Big Data in Psychology:
Statistical methods for linked high-dimensional with traditional data
Katrijn Van Deun, Tilburg University
LET’S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
Outline

• Background: Psychological research in a time of Big Data
• Challenges for data analysis
• Part 1: Exploration
• Part 2: Prediction
• Illustrations
  • Exploration: 500 families
  • Prediction: Systems vaccinology
• Playtime: RegularizedSCA
Background:
Psychological research in a time of Big Data
Big Data in the (Behavioral) Sciences

• Everything is measured
  • What we think and do: Social media, Web browsing behavior
  • Where we are and with whom: GPS tracking, cameras
  • At a very detailed level: Experience sampling, neuron, DNA

• Data are shared
  • Open data: in science, governments (open government data)

• Data are linked
  • Government
  • Science: multi-disciplinary
  • Linked Data web-architecture
• Traditional data: Health & Retirement Study
• Health & Retirement Study: traditional + novel type of data
  
  • Multiple sources, **heterogeneous** in nature
  • **High-dimensional** ($p \gg n$)

Well being

Big Data

Questionnaire data

Novel kind of data: GPS, DNA, ...
• Illustration II: ALSPAC household panel data

• Multiple sources / multi-block (heterogeneous)

• High-dimensional

• (no outcome)
• Illustration III: ADNI data

• Multi-block data

• (Outcome)
The Promises of linked Traditional with novel data

⇒ Extremely information-rich data

• 1) Gives insight in the interplay between multiple factors / system or multi-disciplinary point of view (Exploratory; hypothesis generating)

  Eg. gene-environment interactions: Link susceptibility genes with protective/risk-provoking environmental conditions

• 2) Adds context, detail => deeper understanding + more accurate prediction

  Eg. Similar income, social network, health but difference in well-being?

However, statistical tools fall short ... how to find the linked variables?
Non-theory driven nature of Big Data

• Novel / Big data:
  
  • NOD: Naturally Occurring Data
  
  • High-throughput screening (eg, GWAS)

  => untargeted

  Furthermore, little theory available on these novel sorts of data.

• Relevant variables? Introduces a variable selection problem.
The hidden link

• Traditional and novel data are very different:
  • Different sizes of data blocks, different noise/measurement levels

  • Stronger correlations of variables within blocks than between blocks
    • Traditional: Eg, response tendencies, general psychological processes
    • Novel: Eg, (general) biological processes

=> Link variables / shared mechanisms are hidden while block-specific mechanisms dominate
The hidden link

• Popular but bad ideas for analysing linked (traditional with) big data:
  
  • 1) glueing the data together and analysing them as one block
  
  • 2) analysing each block separately and comparing the results
Furthermore ...
This may take some time; go for a walk in the mountains

WARNING: Ill-conditioned matrix

Best Predictor: Nose length

$R^2(\text{new case}) = 0$
In sum: The challenges

• Automated selection of variables needed, in the (ultra-) high-dimensional setting (=many more variables than observations)

• Integrative approach: find shared mechanisms, even if hidden / dominated by source-specific sources of variation

• Computational efficiency (time + memory)

• Avoid capitalization on chance
PART I:
Sparse Common Components
Revealing the linked variables

Questionnaire data

Genetic data

Gene-environment interaction

Respondents

1 1000... 500 501... 50 000
Sparse CoCo: Basic principles

• Detection of relevant variables

• Find shared / linked mechanisms throughout the data blocks

=> Selection of linked variables (between blocks)
Solving the challenges

• Exploratory approach needed

• Find structural sources of variation / shared mechanisms

⇒ Usual suspect: Principal component analysis (Exploratory Factor Analysis)

⇒ Sparse CoCo is an extension of PCA
The first suggestion I would have for the authors would therefore be to present what they are doing in terms that will be more recognizable to readers. In the discussion they imply that their method could be applied to factor analysis as well as PCA. If this is truly the case, then I'd suggest that they rework their presentation and present this as an extension of factor analysis instead of PCA. This would dramatically increase reader interest in the paper.

