Recent advances on Regularized Generalized Canonical Correlation Analysis

Arthur Tenenhaus  
2014/04/02
Part I: multiblock data analysis

• **SETTINGS:** Several set of variables are measured on the same individuals.

• **OBJECTIVE:** Investigate the relationships between blocks.
Part II: multigroup data analysis

- **SETTINGS**: The same set of variables are measured on individuals structured in several groups.

- **OBJECTIVE**: investigate the relationships among variables within the various groups.
# Glioma Cancer Data

(Department of Pediatric Oncology of the Gustave Roussy Institute)

## Transcriptomic data ($X_1$)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>...</th>
<th>Gene 15201</th>
<th>CGH1</th>
<th>...</th>
<th>CGH 1909</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0.18</td>
<td>-0.21</td>
<td>...</td>
<td>-0.73</td>
<td>0.00</td>
<td>...</td>
<td>-0.55</td>
<td>Hemisphere</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1.15</td>
<td>-0.45</td>
<td>...</td>
<td>0.27</td>
<td>-0.30</td>
<td>...</td>
<td>0.00</td>
<td>Midline</td>
</tr>
<tr>
<td>Patient 3</td>
<td>1.35</td>
<td>0.17</td>
<td>...</td>
<td>0.22</td>
<td>0.33</td>
<td>...</td>
<td>0.64</td>
<td>DIPG</td>
</tr>
<tr>
<td>Patient 53</td>
<td>1.39</td>
<td>0.18</td>
<td>...</td>
<td>-0.17</td>
<td>0.00</td>
<td>...</td>
<td>0.43</td>
<td>Hemisphere</td>
</tr>
</tbody>
</table>

## CGH data ($X_2$)

## Outcome ($X_3$)
Glioma Cancer Data: from a multi-block viewpoint
(Department of Pediatric Oncology of the Gustave Roussy Institute)

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>…</th>
<th>Gene 15201</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0.18</td>
<td>-0.73</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1.15</td>
<td>0.27</td>
</tr>
<tr>
<td>Patient 3</td>
<td>1.35</td>
<td>0.22</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Patient 53</td>
<td>1.39</td>
<td>-0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CGH1</th>
<th>…</th>
<th>CGH 1909</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0.00</td>
<td>-0.55</td>
</tr>
<tr>
<td>Patient 2</td>
<td>-0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Patient 3</td>
<td>0.33</td>
<td>0.64</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Patient 53</td>
<td>0.00</td>
<td>0.43</td>
</tr>
</tbody>
</table>

\[ C_{13} = 1 \]
\[ C_{12} = \emptyset \]
\[ C_{23} = 1 \]
Block components

\[ y_1 = X_1 a_1 = a_{11} \text{Gene}_1 + \cdots + a_{1,15201} \text{Gene}_{15201} \]

\[ y_2 = X_2 a_2 = a_{21} \text{CGH}_1 + \cdots + a_{2,1909} \text{CGH}_{1909} \]

\[ y_3 = X_3 a_3 = a_{31} \text{Hemisphere} + a_{32} \text{DIPG} \]

Block components should verified two properties at the same time:

(i) Block components well explain their own block.

(ii) Block components are as correlated as possible for connected blocks.
Some multi-block methods

**SUMCOR** (Horst, 1961)

\[
\text{maximize } \sum_{j,k} \text{cor}(X_j a_j, X_k a_k)
\]

**GENERALIZED CANONICAL CORRELATION ANALYSIS**

**SABSCOR** (Mathes, 1993; Hanafi, 2004)

\[
\text{maximize } \sum_{j,k} |\text{cor}(X_j a_j, X_k a_k)|
\]
Covariance-based criteria

\[ c_{jk} = 1 \text{ if blocks are linked, } 0 \text{ otherwise and } c_{jj} = 0 \]

**SUMCOR:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} \text{cov}(X_ja_j, X_ka_k) \\
\text{all } \text{var}(X_ja_j) = 1
\]

**SSQCOR:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} \text{cov}^2(X_ja_j, X_ka_k) \\
\text{all } \text{var}(X_ja_j) = 1
\]

