

Interactive Dimension Reduction Through User-defined Combinations of Quality Metrics

Sara Johansson & Jimmy Johansson

Linköping University, Sweden

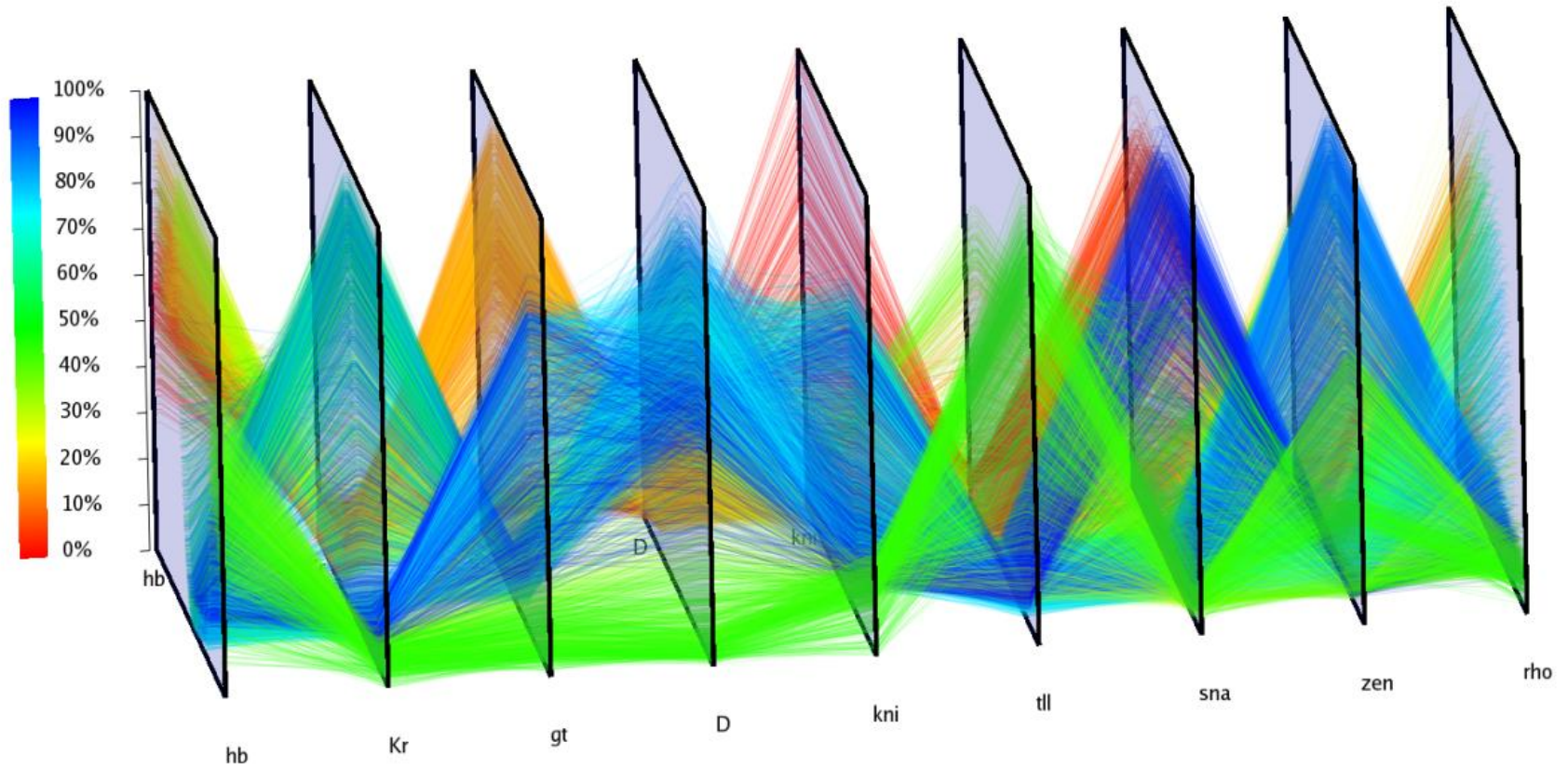
outline

- Multivariate data visualization
- Dimension Reduction
- Dimension Reduction with user interaction

Multivariate data visualization

- Scatter plot matrix
- Table lens
- Parallel coordinates
- Pixel-oriented display
- Value and Relation

Parallel coordinates



Pixel-Oriented display

- Represent data objects as pixels on the screen
- Display as many data objects as possible
- Colors, sub-window, shape and arrangements of sub-windows, order of dimension
- “Designing Pixel-Oriented Visualization techniques: Theory and Applications”, Daniel Keim, TVCG 2000.