An anonymous reviewer
are equivalent in (14.81) for any $q \times q$ orthogonal $\mathbf{R}$. This leaves a certain subjectivity in the use of factor analysis, since the user can search for rotated versions of the factors that are more easily interpretable. This aspect has left many analysts skeptical of factor analysis, and may account for its lack of popularity in contemporary statistics. Although we will not go into

*The Elements of Statistical Learning* (2009)
• Notation and naming conventions

- **data block**: denotes the different data sources forming the *multiblock* data

- **$X_k$**: data block $k$ (with $k=1,...,K$); the outcome(s) - if present - is denoted by $Y$ ($y$ if univariate)

- Each of the data blocks: same set of observation units (respondents)
The many faces of PCA

• 1. Linear combination with maximal variance
   \[ \max_w \mathbf{w}^\prime \mathbf{X} \mathbf{X} \mathbf{w} \quad \text{with} \quad \mathbf{w}^\prime \mathbf{w} = 1 \quad \text{and} \quad \mathbf{w}_r \mathbf{w}_q = 0 \quad \text{for} \quad r \neq q \]

• 2. Data reconstruction/dimensionality reduction
   \[ \min_{\mathbf{T}, \mathbf{P}} \left\| \mathbf{X} - \mathbf{T} \mathbf{P}^\prime \right\|^2 \quad \text{s.t.} \quad \mathbf{T}^\prime \mathbf{T} = \mathbf{I} \]

• 3. Data reconstruction/dimensionality reduction
   \[ \min_{\mathbf{W}, \mathbf{P}} \left\| \mathbf{X} - \mathbf{X} \mathbf{W} \mathbf{P}^\prime \right\|^2 \quad \text{s.t.} \quad \mathbf{P}^\prime \mathbf{P} = \mathbf{I} \]

• 4. And the SVD:
   \[ \mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^\prime + \mathbf{E} \quad \text{with} \quad \mathbf{U}^\prime \mathbf{U} = \mathbf{I}, \ \mathbf{V}^\prime \mathbf{V} = \mathbf{I}, \ \text{and} \ \mathbf{S} \ \text{diagonal} \]

Do the distinctions matter? Aren’t they all the same/equivalent?
Focus on loadings $\mathbf{P}$: Indicators reflect the concept

**Concept = construct**

**Measure**: scale

**Indicators**
- Name the last five presidents of the USA
- Repeat after me: 7 5 3 2 1 6
- What is the advantage of keeping money in a bank?

**Scale**: measures a concept that is latent, non-observable
Indicators reflect the “effects” of the concept
Focus on weights $W$: Indicators form the concept

**Concept**

Measure: index

Indicators: SNPs

Risk depression

Polygenic risk score

Index: measures a concept that depends on **observable** characteristics

In this case indicators “make up” the concept; reification
Here: focus on data reconstruction, considering both the weight (formative) and loading (reflective) PCA model
• **Principal component models**

• Weight based variant:

\[ X_k = X_k W_k P_k^T + E_k \quad \text{s.t. } W_k^T W_k = I, \quad (1) \]

\[ = T_k P_k^T + E_k \]

with \( W_k \) (\( J_k \times R \)) the component weights, \( T_k \) (\( I \times R \)) the component scores, and \( P_k \) (\( J_k \times R \)) the component loadings

Interpretation of component \( t_{rk} \) based on \( J_k \) (!) regression weights:

\[ t_{irk} = \sum_j w_{jrk} x_{ijk} \]

\((T \sim X_k \rightarrow \text{high-dimensional regression})\)
• Principal component models

• Loading based variant

\[ X_k = T_k P_k^T + E_k \quad \text{s.t.} \quad T_k^T T_k = I, \]  \hspace{1cm} (2)

Interpretation of component \( t_{rk} \) based on \( J_k \) (!) correlations:

\[ r(x_{jk}, t_{rk}) = \rho_{jrk} \quad \text{(} X_{jk} \sim T_{rk} \text{ --> low-dim. regression, indep. pred.)} \]

• Note: In a least squares approach subject to \( P_k^T P_k = I \), we have \( W_k = P_k \)
Simultaneous component analysis

For all $k$:
$$X_k = TP_k^T + E_k \text{ s.t. } T^TT = I \quad (3)$$

$\rightarrow$ same component scores for all data blocks!

