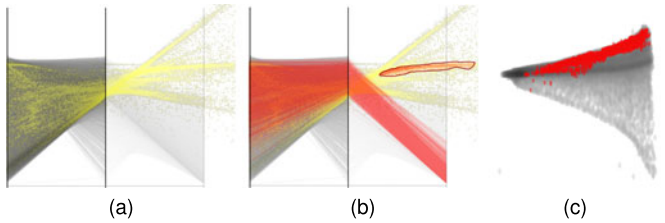
5.6 Data Brushing and Linking

Data brushing and linking allow the user to drill down and analyze specific clusters of polylines and p-flats in parallel coordinates. Range brushes are combined with AND-operation on parallel axes for original parallel coordinates brushing. Rectangular and lasso brushes are used to enable selection for indexed points. Original parallel coordinates, 1- and 2-flat indexed points are linked together and highlighted when a brush is placed.

To accelerate sample queries, a balanced kd-tree is built for each layer. The kd-trees for 1- and 2-flat indexed points are 2D trees recording the position of indexed points in 2D parallel coordinates plane. Additional information, including the ID of the associated raw data point, the subspace ID of where the indexed point resides, the strength of the indexed point, and the percentile in the whole population, is recorded in the node of a tree. For original parallel coordinates, the m-dimensional kd-tree that is used to find nearest neighbors (Section 4.1) is reused.

5.7 The SPLOM View

Scatterplots are good for visualizing correlations [4], and therefore we include a linked SPLOM view in our visualization tool. In order to support the identification of multivariate linear regression, we adopt the dot-line representation [25], [26] and draw fitted 1-flats as short line segments (Figs. 6 and 13a). The SPLOM view is linked with the parallel coordinates view through the kd-trees explained in the previous section.



(a) (b) (c)

Fig. 10. (a) 1-flat indexed points in the subspace of attributes "pressure" and "temperature" of the hurricane Isabel data, overlaid on top of regular parallel coordinates. (b) Lasso brush (top right) applied to a nonlinear, high-density pattern of indexed points; the corresponding parallel coordinates polylines are highlighted. (c) The brushed data points exhibit a corresponding nonlinear regression in the linked scatterplot.

5.8 Implementation

Our method was implemented using Qt and C++, and tested on a machine with 3.5 GHz i7 CPU, 16 GB main memory, and an nVidia GTX960 graphics card. The Eigen library [42] was used to aid the computation of local fitting. Dimension brushing is realized through a drop box where the user selects the first axis for the subspace or chooses to show all axes ((3) in Fig. 5). Alternatively, subspaces can be chosen by clicking on the corresponding bins in the color map for indexed points (color bar in (2) in Fig. 5). For a given first axis, two adjacent attributes are chosen for 1-flats, and three adjacent attributes are selected for 2-flats. Other user interface elements in Fig. 5 have been discussed in previous sections. The computation and plotting of indexed points is performed once the data is loaded, and the timing of the process depends on the size of the sampled data and is typically less than a minute. Thanks to kd-trees, full interactivity is achieved for all datasets for brushing and linking in the paper except for the hurricane Isabel data, which takes up to around 5 seconds to process querying and show brushed result.

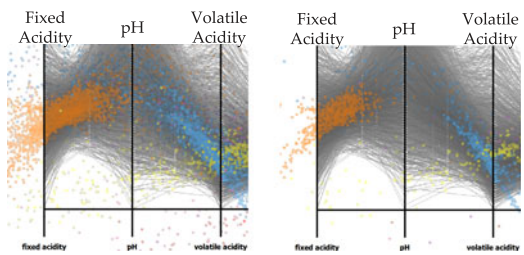
6 Examples

In this section, we demonstrate the usefulness of our method for visual analysis of high-dimensional datasets. We cover a range of data types: two high-dimensional information visualization datasets of different fields and a multivariate volume flow simulation data.