Value and Relation display

- Build a distance matrix between pairs of dims
- Compute dimension position in a 2D space using MDS
- Create a glyph for each dimension that reveals patterns in that dimension
- Place glyphs in their 2D space positions.
- “Value and Relation Display for Interactive Exploration of High Dimensional Data”, J. Yang, et.al. Infovis2004.

Value and Relation display

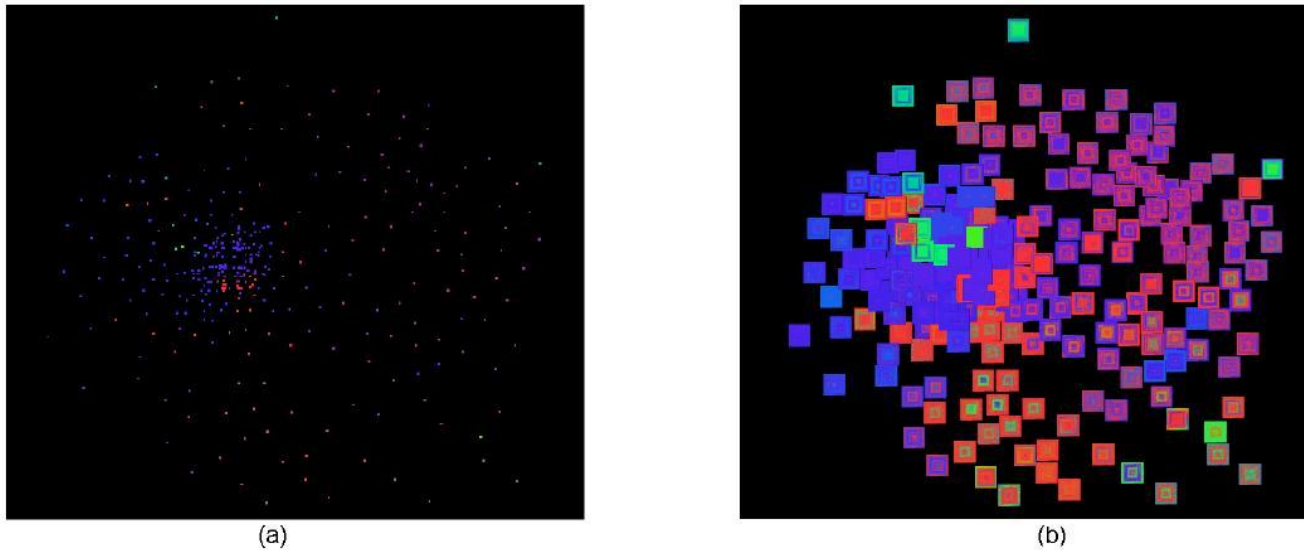
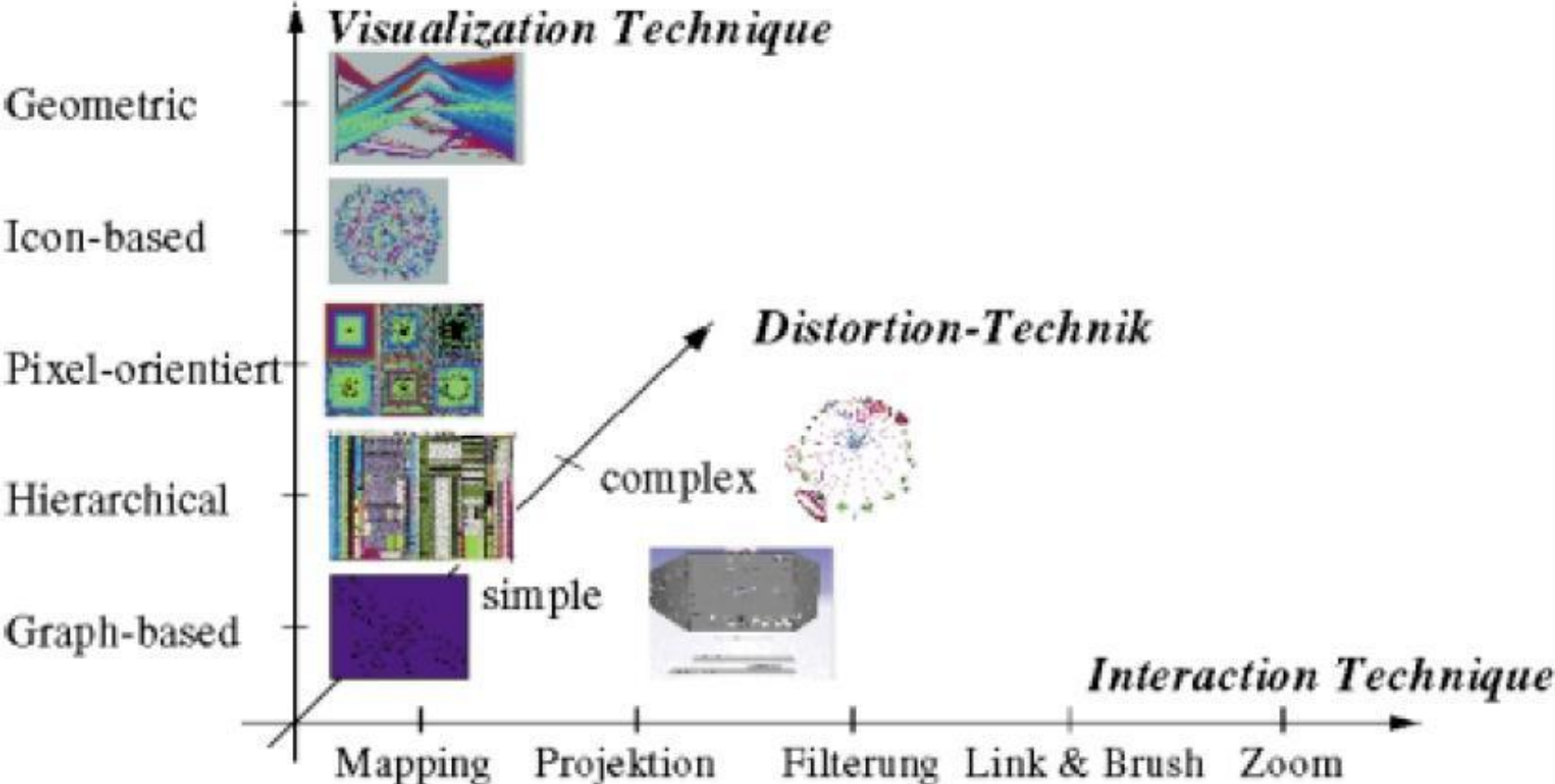


