Big Data in the Social Sciences

Statistical methods for multi-source high-dimensional data

Katrijn Van Deun, Tilburg University
Outline

• Introduction & Motivation
• Methods
  • Model
  • Objective function
  • Estimation
  • Model Selection
• Conclusion & Discussion
Introduction
Big Data in the Social 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: neuron, DNA

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

• Data are linked
  • Government
  • Science: multi-disciplinary
  • Linked Data web-architecture
• Illustration: Health & Retirement Study; traditional data
• Illustration: Health & Retirement Study; traditional + novel type of data

• Multiple sources, \textbf{heterogeneous} in nature
• \textbf{High-dimensional} \((p>>n)\)

\textbf{Big Data}
• Illustration: ALSPAC household panel data

• Multiple sources / multi-block (heterogeneous)

• High-dimensional
⇒ Extremely information-rich data

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

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

• 2) Gives insight in the interplay between multiple factors

  Eg. gene-environment interactions: find (epi-)genetic markers that make someone susceptible to obesity together with the protective/risk-provoking environmental conditions associated to these markers

However, statistical tools fall short ...
This may take some time; go for a walk in the mountains

**WARNING: Ill-conditioned matrix**

Best Predictor: Nose length

$R^2_{\text{new case}} = 0$
• First challenge = Data fusion?

• How to guarantee that the different sources of variation complement each other?

  • Find common sources of structural variation?

• Yet, heterogeneous data sources; often common sources of variation are subtle while source-specific variation is dominant
• Second challenge = Relevant variables?

• Find relevant variables

• Information may be hidden in a bulk of irrelevant variables e.g., genetic markers for obesity hidden in a bulk of irrelevant markers

• Interpretation based on many variables is not very insightful

• Note: in general, we do not know which variables are relevant and which are not

• S-O-A: penalties, eg lasso
  • However: selects only one variable out of a group of correlated variables
    • (i) have a high risk of not selecting the most relevant variable,
    • (ii) are highly instable, and
    • (iii) tend to also select variables of irrelevant groups
Method:
Sparse common and specific components
Method: Sparse common & specific components

• Structured analysis
  • Data fusion: **Common components**

• Detection of relevant variables: **Penalties** (eg lasso)

=> Selection of **linked** variables (between blocks)
• 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) is denoted by \( Y \) (\( y \) if univariate)

• Each of the data blocks: same set of observation units (respondents)
Method: Data fusion

- Point of departure = Simultaneous component analysis (SCA)
  - Extension of PCA to the multiblock case
  - Promising method for data integration
• **Principal component models**

• Weight based variant:

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

\[ = T_k P_k^T + E_k \]

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:

\[ t_{irk} = \sum_j w_{jrk} x_{ijk} \]
• Principal component models

  • Loading based variant

  \[
  \mathbf{X}_k = \mathbf{T}_k \mathbf{P}_k^T + \mathbf{E}_k \quad \text{s.t.} \quad \mathbf{T}_k^T \mathbf{T}_k = \mathbf{I}, \quad (2)
  \]

  Interpretation of component \( \mathbf{t}_{rk} \) based on \( J_k \) (!) correlations:

  \[
  r(x_{jk}, t_{rk}) = p_{jrk}
  \]

  • Note: In a least squares approach subject to \( \mathbf{P}_k^T \mathbf{P}_k = \mathbf{I} \), we have \( \mathbf{W}_k = \mathbf{P}_k \)
• Simultaneous component analysis

For all $k$:

$$X_k = TP_k^T + E_k \text{ s.t. } T^TT = I$$  \hspace{1cm} (3)

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

1. Weight based variant

$$[X_1 \ldots X_K] = T \left[ P_1^T \ldots P_K^T \right] + [E_1 \ldots E_K^T]$$

$$= [X_1 \ldots X_K] \left[ W_1^T \ldots W_K^T \right] \left[ P_1^T \ldots P_K^T \right] + [E_1 \ldots E_K^T]$$

or, in shorthand notation:

$$X_{conc} = X_{conc} W_{conc} P_{conc}^T + E_{conc}$$  \hspace{1cm} (4)

2. Loading based variant

$$X_{conc} = T P_{conc}^T + E_{conc}$$  \hspace{1cm} (5)
• Simultaneous components

