CS-E4840 Information Visualization Lecture 10: Dimensionality reduction & student presentations

Tassu Takala <tapio.takala@aalto.fi>

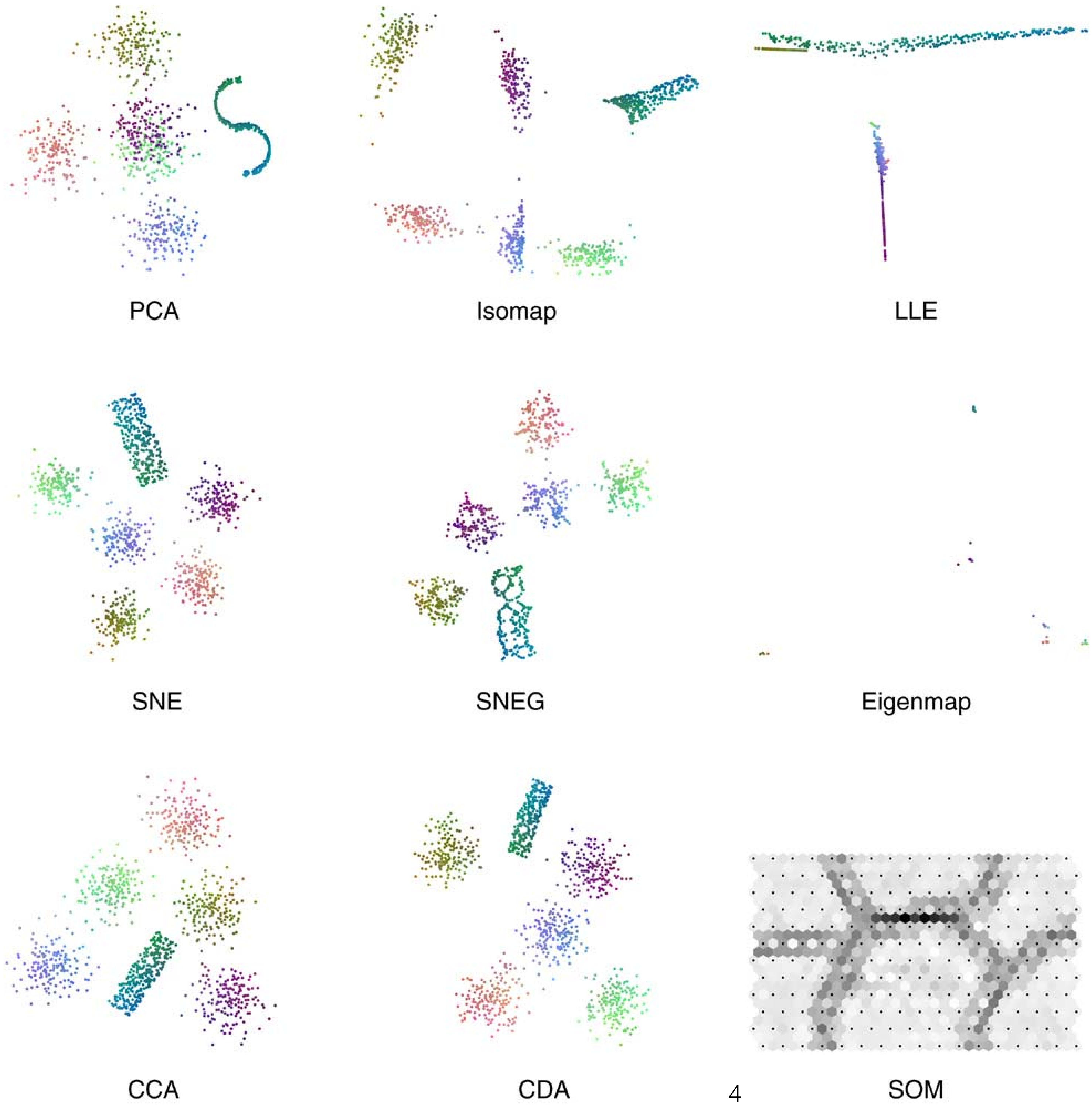
1 April 2019

Go to http://www.iki.fi/kaip/p/dr2.nb.html

Reminder: Give feedback, fill Webropol questionnaire before 24 April

Literature on dimensionality reduction for visualisation

- MDS: Borg, Kroenen, Modern multidimensional scaling: theory and applications. Springer, 1997.
- PCA: any book on matrix algebra.
- Jarkko Venna 2007, Academic Dissertation, http://lib.tkk.fi/Diss/2007/isbn9789512287529/
- Lee & Verleysen, 2007. Nonlinear dimensionality reduction. Springer.
- For a reasonably recent brief review see Verleysen & Lee, 2013 (recommended reading before exam!). https://doi.org/10.1007/978-3-642-42054-2_77
- See the references in the slides! Notice that most doi.org links can be accessed from within Aalto network (but usually not from home).
- (Not to be confused with dimensionality reduction for machine learning where target dimensionality is often higher!)
- Go to http://www.iki.fi/kaip/p/dr2.nb.html



Venna et al. 2006, https://doi.org/10.1016/j.neunet.2006.05.014

Recap

- PCA and MDS variants will struggle with non-linear manifolds
- PCA/Torgerson scaling is a linear projection
- techniques specifically designed to flatten manifolds
    - ISOMAP
    - LLE
    - Laplacian eigenmap
    - local multidimensional scaling
    - many more exist...
- large distances dominate the cost function in MDS methods
- either redefine the distance or look only at the vicinity of individual points
- practical issues: distortions, may be computationally expensive

Problem with lack of guidance

- The previous methods have one major problem: they produce an embedding given some (technical) criteria. The result may or may not be what user wants.