Figure 1: The VaR Display. (a) A star field display where each dimension is mapped to a dot and positioned using MDS according to the correlation among the dimensions. (b) The dots in (a) are replaced by glyphs that present values of the data items to form a VaR display. The dataset is the SkyServer dataset (361 dimensions, 50,000 data items), which was extracted from the Sloan Digital Sky Server (SDSS) data [8].

Multivariate data visualization



outline

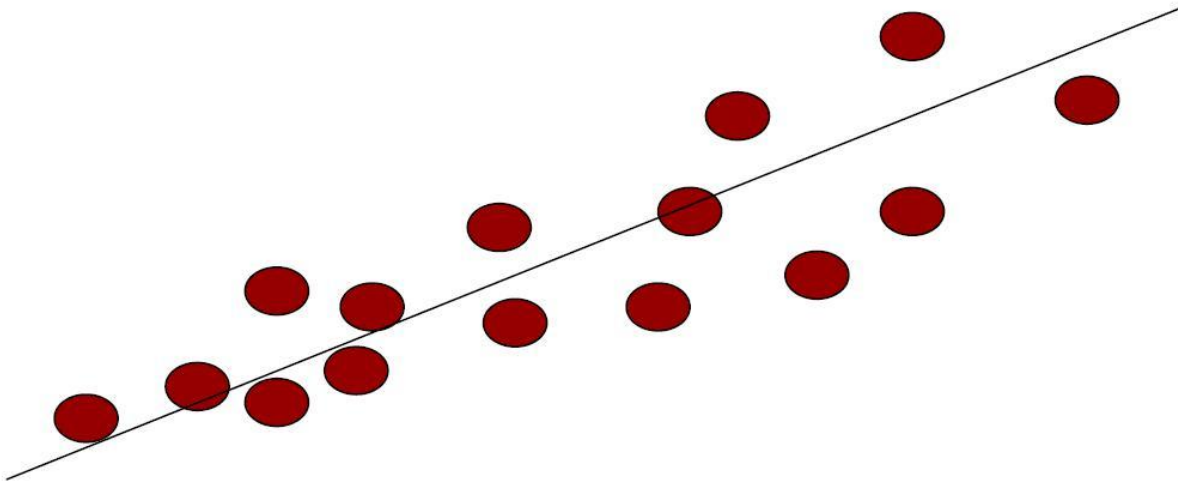
- Multivariate data visualization
- Dimension Reduction
- Dimension Reduction with user interaction

