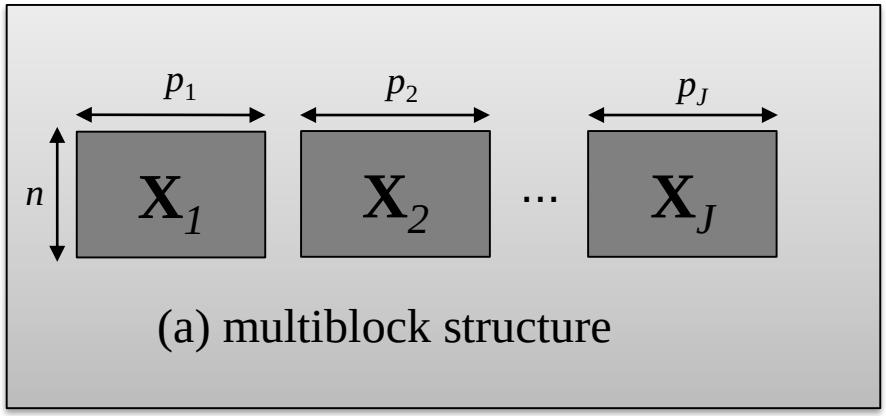
Structured data analysis with RGCCA

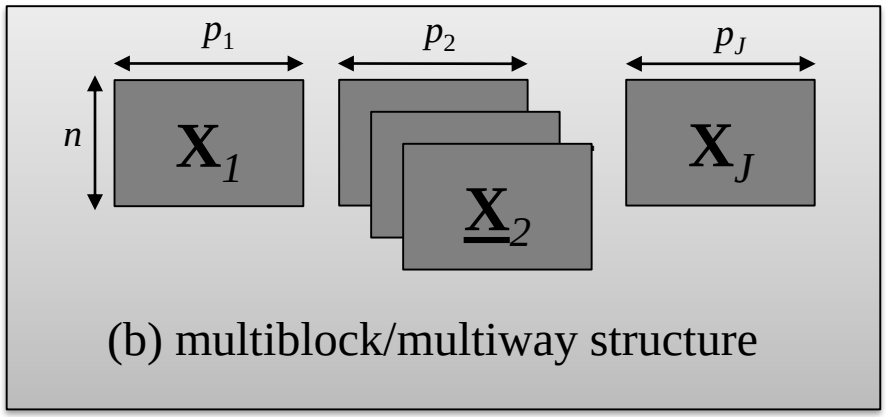
Arthur Tenenhaus

Séminaire de Statistique appliquée du CNAM, 16/03/2018

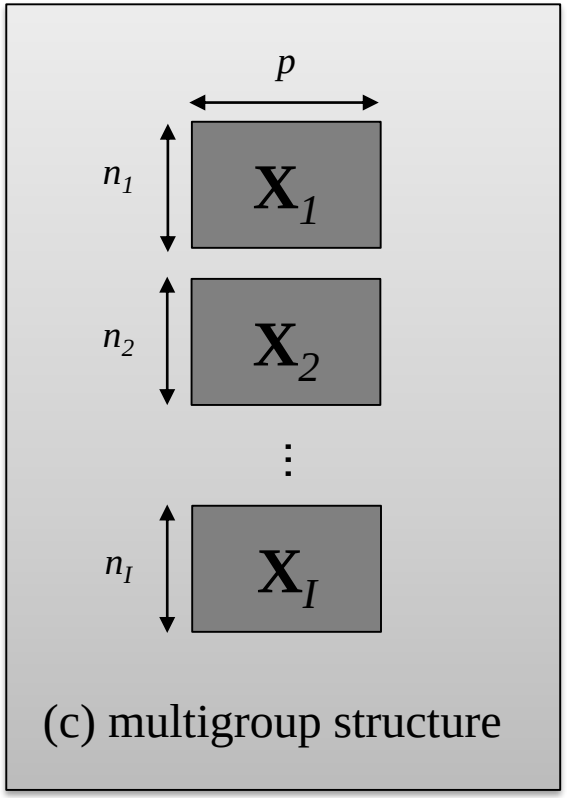
CentraleSupélec



(a) multiblock structure



(b) multiblock/multiway structure



(c) multigroup structure

Spinocerebellar Ataxia (SCA) dataset

Brain and Spine Institute

p_1 ~20000 | p_2 ~20 | p_3 ~50 | p_4 ~10

MET_1 MET_2 MET_3 … $MET\ p_1$ | MRS_1 MRS_2 … $MRS\ p_2$ | VOL_1 VOL_2 … $VOL\ p_3$ | CLI_1 … $CLI\ p_4$

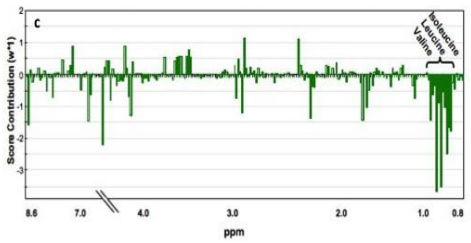
n=83

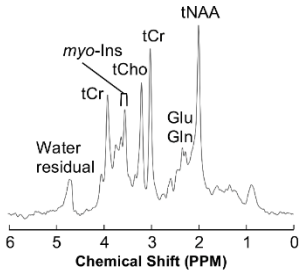
Metabolomic data (X_1)

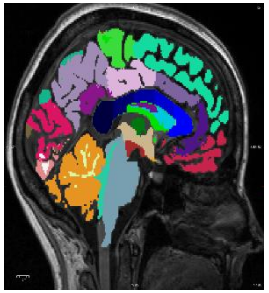
Magnetic Resonance Spectroscopy (X_2)

Volumetric data (X_3)

Clinical data (X_4)





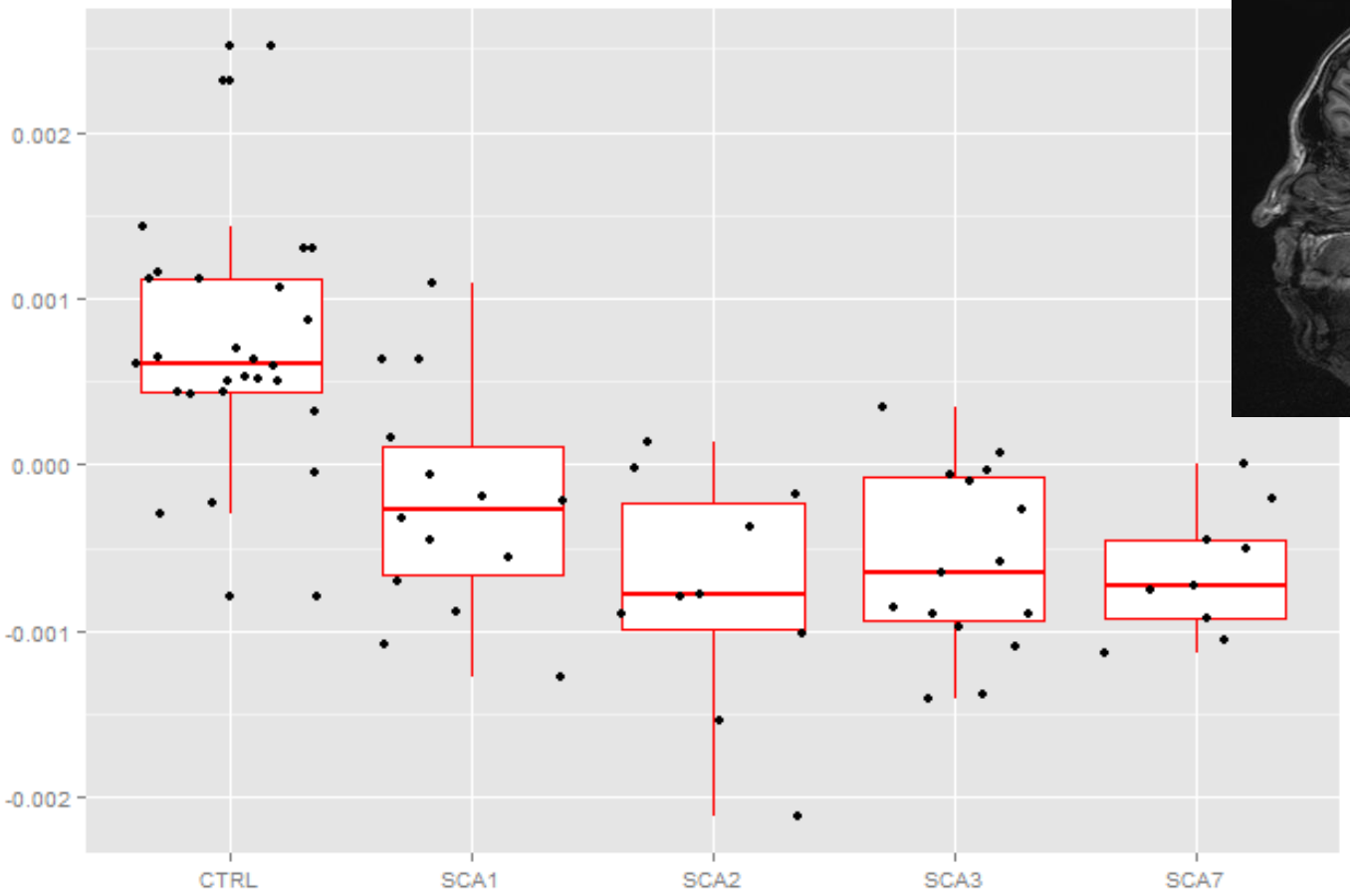


Free surfer segmentation and manual segmentation for vermis and pons

Status, Age of onset, Calorimetry, BMI, CAG length, SARA score, CCSF score age, sex, etc…

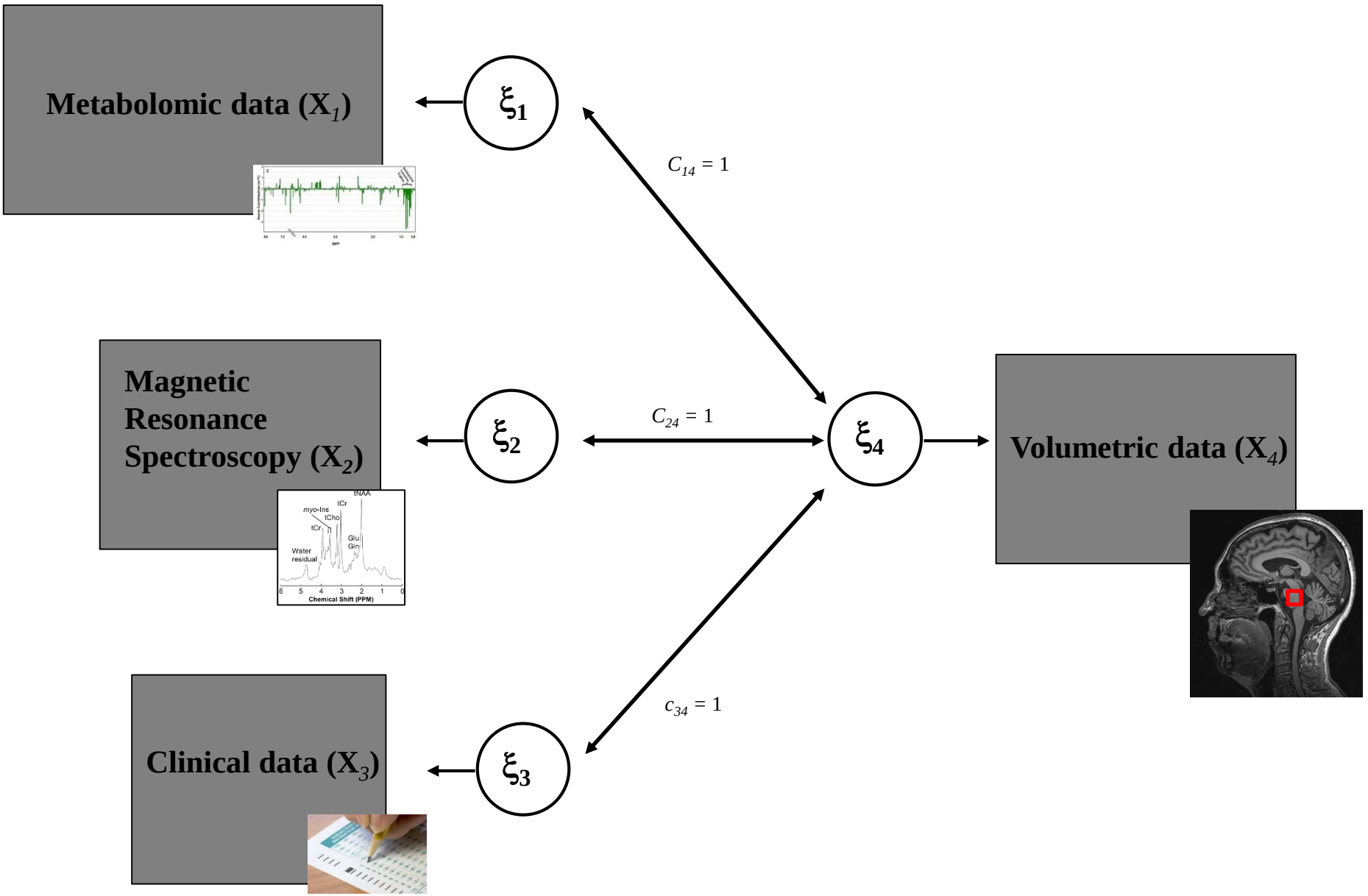
The objective of multiblock component methods is to find block components summarizing the relevant information between and within the blocks.