1. Weight based variant:
$$[X_1 \ldots X_K] = T \begin{bmatrix} P_1^T & \ldots & P_K^T \end{bmatrix} + [E_1 \ldots E_K^T]$$
$$= [X_1 \ldots X_K] \begin{bmatrix} W_1^T & \ldots & W_K^T \end{bmatrix} \begin{bmatrix} P_1^T & \ldots & P_K^T \end{bmatrix} + [E_1 \ldots E_K^T]$$

2. Loading based variant:
$$X_{\text{conc}} = T \ P_{\text{conc}}^T + E_{\text{conc}} \quad (4)$$
• Simultaneous components

  • Model the largest source of variation in the concatenated data, often this is source-specific variation or a mix of common and specific variation

  • Guarantee same source of variation for the different data blocks

  • But ... do not guarantee that the components are common!

  • See also Schouteden et al. (BRM, 2013): Rotation to common & specific components
• **Common component analysis (CoCo)**

  – Account for dominating source specific variation

  – Common component model:

    \[ X_{\text{conc}} = TP_{\text{conc}}' \] such that \( T'T=I \) and \( P_{\text{conc}} \) has *common / specific structure*

  \[
  \begin{array}{c|ccc|ccc}
  & P1 & & & P2 & & \\
  \hline
  \text{Common} & x & x & x & x & x & x \\
  \text{Spec1} & x & x & x & 0 & 0 & 0 \\
  \text{Spec1} & 0 & 0 & 0 & x & x & x \\
  \end{array}
  \]

  – Structure can be imposed (*constrained analysis, if \( K \) small / structure known*)

  – Or: \( X_k = X_{\text{conc}}W_{\text{conc}}P_k' \) with the \( W_{\text{conc}} \) having *common / specific structure*
• So far, so good ...

• Yet:
  
  • Interest in relevant / important variables (social factors, genetic markers)?
  
  • Interpretation of the components based on 1000s of loadings/weights is infeasible
  
  • PCA estimates are not consistent in case of high-dimensional data (Johnstone & Lu)

⇒Need to **automatically** select the “relevant” variables!
⇒**Sparse common components** are needed.
• Structured **sparsity**:  

  • Sparse common components: few non-zero loadings in each data block $\mathbf{X}_k$ 