Dimension Reduction

- Principle Component Analysis (PCA)
- Principle component variable grouping (PCVG)
- Multi-Dimensional Scaling (MDS)
- Self Organizing Maps (SOM)
- Factor analysis
- Locality Preserving Projection

PCA

- Algebra interpretation
- Statistics interpretation
- Linear Algebra (Eigen Vector)



PCVG

- An unsupervised method that assigns a large number of variables to a smaller number of groups
- Choose one representative variable from each group to visualize
- “Dimensionality reduction and visualization in Principle Component Analysis”, G. Lvisev, et. al., Analytical Chemistry

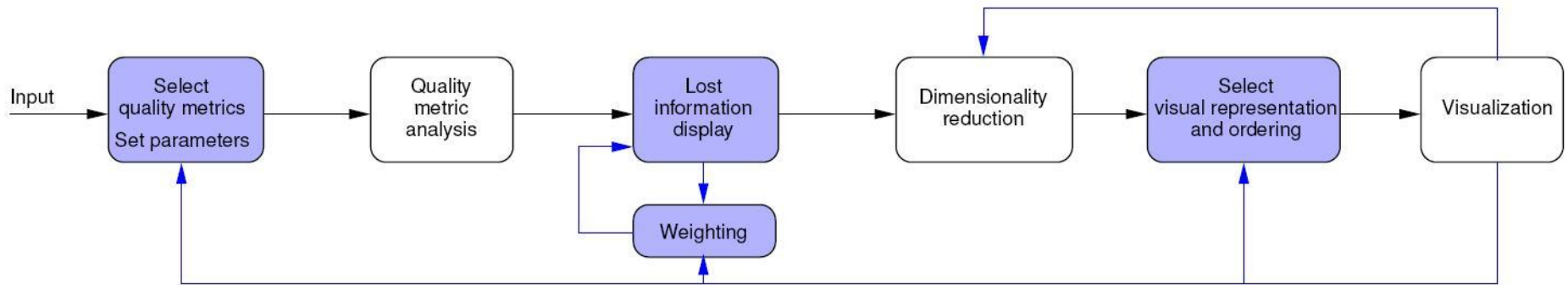
outline

- Multivariate data visualization
- Dimension Reduction
- Dimension Reduction with user interaction

Dimension Reduction with user interaction

- Existing systems
 - Visual hierarchical dimension reduction (VHDR)
 - Dimension ordering spacing and filtering (DOSFA)
- Interactive dimensionality reduction through user-defined combinations of quality metrics
 - Dimension reduction based on user-defined and weighted quality metrics
 - Provide intuitive display of the trade-off of dimension reduction and information loss
 - Automatic variable ordering