- One way to tune the embedding is to add **guidance**: find embedding such that it maximises dependency with respect to some particular variable(s)

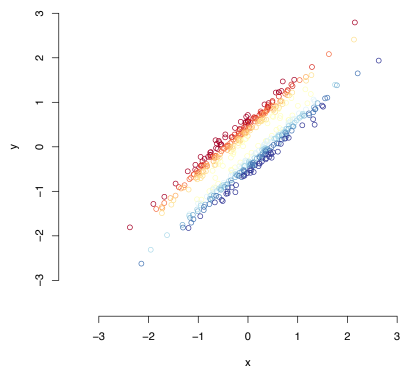
- Assume that in the original (high-dimensional) data consists of pairs of variables (x,y), where x is data variable and y is response variable (e.g., class).
- **Problem:** Find embedding X such that y depends mainly on X.

Supervised PCA

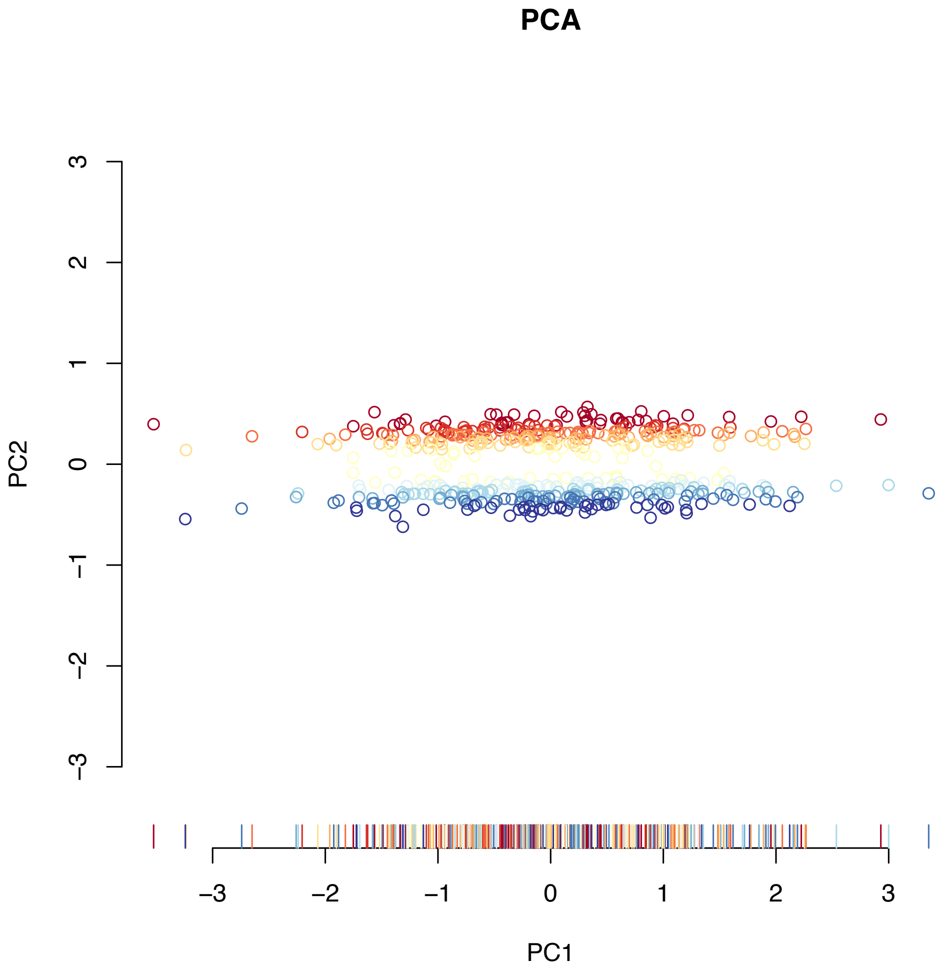
- At simplest,
  let X be $n \times m$ data matrix (with zero mean columns) and Y be $n \times m'$ matrix of response variables.
  - Use largest eigenvectors of $X^T Y Y^T X$ to project into lower dimensions
  - If $YY^T = \mathbf{1}$ this reduces to PCA

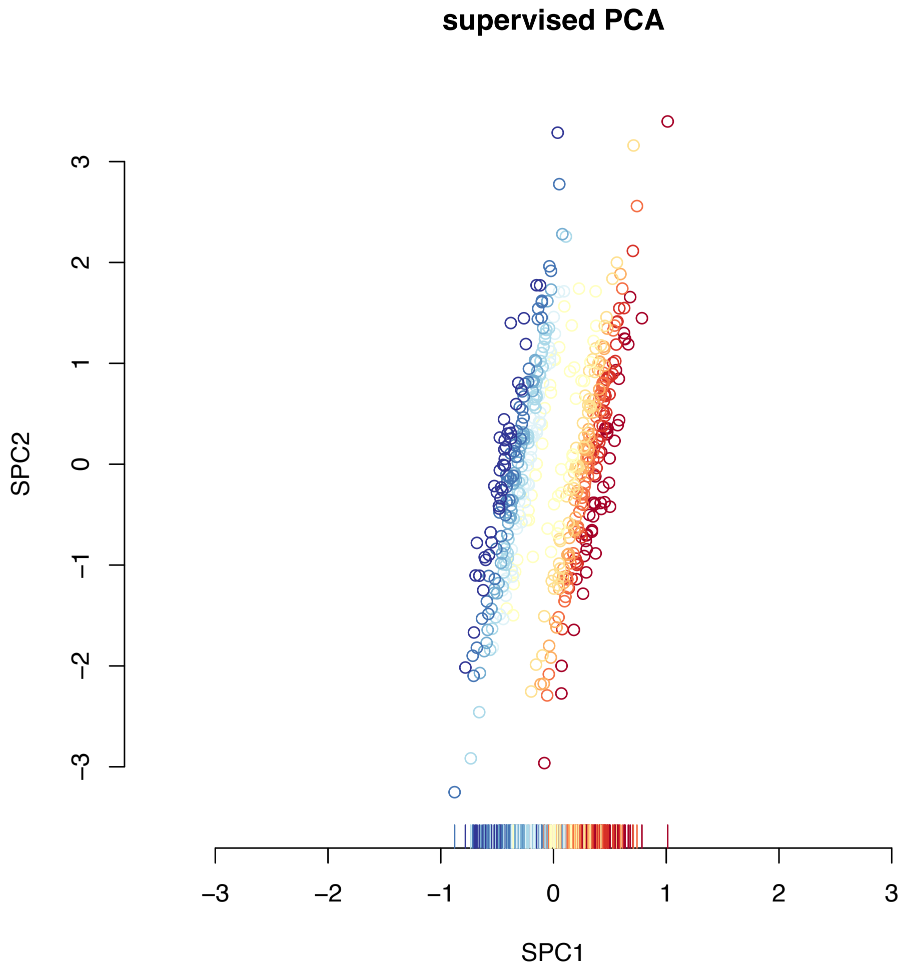
- For details and fancier variants see Barshan et al. 2011, https://doi.org/10.1016/j.patcog.2010.12.015