  • (Sparse) specific components: (few) non-zero loadings only in one/few data blocks $\mathbf{X}_k$ 

```

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th></th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common</strong></td>
<td>x</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Dist1</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>Dist1</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Variable selection: How to?

• Sparse analysis
  – Impose restriction on the loadings/weights: many should become zero
  – S-O-A: add penalties (e.g., lasso) to the objective function
Sparse SCA: Objective function

Add penalty known to have variable selection properties to SCA objective function:

Minimize over $T$ and $P_{\text{Conc}}$ and such that $T'T = I$ and $P_{\text{Conc}}$ constrained

$$\|X_{\text{Conc}} - TP'_{\text{Conc}}\|^2 + \sum_{r,k} \lambda_{r,k} |p_{r,k}|_1$$

Fit / SCA

Penalty

with $|p_{r,k}|_1 = \sum_j |p_{jkr}|$ the $L_1$ penalty or lasso tuned by $\lambda_{r,k} \geq 0$

$\to$ shrinks and selects variables
$\to$ penalty can also be applied to the weights:

$$\|X_{\text{Conc}} - X_{\text{Conc}} W_{\text{Conc}} P'_{\text{Conc}}\|^2 + \sum_{r,k} \lambda_{r,k} |w_{r,k}|_1$$
Why not rotation to simple structure?

- **Must read**: Cadima & Jolliffe (1995)

- Ordinary simultaneous component analysis, common interpretation practice (VARIMAX + thresholding)
  
  - %VAF: .31 (if calculated with neglected loadings set equal to zero)

- Sparse SCA, lasso in action
  
  - %VAF: .35 (> .31, VARIMAX!!!!)
• In-house Algorithm: Alternating procedure

Given fixed tuning parameters and number of common and distinctive components, do

0. Initialize $\mathbf{P}_{\text{conc}}$
1. Update $\mathbf{T}$ conditional upon $\mathbf{P}_{\text{conc}}$
   
   Closed form: $\mathbf{T} = \mathbf{U} \mathbf{V}'$ with $\mathbf{U}$ and $\mathbf{V}$ from the SVD of $\mathbf{X}_c \mathbf{P}$ ($l \times R \rightarrow$ small for H-D data)

2. Update $\mathbf{P}_{\text{conc}}$ conditional upon $\mathbf{T}$
   
   **Coordinate descent** (see next)

3. Check stop criteria (convergence of the loss, maximum number of iterations) and return to step 1 or terminate
Coordinate descent

\[ L = \|X_{\text{Conc}} - TP'_{\text{Conc}}\|^2 + \sum_{r,k} \lambda_{r,k} |p_{r,k}|_1 \]

\[ = k + \sum_{j,k,r} \left\{ \left( \sum_i (x_{ij} - t_{ir}p_{jk})^2 \right) + \lambda_{r,k} |p_{jk}| \right\} \]

For which the root can be found using subgradient techniques (this is a standard lasso regression problem)

Note that the problem is separable over \( j \) and that \( T \) is orthogonal

\[ => \text{all } p_{jk} \text{ can be updated simultaneously (fast!!!)} \]

\[ => \text{fixing of loadings / constrained analysis is } \square \]
• Hence, the following soft thresholding update of the loadings:

\[
p_{jr} = \begin{cases} 
\frac{\sum_i x_{ijk} t_{ir} - \lambda_r/2}{1 + d_{jr}} & \text{if } p_{jr} > 0 \\
\frac{\sum_i x_{ijk} t_{ir} + \lambda_r/2}{1 + d_{jr}} & \text{if } p_{jr} < 0 \\
0 & \text{else}
\end{cases}
\]

which can be calculated for all loadings of component \( r \) simultaneously using simple vector and matrix operations!!!

\( \Rightarrow \) **Highly efficient (time+memory), scalable to large data**
• Algorithm: Weight based variant (sparseness on weights)

• Similar type of algorithm can be constructed

• Expression to estimate $w_{jkr}^*$ using coordinate descent (cycle over all $R \sum_k J_k$ coefficients); sparse group lasso case

$$w_{jkr}^* = \begin{cases} 
\frac{\sum_{i,k,jk} x_{ij}^* r_{ijk} p_{jk}^r - \lambda_1 r/2}{1 + \lambda_2 r} & \text{if } w_{jkr}^* > 0 \\
\frac{\sum_{i,k,jk} x_{ij}^* r_{ijk} p_{jk}^r + \lambda_1 r/2}{1 + \lambda_2 r} & \text{if } w_{jkr}^* < 0 \\
0 & \text{else}
\end{cases}$$

Ridge penalty has to be included

The expression for an individual coefficient is not very expensive but has to be calculated many times. Also here adding elementwise constraints is
• Algorithm

• Coordinate-wise approach allows to fix coefficients

• This can be used to define the specific components by fixing to zero those coefficients corresponding to the block(s) that is not accounted for by the component

• Often no prior knowledge on number of specific components; for which block(s) they are specific => we also included a group lasso penalty (performs selection at the block level)

• Is run with input of the nr of components; their status (common/specific) and/or values for the (group) lasso tuning parameter => MODEL SELECTION