System overview



System overview

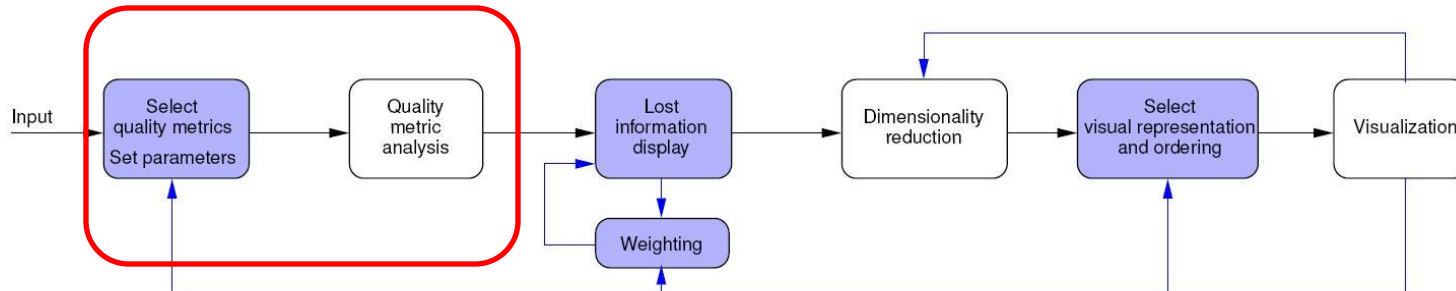
- Data

- $X = \{x_{ij}\}^{m \times n}$; $x_{i.}$: data item; $x_{.j}$: variable

- Quality metrics

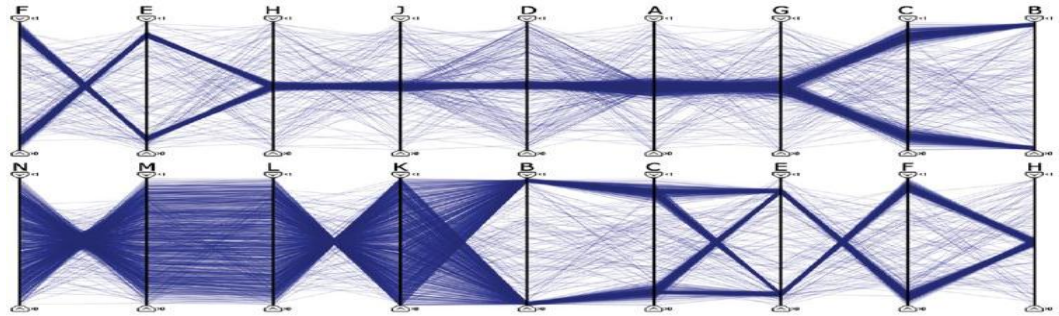
- Correlation analysis
 - Outlier detection
 - Clustering detection

$$I(\vec{x}_j) = w_{\text{corr}}I_{\text{corr}}(\vec{x}_j) + w_{\text{out}}I_{\text{out}}(\vec{x}_j) + w_{\text{clust}}I_{\text{clust}}(\vec{x}_j)$$

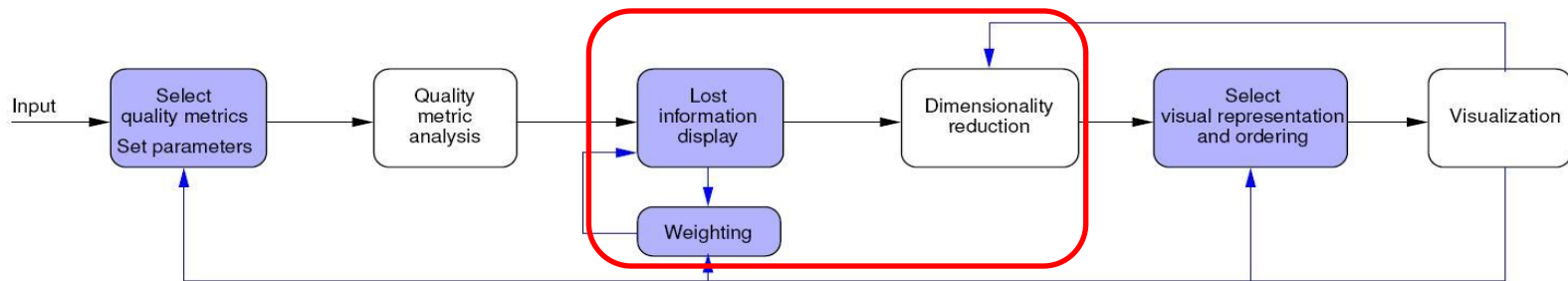
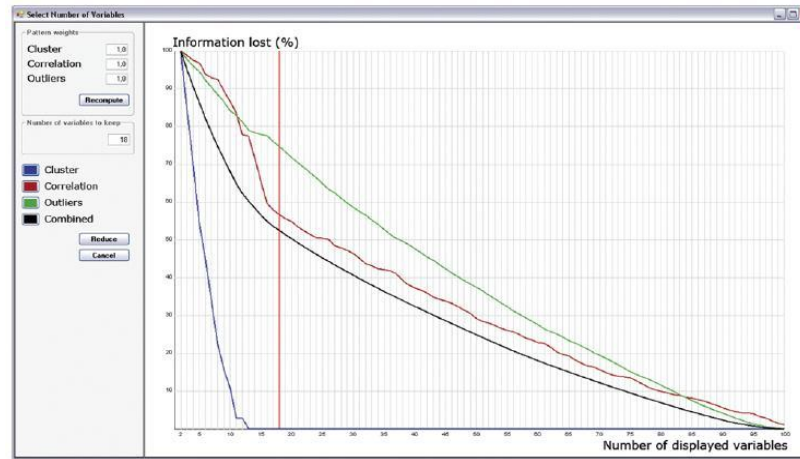