6.1 White Wine Quality

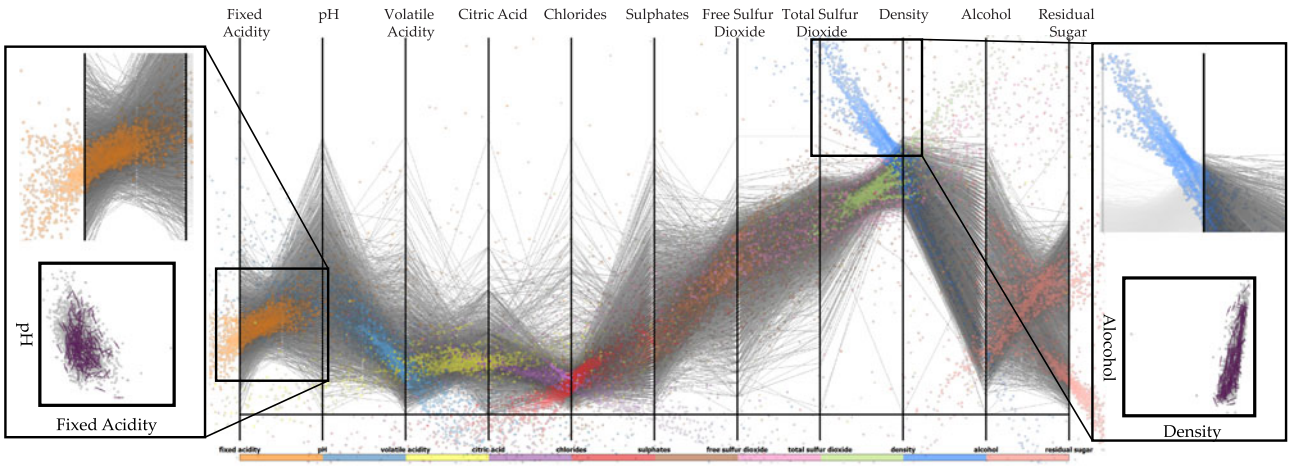
The white wine quality data [38] contains 4,897 samples of 11 chemical properties as continuous variables and a subjective quality attribute as ordinal variable. We discard the quality attribute, rearrange axes orders, and flip the density attribute.

It can be seen from the overview (Fig. 5) that the data follows our intuition: The "fixed acidity" seems to be negatively

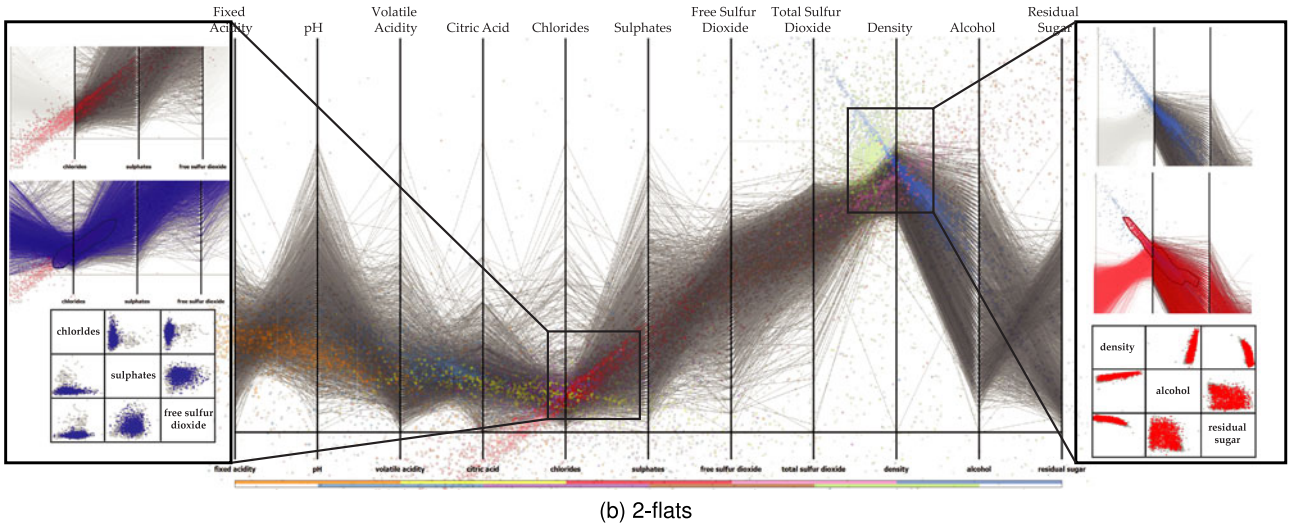


(a) (b)

Fig. 11. Percentile filtering of 1-flat indexed points: (a) before filtering, and (b) after filtering.



