3-Dimensional Display for Clustered Multi-Relational Parallel Coordinates

Jimmy Johansson, Matthew Cooper and Mikael Jern
NVIS — Norrköping Visualization and Interaction Studio
Linköping University
Sweden
{jimjo, matco, mikje}@itn.liu.se

Abstract

Analysing multivariate data is a difficult task. Extensive interaction with the data is often necessary and, hence, the analysis can be quite time consuming. In this paper, we introduce a method to allow the user to simultaneously examine the relationships of a single dimension with many others in the data. The single dimension can then be interactively changed to allow the user to quickly examine all possible combinations. This method is achieved by extending the standard parallel coordinate approach to a 3-dimensional clustered multi-relational parallel coordinate representation (CMRPC). To aid this method, we use a technique called relation spacing which is used to position the axes according to how 'interesting' the different relations are. We also propose a number of interaction techniques to further facilitate the analysis process.

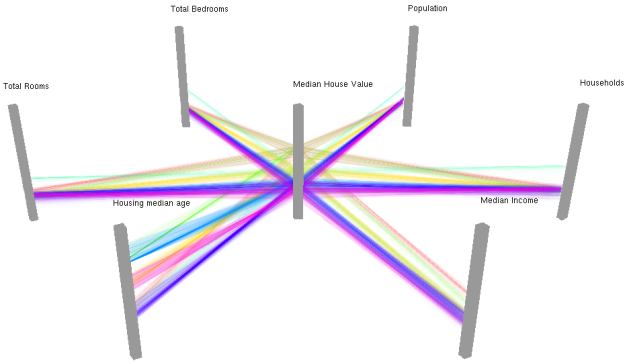


Figure 1. The clustered multi-relational parallel coordinates technique (CMRPC) allows for a simultaneous one-to-one relation analysis between the 'focus' dimension in the centre and all other dimensions.

1 Introduction

Exploratory data analysis is the process of examining data without knowing exactly what relationships or anomalies will be found. Since this often requires the user to try out several different approaches, this task is typically supported by a number of techniques for visualizing and interacting with the data. Such techniques exist in a wide range of areas, see for example, [6, 16, 15]. For data sets of up to 3 dimensions there are numerous standard methods that can be applied such as scatter plots, line plots, bar plots, pie charts, etc. When dealing with data of higher dimensions, (typically 6 dimensions and above) other visualization techniques must be applied (see [2] for an overview). Of the existing multivariate visualization techniques, parallel coordinates [14, 13] is one of the most frequently used and is the technique we also have chosen to build upon. Parallel coordinates transforms N-dimensional data into a 2-dimensional representation making it possible to simultaneously perceive the relationships present between the dimensions mapped to adjacent axes.

Despite the popularity of parallel coordinates, the technique is not without limitations. One disadvantage concerns the number of data items that it is possible to simultaneously display: visualizing a medium or large data set will inevitably cause cluttering. To get an overview of the data a commonly used approach is to perform an initial clustering of the data and, instead of visualizing each single data item, each cluster is visualized. Another disadvantage, which is the focus of this paper, is that the parallel axes configuration only permits analysis of correlations between adjacent axes. To examine all possible combinations can therefore be time consuming and quite tedious work.

In this paper we extend traditional 2-dimensional parallel coordinates to 3-dimensional clustered multi-relational parallel coordinates (CMRPC) (figure 1). This technique allows for complementary analysis of, and interaction with, larger multivariate data sets compared to standard parallel coordinates. To aid this visualization technique, we also in-

troduce a method called relation spacing. This method is used to position the axes according to how 'interesting' the different relations are. In this context, a relation means the relationship between a single pair of dimensions. We also propose a number of interaction techniques to further facilitate the analysis process.

The remainder of this paper is organized as follows. In section 2 we review the related work concerning clustering in parallel coordinates as well as previous efforts made to extend the parallel coordinates technique itself. Section 3 deals with the concept of clustering and how a cluster is represented in the parallel coordinates display. Section 4 describes how the CMRPC visualization technique is created and the supported interaction techniques. In section 5 we describe our method for relation spacing. In Section 6, we present our conclusions and discuss our future research work.

2 Related Work

Several efforts have been made to extend the standard parallel coordinates technique to display and analyse the results produced by different clustering algorithms.