System overview

- weighting



- Information loss



Quality metrics - correlation

- Pearson correlation

$$r(x_{.j}, x_{.k}) = \frac{\text{cov}(x_{.j}, x_{.k})}{\sigma_{x_{.j}} \sigma_{x_{.k}}} = \frac{E[(x_{.j} - \mu_{x_{.j}})(x_{.k} - \mu_{x_{.k}})]}{\sigma_{x_{.j}} \sigma_{x_{.k}}}$$

- Variable correlation metrics

$$I(x_{.j}) = \sum_{k=1, k \neq j, |r(x_{.j}, x_{.k})| > \varepsilon}^n |r(x_{.j}, x_{.k})|$$

Quality metrics - Outlier

- Density and grid based approach
 - Radius r
 - Threshold h
- Two dimensional outliers $\rho_i(j,k) < h$
 - 2D grid length = $\frac{r}{\sqrt{2}}$
 - Candidate outlier
 - Count surrounding points within a circle
- higher dimensional outliers

Quality metrics - Outlier

- Item outlier quality value
 - Defined over 2D outlier pairs

$$o_i = \sum_{j \neq k, \rho_i(j,k) < h} \frac{1}{\rho_i(j,k) + 1}$$

- Variable outlier quality value

$$I_{out}(x_{\cdot j}) = \sum_{i=1, o_i > \zeta, j \in V(i)}^m o_i$$

Quality metrics – cluster

- Mafia clustering algorithm (Nagesh, ICDM01)
 - Density and grid based
 - Bottom-up approach: k -dimensional subspace cluster is obtained by merging two $(k-1)$ -dim clusters sharing with $(k-2)$ dimensions
 - Adaptive grid size based on data distribution