• Guarantee same source of variation for the different data blocks

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

• Explanation: SC methods model the largest source of variation in the concatenated data, often this is source-specific variation as common variation is subtle (e.g. response tendencies and general socio-behavioral or genetic processes vs subtle gene-environment interactions)
Common component analysis (CCA)

- Account for dominating source specific variation

- Common component model:

\[ X_k = T P_k' \] such that \( T'T = I \) and \( P_k \) has common / specific structure

### Example Table

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Spec1</td>
<td>x</td>
<td>0</td>
</tr>
<tr>
<td>Spec1</td>
<td>0</td>
<td>x</td>
</tr>
</tbody>
</table>

- Structure can be imposed (constrained analysis)

- Or: \( X_k = X_{conc} W_{conc} P_k' \) with the \( W_{conc} \) 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 is infeasible

⇒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 $X_k$

• (Sparse) distinctive components: (few) non-zero loadings only in one/few data blocks $X_k$

\[
\begin{array}{c|cc|c|cc}
 & X_1 & & X_2 & \\
\hline
\text{Common} & x & 0 & 0 & x & x & 0 & x \\
\text{Dist1} & x & x & x & 0 & 0 & 0 & 0 \\
\text{Dist1} & 0 & 0 & 0 & x & 0 & x & x \\
\end{array}
\]
Variable selection: How to?

• Sparse analysis
  – Impose restriction on the loadings: 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 $\mathbf{T}$ and $\mathbf{P}_C$ and such that $\mathbf{T}' \mathbf{T} = \mathbf{I}$

$$\|\mathbf{X}_{Conc} - \mathbf{T} \mathbf{P}'_{Conc}\|^2 + \sum_{r,k} \lambda_{r,k} |\mathbf{p}_{r,k}|_1$$

Fit / SCA

Penalty

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

$\rightarrow$ shrinks and selects variables
• (adaptive) Lasso

• Oracle properties (under some conditions)

• Estimation: soft thresholding operator $S(\beta_{OLS}, \lambda/2)$

\[
\beta_{LASSO} = \begin{cases} 
\beta_{OLS} - \lambda/2 & \text{if } \beta_{OLS} > \lambda/2 \\
0 & \text{if } -\lambda/2 < \beta_{OLS} < \lambda/2 \\
\beta_{OLS} - \lambda/2 & \text{if } \beta_{OLS} < -\lambda/2
\end{cases}
\]
• Sparse common and specific components: Objective function

Add penalties and/or constraints to obtain **structured sparsity**

\[
\|X_{\text{Conc}} - TP'_{\text{Conc}}\|^2 + \sum_{r,k} \lambda_{r,k} |p_{r,k}|_1 + \sum_k \left( \sum_s \gamma_{r,s} |p_{r,s,k}|_2 + \sum_c \varepsilon_{r,c} |p_{r,c,k}|_{1,2} \right)
\]

**Penalties:**

**Lasso:** \(|p_{r,k}|_1 = \sum_j |p_{j,r,k}|\) and \(\lambda_{r,k} \geq 0\) a tuning parameter

**Group lasso:** \(|p_{r,k}|_2 = \sqrt{\sum_j (p_{j,r,k})^2}\) and \(\gamma_{r,s} \geq 0\) a tuning parameter

**Elitist lasso:** \(|p_{r,k}|_{1,2} = (\sum_j |p_{j,r,k}|)^2\) and \(\varepsilon_{r,c} \geq 0\) a tuning parameter

*Statistical methods for multi-source high-dimensional data*
Generic objective function:

- Allows for combinations of penalties, some known

### Table 1 - Sparse approaches

<table>
<thead>
<tr>
<th>Norm</th>
<th>Properties</th>
<th>Lasso</th>
<th>Elastic net</th>
<th>Sparse approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_1$</td>
<td>selection and shrinkage at the level of the concatenated data</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>$l_2$</td>
<td>shrinkage, groups correlated variables</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>$l_{2,1}$</td>
<td>selection and shrinkage of entire blocks</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>$l_{1,2}$</td>
<td>selection and shrinkage within each block</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

- Penalties can be imposed per component
- Useful to find sparse common and specific components