Volume of the pons vs groups

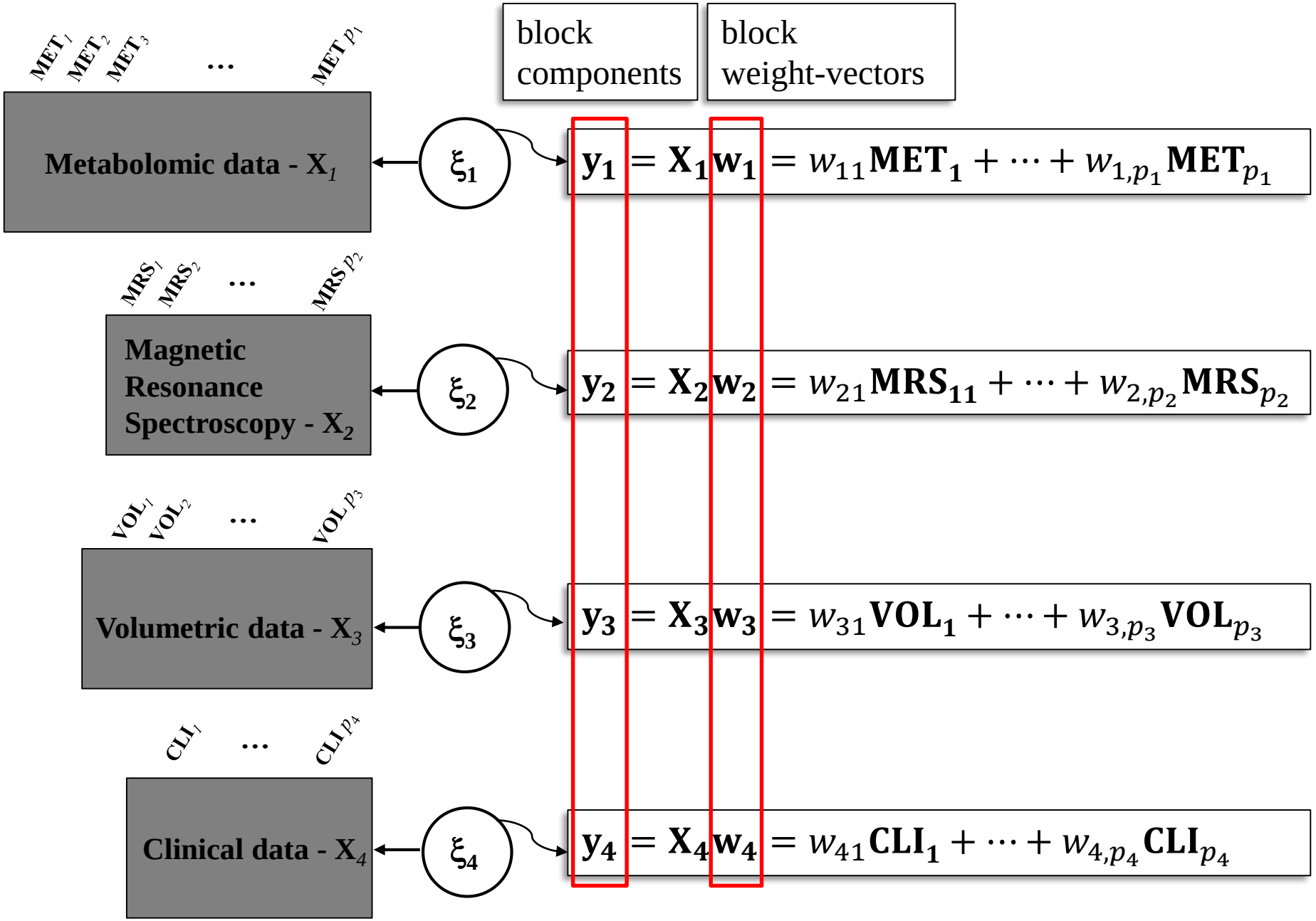


Objective. Identify a set of variables within each block that influence the volume of the pons while accounting for the multiblock structure of the SCA dataset.

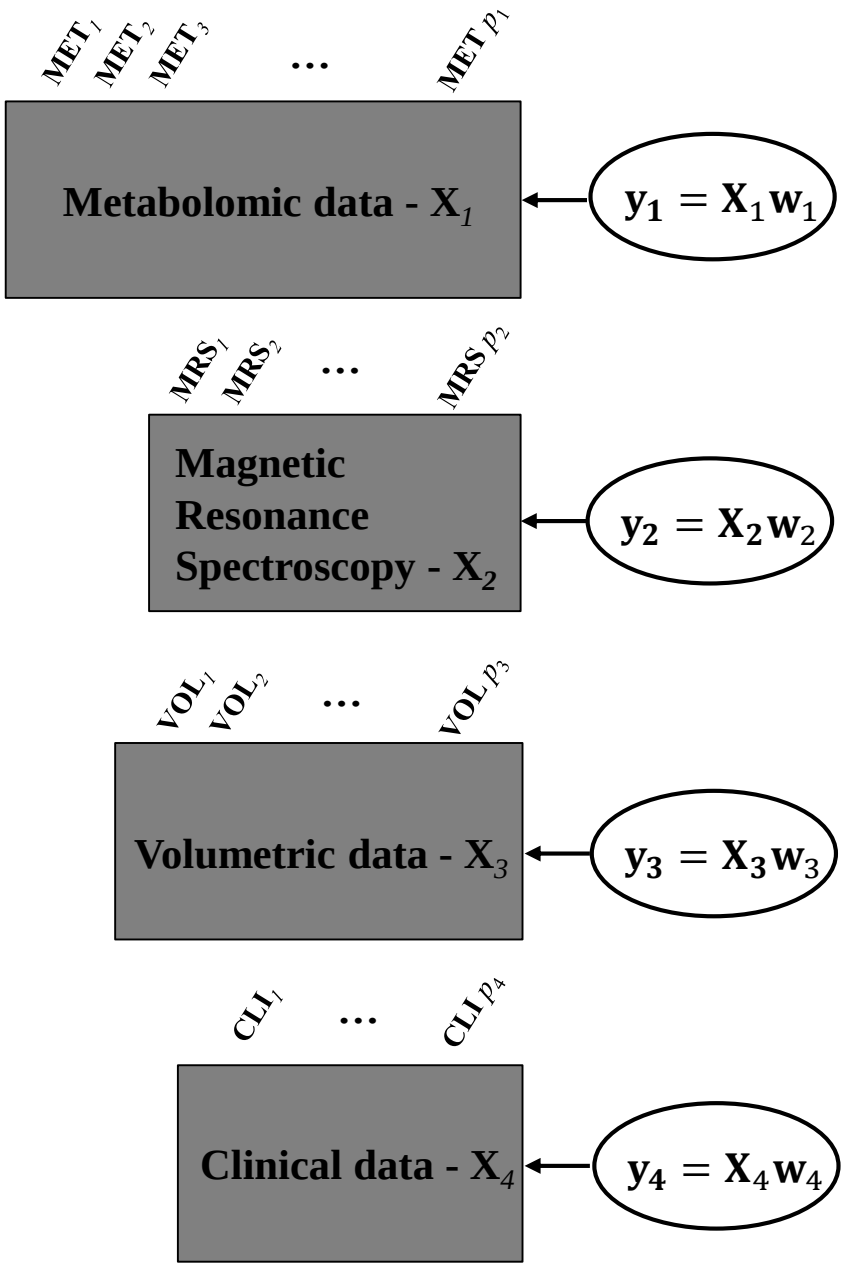
The SCA dataset from an RGCCA point of view



The philosophy of multiblock component methods



The philosophy of multiblock component methods

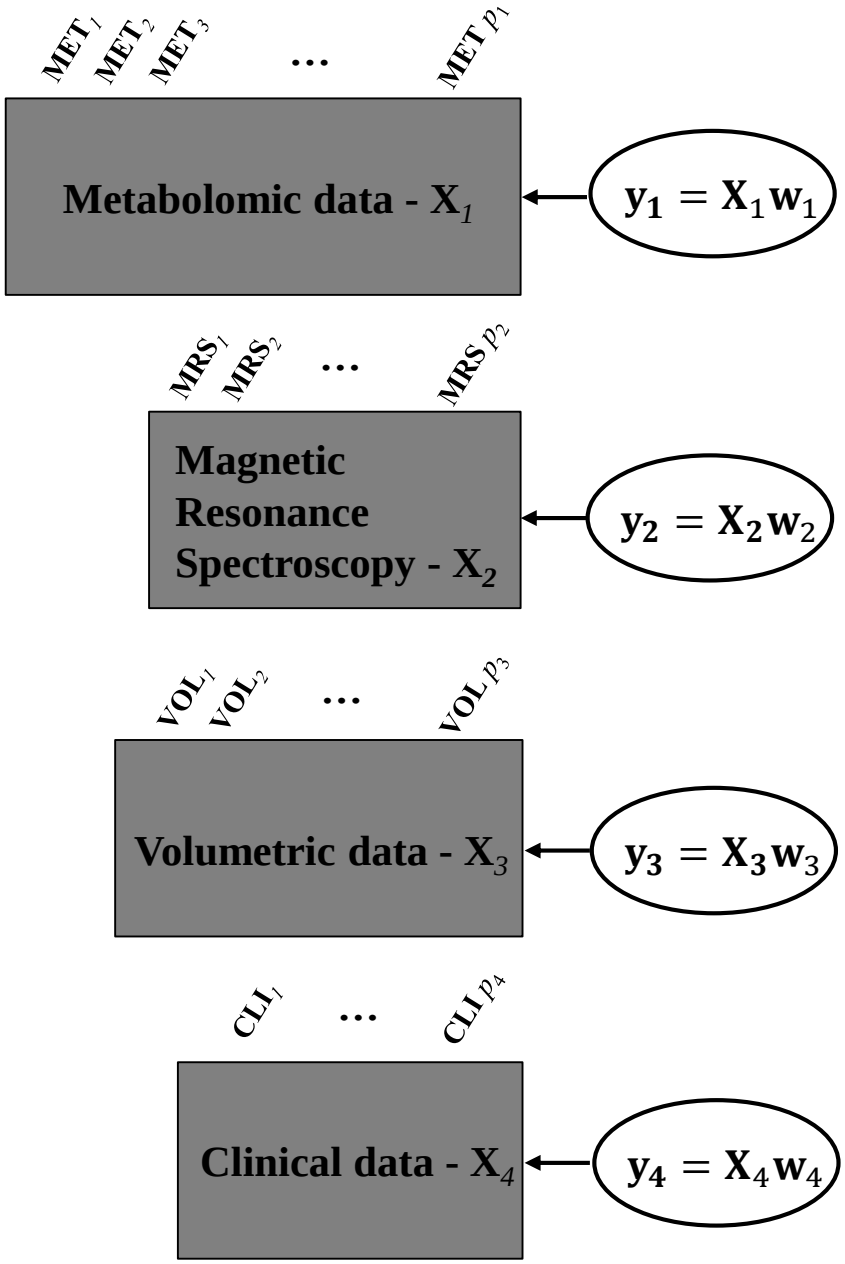


Block components should verified two properties at the same time:

1. Block components well explain their own block.

2. Block components are as correlated as possible for connected blocks.

The philosophy of multiblock component methods



Correlation based methods

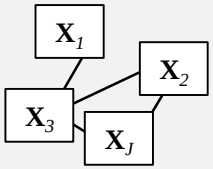
Find block-weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_J$ maximizing a function of $\mathbf{\Phi} = \{\text{cor}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)\}$.