Fua et al. [7] propose a multiresolutional view of the data via hierarchical clustering. Each cluster is visualized as a band faded from a completely opaque centre to a transparent edge. Berthold and Hall [3] use fuzzy rules to first cluster the data and then use parallel coordinates for displaying and analysing the result and their visualization is similar to the one presented in [7]. They use a solid line to represent the centre of each cluster and use the centroid of the cluster as the centre value. They also use a fading region but this shows the decline in membership of each data item. Andrienko and Andrienko [1] suggest "striped" envelopes and ellipse plots as two methods for displaying properties and structure of subsets in parallel coordinates. Both of these methods are based on dividing the value range of each axis into equal frequency intervals. A disadvantage of both methods is that they convey information about each variable independently of each other, hence it is not possible to investigate relationships between pairs of attributes. Another approach, based on the concept of representing each cluster as an envelope or polygon, is presented by Novotny [17]. He also uses a striped texture to further help the user distinguish between the different clusters.

Besides the efforts made to display large data sets in parallel coordinates, techniques have been proposed to extend the parallel coordinates technique itself. Hoffman et al. [12] proposed a variation of parallel coordinates with radial axes. This technique has the advantage that no axis is at the end, making analysis easier. Wegenkittl et al. [19] extended standard parallel coordinates to 3 dimensions for analysing and visualizing the behaviour of trajectories of high-dimensional dynamic systems. They introduced "extruded" parallel coordinates and 3-dimensional parallel coordinates. Extruded parallel coordinates are constructed by moving the parallel coordinate system in the third spatial axis, hence it is possible to have different parallel coordinate systems for each data item. The 3-dimensional parallel coordinates is based on linking parallel planes instead of parallel lines. Falkman builds on the above work in [5] and also uses parallel planes instead of parallel lines to analyse large amounts of clinical data. He also introduces methods for arranging the planes, as well as the lines to reduce the clutter when dealing with larger data sets.

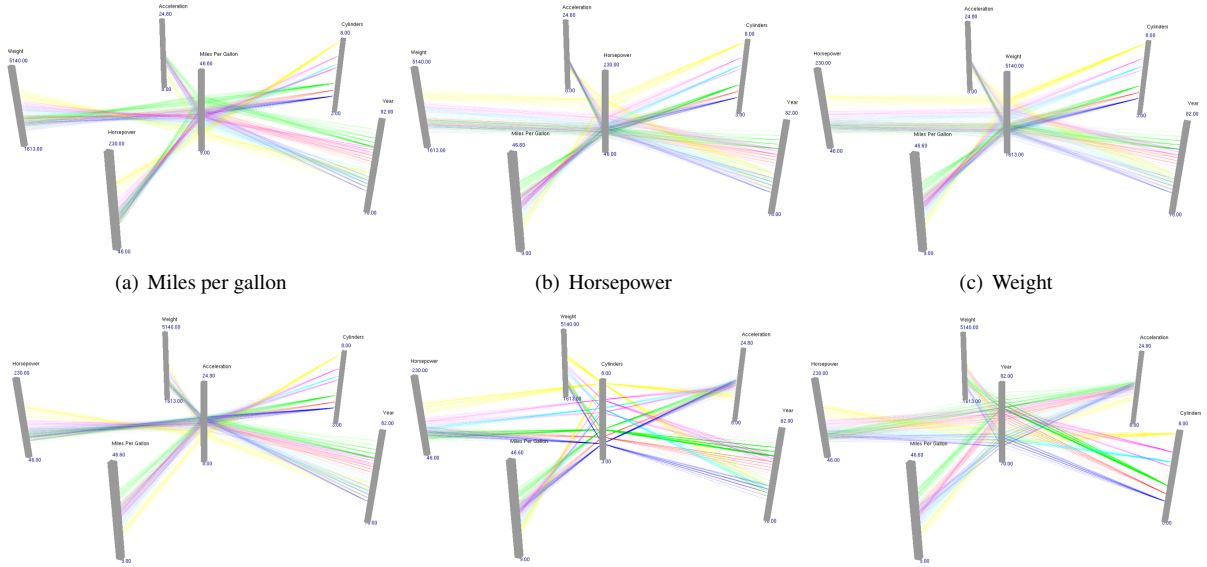
3 Clustering in Parallel Coordinates

A clustering algorithm aims at grouping data items so that data items in a cluster are as similar as possible and as different as possible from data items in the other clusters. We choose to use the K-means algorithm [10, 11], a simple and very well-known partitioning clustering algorithms, to cluster the data prior to visualization.

