Network PCA
A versatile multivariate tool for exploring data structures in metabolomics

S. Codesido¹, M. Hanafi², Y. Gagnebin¹, J. Boccard¹, S. Rudaz¹
Chimiométrie, Montpellier, 01/02/2019

¹Biomedical analysis and metabolomics
Université de Genève

²Statistique, sensométrie et chimiométrie
Oniris, Nantes
Outline

• Structured data
  More than just data tables

• Network PCA
  Unsupervised multivariate analysis for arbitrary data structures

• Application
  Chronic kidney disease metabolite identification

• Conclusions & outlook
Structured data

• Back in the easy days:

<table>
<thead>
<tr>
<th></th>
<th>Substance A</th>
<th>Substance B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control patient</td>
<td>10 mg/L</td>
<td>20 μg/L</td>
</tr>
<tr>
<td>Treated patient</td>
<td>30 mg/L</td>
<td>5 μg/L</td>
</tr>
</tbody>
</table>

• Experimental structure (control vs treated, subs A vs subs B) directly reflected in the data
Structured data: metabolomics workflow

- Biological Question
  - Study design

- Sample collection and preparation

- Multivariate Data Analysis

- Multiple techniques required!
  - Data acquisition
    - LC – CE - IMS / MS

- Features
  - Raw data
  - Preprocessing
  - Fingerprinting
  - Annotation
Structured data: blocks

• Large amounts of data \(\rightarrow\) multivariate analysis

\[
\begin{array}{cccccc}
\text{LC-MS 1} & \cdots & \text{LC-MS 9475} & \text{NMR 1} & \cdots & \text{NMR 1372} \\
\hline
\text{Control 1} & \cdots & \cdots & \cdots & \cdots & \cdots \\
\text{Control 73} & \cdots & \cdots & \cdots & \cdots & \cdots \\
\text{Treated 1} & \cdots & \cdots & \cdots & \cdots & \cdots \\
\text{Treated 94} & \cdots & \cdots & \cdots & \cdots & \cdots \\
\end{array}
\]

• Structured data requires a multi-index description: \(X_{i,j}^{a,b}\)
  • Multiple techniques
  • Multiple sample groups

• How to deal with this?
  • Usual answer: concatenate blocks into a bigger table \(\rightarrow\) lose structure

\[
X_{ij} = \sigma_1 U_{1,i} V_{1,j} + \sigma_2 U_{2,i} V_{2,j} + \cdots
\]
Structured data: common approaches

• Computing loadings and weights \((U_{k,i}, V_{k,j}, \sigma_k)\):
  
  • Unsupervised (PCA)
    
    Maximize global variance
    
    (+) Clear interpretation & significance
    
    (-) Mixes inter-block and intra-block

  
• Supervised (PLS-DA)
  
    Maximize inter-block variance
    
    (+) Loadings: inter-block structure
    
    (-) Cross-validation, harder to assess statistical significance

• Even worse: what if blocks cannot be joined into a single table?
Structured data: networks of tables

- What if blocks cannot be joined in a single table? (missing blocks, longitudinal studies...)

![Diagram showing networks of tables with shared observations and variables over time.]

- Shared observations
- Shared variables
- Time
Network PCA

- Standard PCA:
  - \((\sigma_k, U_{k,i}, V_{k,j}) = \text{argmin} \text{ var}(R_{i,j})\)
  - Min. global residual variance
    - global weights

- Network PCA:
  - \((\sigma_k^{a,b}, U_{k,i}^a, V_{k,j}^b) = \text{argmin} \text{ var}(R_{i,j}^{a,b})\)
    - In practice: variation on NIPALS
  - Min. global residual variance
    - per-block weights

\[ X_{i,j} = \sum_k \sigma_k \cdot U_{k,i}^a \cdot V_{k,j}^b + R_{i,j}^{a,b} \]

(+) Unsupervised minimization goal! (same as standard PCA)
  - Compare i.e. PLS-DA: given block structure is an integral part of minimization goal

(+) Structure preserved in the weights
  - Generalization of multiblock methods, partial diagonalization of the tensor \(X_{i,j}^{a,b}\)
Network PCA

• What do we get in return?

• Scores/loadings for observations/variables (as usual)

• Matrix of block weights
  Recover simple interpretation of experimental conditions

Comp. 1 | Comp. 2 | … | Comp. 3286
---|---|---|---
0.1415 | 0.0594 | … | -0.0537

\[
\sigma^{a,b} \begin{array}{|c|c|}
\hline
\text{LC-MS} & \text{NMR} \\
\hline
\text{Control} & 1.0 & 6.0 \\
\hline
\text{Treated} & 3.0 & 3.5 \\
\hline
\end{array}
\]
Network PCA

• But also: networks!
  • Ability to deal with non-concatenable data structures
  • Weights per-block
  • Structure is enforced through shared loadings

\[ X_{i,j}^{\text{block}} = \sum_k \sigma_k^{\text{block}} \cdot U_{k,i}^{a(\text{block})} \cdot V_{k,j}^{b(\text{block})} + R_{i,j}^{\text{block}} \]

\[ a: \{\text{blocks}\} \mapsto \{\text{observation groups}\} \quad b: \{\text{blocks}\} \mapsto \{\text{variable groups}\} \]

\[ a \left( \square \right) = \text{Control patients} \quad b \left( \square \right) = \text{RPLC measures} \]
Application: Chronic Kidney Disease (CKD)

• Major public health issue, prevalence increasing worldwide
  • >10% in 2016 [Hill NR et al. 2016]

• Progressive loss of kidney function

• Kidney performance has heavy impact on metabolome

• Evolves along several stages
  • Perfect test case for multiblock structure
**CKD: data structure**

- **3 types of patient**
  - Control (healthy)
  - Intermediate stage CKD
  - End stage (kidney failure)
    - Before dialysis
    - After dialysis

- **4 experimental techniques**
- **328 identified compounds**
  - RPLC ESI+ (118)
  - RPLC ESI- (84)
  - HILIC ESI+ (66)
  - z-HILIC ESI- (60)

- Data can be forced onto a single supertable, but structure is lost.