Covariance based methods

Find block-weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_J$ maximizing a function of $\mathbf{\Psi} = \{\text{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)\}$.

BLOCKS ARE PARTIALLY CONNECTED

$c_{jk} = 1$ if $\mathbf{X}_j \leftrightarrow \mathbf{X}_k$, 0 otherwise



SUMCOR	$\max_{\text{var}(\mathbf{X}_j\mathbf{w}_j)=1} \sum_{j,k} c_{jk}\text{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)$
SSQCOR	$\max_{\text{var}(\mathbf{X}_j\mathbf{w}_j)=1} \sum_{j,k} c_{jk}\text{cov}^2(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)$
SABSCOR	$\max_{\text{var}(\mathbf{X}_j\mathbf{w}_j)=1} \sum_{j,k} c_{jk}\left\lvert\text{cov}(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k)\right\rvert$

RGCCA for multiblock analysis

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k}^{J} c_{jk} g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

$$\text{s.t.}\left(1-\tau_j\right)\mathrm{var}\left(\mathbf{X}_j\mathbf{w}_j\right) + \tau_j\left\|\mathbf{w}_j\right\|_2^2 = 1\,, j = 1,\ldots,J$$

- $c_{jk} = 1$ if $\mathbf{X}_j \leftrightarrow \mathbf{X}_k, 0$ otherwise
- $g =$ any convex function $-$ e.g. $\begin{cases} g(x) = x & \text{(Horst sheme)} \\ g(x) = x^2 & \text{(Factorial scheme)} \\ g(x) = |x| & \text{(Centroid scheme)} \end{cases}$
- $0 \le \tau_j \le 1$ continuum between correlation and covariance

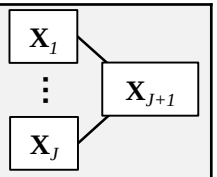
Tenenhaus M, Tenenhaus A, Groenen PJF, (2017) Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods, Psychometrika, vol. 82, no. 3, 737–777

Tenenhaus A. and Guillemot V. (2017): RGCCA Package. http://cran.project.org/web/packages/RGCCA/index.html

Tenenhaus A, Philippe C, Frouin V (2015) Kernel generalized canonical correlation analysis, Computational Statistics & Data Analysis, vol. 90, pp. 114-131.

Tenenhaus A, Tenenhaus M (2011) Regularized generalized canonical correlation analysis, vol. 76, pp. 257-284, Psychometrika.

BLOCKS ARE CONNECTED TO THE SUPERBLOCK $\mathbf{X}_{J+1} = [\mathbf{X}_1, \ldots, \mathbf{X}_J]$



MAXVAR

Horst (1961, 1965)
Carroll (1968)
Kettenring (1971)

$$\max_{\mathbf{w}_j} \sum_{j=1}^{J} \operatorname{cor}^2\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_{J+1}\mathbf{w}_{J+1}\right) \quad \left(= \max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \lambda_1(\boldsymbol{\Phi})\right)$$

MAXVAR-A

Covariance criterion (Carroll, 1968b)
Split principal component (Lohmöller, 1989)
Multiple factor analysis (Escofier, Pagès, 1994)
Multiple Co-Inertia Analysis (Chessel, Hanafi, 1996)
Consensus PCA (Westerhuis, Kourti, MacGregor, 1998)
CPCA-W (Smilde, Westerhuis, de Jong, 2001)
SUM-PCA (Smilde, Westerhuis, de Jong, 2001)

$$\max_{\substack{\|\mathbf{w}_j\|=1 \\ \operatorname{var}(\mathbf{X}_{J+1}\mathbf{w}_{J+1})=1}} \sum_{j=1}^{J} \operatorname{cov}^2\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_{J+1}\mathbf{w}_{J+1}\right)$$

$$\left(= \max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \lambda_1(\boldsymbol{\Psi})\right)$$

HIERARCHICAL-PCA

Wold, Kettaneh, Tjessem (1996)
Hanafi, Kohler and Qannari (2010)

$$\max_{\substack{\|\mathbf{w}_j\|=\cdots=\|\mathbf{w}_J\|=1 \\ \operatorname{var}(\boldsymbol{X}_{J+1}\mathbf{w}_{J+1})=1}} \sum_{j=1}^{J} \operatorname{cov}^4(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_{J+1}\mathbf{w}_{J+1})$$

2-block special cases

$$\max_{\mathbf{w}_1,\mathbf{w}_2} \operatorname{cov}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) \quad \text{s.t.} \quad (1-\tau_j)\operatorname{var}(\mathbf{X}_j\mathbf{w}_j) + \tau_j \|\mathbf{w}_j\|_2^2 = 1\,, j = 1,2$$

Methods	Criterion	Constraints
PLS regression	$\max \operatorname{cov}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$	$\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$
Canonical Correlation Analysis	$\max \operatorname{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2)$	$\operatorname{var}(\mathbf{X}_1\mathbf{w}_1) = \operatorname{var}(\mathbf{X}_2\mathbf{w}_2) = 1$
Redundancy analysis of X1 with respect to X2	$\max \operatorname{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) \times \operatorname{var}(\mathbf{X}_1\mathbf{w}_1)$	$\operatorname{var}(\mathbf{X}_2\mathbf{w}_2) = 1$ $\|\mathbf{w}_1\| = 1$

Components $\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{X}_2\mathbf{w}_2$ $\mathbf{X}_2\mathbf{a}_2$ are well correlated.

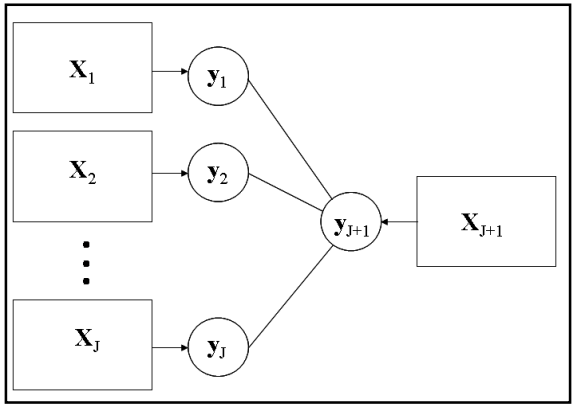
1st component is stable

No stability condition for 2nd component

Hierarchical PCA with RGCCA

Hierarchical PCA optimization problem: $\max_{\substack{\|\mathbf{a}_j\|=\cdots\|\mathbf{a}_J\|=1 \\ \mathrm{var}(\boldsymbol{X}_{J+1}\mathbf{a}_{J+1})=1}} \sum_{j=1}^{J} \mathrm{cov}^4(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_{J+1}\mathbf{w}_{J+1})$

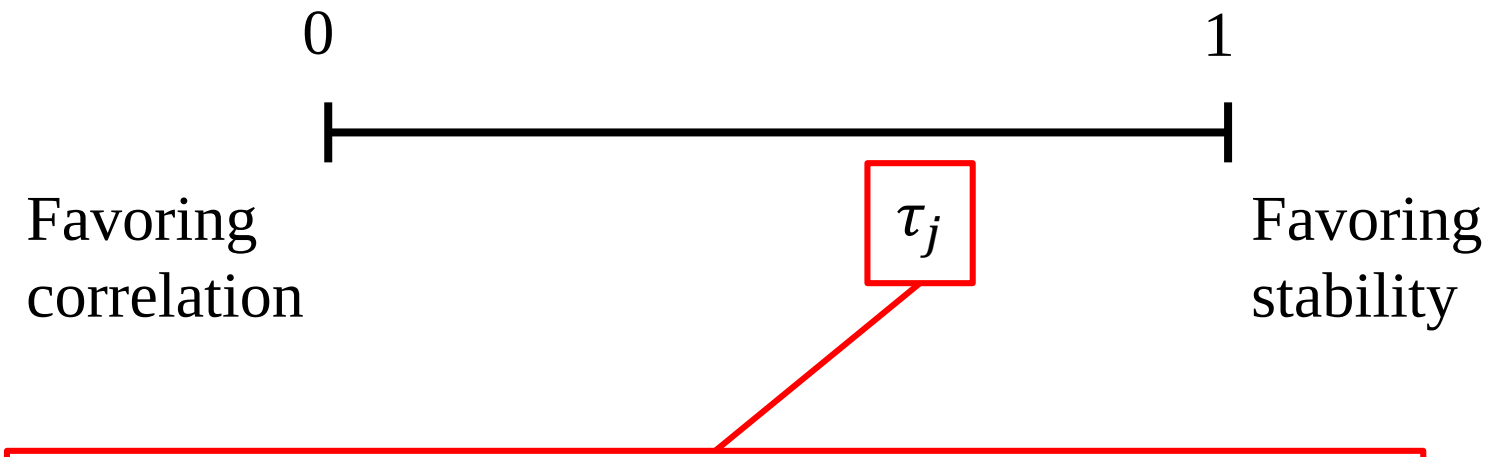
- Scheme function: $g(x) = x^4$
- Shrinkage parameters: $\tau_1 = \cdots = \tau_J = 1$ and $\tau_{J+1} = 0$
- Design matrix: $\mathbf{C} = \begin{array}{c} \\ \mathbf{X}_1 \\ \\ \mathbf{X}_J \\ \mathbf{X}_{J+1} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{X}_1 & & \mathbf{X}_J & \mathbf{X}_{J+1} \end{array} \\ \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \end{array}$



Choice of the shrinkage constant τ_j

$$\max_{\mathbf{a}_1,\ldots,\mathbf{a}_J} \sum_{j,k}^{J} c_{jk} g\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

$$\text{s.t. } \mathbf{w}_j^\top \left((1-\tau_j) n^{-1}\mathbf{X}_j^\top\mathbf{X}_j + \tau_j \mathbf{I}_{p_j}\right)\mathbf{w}_j = 1\,, j = 1, \ldots, J$$



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

RGCCA for multiblock analysis

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k}^{J} c_{jk} g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

$$\text{s.t.}\left(1-\tau_j\right)\mathrm{var}\left(\mathbf{X}_j\mathbf{w}_j\right) + \tau_j\left\|\mathbf{w}_j\right\|_2^2 = 1\,, j = 1,\ldots,J$$

Two key ingredients:

(i) Block relaxation

(ii) Majorization by Minorization (MM)

Tenenhaus M, Tenenhaus A, Groenen PJF, (2017) Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods, Psychometrika, vol. 82, no. 3, 737–777

Tenenhaus A. and Guillemot V. (2017): RGCCA Package. http://cran.project.org/web/packages/RGCCA/index.html

Tenenhaus A, Philippe C, Frouin V (2015) Kernel generalized canonical correlation analysis, Computational Statistics & Data Analysis, vol. 90, pp. 114-131.

Tenenhaus A, Tenenhaus M (2011) Regularized generalized canonical correlation analysis, vol. 76, pp. 257-284, Psychometrika.