(a) 1-flats



(b) 2-flats

Fig. 12. Analysis of *p*-flats of the white wine quality dataset. Indexed points of 1-flats are shown in (a), whereas (b) shows the analysis of 2-flats. Local linear regression is based on kNN with $n = 100$ nearest neighbors.

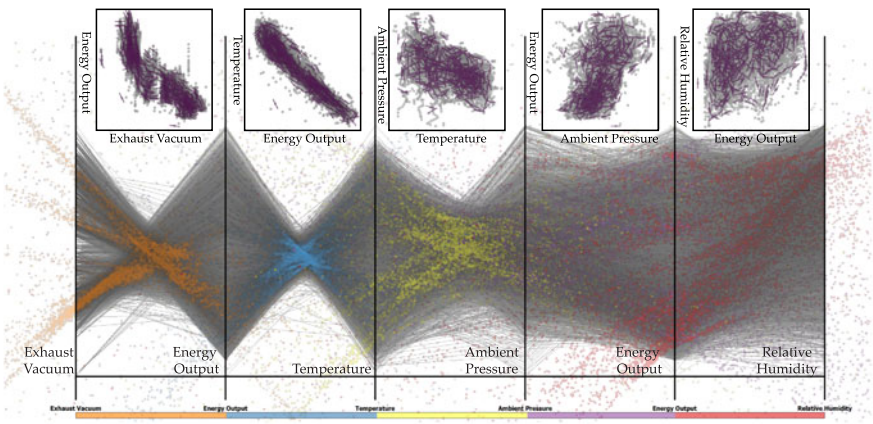
correlated to "pH", and "alcohol" is positively correlated to the flipped "density". These follow our expectation as we know that pH measures acidity and higher alcohol leads to lower density as alcohol is less dense than water. To examine the actual correlations, we use the barycentric widget to focus on 1-flats over the original parallel coordinates and take a closer look at these relationships. We apply dimension brushing to focus on pair-wise correlations as shown in insets in Fig. 12a. It is clear that "fixed acidity" is negatively correlated with "pH" (shown in orange), flipped "density" is proportional to "alcohol" (shown in dark blue).

Next, we explore the 2-flats. Giving more weight to 2-flats in the barycentric widget leads to the visualization shown in Fig. 12b. Using dimension brushing, we check 2-flats of all three adjacent attributes. Only two 3D subspaces are shown with obvious patterns: The subspace formed by chlorides, sulphates, and free sulfur dioxide attributes (the left inset), and the subspace of density, alcohol, and residual sugar (the right inset). We analyze details of the patterns of 2-flats using lasso brushes. From the associated SPLOM, we can see that the brushed area of the chlorides-sulphates-free sulfur dioxide subspace is shown as a plane (Fig. 12b left). Sulphates and free sulfur dioxide attributes are on a plane spanned by a small range of low chlorides. Samples in the density-alcohol-residual sugar subspace form a plane as seen in Fig. 12b right. Therefore, density, alcohol, and residual sugar attributes have a strong linear relationship.