Supervised PCA



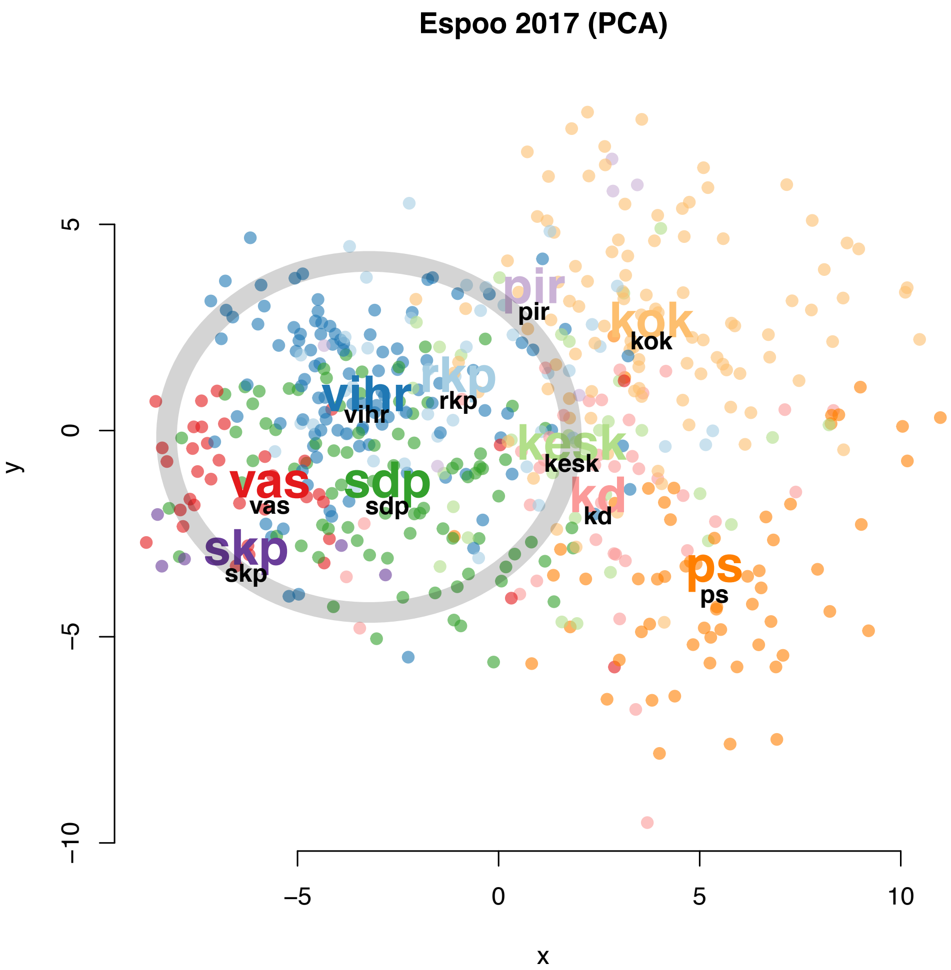
- Same 2-cluster data as before
- Y is $n \times 1$ matrix and Y_{i1}=-1 or 1 if i is in red or blue cluster, respectively
  (i.e. Y gives a classification of the data)