Block relaxation: from $\mathbf{w}^s$ to $\mathbf{w}^{s+1}$

$$\boldsymbol{w}^s = \left(\mathbf{w}_1^s, \mathbf{w}_2^s, \ldots, \mathbf{w}_J^s\right) \qquad \underset{\mathbf{w}_1, \mathbf{w}_1^\top \mathbf{M} \mathbf{w}_1 = 1}{\operatorname{argmax}} \; h\left(\mathbf{w}_1, \mathbf{w}_2^s, \ldots, \mathbf{w}_J^s\right) \qquad \rightarrow \; \mathbf{w}_1^{s+1}$$

$$\vdots$$

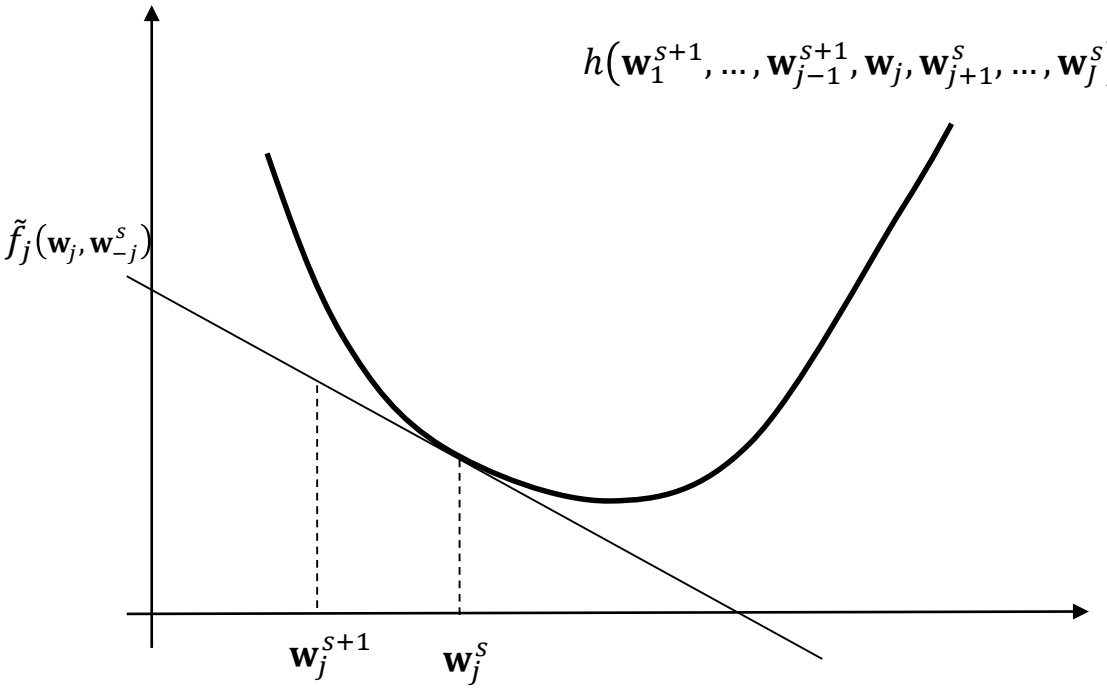
$$\rightarrow \; \mathbf{w}_j^{s+1}$$

$$\vdots$$

$$\rightarrow \; \mathbf{w}_J^{s+1}$$

MM approach for RGCCA

$$\max_{\mathbf{w}_j} h\left(\mathbf{w}_1^{s+1}, \ldots, \mathbf{w}_{j-1}^{s+1}, \mathbf{w}_j, \mathbf{w}_{j+1}^{s}, \ldots, \mathbf{w}_J^{s}\right) \quad \text{s.t.} \quad \mathbf{w}_j^\top \mathbf{M}_j \mathbf{w}_j = 1$$



$$\mathbf{w}_j^{s+1} = \underset{\mathbf{w}_j, \mathbf{w}_j^\top \mathbf{M}_j \mathbf{w}_j = 1}{\operatorname{argmax}} \tilde{f}_j\left(\mathbf{w}_j, \mathbf{w}_{-j}^{s}\right)$$

primal algorithm $n > p_j$

dual algorithm $n < p_j$

- Tenenhaus A. and Tenenhaus M., Regularized Generalized Canonical Correlation Analysis, Psychometrika, vol. 76, Issue 2, pp. 257-284, 2011
- Tenenhaus M, Tenenhaus A, Groenen P.J.F, (2017) Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods, Psychometrika, doi: 10.1007/s11336-017-9573-x
- Tenenhaus, A., Philippe, C., Frouin, V. (2015). Kernel Generalized Canonical Correlation Analysis. Computational Statistics & Data Analysis, 90, 114-131.
- Tenenhaus A. and Guillemot V. (2017): RGCCA Package. http://cran.project.org/web/packages/RGCCA/index.html

Properties of the RGCCA algorithm for multiblock data

► Monotone convergence: $h\left(\mathbf{w}_1^{s+1}, \ldots, \mathbf{w}_J^{s+1}\right) \geq h\left(\mathbf{w}_1^{s}, \ldots, \mathbf{w}_J^{s}\right)$.

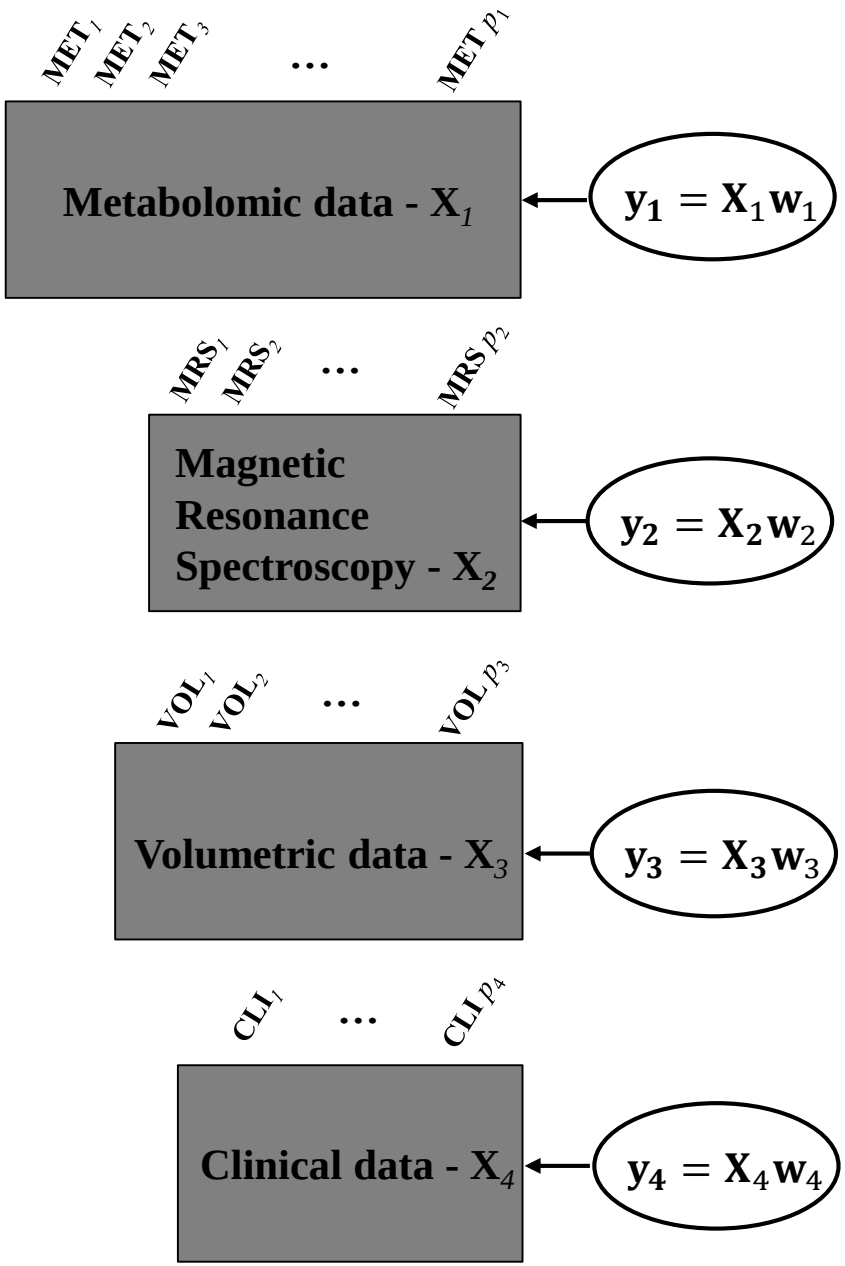
In addition, assuming uniqueness of the solution of the MM step, the following properties hold:

► The sequence $\{\mathbf{w}^s\}$ is asymptotically regular: $\lim_{s\to\infty} \|\mathbf{w}^{s+1} - \mathbf{w}^s\| = 0$.

► At convergence, a stationary point is obtained.

- Tenenhaus M, Tenenhaus A, Groenen P.J.F, (2017) Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods, Psychometrika, doi: 10.1007/s11336-017-9573-x
- Tenenhaus A. and Guillemot V. (2017): RGCCA Package. http://cran.project.org/web/packages/RGCCA/index.html

multiblock component methods with sparsity



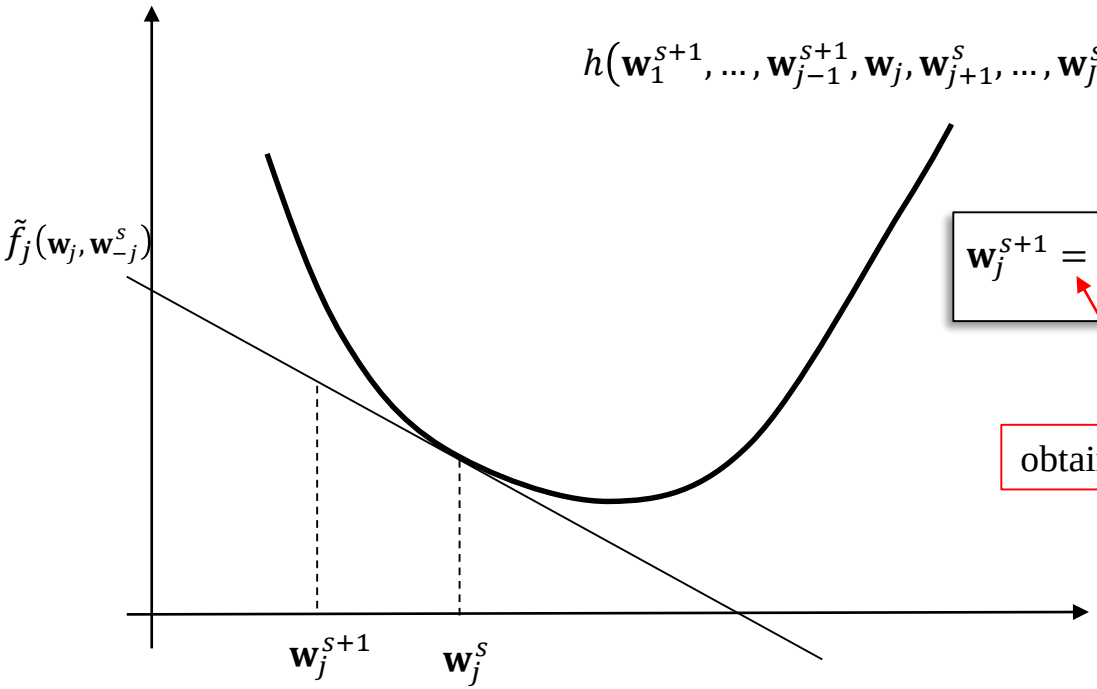
Block components should verified two properties at the same time:

1. Block components well explain their own block.
2. Block components are as correlated as possible for connected blocks.
3. Block components are built from sparse $\mathbf{a}_j$

extension of RGCCA for sparse solution

$$\max_{\mathbf{w}_j} h\left(\mathbf{w}_1^{s+1}, \ldots, \mathbf{w}_{j-1}^{s+1}, \mathbf{w}_j, \mathbf{w}_{j+1}^{s}, \ldots, \mathbf{w}_J^{s}\right) \quad \text{s.t.} \quad \left\|\mathbf{w}_j\right\|_2^2 = 1 \;\&\; \left\|\mathbf{w}_j\right\|_1 \leq s_j$$

sparsity constant



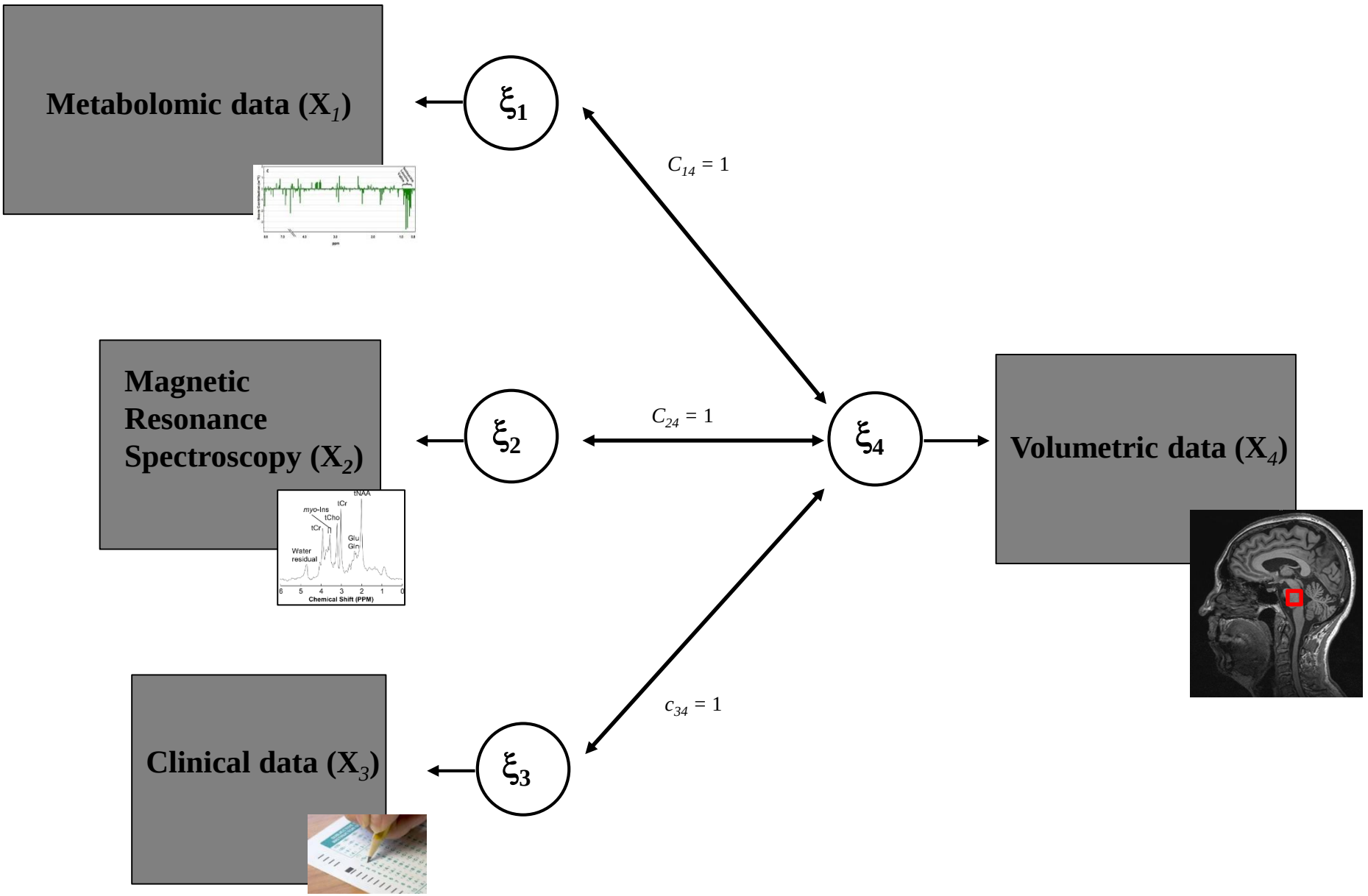
$$\mathbf{w}_j^{s+1} = \underset{\mathbf{w}_j}{\operatorname{argmax}}\, \tilde{f}\left(\mathbf{w}_j, \mathbf{w}_{-j}^s\right) \text{ s.t. } \left\|\mathbf{w}_j\right\|_2 = 1 \text{ and } \left\|\mathbf{w}_j\right\|_1 \leq s_j$$

obtained by soft-thresholding

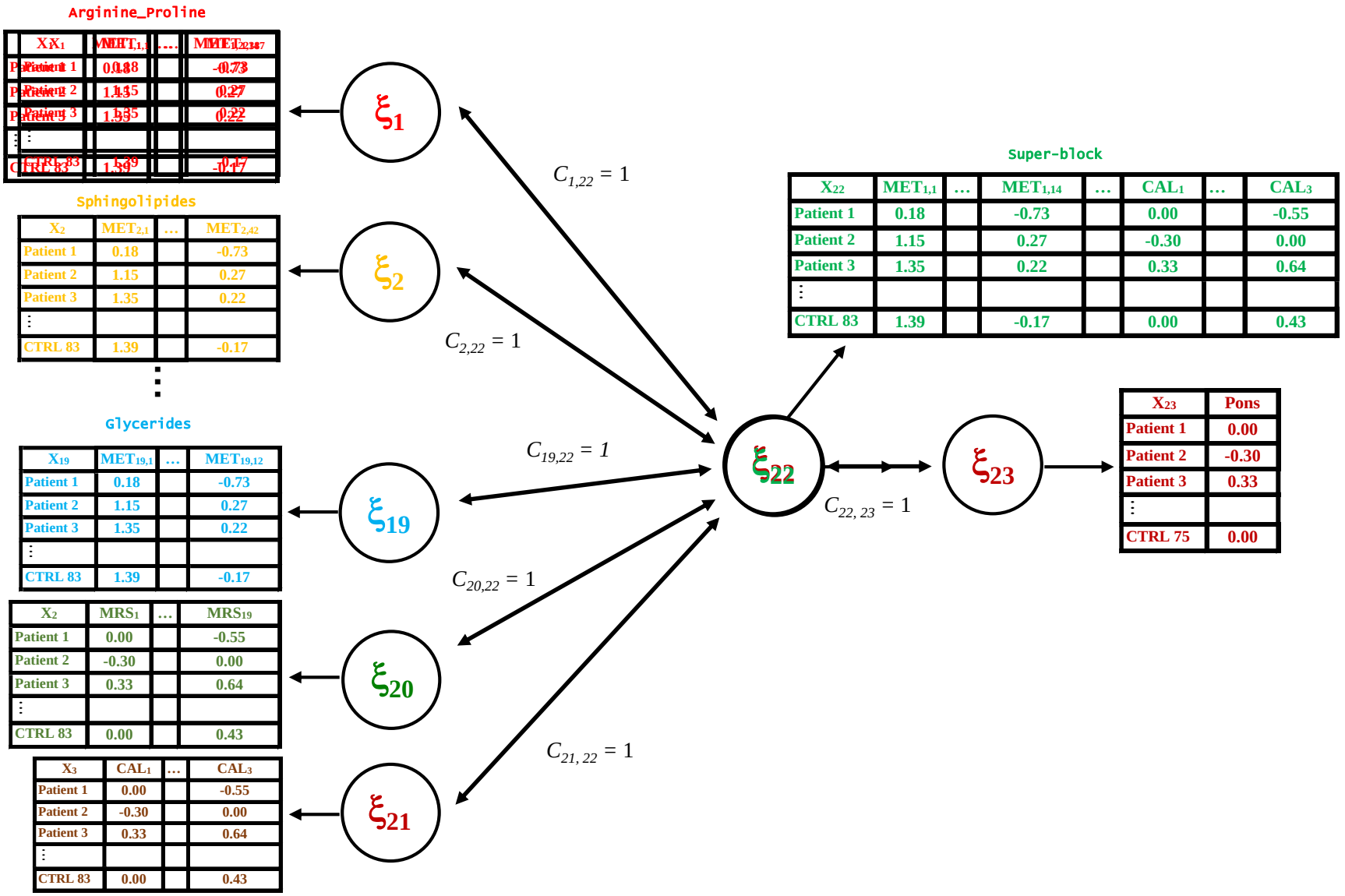
Tenenhaus A., Philippe C., Guillemot V, et al..(2014). Variable Selection for Generalized Canonical Correlation Analysis, Biostatistics, 15 (3) : 569-583

Tenenhaus A. and Guillemot V. (2017): RGCCA Package. http://cran.project.org/web/packages/RGCCA/index.html

The SCA dataset from an RGCCA point of view



Divide and conquer strategy with SGCCA



X_1	$MET_{1,1}$	…	$MET_{1,147}$
Patient 1	0.18		-0.73
Patient 2	1.15		0.27
Patient 3	1.35		0.22
⋮			
CTRL 83	1.39		-0.17

X_2	$MET_{2,1}$	…	$MET_{2,42}$
Patient 1	0.18		-0.73
Patient 2	1.15		0.27
Patient 3	1.35		0.22
⋮			
CTRL 83	1.39		-0.17

X_{19}	$MET_{19,1}$	…	$MET_{19,12}$
Patient 1	0.18		-0.73
Patient 2	1.15		0.27
Patient 3	1.35		0.22
⋮			
CTRL 83	1.39		-0.17

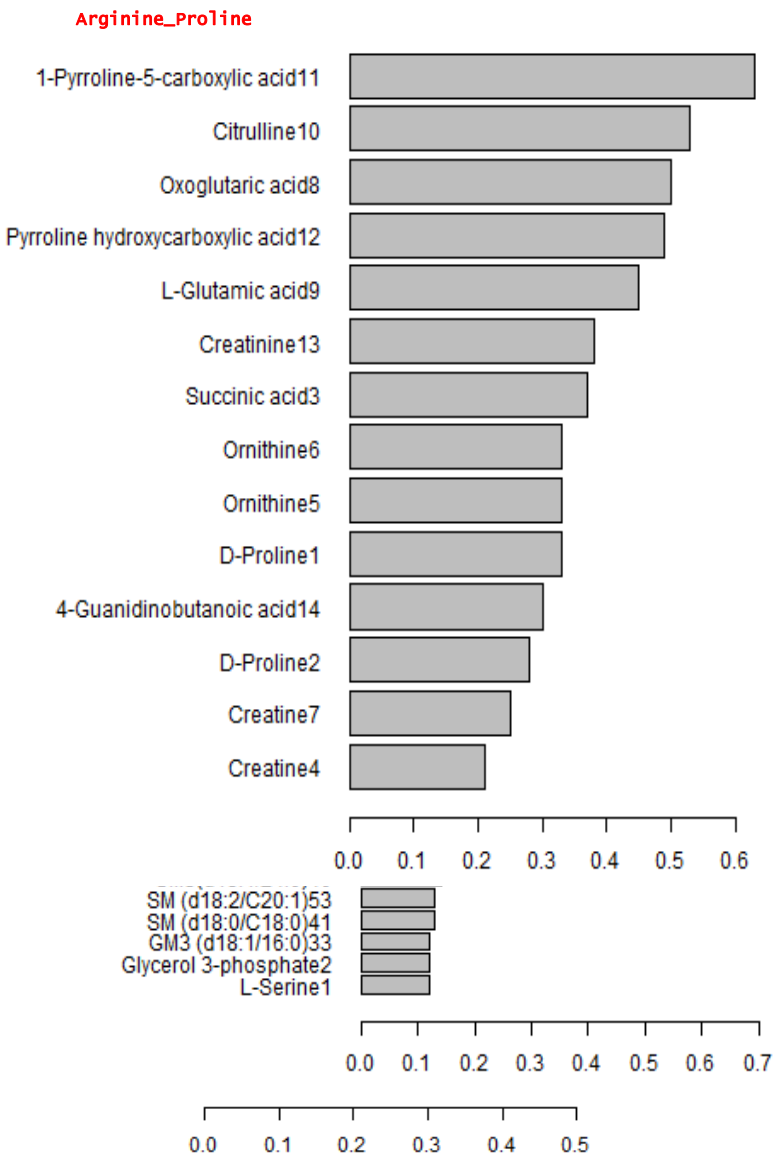
X_2	MRS_1	…	MRS_{19}
Patient 1	0.00		-0.55
Patient 2	-0.30		0.00
Patient 3	0.33		0.64
⋮			
CTRL 83	0.00		0.43