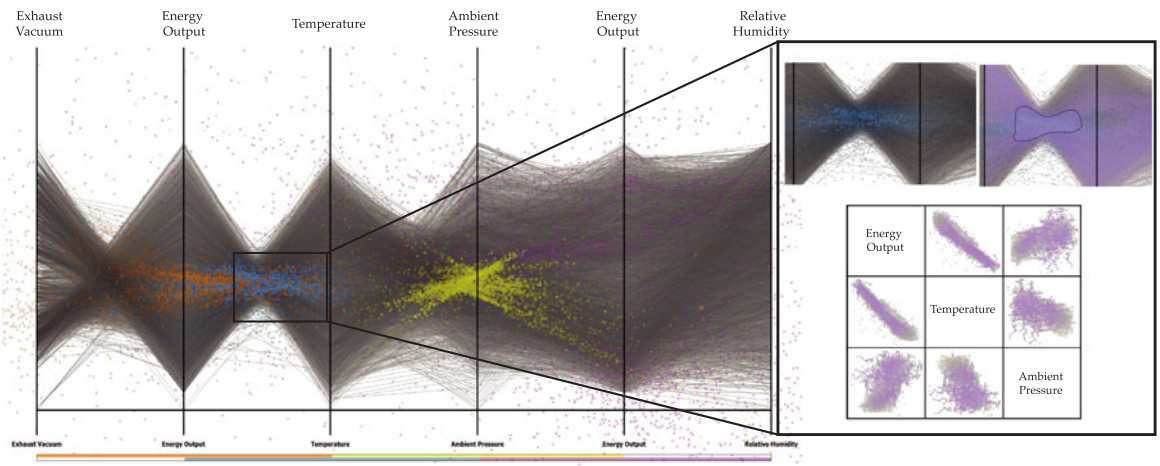
6.2 Combined Cycle Power Plant

The power plant dataset [43] contains 9,568 samples that record hourly average variables: "temperature", "ambient pressure", "relative humidity", "exhaust vacuum", and the net hourly electrical "energy" output of a combined cycle power plant over 6 years.

The goal is to examine the relationships between energy and other variables. We reorder the axes and make another copy of the energy attribute to better examine the correlations between energy and all other variables. The 1-flats are shown in Fig. 13a where we can clearly observe strong negative correlations between exhaust vacuum and energy (orange) as well as temperature versus energy (blue). We are able to visually determine that the negative correlation between temperature and energy is stronger than that of vacuum and energy. It can be seen that blue points form a tighter cluster than red points, also that fewer blue points are outside the temperature-energy pair than red points outside the vacuum-energy pair. Less strong negative correlations are also found between



(a) 1-flats



(b) 2-flats

Fig. 13. Visualizations of 1-flats (a) and 2-flats (b) of the power plant dataset. Local linear regression is based on kNN with $n = 190$ nearest neighbors.

temperature-ambient pressure, ambient pressure-energy, and energy-relative humidity.

The exploration of 2-flats can be seen in Fig. 13b. Tight patterns can be seen in subspaces energy-temperature-ambient pressure, shown by the blue indexed points (the yellow 2-flat pattern is also salient, which is a different permutation of the same attributes). The core of the pattern is highlighted by brushing, shown in the inset in which the pattern is associated with a plane. We can identify a strong linear relationship between energy-temperature-ambient pressure attributes, which can be attributed to the strong negative correlation between energy and temperature.

6.3 Hurricane Isabel

The hurricane Isabel simulation dataset is a widely used multivariate volume test data that contains $500 \times 500 \times 100$ spatial samples. We examine time step 25 using the "pressure", "temperature", "vapor", "height", and "speed" attributes and create the data domain by randomly sampling 1 percent of the original dataset, i.e., 250,000 samples, which well represents the original data domain. The "height" attribute records the vertical coordinate of the data, and the "speed" is calculated as the magnitude of the velocity attributes "U", "V", and "W". Unlike the simple geometry and therefore patterns in data domain of previous examples, the hurricane Isabel data is much more complicated and more local linear patterns can be observed in the data domain.

Fig. 14a shows the result of 1-flats over original parallel coordinates. Strong negative correlations are shown between height and temperature (blue)—in general, the temperature drops with increasing height. The pattern is not a single point since the temperature is not perfectly aligned with height. The most significant patterns of 1-flats for the temperature-pressure and pressure-vapor subspaces are both diagonal lines (yellow and purple). These main patterns consist of the majority of samples and therefore are less interesting. Noticeable, however, are several tilted lines off this main-trend diagonal line. These fine structures are closely examined in the inset to the right by selection using lasso brush. Each selection results in a cluster shown in the SPLOM view. These clusters are positively correlated stripes in the temperature-pressure subspace with slightly different characteristics. According to the ideal gas law, pressure and temperature are positively correlated. This may explain the behavior of these 1-flat patterns.

The analysis of 2-flats is shown in Figs. 14b–14e. In each 3D subspace, distinct patterns exist, but due to the large amount of samples in this data, percentile filtering is applied to analyze only indexed points with good projections. In Fig. 14c, after removing 65 percent of the indexed points via percentile filtering, a single cluster remains. Selecting the cluster yields a slanted plane in the speed-height-temperature subspace as shown in cyan in the SPLOM view. Two clusters exist after percentile filtering

