Reordering and selecting continuous variables

Katrin Grimm

Department of Computer-oriented Statistics and Data Analysis
University of Augsburg

July 02, 2015
Handling big, unknown data sets with a huge number of numerical variables is challenging.

Overviewing the correlation structure or finding more general two-dimensional anomalies can be quite difficult, **but**

- two-dimensional structures are a good starting point for discovering interactions
- interactions between two variables are visualizable and can be easily interpreted in scatterplots

⇒ It’s useful to be able to handle them

The talk tries to help with two concepts:

- Present a convenient graphic to overview the whole correlation structure
- Select a scatterplot matrix to allow a first deeper insight in the real data
German election

- 299 cases (Germans electoral districts)
- 70 variables
  - vote shares of the parties in 2009 and 2005
  - demographic and economic information about the districts (e.g. unemployment rate, population density, birth rate)

68 numeric variables $\Rightarrow$ 2278 possible two-dimensional plots

How can we overview the correlation structure visually?
Overview of all bivariate structures: A scatterplot matrix
A corrgram\(^1\) with all numerical variables in random order

- Visualizes the correlation matrix
- Direction of correlation is visualized by the color (blue for positive correlation, red for negative)
- Strength is visualized by the color intensity

More to see, but still unclear $\implies$ we need a new order

\(^1\)Michael Friendly (2002), implemented in the package corrgram
Reordering based on angles of principal components

Based on eigenvectors $v_1$, $v_2$ of the correlation matrix we consider for each variable the loading on the first two principal components $(v_{i1}, v_{i2})^T$. The angles to the $x$-axis of these vectors can be used as a measure of similarity:

$$\alpha_i = \begin{cases} \arctan \left( \frac{v_{i2}}{v_{i1}} \right) & \text{if } v_{i1} > 0, \\ \arctan \left( \frac{v_{i2}}{v_{i1}} \right) + \pi & \text{else} \end{cases}$$
For a comparison: Optimal leaf ordering (OLO)

Approach:
- Hierarchical clustering (average linking) of the variables based on correlation matrix
- Reordering of the leaves to maximize the sum of the correlations of adjacent variables

from: Fast optimal leaf ordering for hierarchical clustering (Ziv Bar-Joseph et al)

A corrgram with variables in OLO based order

- Seems even more ordered (over the PCA based ordering)
- Can also help to make a choice for the number of clusters (alternative to dendrograms)
Selecting $q$ variables to show in a scatterplot matrix

**Goal:** Show $q$ variables with maximal correlation in terms of:

$$\max g(i_1, i_2, \ldots, i_q) = \sum_{j=1}^{q} \sum_{k=j+1}^{q-1} \text{Cor}(X_{ij}, X_{ik})$$

where $i_1 < i_2 < \ldots < i_q$ and $i_j \in \{1, \ldots, p\} \forall i_j$ (1)

Easiest approach: Checking all possible combinations, which is computationally intensive.

- There are more than 10 million possibilities to select 5 variables from the *German election* dataset and
- more than 500 million to select 8 variables.

How can the reordering help?
Approach to find $q$ variables with maximal correlation

Based on the new order of the variables $X_1, \ldots, X_p$ both goals are (approximatively) reachable with the following steps:

1. Choose $r \geq q$

2. For all $i \in \{1, \ldots, p\}$ calculate:

   $sumCor_i = \max g(i_1, i_2, \ldots, i_q)$

   where $i_1 < i_2 < \ldots < i_q$

   and $i_j \in \{i, (i + 1) \mod p, \ldots, (i + r - 1) \mod p\}$

   \forall \ i_j \text{ with } (p \mod p) := p

3. Find $\max_i sumCor_i$
Illustration of the approach

\[ r = 10, \ q = 5 \]
Consider randomly chosen samples (30 variables) from dataset *German election* (500 times)

Compare the results from the PCA based and from the optimal leaf reordering with the real optimum

Vary $r$

Of interest is for different numbers of $r$:

- How many variables from real optimal matrix are found?
- Which rank has the approximative matrix under all possible in terms of the maximum correlation?
$r = 5$ (check only adjacent variables)

Checking 30 combinations instead of all $\binom{30}{5} = 142506$
Checking $30 \binom{10}{5} = 7560$ combinations instead of $\binom{30}{5} = 142506$
$r = 15$

Checking $30 \binom{15}{5} = 90090$ combinations instead of $\binom{30}{5} = 142506$
What’s the result with the whole dataset?

- $r = 10$: PCA based reordering finds 7 from the 8 variables of the optimal matrix, OLO already delivers the optimal matrix
- $r = 15$: Both methods find the optimal matrix
Possible extensions to select scatterplot matrices

1. Ignore really high correlations (in a second step)
2. Substitute the correlation with a more general dependency measure
3. Based on the *scagnostics* idea, that scatterplots are describable through a few measures, a more complex approach is conceivable:
   - Define a *relevance measure* based on scagnostics measures
   - Use the same approach to find the matrix with “highest relevance”

---

3 Some of the measures from package *scagnostics* are used for the example
Conclusions

- Reordering in connection with corrgrams satisfies the goal to offer a good first overview of the correlation structure.
- The idea of studying subgroups can bring good insights in the data.
- Especially the *scagnostics* approach can offer a general and interesting extension of that idea.
References