**SABSCOR:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} |\text{cov}(X_ja_j, X_ka_k)| \\
\text{all } \text{var}(X_ja_j) = 1
\]

**SUMCOV:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} \text{cov}(X_ja_j, X_ka_k) \\
\text{all } \|a_j\| = 1
\]

**SSQCOCV:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} \text{cov}^2(X_ja_j, X_ka_k) \\
\text{all } \|a_j\| = 1
\]

**SABSCOCV:**

\[
\text{maximize} \quad \sum_{j,k} c_{jk} |\text{cov}(X_ja_j, X_ka_k)| \\
\text{all } \|a_j\| = 1
\]

\[
cov^2(X_ja_j, X_ka_k) = \text{var}(X_ja_j) \text{cor}^2(X_ja_j, X_ka_k) \text{var}(X_ja_j)
\]
RGCCA optimization problem

\[
\text{argmax}_{a_1, a_2, \ldots, a_J} \sum_{j \neq k}^J c_{jk} g \left( \text{cov}(X_j a_j, X_k a_k) \right)
\]

Subject to the constraints

\[
(1 - \tau_j)\text{var}(X_j a_j) + \tau_j \|a_j\|^2 = 1, j = 1, \ldots, J
\]

where:

\[
g = \begin{cases} 
\text{id} \text{entity} & \text{(Horst scheme)} \\
\text{square} & \text{(Factorial scheme)}
\end{cases}
\]

A monotone convergent algorithm related to this optimization problem will be described.

Schäfer and Strimmer formula can be used for an optimal determination of the shrinkage constants and:

\[
\tau_j = \text{Shrinkage constant between 0 and 1}
\]
Choice of the shrinkage constant $\tau_j$ (part 1)

\[
\arg\max_{a_1, a_2} \text{cov}(X_1 a_1, X_2 a_2)
\]
Subject to the constraints \((1 - \tau_j) \text{var}(X_j a_j) + \tau_j \|a_j\|^2 = 1, j = 1, 2\)

### Special cases

<table>
<thead>
<tr>
<th>Method</th>
<th>Criterion</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS regression</td>
<td>[\text{Maximize} \ \text{Cov}(X_1 a_1, X_2 a_2)]</td>
<td>$|a_1| = |a_2| = 1$</td>
</tr>
<tr>
<td>Canonical Correlation Analysis</td>
<td>[\text{Maximize} \ \text{Cor}(X_1 a_1, X_2 a_2)]</td>
<td>$\text{Var}(X_1 a_1) = \text{Var}(X_2 a_2) = 1$</td>
</tr>
<tr>
<td>Redundancy analysis of $X_1$ with respect to $X_2$</td>
<td>[\text{Maximize} \ \text{Cor}(X_1 a_1, X_2 a_2) \ \text{Var}(X_1 a_1)^{1/2}]</td>
<td>$|a_1| = 1$, $\text{Var}(X_2 a_2) = 1$</td>
</tr>
</tbody>
</table>

Components $X_1 a_1$ and $X_2 a_2$ are well correlated.  
1st component is stable.  
No stability condition for 2nd component.
Choice of the shrinkage constant $\tau_j$ (part 2)

$$\arg\max_{a_1, a_2, \ldots, a_J} \sum_{j \neq k}^J c_{jk} g \left( \text{cov}(X_j a_j, X_k a_k) \right)$$

Subject to the constraints

$$(1 - \tau_j) \text{var}(X_j a_j) + \tau_j \|a_j\|^2 = 1, \ j = 1, \ldots, J$$

Schäfer and Strimmer formula can be used for an optimal determination of the shrinkage constants.
Choice of the design matrix C