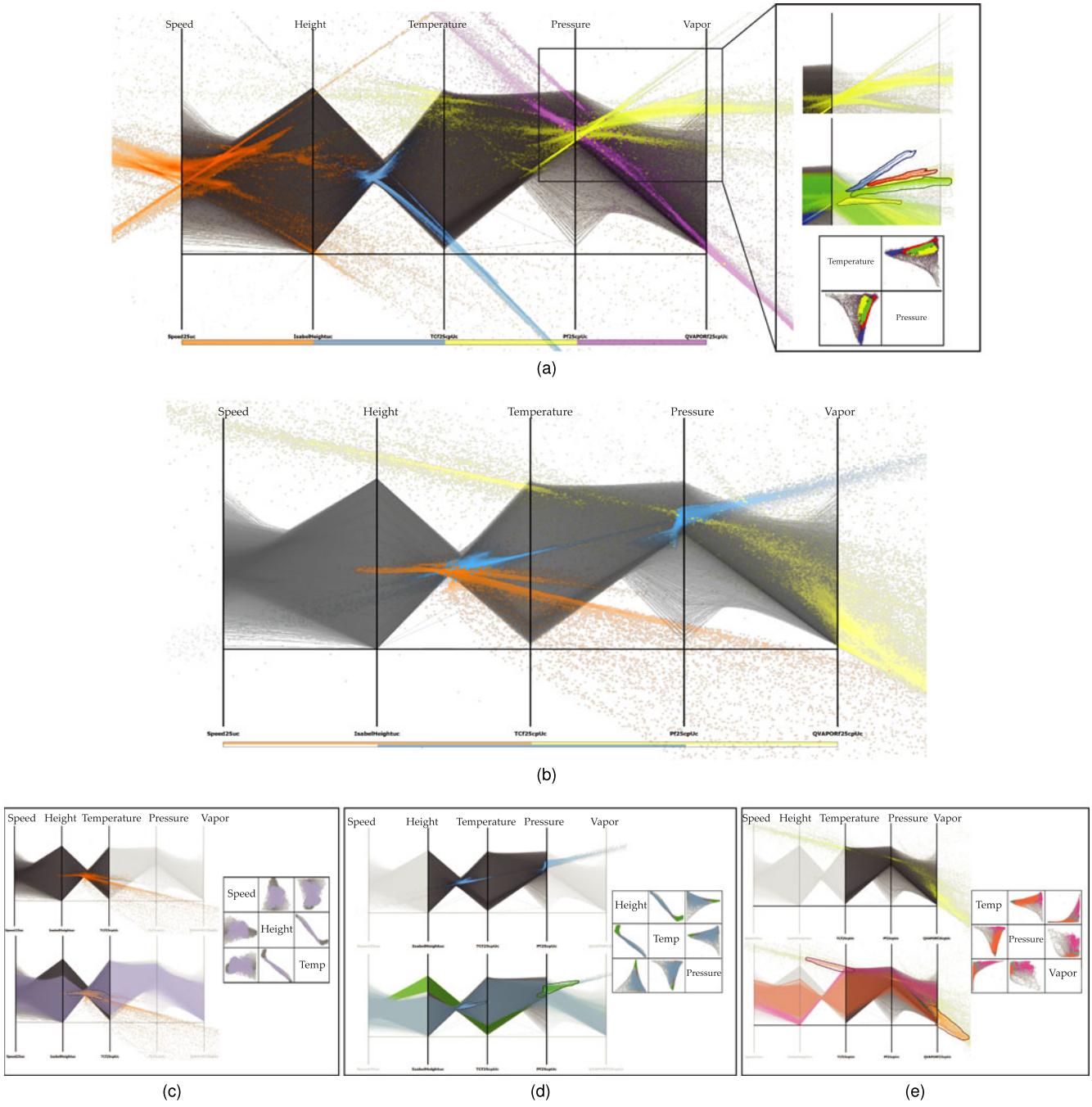


Fig. 14. Visualizations of the hurricane Isabel dataset. The five attributes from left to write are speed, height, temperature, pressure, and vapor. The analysis of 1-flats is shown in (a), 2-flats and their analysis are shown in (b)–(e). Local linear regression is based on a neighborhood within a ball of fixed radius 0.1.

out 70 percent of the indexed points in the height-temperature-pressure subspace (Fig. 14d). Each cluster is associated with a local plane: A small one in green and a larger one in blue. The green cluster covers a small range of the lowest temperature and the largest height at the tip of the structure in data domain. In Fig. 14e, two clusters remain after removing 70 percent of the indexed points. One cluster (the orange cluster) is close to the origin of the "vapor" axis, i.e., the intersection point of the vertical "vapor" axis and the horizontal axis. The region in data domain covers a very small range of low vapor, a medium range of pressure and a big range of temperature which forms a plane. It is clear that the plane is associated with the lowest vapor. The lowest vapor in this data is 0 indicating dry air, and therefore the associated regions in spatial domain should be unaffected by the hurricane. Besides, we can observe the positive correlation between temperature and pressure in the SPLOM view. In contrast, the other long and thin cluster (in purple) corresponds to a set of closely located small planes with highest temperature, high pressure and vapor, and lowest height. We can conclude from this combination that the cluster is associated with the area at the bottom of the data in the surrounding of the hurricane eye.