Some illustrations
500 Family Study (Schneider & Linda, 2008)

• Three data blocks: 195 families
  • Father -> 8 variables
  • Mother -> 8 variables
  • Child -> 7 variables
  Here: part of questionnaires

• Analyzed with the RegularizedSCA package (available from CRAN)
  • This is the sparse loading variant
  • Several options for model selection; includes the DISCO rotation

• Exploratory analysis
• Model Selection Step 1: Number of components?

• VAF method => 5 components

Proportion of VAF for each component of each block:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1902580</td>
<td>0.08505923</td>
<td>0.08672370</td>
<td>0.07153828</td>
<td>0.06447495</td>
<td>0.05526066</td>
<td>0.02852091</td>
<td>0.03936437</td>
<td>0.05352701</td>
<td>0.05906965</td>
</tr>
<tr>
<td>0.1366034</td>
<td>0.09455771</td>
<td>0.06367872</td>
<td>0.08605901</td>
<td>0.10354762</td>
<td>0.08630231</td>
<td>0.03933645</td>
<td>0.04906133</td>
<td>0.04728068</td>
<td>0.02230673</td>
</tr>
<tr>
<td>0.1940918</td>
<td>0.10392886</td>
<td>0.11339839</td>
<td>0.06370639</td>
<td>0.02089805</td>
<td>0.03280644</td>
<td>0.09052555</td>
<td>0.03910193</td>
<td>0.01721554</td>
<td>0.02715787</td>
</tr>
</tbody>
</table>
• Model selection Step 2: Status components and degree of sparseness?

• Cross-validation for both group lasso and lasso tuning parameters

Recommended tuning parameter values are:
Lasso Group Lasso
[1,] 2.82068   1.28369
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M: Relationship with partners</td>
<td>0.000000</td>
<td>12.063380</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Argue with partners</td>
<td>-5.656729</td>
<td>5.606819</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Childs bright future</td>
<td>-8.583745</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Activities with children</td>
<td>-4.462827</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-9.015597</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Feeling about parenting</td>
<td>-8.757033</td>
<td>2.750746</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Communation with children</td>
<td>-9.077420</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-3.305209</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Argue with children</td>
<td>-9.268780</td>
<td>0.000000</td>
<td>0.000000</td>
<td>3.193377</td>
<td>0.000000</td>
</tr>
<tr>
<td>M: Confidence about oneself</td>
<td>-6.680594</td>
<td>7.242076</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Relationship with partners</td>
<td>0.000000</td>
<td>11.903001</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Argue with partners</td>
<td>0.000000</td>
<td>5.104832</td>
<td>-9.236275</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Childs bright future</td>
<td>-3.171864</td>
<td>0.000000</td>
<td>-5.817937</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Activities with children</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-11.016351</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Feeling about parenting</td>
<td>-3.940123</td>
<td>0.000000</td>
<td>-6.861383</td>
<td>-2.153304</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Communation with children</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-8.373867</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Argue with children</td>
<td>-5.098655</td>
<td>0.000000</td>
<td>-9.855232</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>D: Confidence about oneself</td>
<td>0.000000</td>
<td>5.561253</td>
<td>-8.181819</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>Self confidence/esteem</td>
<td>-5.963305</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-8.467937</td>
<td>0.000000</td>
</tr>
<tr>
<td>Academic performance</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-7.091520</td>
</tr>
<tr>
<td>Social life and extracurricular</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-4.260376</td>
</tr>
<tr>
<td>Importance of friendship</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-9.715712</td>
</tr>
<tr>
<td>Self Image</td>
<td>-2.560587</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-10.153244</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-9.424702</td>
</tr>
<tr>
<td>Confidence about the future</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-7.516765</td>
</tr>
</tbody>
</table>