Statistical methods for multi-source high-dimensional data
• **Algorithm: Alternating procedure**

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

0. Initialize $P_{\text{conc}}$

1. Update $T$ conditional upon $P_{\text{conc}}$
   
   Closed form: $T=UV'$ with $U$ and $V$ from the SVD of $X_C P$ ($I \times R \rightarrow small for H-D data$)

2. Update $P_{\text{conc}}$ conditional upon $T$
   
   MM+UST: Introduce surrogated objective function and apply univariate soft thresholding to surrogate (see next)

3. Check stop criteria (convergence of the loss, maximum number of iterations) and return to step 1 or terminate
• Algorithm: Conditional estimation of $P_{\text{conc}}$ given $T$

Complicated optimization problem because of penalties $\Rightarrow$ MM procedure
• MM in brief
Find min of $f(x)$ via surrogate function $g(x,c)$ (c: constant):
  
  – $g(x,c)$ easy to minimize
  – $g(x,c) \geq f(x)$
  – $g(c,c) = f(c)$ (supporting point)

$$f(x_{\text{min}}) \leq g(x_{\text{min}},a) \leq g(a,a) = f(a)$$

$\Rightarrow$ min of prev. iter. = supp. point current it.
Yields non-increasing series of loss values
Example: constructing a majorizing function for the elitist lasso penalty

\[
\left( \sum_{j_k} |p_{jkr}| \right)^2 \leq \left( \sum_{j_k} |p_{jkr}^{(0)}| \right) \sum_{j_k} \frac{p_{jkr}^2}{|p_{jkr}^{(0)}|}
\]

\[= p'D_1p \text{ with } D_1 \text{ a diagonal matrix of } \frac{(\sum_{j_k} p_{jkr}^{(0)})}{|p_{jkr}^{(0)}|}
\]

Applied to group, elitist and regularly lasso we obtain as a surrogate function

\[
L \leq k + \|X_{\text{conc}} - TP'_{\text{conc}}\|^2 + \text{vec}(P_{\text{conc}})'D \text{ vec}(P_{\text{conc}})
\]

\[= k + \|\text{vec}(X_{\text{conc}}) - (I \otimes T)\text{vec}(P_{\text{conc}})\|^2 + \text{vec}(P_{\text{conc}})'D \text{ vec}(P_{\text{conc}})
\]

For which the root can be found using standard techniques
Hence, the problem is now to find the minimum of

$$\| \text{vec}(X_{\text{conc}}) - (I \otimes T)\text{vec}(P_{\text{conc}}) \|^2 + \text{vec}(P_{\text{conc}})'D \text{ vec}(P_{\text{conc}}) $$

which can be found using standard techniques:

$$\text{vec}(P_{\text{conc}}) = [D + I]^{-1}\text{vec}(T^T X_{\text{conc}})$$

Note that $D+I$ is diagonal
Alternative approach: Coordinate descent

\[ L \leq k + \|X_{\text{conc}} - TP_{\text{conc}}\|^2 + \text{vec}(P_{\text{conc}})'D^* \text{vec}(P_{\text{conc}}) + \sum_{r,k} \lambda_{r,k} |p_{r,k}|_1 \]

\[ = k + \|\text{vec}(X_{\text{conc}}) - (I \otimes T)\text{vec}(P_{\text{conc}})\|^2 + \text{vec}(P_{\text{conc}})'D^*\text{vec}(P_{\text{conc}}) + \sum_{r,k} \lambda_{r,k} |p_{r,k}|_1 \]

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

For which the root can be found using subgradient techniques
• Hence, the following soft thresholding update of the loadings:

\[
p_{jr} = \left\{ \begin{array}{ll}
\frac{\sum_i x_{ijr} t_{ir} - \lambda_r/2}{1 + d_{jr}} & \text{if } p_{jr} > 0 \\
\frac{\sum_i x_{ijr} t_{ir} + \lambda_r/2}{1 + d_{jr}} & \text{if } p_{jr} < 0 \\
0 & \text{else}
\end{array} \right.
\]

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

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

• Similar type of algorithms can be constructed

Standard MM: $\text{vec}(W_{\text{conc}}) = [D + I \otimes (X_{\text{conc}}^T X_{\text{conc}})]^{-1} \text{vec}(T^T X_{\text{conc}})$