Selecting the aforementioned clusters would be very difficult if not impossible in this data without our method, since the associated structures are not distinct patterns in

neither basic parallel coordinates nor the SPLOM. Furthermore, from the exploration, we infer that the simulation uses gas laws as its computational model.

The above datasets could also be analyzed by data mining [44] and clustering techniques [45], which might lead to the same insights we obtained. However, our goal was to enable visualization to arrive at these results. Other visualization methods would have limitations analyzing these datasets. For 1-flats, the flow-based scatterplot approaches [25], [26] could lead to misinterpretation of regression information as the overdrawn streamlines might occlude each other. In contrast, our method maps regression parameters to indexed point locations, which allows users to interpret regression information regardless of the high density of the data—a benefit shared with another local correlation visualization technique for parallel coordinates [27]. Moreover, the 2-flats could be visualized by a 3D scatterplot, yet subject to the inherent perceptual issues of 3D visualization, whereas our method encodes 3D structures in a 2D static visualization. The effectiveness of our interaction techniques are illustrated in the supplementary video, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2017.2698041.

7 Limitations

Although our approach comes with several benefits outlined above, there are some limitations and characteristics that should be observed. First, the indexed points are based on combinations of data dimensions. For 1-flats, these dimensions are given by the neighboring axes of the parallel coordinates plot. As for regular parallel coordinates, we have to choose an appropriate combination of pairs of axes, i.e., existing methods of axis ordering could be applied [46]. For 2-flats, we need to find a good ordering of triples of dimensions (i.e., 3 subsequent dimension axes); to this end, axes ordering methods will need to be extended.

Second, further studies are needed to assess how well users can understand and identify the different visual patterns of 1-flats and 2-flats. We have included some basic interpretations in the paper as well as in the supplementary materials, available online, but this interpretation of the patterns has to be learned—similar to the learning curve involved in the first-time use of traditional parallel coordinates.

Third, the effectiveness of the visual representation of general p-flats ($p > 2$) is unclear. This issue is left for future work. It should be pointed out that such complex multivariate correlation poses a hard visualization problem in general; even advanced techniques like [47], [48] do not facilitate these kinds of correlation. At least, we know that our computation of the local regression and indexed point representation are both conceptually and mathematically independent of the dimensionality. Our method preserves the $O(N)$ computational complexity for general p-flats and keeps a visual point-to-point mapping. Therefore, there is some indication that the generalization might be feasible.

Fourth, by the definition of the 1-flat indexed point, its location approaches infinity if the slope of the linear regression approaches 1, making it impossible to be shown on a finite region on the 2D image plane.

8 Conclusion

We have presented a method for visualizing local multivariate correlations in parallel coordinates using indexed points of p-flats. Indexed points of p-flats are designed to represent generalized flat surface of dimension p in high-dimensional space in the same 2D domain of parallel coordinates [3]. We utilize indexed points to represent linear fittings of the local neighborhood of data points in the data domain. Specifically, lines and planes are calculated for 2D and 3D subspaces in local neighborhoods in data domain via principal component analysis. Then, indexed points of 1- and 2-flats are plotted for lines and planes respectively in parallel coordinates. This enables the visualization of local multivariate correlations in parallel coordinates, which was previously impossible. Another important benefit of our method is that it enables the visualization of positive and negative correlations with clear visual patterns. We have shown examples using different types of datasets to demonstrate the usefulness of our method. These examples demonstrate that the proposed method allows visual analysis that was not possible with basic parallel coordinates.

For future work, we would like to improve our method in several directions. First, we want to explore to which degree general p-plats can be effectively visualized. Second, strategies for the placement of dimensions should be investigated for 2-flats and higher-dimensional p-flats. Third, a mapping technique should be devised to show all indexed points in a limited 2D area. Fourth, the visual analysis process of indexed points could be aided by machine learning methods that automatically cluster indexed points before fine-tuning by the user. For example, clustering approaches [27] could be extended naturally to support the visual analysis of p-flats. Finally, our approach could be combined with other methods, for example, correlation summarization methods for higher-dimensional data, to analyze datasets with much higher dimensionality.