Representing a cluster with a coloured polygon in the parallel coordinates display gives, for a large data set, a huge performance advantage. This gives a good overview of the data set but the structure within the clusters is lost. On the other hand, colouring all data items that belong to the same cluster with the same colour makes it possible to see the individual lines and at the same time get information about their cluster membership. For a large data set, however, this would require far too large a number of lines to be rendered resulting in a non-interactive visualization. Selection and manipulation of the clusters could therefore not be interactively performed. To deal with this problem, we use a two-dimensional alpha-texture which is created as a pre-processing step using graphics hardware. As done in [20] we draw each line with a user-defined transparency value and use additive blending. This reveals the structure since high density regions will be more opaque than sparse regions. The final result is produced by applying the alpha-texture to a coloured polygon.

In the parallel coordinates display, the polygon used to represent each cluster can either be a uniform band displayed at the position of the cluster centroid or be of the clusters true size. For the uniform band, the relative width W is calculated according to $W = \frac{C}{C_{max}}\kappa$, where C is the population of the current cluster, C_{max} is the population of the largest cluster and κ is a scaling factor. To maximize the visual separation between the colours of each cluster we use the hue, saturation and intensity (HSI) colour model [8]. The saturation and intensity component are set to fixed values and the angle, ϕ, between each hue component is calculated as $\phi = \frac{2\pi}{P}$, where P is the number of clusters.

The advantage of this method is that we now only need to



(a) Miles per gallon (b) Horsepower (c) Weight

(d) Acceleration (e) Cylinders (f) Year

Figure 2. Analysing all possible relations amongst N dimensions. Figures (a–f) show the corresponding focus dimension of the cars data set.

render one polygon for each axis pair and cluster, resulting in a highly interactive visualization environment. In order to fine-tune the visual appearance of the cluster structures, the transparency and width of each line that constitutes the structure within each cluster together with the cluster width can be interactively changed during runtime.

4 3-Dimensional Multi-Relational Parallel Coordinates

In contrast with standard parallel coordinates, where all axes are mapped to a plane, CMRPC has the axes mapped onto a cylinder with one axis at its centre (hereinafter referred to as the 'focus' dimension). This visualization technique allows a simultaneous one-to-one relation analysis between the focus dimension and the other $N-1$ dimensions mapped to the cylinder. The axes are positioned according to a user-defined radius and a fixed angle, that is equal spacing between all axes. The angle, α, is simply calculated as $\alpha = \frac{2\pi}{N-1}$.

Any of the other dimensions can, with a single mouse click, be made the focus dimension. This technique makes it possible to analyse all possible relations in only $N-1$ visualizations. An illustration of this can be seen in figure 2 where the well-known cars data set [18] is visualized.

The dilemma of simultaneously perceiving a high number of dimensions is also a problem in the CMRPC display. We have found the upper limit for the number of dimensions that easily can be perceived and distinguished to be somewhere between 15-20. However, this upper limit is extremely dependent on the structure of the data and the number of clusters chosen. It is also dependent on the radius of the cylinder and, hence, the resolution of the screen. When visualizing a data set with a larger number of dimensions, we apply relation fading which allows the user to interactively fade relations and an illustration of this is shown in figure 3. In this example, 6 of the 12 relations between the focus axis and the other axes of a data set have been faded. The relation fading is performed with a single mouse-click on the axis and each relation can individually be faded or switched back to full opacity. Since the relations are faded, rather than completely turned off, they are always visible to the user and hence the overview is at all times preserved. Fading relations is something that would not serve any purpose in standard parallel coordinates since no more space would be released. In standard parallel coordinates, the alternative to the relation fading is to either completely remove an axis or to decrease the spacing between adjacent axes. This first alternative does not preserve the overview which is something the second technique does but at the cost of increased cluttering.

By displaying the intra-cluster structure we always have access to the original data items and, at the same time, have an overview of the data. Thus, we provide a kind of focus+context concept [4, 9]. For a more detailed analysis it

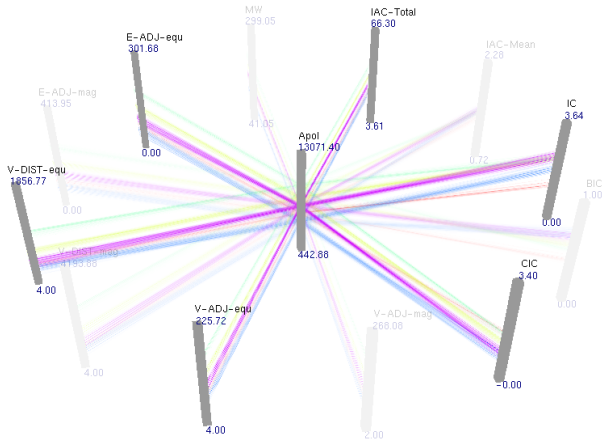


Figure 3. Interactive fading of every second relation helps prevent cluttering.