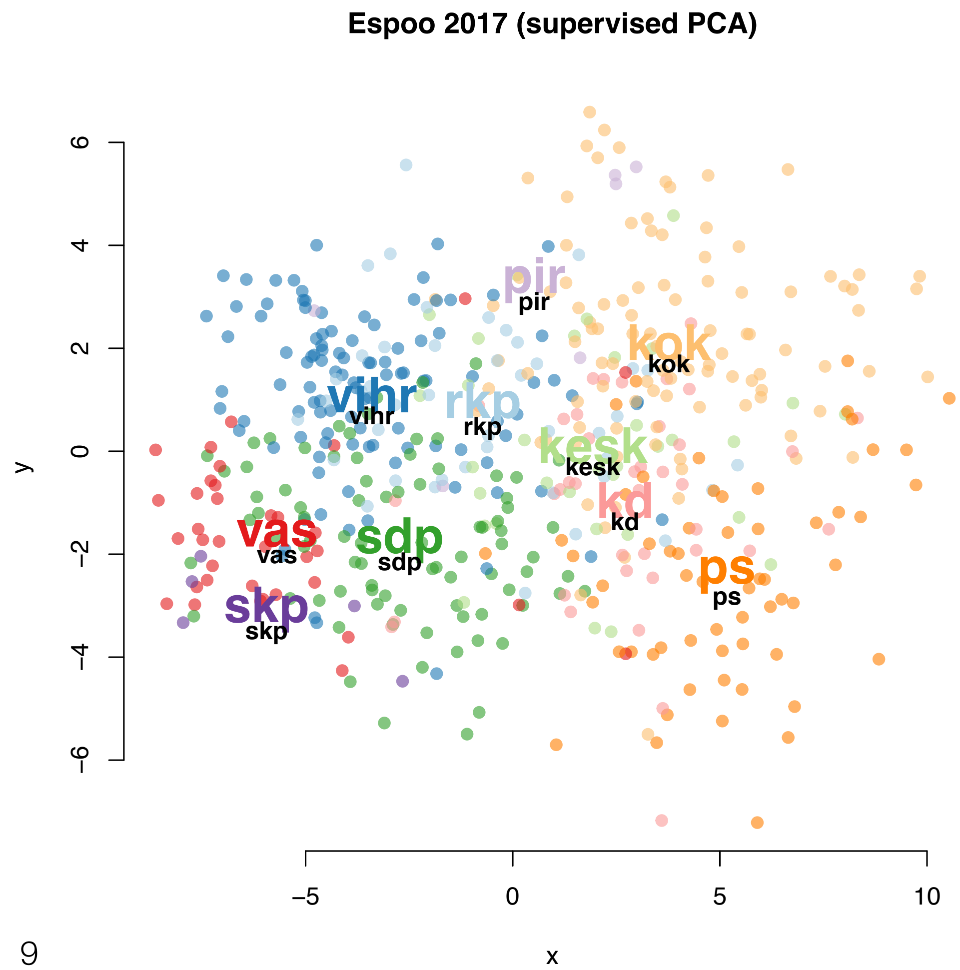




Supervised PCA

- Supervise PCA to **separate the following parties**: *vihr, rkp, sdp, vas*
- Y is 515 x (4+515) matrix where
  $Y_{i1} = 1$ if candidate i is in *virh*, $Y_{i2} = 1$ if candidate i is in *rkp*,
  $Y_{i3} = 1$ if candidate i is in *sdp*, $Y_{i4} = 1$ if candidate i is in *vas*, otherwise $Y_{ij} = 0$ *for* $j<5$.
- In addition, we set $Y_{i(j+4)}=0.01$ if i=j and 0 otherwise (this is to guide PCA to find some structure even within points in the same class)





Guided locally linear embedding

- It is possible to guide also other methods such as locally linear embedding (LLE)
- The principles in **guided** LLE (GLLE) are similar as for supervised PCA
- For details see Alipanahi et al. 2011, https://doi.org/10.1016/j.patrec.2011.02.002

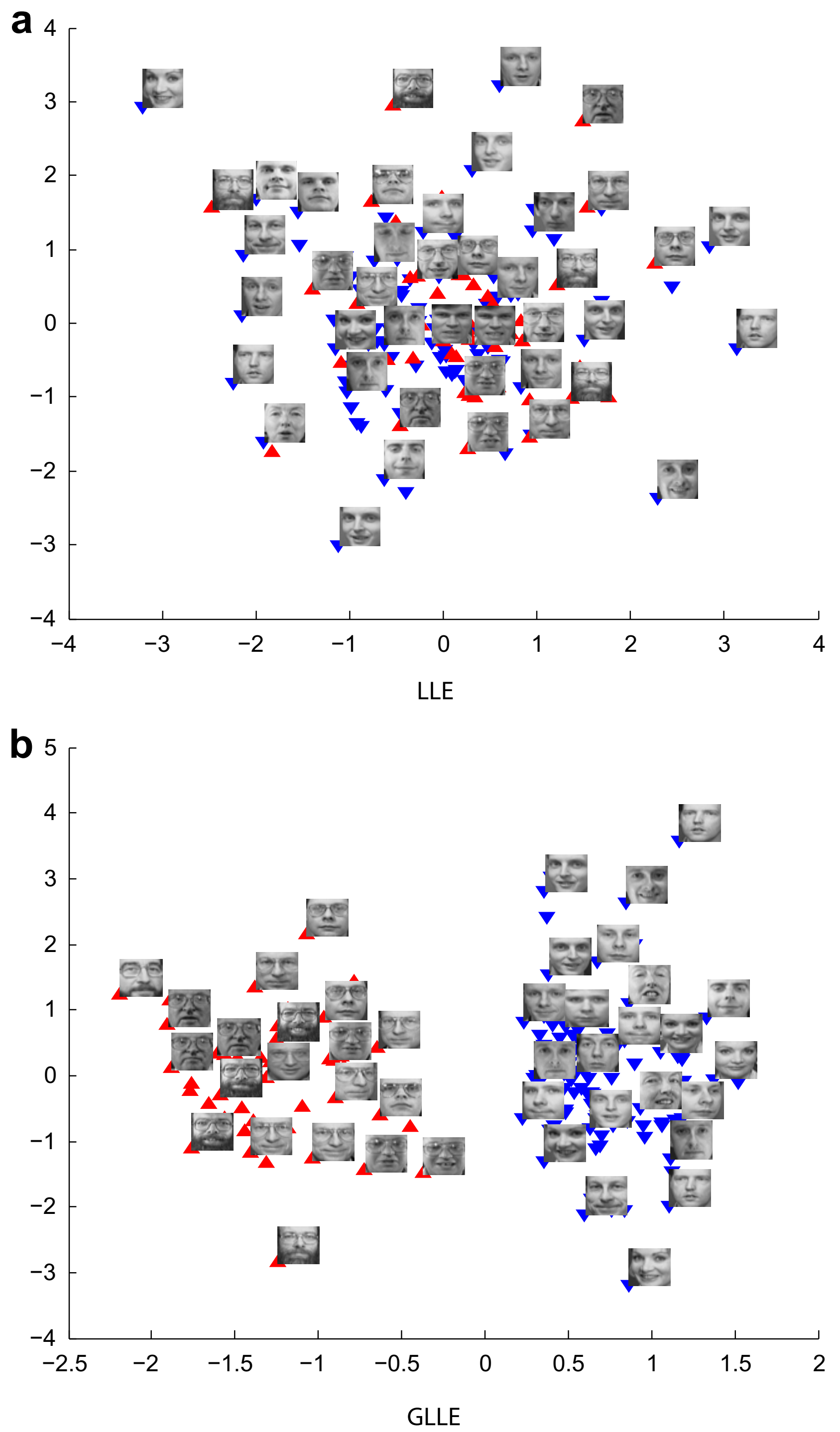


Fig. 2. Comparison of visualizations acquired by LLE and 0.5-GLLE ($k = 50$). There are two groups: persons with and without glasses.