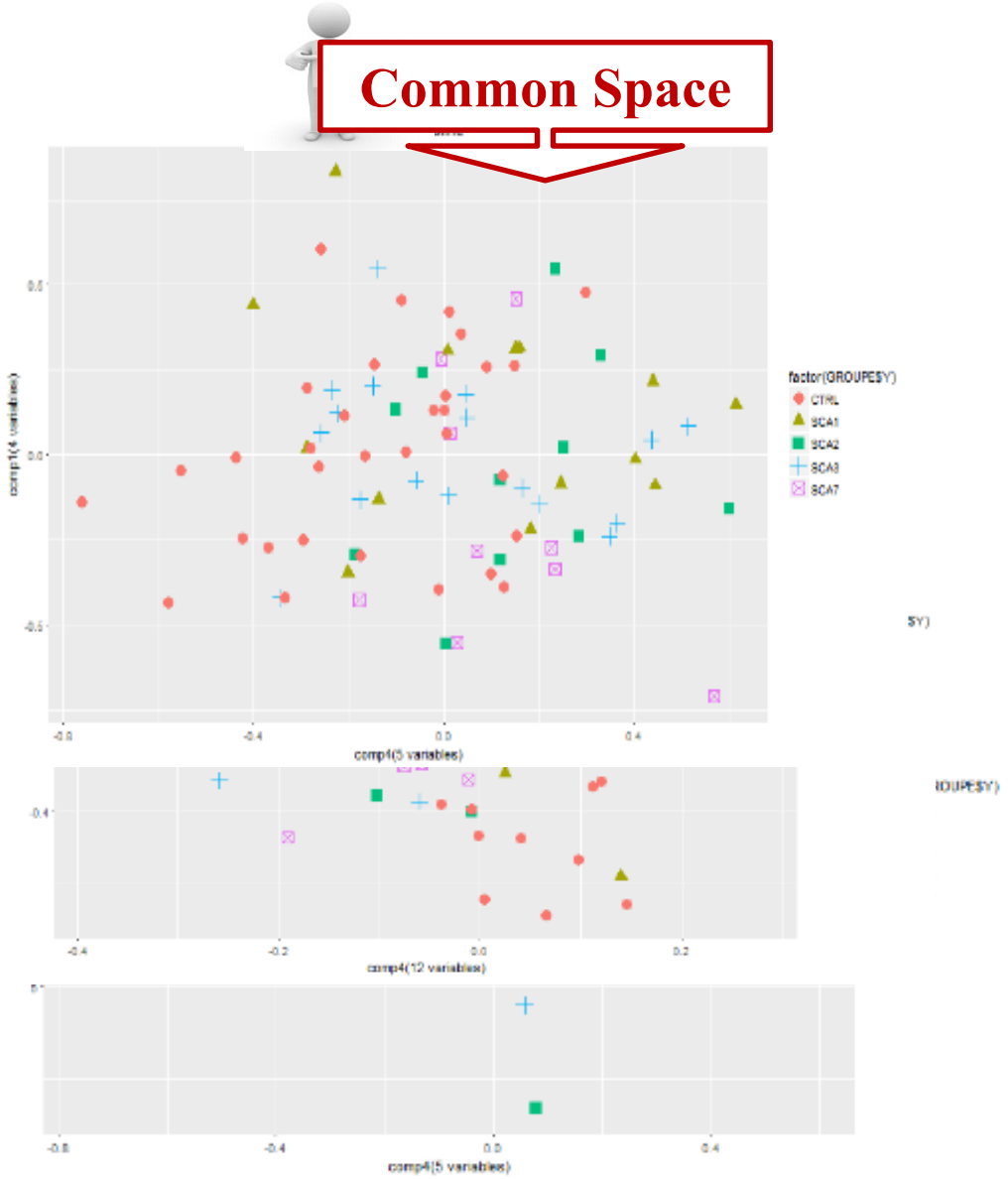
X_3	CAL_1	…	CAL_3
Patient 1	0.00		-0.55
Patient 2	-0.30		0.00
Patient 3	0.33		0.64
⋮			
CTRL 83	0.00		0.43

X_{22}	$MET_{1,1}$	…	$MET_{1,14}$	…	CAL_1	…	CAL_3
Patient 1	0.18		-0.73		0.00		-0.55
Patient 2	1.15		0.27		-0.30		0.00
Patient 3	1.35		0.22		0.33		0.64
⋮							
CTRL 83	1.39		-0.17		0.00		0.43

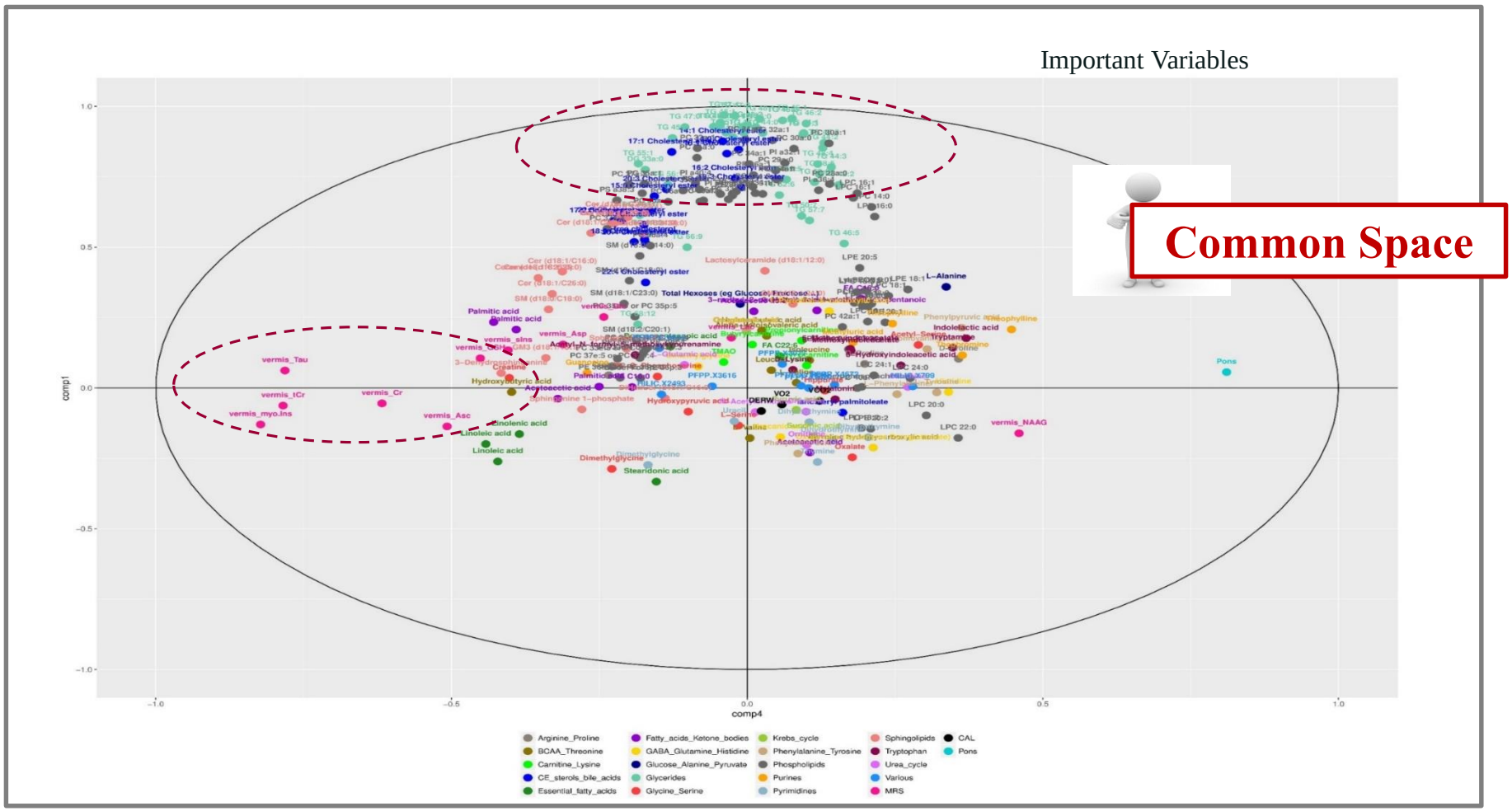
X_{23}	Pons
Patient 1	0.00
Patient 2	-0.30
Patient 3	0.33
⋮	
CTRL 75	0.00