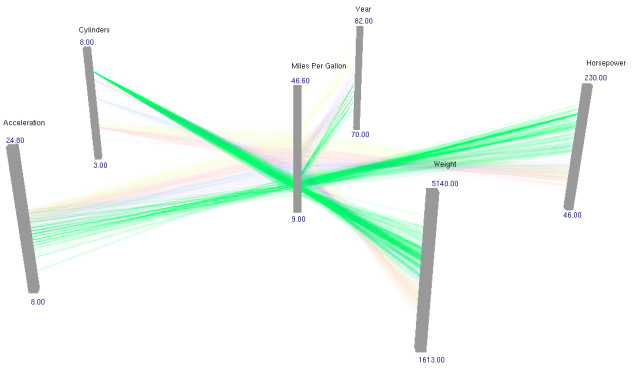


Figure 4. Selecting a cluster in the CMRPC display. The remaining cluster bands are made highly transparent. In this example, the clusters are represented at their true size.

is possible to highlight a single cluster to further examine its structure. The remaining cluster bands are then made highly transparent. Figure 4 shows the selected cluster represented at its true size in CMRPC.

5 Spacing of Relations

Instead of positioning each axis according to a fixed angle, the position can be used to represent a certain feature of the relation, referred to as relation spacing. We use correlation analysis to define how 'interesting' a relation is and calculate the correlation coefficient between the focus dimension, x, and all other dimensions, $y_{1,..,N-1}$. We use the absolute value of Pearson's correlation coefficient

$$\rho_j = \left| \frac{\sum_{i=1}^{M}((x_i - \overline{x})(y_{ij} - \overline{y}_j))}{\sqrt{\sum_{i=1}^{M}(x_i - \overline{x})^2 \sum_{i=1}^{M}(y_{ij} - \overline{y}_j)^2}} \right|, \quad (1)$$

where $\overline{x}$ is the sample mean of the x_i values (focus dimension), $\overline{y}_j$ is the sample mean of the y_{ij} values (the j^{th} dimension) and M is the number of data items in the data set. We use the absolute value, since we are only interested in the magnitude, not the sign of the correlation. This means that ρ_j will take on values in the range of $0 \leq \rho_j \leq 1$. We calculate the perceptually adjusted (square root) and normalized ρ'_j according to

$$\rho'_j = \frac{\sqrt{\rho_j}}{\sum_{i=1}^{N-1}\sqrt{\rho_i}}. \quad (2)$$

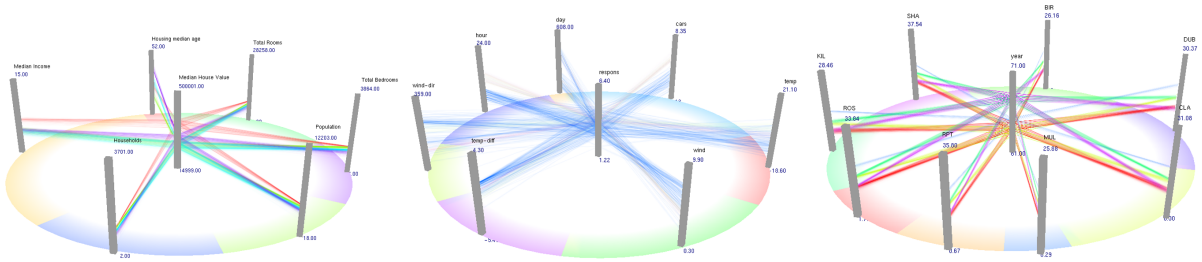
Taking the square root of the correlation gives a better distribution between the different correlation factors. We have found that, without the square root, a relation having a significantly larger correlation than the rest will tend to dominate the display. The angle, α, between dimensions j and $j+1$ is then calculated as

$$\alpha = 2\pi \frac{\rho'_j + \rho'_{j+1}}{2}. \quad (3)$$

The more interesting the relation, the more space is allocated on each side of the corresponding dimension. The method also takes the neighbouring relations into consideration so that two correlated neighbouring relations have more space between them than between one correlated and one uncorrelated. This technique is partly similar with the one proposed by Yang et al. [21]. They use correlation analysis as a method for calculating similarities between dimensions, but for 2-dimensional parallel coordinates. However, they do the opposite, namely putting similar axes close to each other. This approach does convey the measure of correlation between dimension pairs but our aim is to also avoid cluttering of the parallel coordinates display to which their method is prone. To perceive the magnitude of each relation correlation, a multi-coloured circle can be enabled and displayed at the bottom of the CMRPC display. Figure 5 illustrates the result of applying relation spacing when visualizing three different data sets with the CMRPC technique.