Hierarchical models

(a) One second order block

(b) Several second order blocks

Very often:

\[ X_{1}, \ldots, X_{J_{1}} = \text{Predictor blocks} \]

\[ X_{J_{1}+1}, \ldots, X_{J} = \text{Response Blocks} \]

\[
\begin{align*}
\max_{a_1, a_2, \ldots, a_J} & \sum_{j=1}^{J} g \left( \text{cov}(X_j a_j, X_{j+1} a_{j+1}) \right) \\
(1 - \tau_j) \text{var}(X_j a_j) + \tau_j \| a_j \|^2 & = 1, j = 1, \ldots, J + 1
\end{align*}
\]

\[
\begin{align*}
\max_{a_1, a_2, \ldots, a_J} & \sum_{j=1}^{J_1} \sum_{k=J_1+1}^{J} c_{j,k} g \left( \text{cov}(X_j a_j, X_k a_k) \right) \\
(1 - \tau_j) \text{var}(X_j a_j) + \tau_j \| a_j \|^2 & = 1, j = 1, \ldots, J
\end{align*}
\]
Choice of the design for NeuroImaging-Genetic datasets

$X_1$ DNA arrays (SNP)  \hspace{2cm} X_2$ Functional MRI  \hspace{2cm} X_3$

- Developmental disorders - Reading difficulties - Basic numerical knowledge - ... - Visuo-spatial abilities - Visuo-motor abilities

$p_1 \sim 10^6$  \hspace{2cm} p_2 \sim 10^4$  \hspace{2cm} p_3 \sim 10$

$n \sim 100$  \hspace{2cm} c_{12} = 1$  \hspace{2cm} c_{23} = 1$  \hspace{2cm} c_{13} = 0$
special cases of RGCCA (among others)

**two-block case**

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
</table>

**multi-block case**

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
</table>
monotone convergent algorithms for these criteria

\[
\arg\max_{a_1,a_2,\ldots,a_J} \sum_{j \neq k} c_{j,k} g \left( \text{cov} (X_j a_j, X_k a_k) \right)
\]

Subject to the constraints

\[
(1 - \tau_j) \text{var}(X_j a_j) + \tau_j \|a_j\|^2 = 1, \quad j = 1, \ldots, J
\]

- Construct the Lagrangian function related to the optimization problem.
- Cancel the derivative of the Lagrangian function with respect to each \(a_j\).
- Use the Wold’s procedure to solve the stationary equations (\(\approx\) Gauss-Seidel algorithm).
- This procedure is monotonically convergent: the criterion increases at each step of the algorithm.
The RGCCA algorithm (primal version)

**Outer Estimation** (explains the block)

\[
(1 - \tau_j) \operatorname{var} (X_j a_j) + \tau_j \|a_j\|^2 = 1
\]

**Inner Estimation** (explains relation between block)

\[
z_j = \sum_{k \neq j} e_{jk} y_k
\]

**Initial step**

\[
y_j = X_j a_j
\]

**Iterate until convergence of the criterion**

**Choice of weights** \( e_{jh} \):

- **Horst**:
  \[ e_{jk} = c_{jk} \]

- **Centroid**:
  \[ e_{jk} = c_{jk} \operatorname{sign} \left( \operatorname{cor} (y_j, y_k) \right) \]

- **Factorial**:
  \[ e_{jk} = c_{jk} \operatorname{cov} (y_j, y_k) \]

Dimension = \( p_j \times p_j \)

\( c_{jk} = 1 \) if blocks are linked, 0 otherwise and \( c_{jj} = 0 \)
The RGCCA algorithm (dual version)

\[ a_j = X_j^t \alpha_j \]

\[ y_j = X_j^t X_j \alpha_j \]

Initial step

Outer Estimation (explains the block)

\[ \alpha_j^t [X_j^t X_j (\tau_j I + (1 - \tau_j) \frac{1}{n} X_j^t X_j)] \alpha_j = 1 \]

Iterate until convergence of the criterion

Inner Estimation (explains relation between block)

\[ z_j = \sum_{k \neq j} e_{jk} y_k \]

Choice of weights \( e_{jh} \):
- Horst: \( e_{jk} = c_{jk} \)
- Centroid: \( e_{jk} = c_{jk} \text{ sign } (\text{cor} (y_j, y_k)) \)
- Factorial: \( e_{jk} = c_{jk} \text{ cov } (y_j, y_k) \)

Dimension = \( n \times n \)

\( c_{jk} = 1 \) if blocks are linked, 0 otherwise and \( c_{jj} = 0 \)
Glioma Cancer Data: from an RGCCA viewpoint
(Department of Pediatric Oncology of the Gustave Roussy Institute)