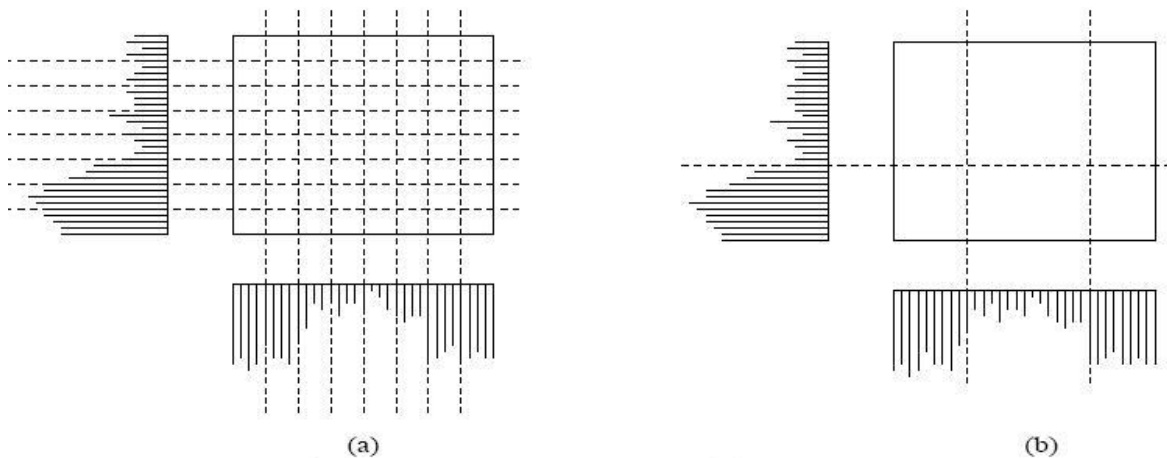


Figure 1. (a) Uniform grid size (b) Adaptive grid size

Quality metrics – cluster

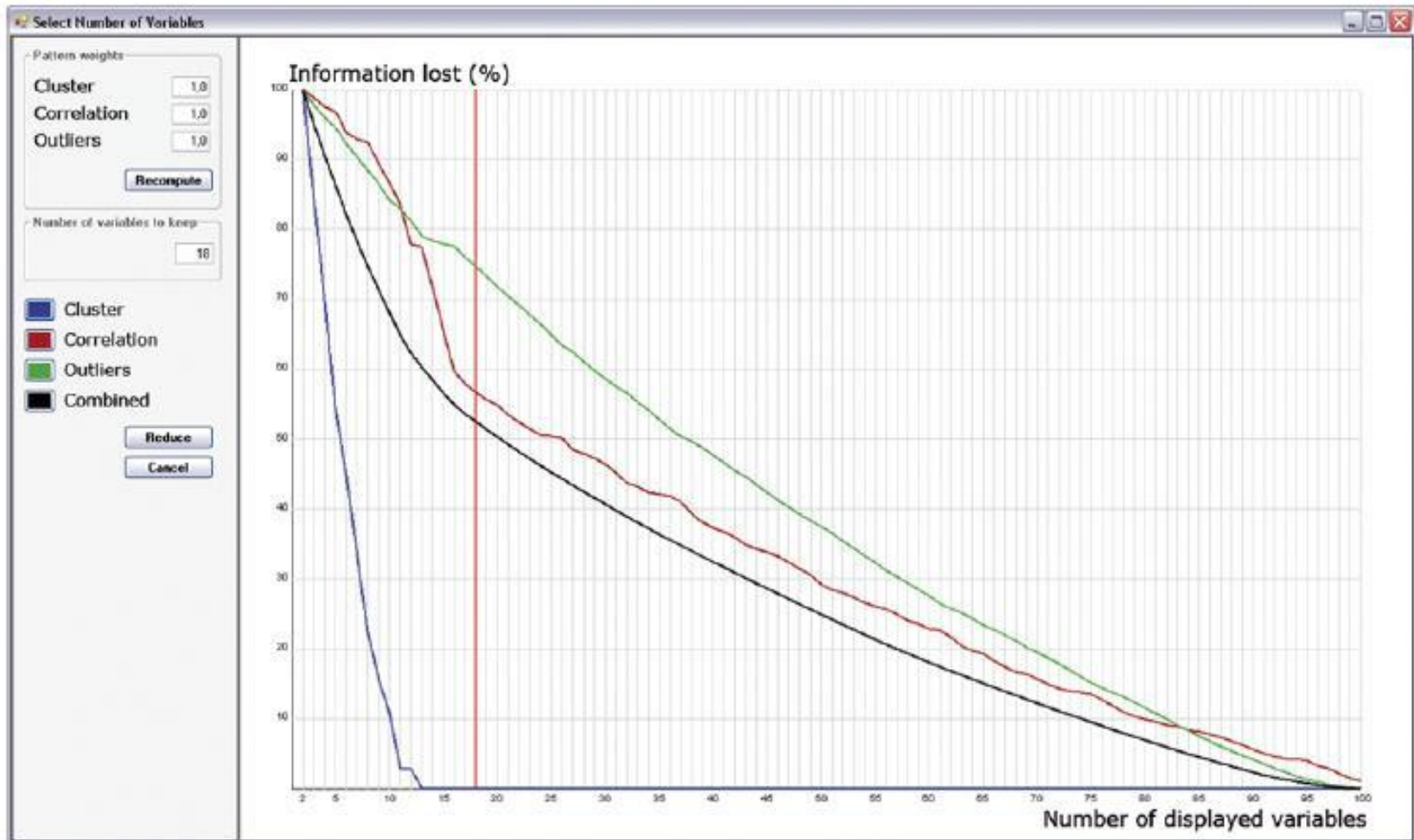
- Cluster quality metrics
 - Density measure
 - Dimensionality measure
 - Coverage measure

$$\sigma_c = d_c k_c f_c$$

- Variable cluster quality metrics

$$I_{cluster}(x_j) = \sum_{c, f_c > \varphi, j \in V(c)} \sigma_c$$

Information loss & select data size



Information loss & select data size

- Order n dimension based on their weighted combination of quality metrics

$$I(\vec{x}_j) = w_{\text{corr}}I_{\text{corr}}(\vec{x}_j) + w_{\text{out}}I_{\text{out}}(\vec{x}_j) + w_{\text{clust}}I_{\text{clust}}(\vec{x}_j)$$