6 Conclusions and Future Work

In this paper, we have extended standard parallel coordinates to 3-dimensional clustered multi-relational parallel coordinates (CMRPC) and proposed a method for relation spacing.



(a) (b) (c)

Figure 5. Relation spacing applied to three different data sets. The greater the space is on each side of an axis, the more correlation exists between that dimension and the focus dimension. The coloured circle at the bottom reveals each relation's space. In figures (a–c), housing, pollution and meteorological data sets are visualized. In (b) a particular cluster has been selected for a more detailed analysis.

The CMRPC visualization technique has the axes mapped to a cylinder, with a 'focus' dimension in the centre. This makes it possible to simultaneously analyse the relationship between the focus dimension and all other dimensions. In CMRPC it is possible to either position all axes with equal spacing, treating all relations between the focus dimension and the other dimensions the same, or to use variable spacing, taking the correlation with the focus dimension into consideration. This method maps each axis with a space to its neighbouring axes according to the magnitude of the correlation. By doing this, the user's attention is attracted to the highly correlated relations providing more space for analysis. By allowing for techniques such as interactive axis rearrangement and relation fading, we provide a highly interactive visualization environment for exploratory analysis of multivariate data.

We find CMRPC to be an excellent extension of standard parallel coordinates because it enables an intuitive multiple one-to-one dimension analysis of multivariate data, which is something that the standard parallel coordinates is unable to do. Analysing all possible relations is done in $N-1$ visualizations, a task that would be much more time consuming to do with standard parallel coordinates. By combining the CMRPC technique with our cluster representation technique, we are able to interactively analyse medium as well as large data sets.

We see no limit to the applications of the CMRPC technique, every time a user needs to investigate the relationship between one variable and many others, this visualization technique can be applied.

For our future work, we would like to investigate if other measures, besides correlation, may be used for spacing the relations. It is also of interest to investigate if other interaction techniques could be supported.

Acknowledgements

This work has been funded by grant A3 02:116, supported by the Swedish Foundation for Strategic Research.

References

[1] G. Andrienko and N. Andrienko. Parallel coordinates for exploring properties of subsets. In *2nd IEEE International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 93–104, 2004.

[2] W. Basalaj. *Proximity Visualization of Abstract Data*. PhD thesis, University of Cambridge, 2000.

[3] M. R. Berthold and L. O. Hall. Visualizing fuzzy points in parallel coordinates. pages 369–374, 2003.

[4] A. Buja, J. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *IEEE Visualization*, pages 156–163, 1991.

[5] G. Falkman. *Issues in Structured Knowledge Representation: A Definitional Approach with Application to Case-Based Reasoning and Medical Informatics*. PhD thesis, Chalmers University of Technology, 2003.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *2nd International Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[7] Y. H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *IEEE Visualization*, pages 43–50, 1999.

[8] R. C. Gonzales and R. E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2002.

[9] M. Graham and J. Kennedy. Combining linking and focusing techniques for a multiple hierarchy visualization. In *5th IEEE International Conference on Information Visualization*, pages 425–432, 2001.

[10] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[12] P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *NPIVM '99: Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation.*, pages 9–16. ACM Press, 1999.

[13] A. Inselberg. The plane with parallel coordinates. pages 69–92, 1985.

[14] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *IEEE Visualization*, pages 361–378, 1990.

[15] R. Kincaid. Vistaclara: an interactive visualization for exploratory analysis of dna microarrays. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 167–174. ACM Press, 2004.

[16] W. Mackay and M. Beaudouin-Lafon. Diva: Exploratory data analysis with multimedia stream. In *Proceedings of the ACM CHI'98 Conference on Human Factors in Computing System*, pages 416–423, 1998.

[17] M. Novotny. Visually effective information visualization of large data. In *8th Central European Seminar on Computer Graphics (CESCG 2004)*, pages 41–48, 2004.

[18] Statlib. http://lib.stat.cmu.edu/datasets/.

[19] R. Wegenkittl, H. Löffelmann, and E. Gröller. Visualizing the behavior of higher dimensional dynamical systems. In *IEEE Visualization*, pages 119–ff. IEEE Computer Society Press, 1997.

[20] E. J. Wegman and Q. Luo. High dimensional clustering using parallel coordinates and the grand tour. Technical Report 124, Fairfax, Virginia 22030, U.S.A., 1996.

[21] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional dataset. In *IEEE 9th International Conference on Information Visualization*, pages 105–112, 2003.