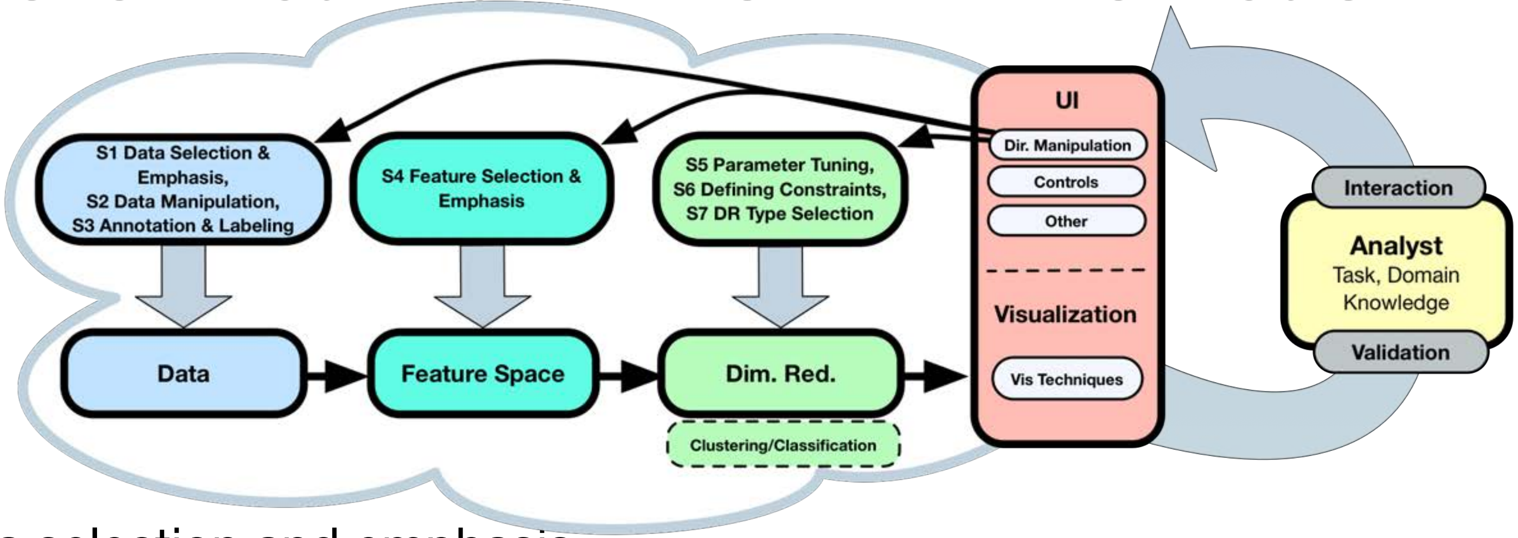
Problem with lack of interaction

- "Controllability and interaction are two concepts that are mostly absent from dimensionality reduction." (Verleysen et al. 2013)
- First papers on interactive DR in 2006 (Sacha et al. 2017)

- The previous methods have one major problem: they produce an embedding given some technical criteria. The result may or may not be what user wants.

- ***New problem:*** How to create efficient interaction such that the user can in an understandable way modify the embedding?
  - (E.g., by noticing cluster structures or outliers and asking to show something different, by must-link or cannot-link constraints etc.)

Scenarios for interaction



- S1 Data selection and emphasis
  - Filter applied to data and DR rerun on the remaining subset
- S2 Annotation and labelling
  - Enrich data with labels etc. and use the annotations to define distance measure
- S3 Data manipulation
  - Analyst manipulate data directly
- S4 Feature selection and emphasis
  - Analysts, e.g., can weight the importance of features
- S5 DR parameter tuning
  - Tune parameters (such as k in k-nearest neighbour)
- S6 Define constraints
  - Analyst directly arranges points in visualisation
- S7 DR type selection
  - Vary DR algorithm
- From Sacha et al. 2017, https://doi.org/10.1109/TVCG.2016.2598495

Interactive knowledge-based kernel PCA

Paurat et al. 2013

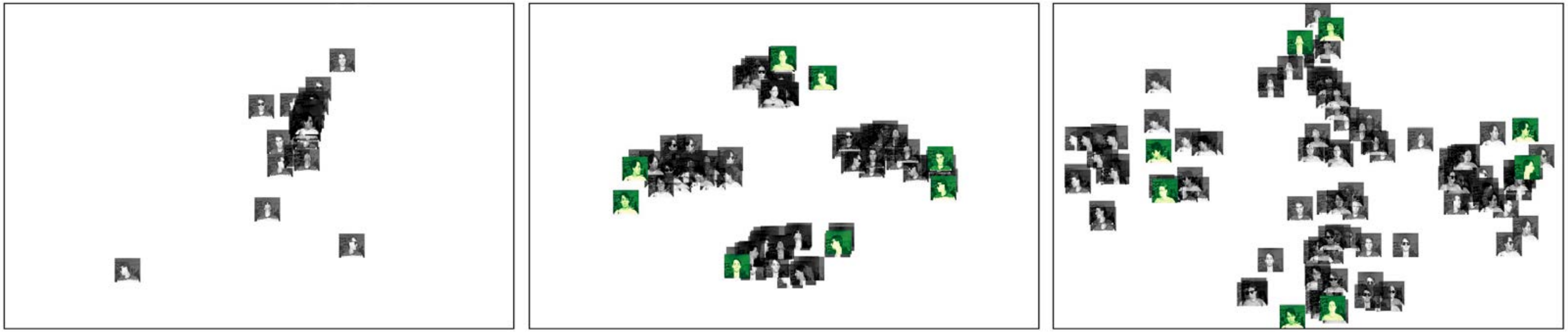


Fig. 3. A dataset of facial images embedded in different ways. The left figure shows a plain PCA embedding, while the other two figures use LSP to group the control points by person and by pose (*looking-straight*, *-up*, *-left* and *-right*), respectively.