![Diagram showing data structure with different stages and techniques]
CKD: standard PCA score plot
CKD: network PCA score plot

Control (healthy)

Intermediate stage

End stage (post-dialysis)

End stage (pre-dialysis)
CKD: loading comparison

RPC +

RPC -

HILIC +

HILIC -

PC 1

PC 2

Concatenated PCA loadings
CKD: superblocks

We can sum over blocks to see the contribution of each “superblock”

\[ R_{k,a} = \frac{\sum_b (\sigma_k^{a,b})^2}{\sum_a \sum_{i,j} (X_{i,j}^{a,b})^2} \]

\[ C_{k,b} = \frac{\sum_a (\sigma_k^{a,b})^2}{\sum_a \sum_{i,j} (X_{i,j}^{a,b})^2} \]
CKD: return to variables

- Define the per-block relative increase of residual variance when neglecting one loading

\[ I^{a,b}_j = \frac{\text{var}(R \text{ without } V^b_j \text{ in } \{a, b\}) - \text{var}(R)}{\text{var}(R)} \]

- Gives “influence” of compound \( j \) in block \( \{a, b\} \) (as PLS VIPs)
- \( \sum_a I^{a,b}_j \propto (V^b_j)^2 \): sum over blocks reproduces loadings
Let us erase some blocks to create an unbalanced, non-tabular design

Useful for longitudinal studies where not all blocks can be measured at every time step
CKD: missing blocks

We still reliably find the right loadings for the compounds!
Conclusions & outlook

• We wanted a MVA
  • with unsupervised goal
  • yet with data structure reflected in output

• Structures: not always tabular and concatenable

• Network PCA: solve both issues by
  • letting each data block have its own effect sizes
  • enforcing structural relations via shared loadings

• Application example: CKD
  • Network PCA: as robust as PCA
  • Extra return on block structure via per-group weights
  • Ability to cope with missing blocks

• Network PCA has been implemented in MATLAB
  • Ports to other languages (Python!) and GUI in development
Network PCA vs PCA: degrees of freedom

Loadings

<table>
<thead>
<tr>
<th>Effect 1, $\sigma^{a,b}$</th>
<th>1.25</th>
<th>1.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>-1.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect 2, $\sigma^{a,b}$</th>
<th>0.8</th>
<th>-0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.3</td>
<td>1.6</td>
<td></td>
</tr>
</tbody>
</table>

Full dataset
Network PCA vs PCA: degrees of freedom

Standard PCA misses one the dark “blobs” because it forces each loading to have the same influence on every block.

It can be fixed by adding more components to the PCA decomposition, but visualization becomes harder!
Application: block influence

Influence of each block:
weight matrix $\sigma^{a,b}$

Component 1
Component 2
Application: compound influence

Compounds in RP-, sorted by PC1 influence on:

End-stage (pre-dial)

Total (~|loading|^2)

Control

- CITRAMALIC ACID
- SUBERIC ACID
- HOMOVANILIC ACID
- ROSMARINIC ACID
- TRANS-ACONITIC ACID
- SEBASIC ACID
- N-ACETYLGlyCINE
- TRYPTAMINE
- AZELAIC ACID
- ETHYLMALONIC ACID

- CITRAMALIC ACID
- HOMOVANILIC ACID
- SUBERIC ACID
- SEBASIC ACID
- ROSMARINIC ACID
- N-ACETYLDL-METHIONINE
- FORMYL-L-METHIONYL PEPTIDE
- AZELAIC ACID
- N-ACETYLGlyCINE

- FORMYL-L-METHIONYL PEPTIDE
- N-ACETYLDL-METHIONINE
- 5'-DEOXYADENOSINE
- TRANS-ACONITIC ACID
- 10-HYDROXYDECANOIC ACID
- 3A,5B-TETRAHYDROCORTICOSTERONE
- AZELAIC ACID
- XYLITOL
- HOMOVANILIC ACID
- CITRAMALIC ACID
Application: compound influence

Explicit formula:

\[ I_j^{A,B} \text{ SS}(R) = \text{var}(R \text{ without } V_j^B \text{ in } \{A, B\}) - \text{var}(R) = \]

\[ = \sum_{a,b} \sum_{i,j} \left( R_{i,j}^{a,b} + \sigma^{a,b} \cdot U_i^a V_j^b \cdot \delta^b \delta^a \delta_j \right)^2 - \sum_{a,b} \sum_{i,j} \left( R_{i,j}^{a,b} \right)^2 = \]

\[ = \left( \sigma^{A,B} V_j^B \right)^2 + 2 \sigma^{A,B} V_j^B \cdot \sum_i R_{i,j}^{A,B} U_i^A \]

Relation between “influence” and loadings, \( \sum_a I_j^{a,b} \propto (V_j^b)^2 \):

---

PC 1 compound loadings, \( |V_j^b| \)
Application: missing blocks

Full Network PCA: zHILIC- PC1 influences

End is most important group for zHILIC-
Even when missing loadings agree on effect sign!