Visualization





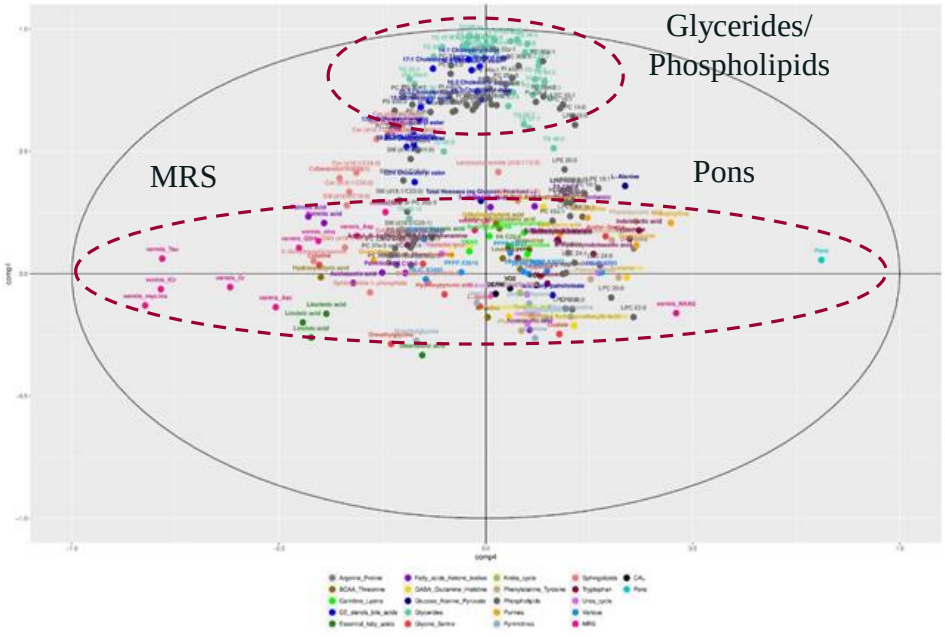
Important Variables

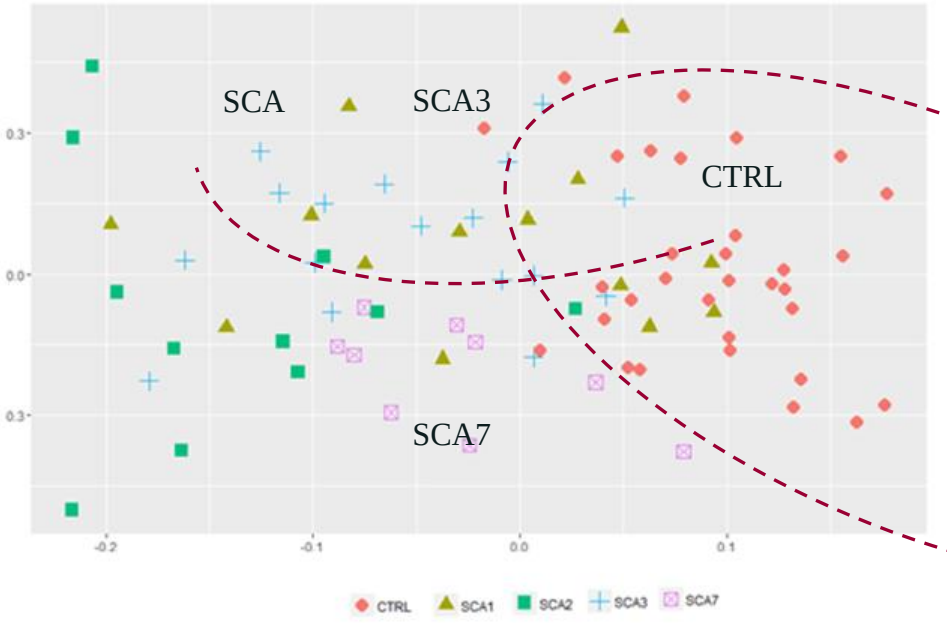


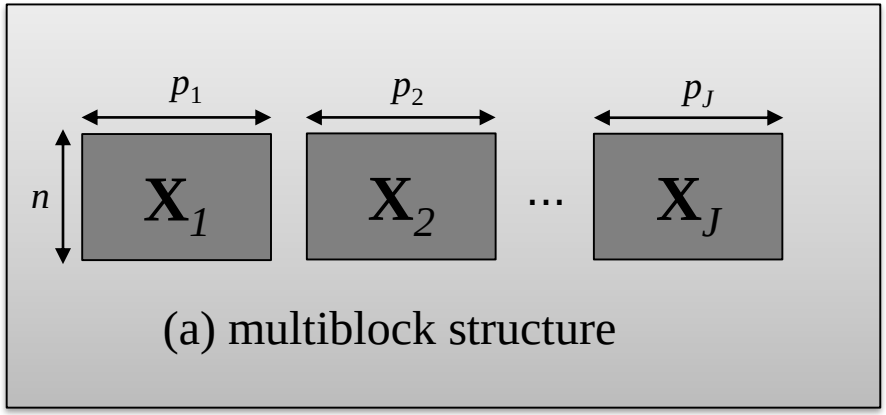
Common Space

Variables Space

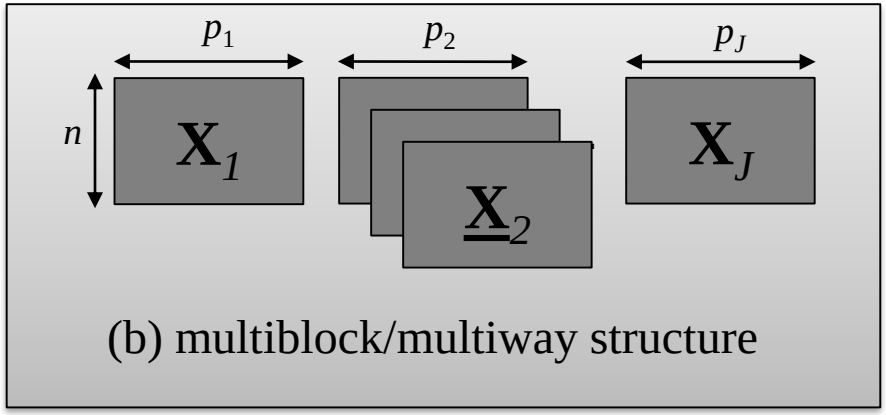
Individuals Space



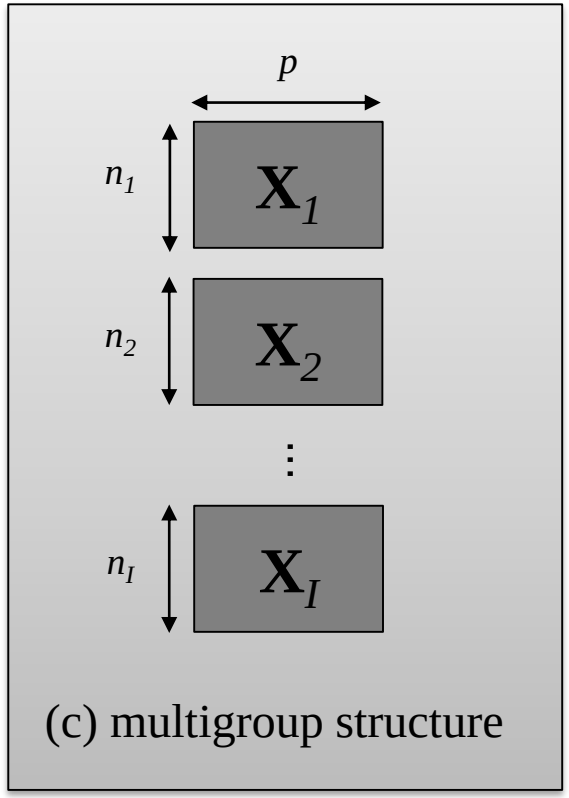




(a) multiblock structure



(b) multiblock/multiway structure

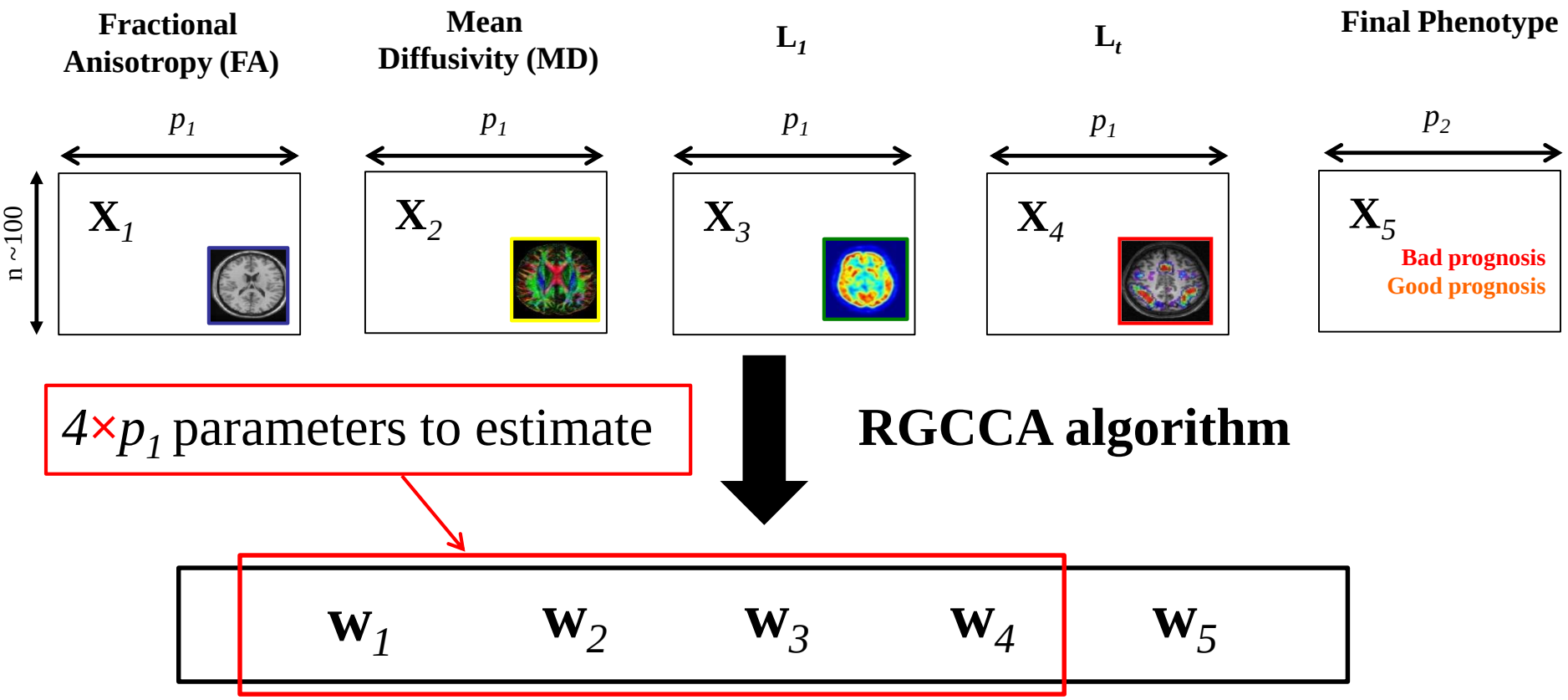


(c) multigroup structure

The COMA project from a multiblock viewpoint

(Brain and Spine Institute, La pitié Salpêtrière Hospital)

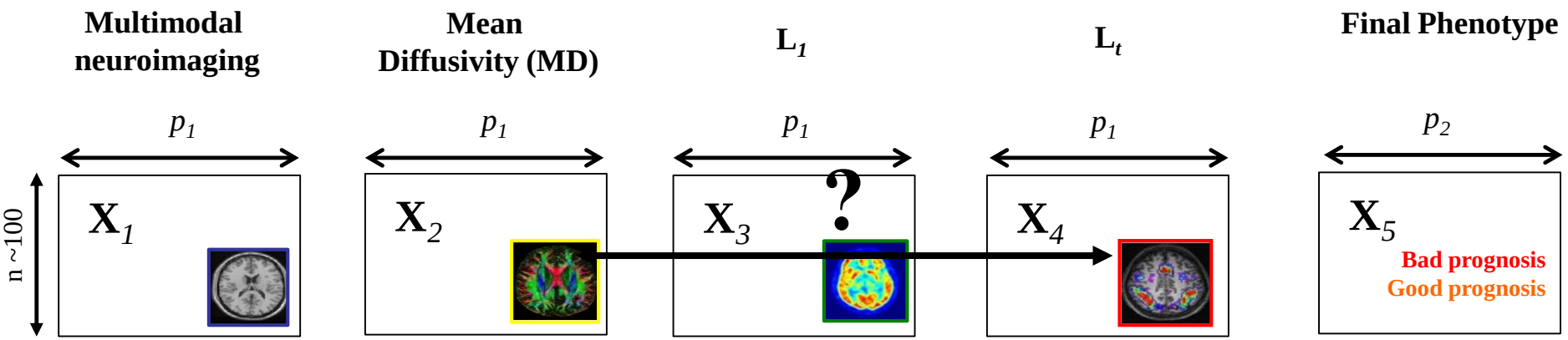
Predict the long term recovery of patients after traumatic brain injury



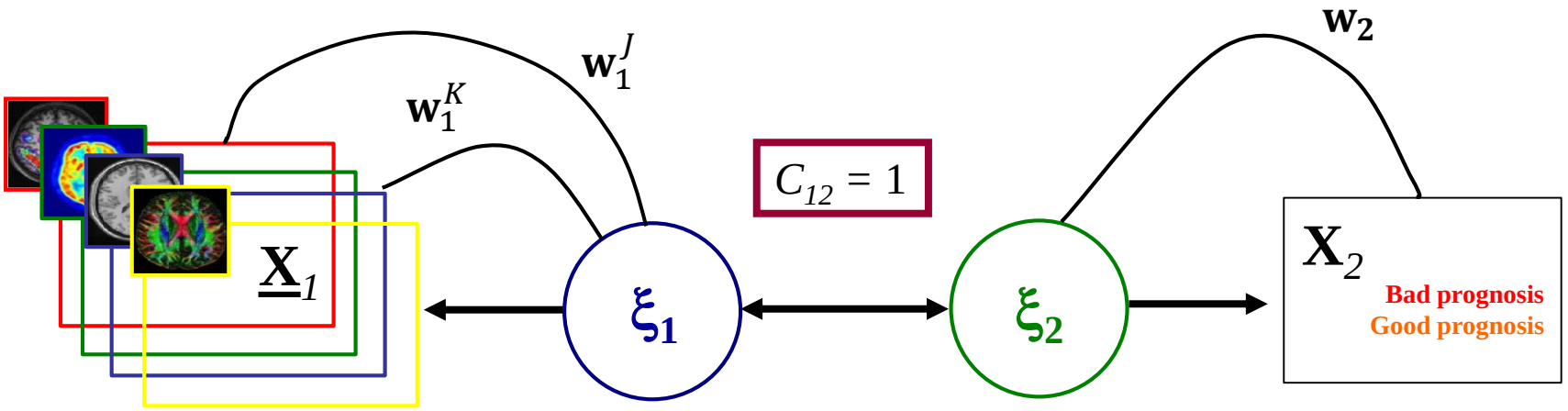
The COMA project from a multiblock viewpoint

(Brain and Spine Institute, La pitié Salpêtrière Hospital)