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

$$w_{jk^{*}r^{*}} = \begin{cases} 
\frac{\sum_{i,k,j_k} x_{ij_k} r_{ij_k} p_{j_k r^{*}} - \lambda_r / 2}{1 + d_{j_k r}} & \text{if } w_{jk^{*}r^{*}} > 0 \\
\frac{\sum_{i,k,j_k} x_{ij_k} r_{ij_k} p_{j_k r^{*}} + \lambda_r / 2}{1 + d_{j_k r}} & \text{if } w_{jk^{*}r^{*}} < 0 \\
0 & \text{else}
\end{cases}$$

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

• Coordinate-wise approach allows to fix coefficients

• This can be used to define the specific components by fixing to zero the coefficients corresponding to the block not accounted for by the component
Model selection

• Number of components $R$, penalty tuning parameters and/or status (common or specific):
  
  • How to select suitable values?
  
  • No definite answer, only suggestions
    • $R$: inspect VAF
    • Use stability selection to tune variable selection ($\lambda_l$)
    • Cross-validation
    • Exhaustive strategy for status if nr blocks & components limited
Some results
Weights or loadings: Simulation results

- Data generated with either sparse $W$ or sparse $P$
- All data analyzed with both a model with sparsity imposed on $W$ and sparsity imposed on $P$
- Recovery best when generated with sparse loadings and analyzed with sparse weights
- More direct link between reproduced data and loadings

Statistical methods for multi-source high-dimensional data
Structured sparsity needed or just lasso?

Tucker congruence between $W^{TRUE}$ and estimated $W$
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse weights</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sparse loadings</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: best recovery for generated with sp.loadings and estimated with sp.weights

• Different algorithmic strategies possible: here MM and coord. Descent considered
• Planned developments / in progress

• Selection of relevant clusters of correlated variables (to deal with issues of high-dimensionality that haunt lasso, elastic net and so on)

• Prediction by extensions of principal covariates regression

\[ L(W, P_X, P_y) = \alpha \frac{\|y - XWP_y \|^2}{\|y\|^2} + (1 - \alpha) \frac{\|X - XWP_X^T\|^2}{\|X\|^2} \]

\[ + \lambda_L \|W\|_1 + \lambda_R \|W\|_2 \]

with 0 ≤ α ≤ 1 (α = 0 -> PCR; α = 1 -> MLR/RRR)
• SPCovR Application: Find early (day 3) genetic signature that predicts flu vaccine efficacy (day 28) (public data: Nakaya et al., 2011)
• 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>
• 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
Thank you!
• I am looking for a PhD candidate!!

<table>
<thead>
<tr>
<th>Marriage vs. The Ph.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Marriage</strong></td>
</tr>
<tr>
<td>7.5 years</td>
</tr>
<tr>
<td>A proposal</td>
</tr>
<tr>
<td>A thesis proposal</td>
</tr>
<tr>
<td><strong>Ph.D.</strong></td>
</tr>
<tr>
<td>7 years</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Typical Length:</strong></td>
</tr>
<tr>
<td><strong>Begins with:</strong></td>
</tr>
<tr>
<td><strong>Culminates in a</strong></td>
</tr>
<tr>
<td>ceremony where you</td>
</tr>
<tr>
<td>walk down an aisle</td>
</tr>
<tr>
<td>dressed in a gown:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Usually entered</strong></td>
</tr>
<tr>
<td>into by:**</td>
</tr>
<tr>
<td>Foolish young people</td>
</tr>
<tr>
<td>in love</td>
</tr>
<tr>
<td>Foolish young people</td>
</tr>
<tr>
<td>without a job</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>50% end in:</strong></td>
</tr>
<tr>
<td>Bitter divorce</td>
</tr>
<tr>
<td>Bitter remorse</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Involves exchange of:</strong></td>
</tr>
<tr>
<td>Vows</td>
</tr>
<tr>
<td>Know-how</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Until death do you part?</strong></td>
</tr>
<tr>
<td>If you're lucky</td>
</tr>
<tr>
<td>If you're lazy</td>
</tr>
</tbody>
</table>