- Variant of kernel PCA where user can add, e.g., must-link constraints to modify the embedding in a computationally efficient way (so that it is usable in interactive systems!)
- Paurat et al. 2013, https://doi.org/10.1007/978-3-642-40994-3_52
  Oglic et al. 2014, https://doi.org/10.1007/978-3-662-44851-9_32

Interactive exploration of coctails

Table 1: Exemplary results of the ten highest quality patterns, delivered by different pattern-mining approaches on the cocktail dataset. Note that here the top-10 frequent item sets are also all closed. The high-lift patterns were sampled according to their *rarity* measure [6]. In case of subgroup discovery, the label indicates whether a cocktail is creamy or not.

Unsupervised pattern-mining methods		Supervised pattern-mining methods	
Frequent (closed) item sets	**Sampled patterns with high lift**	**closed subgroups**	**Δ_1-relevant subgroups**
Vodka	Vodka **&** Cranberry juice	Baileys	Baileys
Orange juice	Vodka **&** Triple sec	Crème de cacao	Crème de cacao
Amaretto	Baileys **&** Kahlúa	Milk	Milk
Pineapple juice	Vodka **&** Gin	Kahlúa	Kahlúa
Grenadine	Vodka **&** Blue curaçao	Baileys **&** Kahlúa	Cream
Gin	Pineapple juice **&** Malibu rum	Cream	Irish cream
Baileys	Vodka **&** Amaretto	Irish cream	Crème de banana
Tequila	Vodka **&** Rum	Vodka **&** Baileys	Butterscotch schnapps
Kahlúa	Orange juice **&** Amaretto	Crème de banana	Whipped cream
Triple sec	Vodka **&** Tequila	Baileys **&** Butterscotch schnapps	Vodka **&** Kahlúa

- Mine patterns, represent them with high-dimensional vectors, and then reduce dimensionality to 2
- Patterns = frequently occurring combinations of ingredients of coctails
- Clusters of patterns represents ~classes of coctails

Paurat et al. IDEA 2014.
https://core.ac.uk/download/pdf/34655536.pdf#page=98

Interactive exploration of coctails

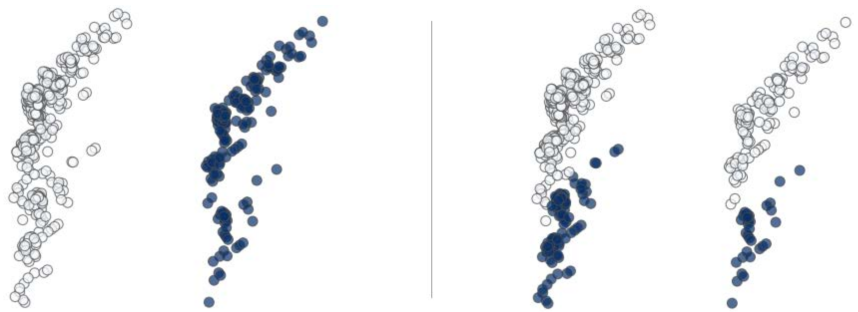


Figure 1: The 1000 most-frequent item sets of the cocktail dataset, embedded onto their first two principal components, labeled by the presence of *Vodka* (left) and *Orange juice* (right).

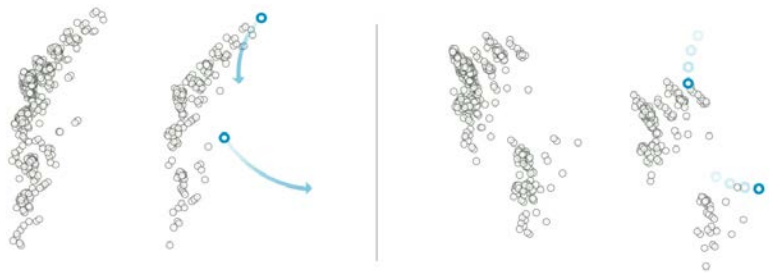


Figure 2: Dragging two control points (emphasized in blue) to new locations, reveals a structure that was previously hidden in the PCA embedding. The four clusters indicate the presence or absence of the two ingredients *Vodka* and *Orange juice*.