- Information loss

$$I_{\text{total}} = \sum_{j=1}^n I(x_{.j})$$

$$I_{\text{removed}} = \sum_{j=k+1}^n I(x_{.j})$$

$$I_{\text{loss}} = \frac{I_{\text{removed}}}{I_{\text{total}}}$$

Variable ordering

- Basic idea
 - Try to put variables belonging to the same cluster together
 - Based on the correlation quality metrics

$S_{reduced} = [1, 2, 3, 7, 10, 12, 16]$		
Cluster	Variables	Quality value
c_0	[3, 6, 7, 10]	0.9
c_1	[2, 3, 10, 17]	0.7
c_2	[1, 2, 7]	0.4

First iteration: [1, 2, 12, 16, 3, 7, 10]
 c_0

Second iteration: [1, 12, 16, 2, 3, 10, 7]
 c_1 c_0

Third iteration: [12, 16, 1, 2, 3, 10, 7]
 c_2 c_1 c_0

Fig. 5. Example of variable ordering algorithm for cluster enhancement. Initially the clusters are ordered according to quality values. For each iteration the reordering is found that results in the longest sequence of connected variables being part of c_i , without traversing the borders of previous clusters (represented by red and pink rectangles)

Variable ordering

- Steps
 - Sort clusters in descending order based on their correlation quality value
 - In the first iteration, all variables of c_0 that are in $S(\text{reduced})$ are positioned next to each other
 - In subsequent iterations, reorder variables in $S(\text{reduced})$ that result in the longest sequence of variables in the current iteration cluster.

Result

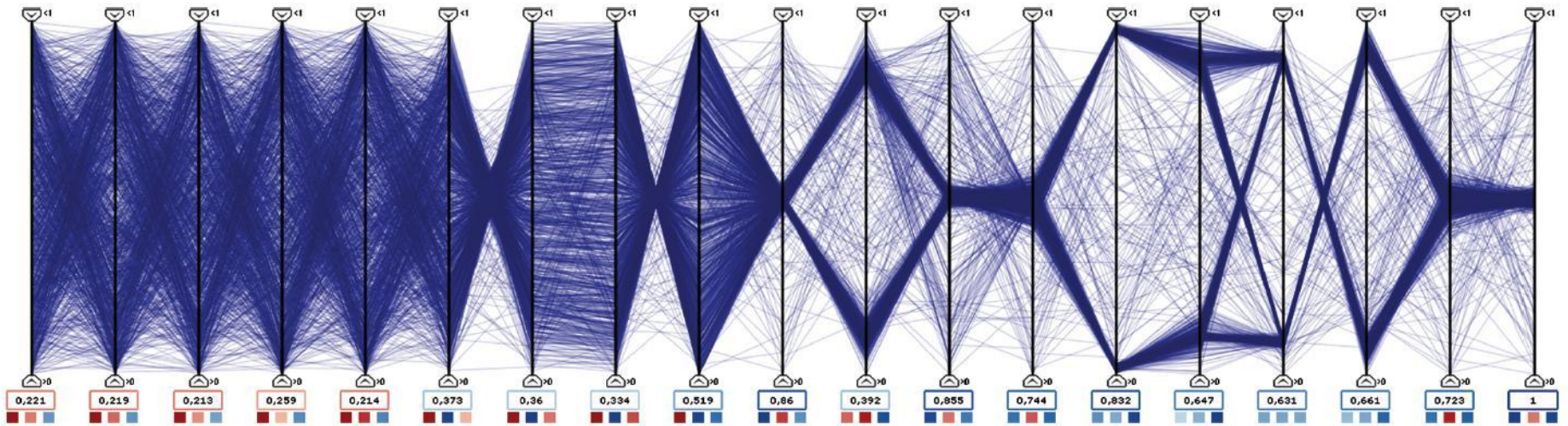


Fig. 6. The synthetic data set reduced to 18 variables, according to figure 4. The additional information at the bottom of the axes indicates by the red colour that the five leftmost variables, which mainly contain noise, have little importance in representing the structures of the data set.

Conclusion

- A Multivariate data visualization System
- Dimensionality reduction with user interaction
 - Weighted quality metrics
 - Reduction set selection
- Automatic variable ordering

Thanks!