Common  Parents  Father  Parents  Child
PART 2:
Prediction in a high-dimensional context, the psychologist’s way
Illustration: An insightful linear model for high-dimensional prediction

• Motivating example: Systems vaccinology (Nakaya et al., 2011)
Sparse Principal Covariates Regression

• Extension of PCovR (de Jong & Kiers, 1992)
• Simultaneous sparse PCA and regression
• Stability selection (Meinshausen & Buhlmann, 2010)

\[
X = X W P_x T + E_x = T P_x T + E_x \text{ s.t. } T^T T = I
\]

\[
y = X W \rho_y + E_y = T \rho_y + E_y \text{ s.t. } T^T T = I
\]
Systems vaccinology example: Comparison of results with sparse PLS

Table 1 Fit of modeled to observed data for three methods: SPCovR, spls, and SGCCA. Displayed are the variance accounted for by the components in the block of covariates and the squared correlation between the modeled and observed outcome for the 2008 and 2007 season. The model was constructed using the 2008 data.

<table>
<thead>
<tr>
<th>Method</th>
<th>VAF</th>
<th>$r(\hat{y}, y)^2$</th>
<th>$r(\hat{y}<em>{2007}, y</em>{2007})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPCovR</td>
<td>0.19</td>
<td>0.42</td>
<td>0.79</td>
</tr>
<tr>
<td>spls</td>
<td></td>
<td>0.99</td>
<td>0.55</td>
</tr>
<tr>
<td>SGCCA</td>
<td>0.11</td>
<td>1</td>
<td>0.53</td>
</tr>
</tbody>
</table>

$r(\hat{y}, y)^2$ for elastic net = 0.06 (200 non-zero regression weights, tuned with CV using glmnet)
• SPCovR: Biological content of selected transcripts

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Nr of genes found</th>
<th>Nr of genes expected</th>
<th>+/−</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rRNA methylation</td>
<td>5</td>
<td>.21</td>
<td>+</td>
<td>2.03E−02</td>
</tr>
<tr>
<td>Cellular macromolecule metabolic process</td>
<td>89</td>
<td>58.65</td>
<td>+</td>
<td>1.68E−02</td>
</tr>
<tr>
<td>Nucleic acid metabolic process</td>
<td>60</td>
<td>34.14</td>
<td>+</td>
<td>2.84E−02</td>
</tr>
<tr>
<td>Cellular component organization or biogenesis</td>
<td>75</td>
<td>47.15</td>
<td>+</td>
<td>3.86E−02</td>
</tr>
<tr>
<td>Gene expression</td>
<td>57</td>
<td>31.88</td>
<td>+</td>
<td>3.30E−02</td>
</tr>
<tr>
<td>Leukocyte activation</td>
<td>18</td>
<td>4.59</td>
<td>+</td>
<td>6.11E−03</td>
</tr>
<tr>
<td>Cell activation</td>
<td>20</td>
<td>5.36</td>
<td>+</td>
<td>2.88E−04</td>
</tr>
<tr>
<td>Immune system process</td>
<td>31</td>
<td>13.13</td>
<td>+</td>
<td>2.79E−02</td>
</tr>
<tr>
<td>Immune effector process</td>
<td>19</td>
<td>5.25</td>
<td>+</td>
<td>9.41E−03</td>
</tr>
<tr>
<td>Negative regulation of metabolic process</td>
<td>32</td>
<td>14.16</td>
<td>+</td>
<td>4.59E−02</td>
</tr>
</tbody>
</table>

• Sparse PLS: no (significant) terms found
DISCUSSION
• We presented a generalization of sparse PCA to the multiblock case

  • Sparsity can be imposed accounting for the block structure

  • Sparsity can be imposed either on the weights or the loadings

  • Several (combinations of) penalties possible

=> Flexible modeling framework that includes several known methods as special cases: E.g., SPCA (Zou et al., 2006); sPCA-rSVD (Shen & Huang, 2007); SCA (Escofier & Pagès, 1983); and PCA
• We also presented an extension of PCovR to the high-dimensional setting

• This method balances prediction with interpretation and – in a sense – presents the best of two worlds (machine learning/data mining and psychometrics)

• Gave highly promising results on worst-case definition of Big Data for the statistician (this is, (ultra-)high-dimensional)

• Stability selection is a key ingredient
• Planned developments / in progress

• Merger of the two projects: Prediction based on sparse common covariates

• Extension to long data and the presence of heterogeneity (Presentation of Shuai Yuan)

• (Further) develop model selection methods

• Causal relations: network approach (Presentation Pia Tio) and beyond

• Improve efficiency R package, add weight based approach, add visualization

• Proof of the pudding: Joint analysis questionnaire + genetic (and other) types of data
Much to learn, we still have
• Software:
  • RegularizedSCA package available from CRAN
  • Sparse Principal Covariates regression available from GitHub:
    https://github.com/katrijnvandeun/SPCovR

• References:

Thx to the cool people!

Mr. Regularized SCA

Mrs. SNAC

Mr. Sparse W

Mr. Clusterwise sparse CoCo

Mr. Compare
Thank you!

&

NWO
Netherlands Organisation for Scientific Research