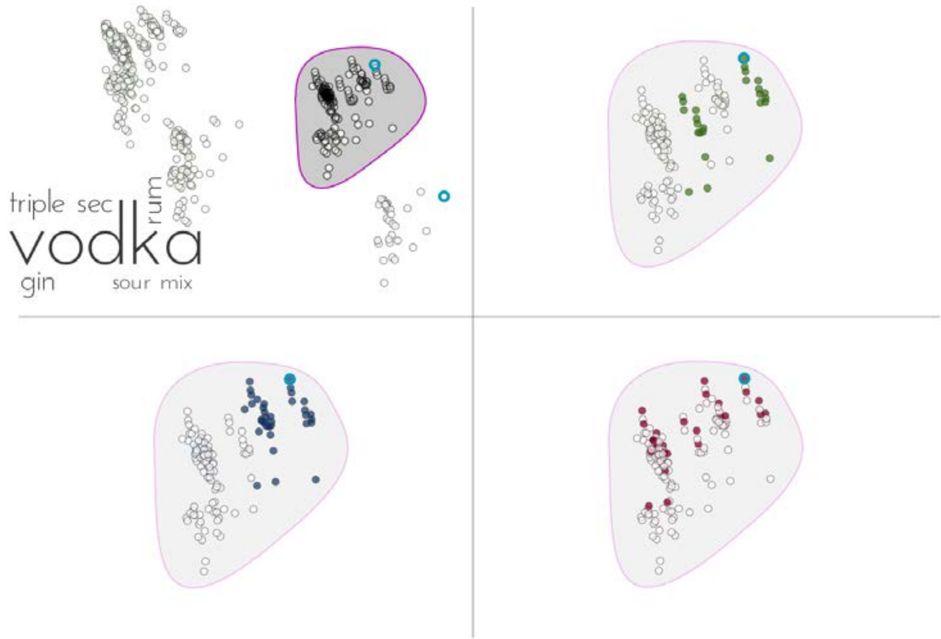
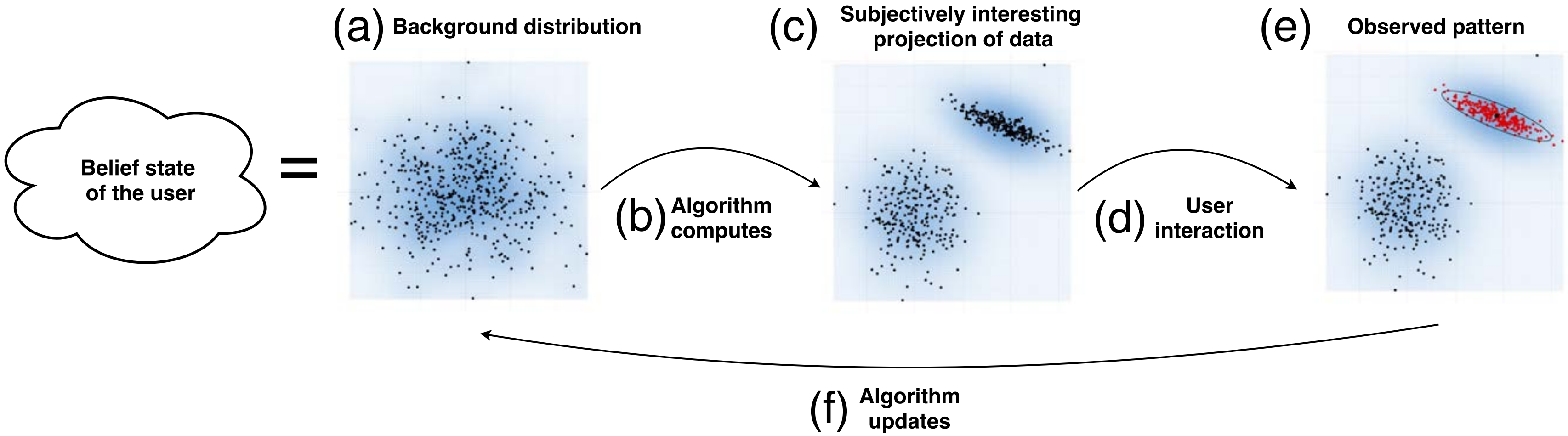


Figure 3: A closer look at the top-right cluster of Figure 2 reveals the ingredients that the patterns from the "*Vodka* and no *Orange juice* cluster" are frequently mixed with (top-left). The other three pictures indicate the presence of *Rum* (highlighted in green), *Gin* (blue), and *Triple sec* (red).

Paurat et al. IDEA 2014.
https://core.ac.uk/download/pdf/34655536.pdf#page=98

Tell the me something I don't know

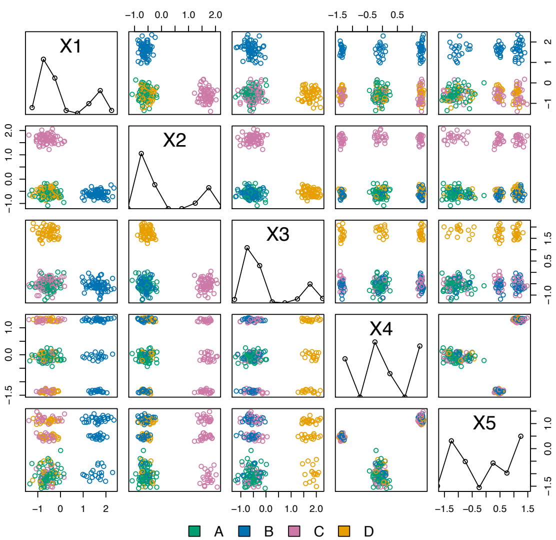


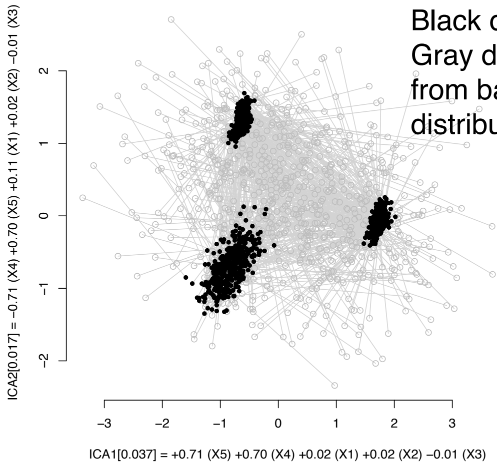
- We model user's knowledge of the data (*background model*)
- We show the user the view in which the data and the background model differs most
- Each time the user observes something marks it (e.g., cluster, outlier) the background model is updated accordingly
- Uses dimensionality reduction to produce views (tuned to show maximal difference between data distributions)
- **Visually controllable data mining.** Extension of Furnas' effective view navigation to the context of having automated analysis. Puolamäki et al. 2010, https://doi.org/10.1109/ICDMW.2010.141
- Demo (implemented by R Shiny) http://www.iki.fi/kaip/sider.html
- Puolamäki et al. 2017, https://arxiv.org/abs/1710.08167

- Data = **real vectors**
- Background distribution **(a)** = **Maximum Entropy distribution satisfying constraints** (initially: no constraints, unit Gaussian spherical distribution with zero mean)

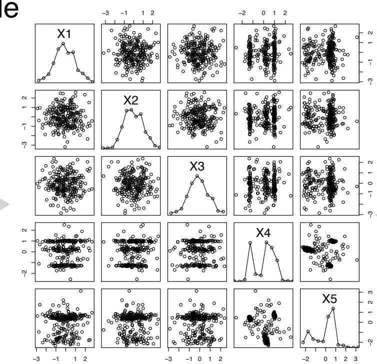
- Direction-preserving **whitening** transformation of the data results in a unit Gaussian spherical distribution, if the data follows the current background distribution
- **PCA/ICA** used to find non-Gaussian directions: **subjectively interesting projection** of data **(b,c)**