Acknowledgments

This work was supported by DFG within SFB 716/D.5.

References

[1] P. C. Wong and R. D. Bergeron, "30 years of multidimensional multivariate visualization," in *Proc. Sci. Vis. Overviews Methodologies Tech.*, 1997, pp. 3–33.
[2] A. Inselberg, "The plane with parallel coordinates," *Vis. Comput.*, vol. 1, no. 2, pp. 69–91, 1985.
[3] A. Inselberg, *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Berlin, Germany: Springer, 2009.
[4] J. Li, J.-B. Martens, and J. J. Van Wijk, "Judging correlation from scatterplots and parallel coordinate plots," *Inf. Vis.*, vol. 9, no. 1, pp. 13–30, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1057/ivs.2008.13
[5] R. A. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*. New York, NY, USA: Springer, 1987.
[6] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," in *Proc. Eurographics Conf. Vis.*, 2015, pp. 115–127.
[7] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates," in *Proc. 34th Annu. Conf. Eur. Assoc. Comput. Graph.—State of the Art Reports*, 2013, pp. 95–116.
[8] E. J. Wegman, "Hyperdimensional data analysis using parallel coordinates," *J. Amer. Statist. Assoc.*, vol. 85, no. 411, pp. 664–675, 1990.
[9] D. Holten and J. J. Van Wijk, "Evaluation of cluster identification performance for different PCP variants," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 793–802, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1467–8659.2009.01666.x

[10] X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu, "Scattering points in parallel coordinates," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1001–1008, Nov./Dec. 2009.

[11] C. Viau, M. J. McGuffin, Y. Chiricota, and I. Jurisica, "The FlowVizMenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1100–1108, Nov./Dec. 2010.

[12] J. H. T. Claessen and J. J. van Wijk, "Flexible linked axes for multivariate data visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2310–2316, Dec. 2011.

[13] H. Theisel, "Higher order parallel coordinates," in *Proc. Conf. Vis. Model. Vis.*, 2000, pp. 415–420.

[14] J. Heinrich, Y. Luo, A. E. Kirkpatrick, and D. Weiskopf, "Evaluation of a bundling technique for parallel coordinates," in *Proc. Int. Conf. Inf. Vis. Theory Appl.*, 2012, pp. 594–602.

[15] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualizations," in *Proc. IEEE Symp. Inf. Vis.*, 2004, pp. 81–88. [Online]. Available: http://dx.doi.org/10.1109/INFVIS.2004.68

[16] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing structure within clustered parallel coordinates displays," in *Proc. IEEE Symp. Inf. Vis.*, 2005, pp. 125–132.

[17] M. Novotny and H. Hauser, "Outlier-preserving focus+context visualization in parallel coordinates," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 5, pp. 893–900, Sep./Oct. 2006.

[18] J. Heinrich and D. Weiskopf, "Continuous parallel coordinates," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1531–1538, Nov. 2009.

[19] Y.-H. Fua, M. Ward, and E. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *Proc. IEEE Vis. Conf.*, 1999, pp. 43 –508.

[20] H. Zhou, W. Cui, H. Qu, Y. Wu, X. Yuan, and W. Zhuo, "Splatting the lines in parallel coordinates," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 759–766, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1111/j.1467–8659.2009.01476.x

[21] M. O. Ward, "XmdvTool: Integrating multiple methods for visualizing multivariate data," in *Proc. IEEE Vis. Conf.*, 1994, pp. 326–333. [Online]. Available: http://dl.acm.org/citation.cfm?id=951087.951146

[22] A. R. Martin and M. O. Ward, "High dimensional brushing for interactive exploration of multivariate data," in *Proc. IEEE Vis. Conf.*, 1995, pp. 271–278.

[23] H. Hauser, F. Ledermann, and H. Doleisch, "Angular brushing of extended parallel coordinates," in *Proc. IEEE Symp. Inf. Vis.*, 2002, pp. 127–130.