RGCCA with factorial scheme - $\tau_1 = 1$, $\tau_2 = 1$ and $\tau_3 = 0$

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gene 1</th>
<th>...</th>
<th>Gene 15201</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0.18</td>
<td></td>
<td>-0.73</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1.15</td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Patient 3</td>
<td>1.35</td>
<td></td>
<td>0.22</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient 53</td>
<td>1.39</td>
<td></td>
<td>-0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient</th>
<th>CGH1</th>
<th>...</th>
<th>CGH 1909</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0.00</td>
<td></td>
<td>-0.55</td>
</tr>
<tr>
<td>Patient 2</td>
<td>-0.30</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Patient 3</td>
<td>0.33</td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient 53</td>
<td>0.00</td>
<td></td>
<td>0.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>DIPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>1</td>
</tr>
<tr>
<td>Patient 2</td>
<td>0</td>
</tr>
<tr>
<td>Patient 3</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Patient 53</td>
<td>1</td>
</tr>
</tbody>
</table>

High dimensional block settings $\Rightarrow$ dual algorithm for RGCCA
Bayesian Discriminant Analysis of localization on $y_1$ and $y_2$
Predictive performance

Table 1. Learning phase

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th>DIPG</th>
<th>Hemispheres</th>
<th>Midline</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIPG</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hemispheres</td>
<td>0</td>
<td>19</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Midline</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy = 82%

Table 2. Testing phase (leave-one-out)

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th>DIPG</th>
<th>Hemispheres</th>
<th>Midline</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIPG</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hemispheres</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Midline</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy = 75%
More details on RGCCA

Regularized Generalized Canonical Correlation Analysis

Arthur Tenenhaus
SUPELEC, GIF-SUR-YVETTE

Michel Tenenhaus
HEC PARIS, JOUY-EN-JOSAS

Regularized generalized canonical correlation analysis (RGCCA) is a generalization of regularized canonical correlation analysis to three or more sets of variables. It constitutes a general framework for many multi-block data analysis methods. It combines the power of multi-block data analysis methods (maximization of well identified criteria) and the flexibility of PLS path modeling (the researcher decides which blocks are connected and which are not). Searching for a fixed point of the stationary equations related to RGCCA, a new monotonically convergent algorithm, very similar to the PLS algorithm proposed by Herman Wold, is obtained. Finally, a practical example is discussed.

Key words: generalized canonical correlation analysis, multi-block data analysis, PLS path modeling, regularized canonical correlation analysis.

Package ‘RGCCA’

October 15, 2010

Type Package
Title Regularized Generalized Canonical Correlation Analysis
Version 1.0
Date 2010-06-08
Author Arthur Tenenhaus
Maintainer Arthur Tenenhaus <arthur.tenenhaus@supelec.fr>
Description Regularized Generalized Canonical Correlation Analysis
Block components

\[ y_1 = X_1 a_1 = a_{11} \text{Gene}_1 + \cdots + a_{1,15201} \text{Gene}_{15201} \]

\[ y_2 = X_2 a_2 = a_{21} \text{CGH}_1 + \cdots + a_{2,15201} \text{CGH}_{15201} \]

\[ y_3 = X_3 a_3 = a_{31} \text{Hemisphere} + a_{32} \text{DIPG} \]

Block components should verified two properties at the same time:

(i) Block components well explain their own block.

(ii) Block components are as correlated as possible for connected blocks.

(iii) Block components are built from sparse \( a_j \)
Variable selection for RGCCA

\[
\arg\max_{a_1, a_2, \ldots, a_J} \sum_{j \neq k} c_{jk} g \left( \text{cov}(X_j a_j, X_k a_k) \right)
\]