- User observes **patterns** and adds respective **constraints (d,e)**
- Background distribution is updated **(f);** the process is **iterative**
- Various constraints based on simple **linear and quadratic constraints**

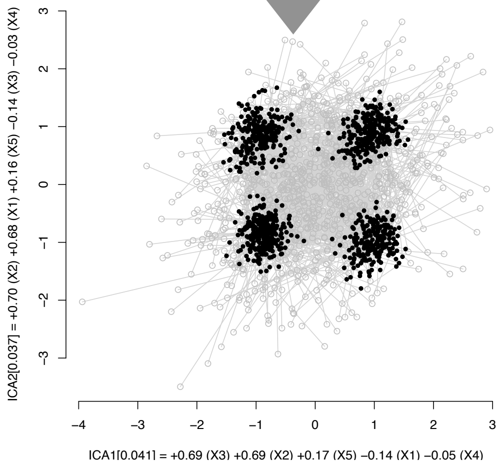


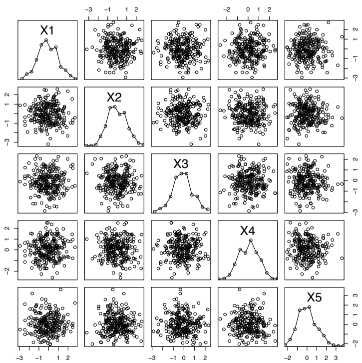


Black dots = data
Gray dots = sample from background distribution



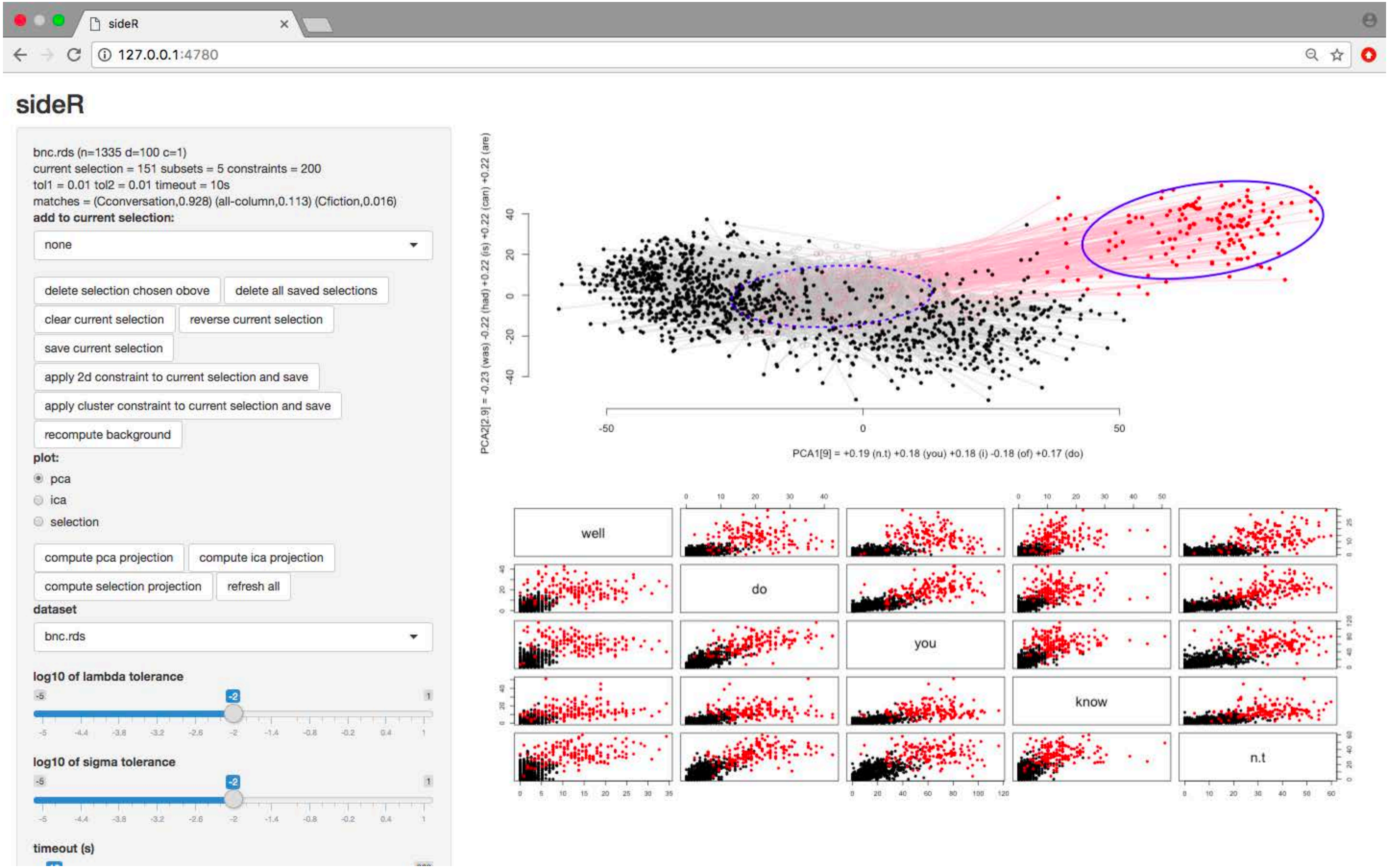
Whitened data after adding cluster constraints for the visible clusters





For details, see: Kai Puolamäki, Emilia Oikarinen, Bo Kang, Jefrey Lijffijt, Tijl De Bie. Interactive Visual Data Exploration with Subjective Feedback: Information-Theoretic Approach, arXiv: 1710.08167, 2017

Download for the SIDER tool



http://www.iki.fi/kaip/sider.html

Student presentations

- **Global warming** (A1E3):
  - Eetu Rantanen
  - Savolainen Eerika

- **Eurostat** (A2E2):
  - Yuan Zheng: *green house gas (GHG) emission*
  - Kévin Selänne: *Timeseries (2000–2014) of patents*
  - Tuomo Kivekäs: *Broadband internet penetration*