[24] H. Sanftmann and D. Weiskopf, "Illuminated 3D scatterplots," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 751–758, Jun. 2009. [Online]. Available: http://dx.doi.org/10.1111/j.1467–8659.2009.01477.x

[25] Y. H. Chan, C. D. Correa, and K. L. Ma, "Flow-based scatterplots for sensitivity analysis," in *Proc. IEEE Symp. Visual Anal. Sci. Technol.*, 2010, pp. 43–50.

[26] Y. H. Chan, C. D. Correa, and K. L. Ma, "The generalized sensitivity scatterplot," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 10, pp. 1768–1781, Oct. 2013.

[27] H. Nguyen and P. Rosen, "DSPCP: A data scalable approach for identifying relationships in parallel coordinates," *IEEE Trans. Vis. Comput. Graph.*, vol. PP, no. 99, 2017.

[28] H. Nguyen and P. Rosen, "Improved identification of data correlations through correlation coordinate plots," in *Proc. Int. Conf. Inf. Vis. Theory Appl.*, 2016, pp. 60–71.

[29] S. Bachthaler and D. Weiskopf, "Continuous scatterplots," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1428–1435, Nov. 2008.

[30] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975. [Online]. Available: http://doi.acm.org/10.1145/361002.361007

[31] E. Liberty, "Lecture notes in data mining: Nearest neighbor search," 2012. [Online]. Available: https://edoliberty.github.io/datamining2012a.html

[32] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2000.

[33] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, Jul./Aug. 2010. [Online]. Available: http://dx.doi.org/10.1002/wics.101

[34] H. Abdi, "Partial least square regression (PLS regression)," in *Proc. Encyclopedia Res. Methods Social Sci.*, 2003, pp. 792–795.

[35] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 1, pp. 1:1–1:58, Mar. 2009. [Online]. Available: http://doi.acm.org/10.1145/1497577.1497578

[36] A. Tatu, et al., "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2012, pp. 63–72.

[37] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2625–2633, Dec. 2013.

[38] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Syst.*, vol. 47, no. 4, pp. 547–553, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2009.05.016

[39] M. Harrower and C. A. Brewer, "ColorBrewer.org: An online tool for selecting colour schemes for maps," *Cartographic J.*, vol. 40, no. 1, pp. 27–37, Jun. 2003.

[40] G. Kindlmann, D. Weinstein, and D. Hart, "Strategies for direct volume rendering of diffusion tensor fields," *IEEE Trans. Vis. Comput. Graph.*, vol. 6, no. 2, pp. 124–138, Apr. 2000.

[41] S. Johansson and J. Johansson, "Interactive dimensionality reduction through user-defined combinations of quality metrics," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 993–1000, Nov. 2009.

[42] G. Guennebaud, et al., "Eigen v3," 2010. [Online]. Available: http://eigen.tuxfamily.org

[43] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061514000908

[44] M. Goebel and L. Gruenwald, "A survey of data mining and knowledge discovery software tools," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 1, pp. 20–33, Jun. 1999. [Online]. Available: http://doi.acm.org/10.1145/846170.846172

[45] P. Berkhin, "Survey of clustering data mining techniques," Accrue Software, Inc., 2002.

[46] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2203–2212, Dec. 2011.

[47] Y.-H. Chan, C. D. Correa, and K.-L. Ma, "Regression cube: A technique for multidimensional visual exploration and interactive pattern finding," *ACM Trans. Interactive Intell. Syst.*, vol. 4, no. 1, pp. 7:1–7:32, Apr. 2014. [Online]. Available: http://doi.acm.org/10.1145/2590349

[48] P. Klemm, et al., "3D regression heat map analysis of population study data," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 81–90, Jan. 2016.

Liang Zhou received the PhD degree in computing from the University of Utah, in 2014. He is a postdoc researcher in the Visualization Research Center (VISUS), University of Stuttgart, Germany. His research interests include scientific and information visualization, and visual analytics.

Daniel Weiskopf received the PhD degree in physics from the University of Tübingen, Germany, in 2001, and the Habilitation degree in computer science from the University of Stuttgart, Germany, in 2005. He is a professor in the Visualization Research Center (VISUS), University of Stuttgart, Germany. His research interests include information and scientific visualization, visual analytics, eye tracking, GPU methods, computer graphics, and special and general relativity. He is a member of the IEEE Computer Society, ACM SIGGRAPH, Eurographics, and the Gesellschaft für Informatik.