Subject to the constraints

\[
\begin{cases}
\|a_j\|_2^2 = 1, & j = 1, \ldots, J \\
\|a_j\|_1 \leq c_j, & j = 1, \ldots, J
\end{cases}
\]

where:

\[
c_{jk} = \begin{cases}
1 & \text{if } X_j \text{ and } X_k \text{ is connected} \\
0 & \text{otherwise}
\end{cases}
\]

\[
g = \begin{cases}
\text{identity} & \text{(Horst scheme)} \\
\text{square} & \text{(Factorial scheme)} \\
\text{absolute value} & \text{(Centroid scheme)}
\end{cases}
\]

and:

\[
\tau_j = \text{Shrinkage constant between 0 and 1}
\]
Sparse GCCA

\[ y_j = X_j a_j \]

Outer Estimation (explains the block)

\[ \|a_j\|_2^2 = 1 \]

Iterate until convergence of the criterion

\[ z_j = \sum_{k \neq j} e_{jk} y_k \]

Inner Estimation (explains relation between block)

Choice of weights \( e_{jh} \):
- Horst: \( e_{jk} = c_{jk} \)
- Centroid: \( e_{jk} = c_{jk} \text{ sign } (\text{cor } (y_j, y_k)) \)
- Factorial: \( e_{jk} = c_{jk} \text{ cov } (y_j, y_k) \)

\( S(a, \lambda) = \text{sign}(a) \max(0, |a| - \lambda) \)

\( c_{jk} = 1 \text{ if blocks are linked, } 0 \text{ otherwise and } c_{jj} = 0 \)
Visualization
Predictive performances

![Box plot showing predictive performances for different methods and designs. The x-axis represents different methods and designs, while the y-axis shows performance values. The box plots compare L1LDA, Supervised CCA, Horst, Factorial, and Centroid across Design 2, Design 1, Design 3, and SGCCA.]
### Signature stability

**Stability and mean of the length of the signatures for SGCCA, \(\ell_1\)-LDA, and supervised CCA across the 10-fold**

<table>
<thead>
<tr>
<th>Method</th>
<th>Fleiss’ (\kappa) (GE)</th>
<th>length of the GE signature</th>
<th>Fleiss’ (\kappa) (CGH)</th>
<th>length of the CGH signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised CCA</td>
<td>0.130</td>
<td>455.3</td>
<td>0.116</td>
<td>36.5</td>
</tr>
<tr>
<td>(\ell_1)-LDA</td>
<td>0.476</td>
<td>9790.0</td>
<td>0.322</td>
<td>480.9</td>
</tr>
<tr>
<td>Horst SGCCA (Complete)</td>
<td>0.103</td>
<td>132.2</td>
<td>0.071</td>
<td>35.9</td>
</tr>
<tr>
<td>Factorial SGCCA (Complete)</td>
<td>0.071</td>
<td>79.0</td>
<td>0.014</td>
<td>59.6</td>
</tr>
<tr>
<td>Centroid SGCCA (Complete)</td>
<td>0.137</td>
<td>73.6</td>
<td>0.105</td>
<td>22.6</td>
</tr>
<tr>
<td>Horst SGCCA (Hierarchical)</td>
<td>0.468</td>
<td>61.1</td>
<td>0.296</td>
<td>33.6</td>
</tr>
<tr>
<td>Factorial SGCCA (Hierarchical)</td>
<td>0.439</td>
<td>42.0</td>
<td>0.343</td>
<td>37.6</td>
</tr>
<tr>
<td>Centroid SGCCA (Hierarchical)</td>
<td>0.478</td>
<td>40.6</td>
<td>0.317</td>
<td>34.8</td>
</tr>
<tr>
<td>Horst SGCCA (Cascade)</td>
<td>0.071</td>
<td>83.6</td>
<td>0.074</td>
<td>40.7</td>
</tr>
<tr>
<td>Factorial SGCCA (Cascade)</td>
<td>0.061</td>
<td>118.3</td>
<td>0.026</td>
<td>49.2</td>
</tr>
<tr>
<td>Centroid SGCCA (Cascade)</td>
<td>0.040</td>
<td>75.5</td>
<td>0.035</td>
<td>40.2</td>
</tr>
</tbody>
</table>
Variable selection for generalized canonical correlation analysis

ARThUR TENENHAUS*
SUPELEC, Plateau de moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France
arthur.tenenhaus@supelec.fr

CATHY PHILIPPE
CNRS-IGR-Paris XI university, UMR8203, 94805 Villejuif cedex, France

VINCENT GUILLEMOT
NEUROSPIN, I2BM, CEA saclay, 91191 Gif-sur-Yvette cedex, France

KIM-ANH LE CAO
Queensland Facility for Advanced Bioinformatics, University of Queensland, 306 Carmody Road, St Lucia, QLD 4072, Australia

JACQUES GRILL
CNRS-IGR-Paris XI university, UMR8203, 94805 Villejuif cedex, France

VINCENT FROUIN
NEUROSPIN, I2BM, CEA saclay, 91191 Gif-sur-Yvette cedex, France

Package ‘RGCCA’

October 15, 2010

Type Package
Title Regularized Generalized Canonical Correlation Analysis
Version 1.0
Date 2010-06-08
Author Arthur Tenenhaus
Maintainer Arthur Tenenhaus <arthur.tenenhaus@supelec.fr>
Description Regularized Generalized Canonical Correlation Analysis
Multiblock conclusions

► Depending on the dimension of the blocks, you can use either the primal or the dual algorithm.

► The dual representation of the RGCCA algorithm allows:

  • Analysing high dimensional blocks.
  • Recovering nonlinear relationships between blocks (choice of the kernel function).
  • Handling any type of data (e.g. histogram) as long as relevant kernel is defined.

► Sparse constraints are useful when the relevant variables are masked by (too many) noisy variables.
Structured variable selection for RGCCA

Functional MRI
Intermediate phenotype

Gene Expression
Genotype

Behavioral data
(Clinic, psychometric)
Final phenotype
Structured variable selection within RGCCA

\[
\text{argmax}_{a_1,a_2,...,a_J} \sum_{j \neq k} c_{jk} g \left( \text{cov}(X_j a_j, X_k a_k) \right)
\]

subject to

\[
\begin{cases}
    a_j^t M_j a_j = 1, j = 1, ..., J \\
    \Omega(a_j) \leq c_j, j = 1, ..., J
\end{cases}
\]

- LASSO: \( \Omega(a_j) = \|a_j\|_1 \)
- Fused LASSO: \( \Omega(a_j) = \sum_{k=1}^{p_j} |a_{jk}| + \lambda \sum_{k=1}^{p_j} |a_{jk} - a_{j,k-1}| \)
- Group LASSO: \( \Omega(a_j) = \sum_{g \in G} \|a_g\|_2 \)
- Graph LASSO \iff\ Group LASSO with overlap

**Part II: Multi-group analysis**

**Setting:** The same set of variables is observed on individuals structured in groups. Usually groups are centered and normalized (unit norm).

**Objective:** Investigate the relationships between variables within the various groups.

$$\arg\max_{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_I} \sum_{i,l,i \neq l} c_{il} g(\langle \mathbf{X}_i^t \mathbf{X}_i \mathbf{a}_i, \mathbf{X}_l^t \mathbf{X}_l \mathbf{a}_l \rangle)$$

s.c. $$(1 - \tau_i) \|\mathbf{X}_i \mathbf{a}_i\|^2 + \tau_i \|\mathbf{a}_i\|^2 = 1, \ i = 1, \ldots, I$$
Invited Review

Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis

Arthur Tenenhaus\textsuperscript{a,}\textsuperscript{*}, Michel Tenenhaus\textsuperscript{b,1}

\textsuperscript{a}SUPELEC, Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France
\textsuperscript{b}HEC Paris, 1 rue de la Libération, 78351 Jouy-en-Josas Cedex, France

\begin{abstract}
This paper presents an overview of methods for the analysis of data structured in blocks of variables or in groups of individuals. More specifically, regularized generalized canonical correlation analysis (RGCCA), which is a unifying approach for multiblock data analysis, is extended to be also a unifying tool for multigroup data analysis. The versatility and usefulness of our approach is illustrated on two real datasets.
\end{abstract}
From Multi-block to Multi-Way analysis

- Multiway Fisher Discriminant Analysis
- Multiway Logistic Regression
- Multiway RGCCA

In collaboration with Laurent Le Brusquet and Gisela Lechuga (PhD Student)