Predict the long term recovery of patients after traumatic brain injury



… to Multiblock / Multiway data

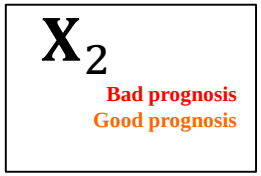


$4 + p_1$ parameters to estimate instead of $4 \times p_1$

$\mathbf{w}_1^K$ $\mathbf{w}_1^J$ $\mathbf{w}_2$

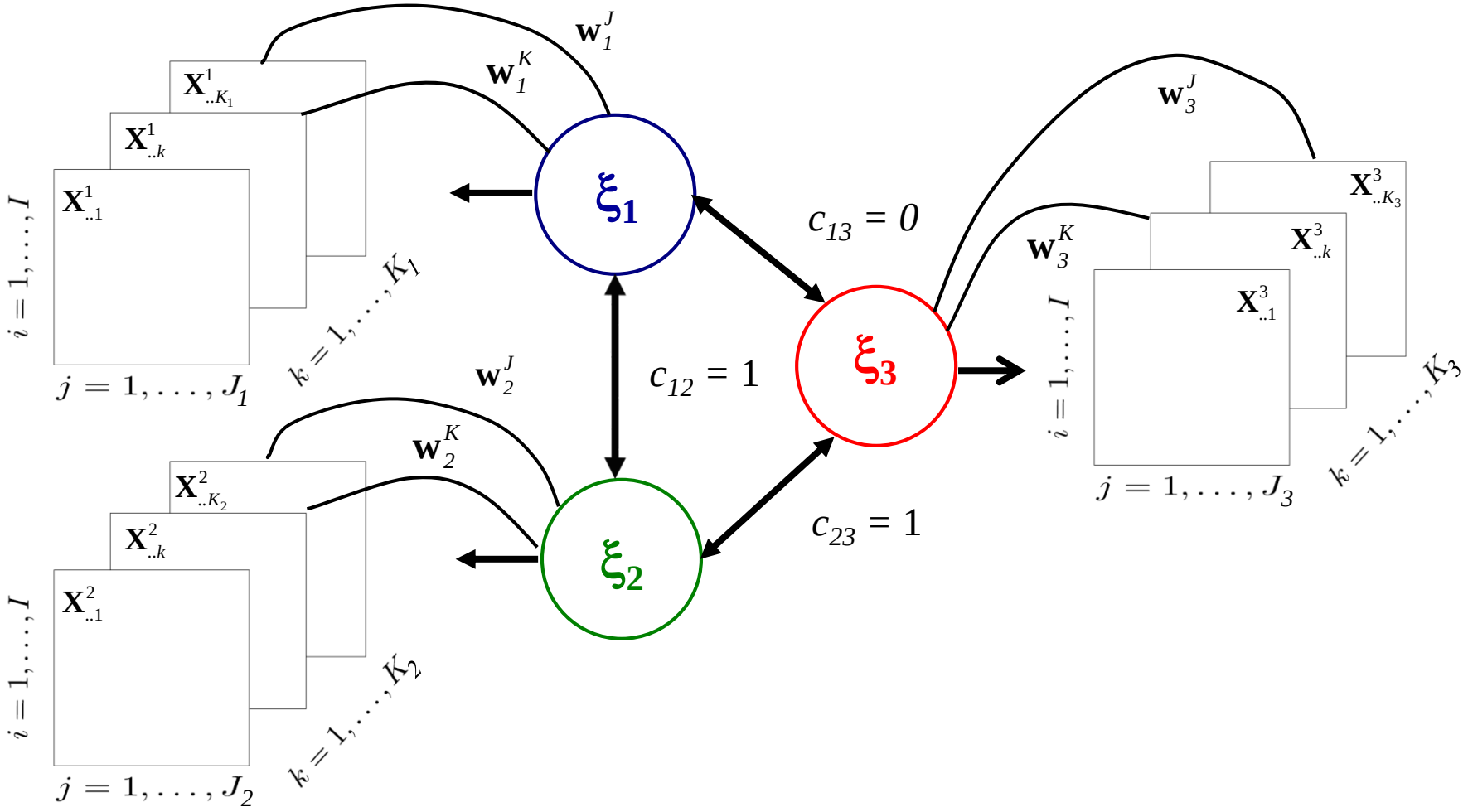
$$\mathbf{w}_1 = \mathbf{w}_1^K \otimes \mathbf{w}_1^J$$

$$\mathbf{X}_1 = \left(\mathbf{X}_{..1} \; \mathbf{X}_{..2} \; \mathbf{X}_{..3} \; \mathbf{X}_{..4} \right) \quad \text{and} \quad \mathbf{X}_2 =$$



$$\max_{\mathbf{w}_1, \mathbf{w}_2} \operatorname{cov}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) \qquad \text{s.t.} \quad \begin{cases} \mathbf{w}_1^\top \mathbf{M}_1 \mathbf{w}_1 = \mathbf{w}_2^\top \mathbf{M}_2 \mathbf{w}_2 = 1 \\ \mathbf{w}_1 = \mathbf{w}_1^K \otimes \mathbf{w}_1^J \end{cases}$$

MGCCA optimization problem

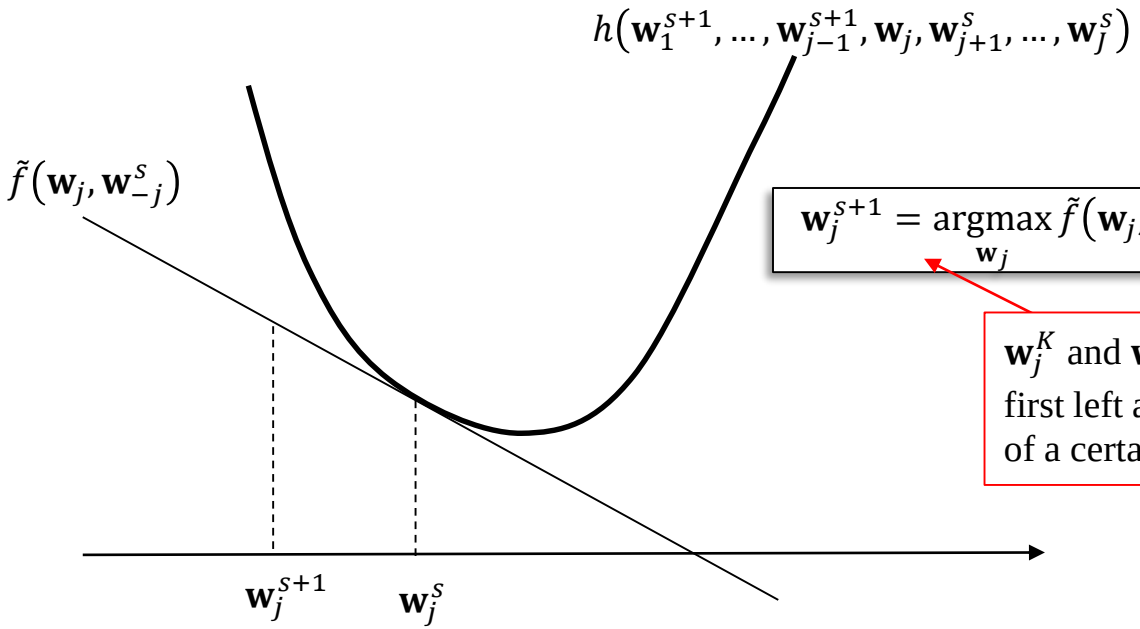


$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right) \quad \text{s.t.} \quad \mathbf{w}_j^\top \mathbf{M}_j \mathbf{w}_j = 1 \text{ and } \mathbf{w}_j = \mathbf{w}_j^K \otimes \mathbf{w}_j^J, j = 1, \ldots, J$$

The MGCCA algorithm

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk} g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right) \quad \text{s.t.} \quad \mathbf{w}_j^\top \mathbf{M}_j \mathbf{w}_j = 1 \text{ and } \mathbf{w}_j = \mathbf{w}_j^K \otimes \mathbf{w}_j^J, j = 1, \ldots, J$$

Two key ingredients: Block relaxation and Majorization by Minorization



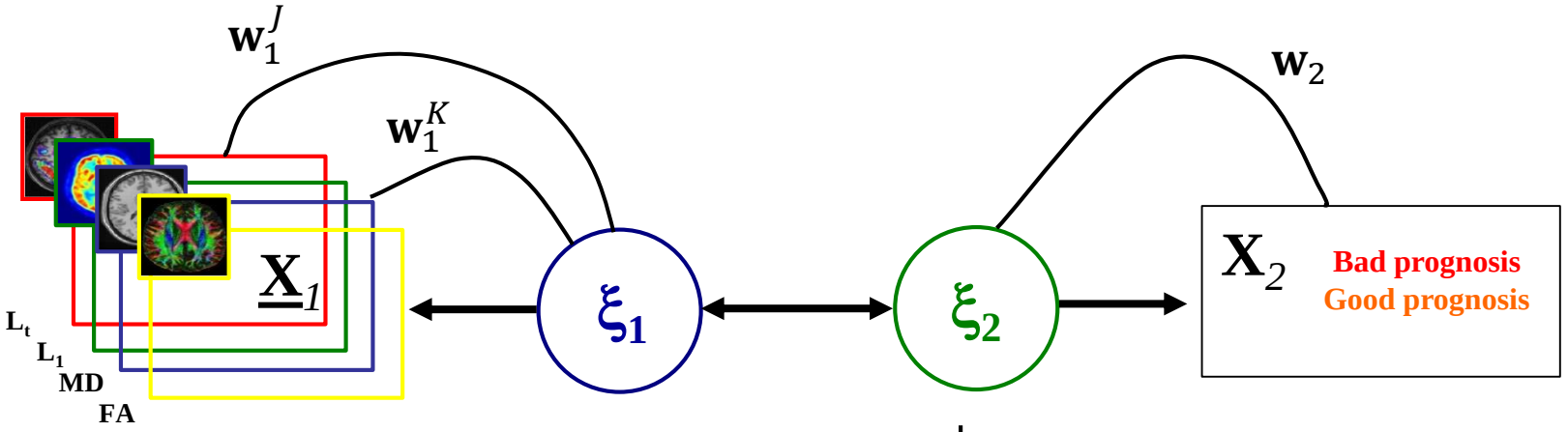
$$\mathbf{w}_j^{s+1} = \underset{\mathbf{w}_j}{\mathrm{argmax}}\, \tilde{f}\left(\mathbf{w}_j, \mathbf{w}_j^{s}\right) \text{ s.c. } \mathbf{w}_j^\top \mathbf{M}_j \mathbf{w}_j = 1 \text{ and } \mathbf{w}_j = \mathbf{w}_j^K \otimes \mathbf{w}_j^J$$

$\mathbf{w}_j^K$ and $\mathbf{w}_j^J$ are obtained as the first left and right singular vector of a certain matrix of dimension $K_j \times J_j$

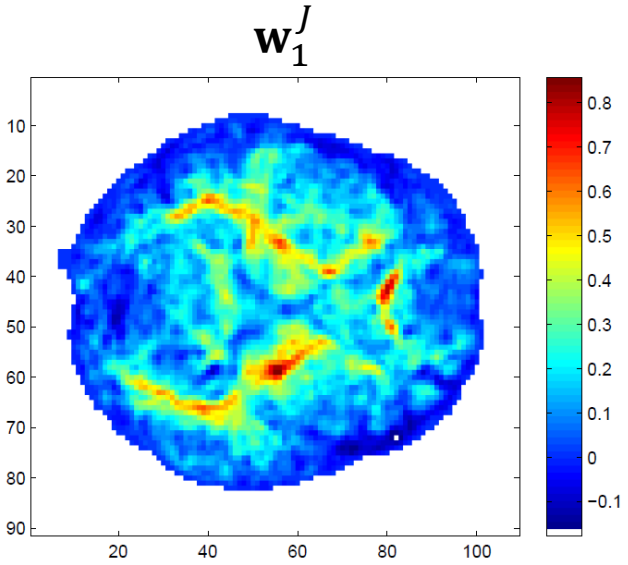
- Tenenhaus, A., Le Brusquet, L., Lechuga, G. (2015). Multiway Regularized Generalized Canonical Correlation Analysis. 47ème Journées de Statistique, Lille, France
- Tenenhaus A., Le Brusquet L. Three-way Regularized Generalized Canonical Correlation Analysis, ThRee-way methods In Chemistry And Psychology, (TRICAP) ,2015

MGCCA results

Predict the long term recovery of patients after traumatic brain injury



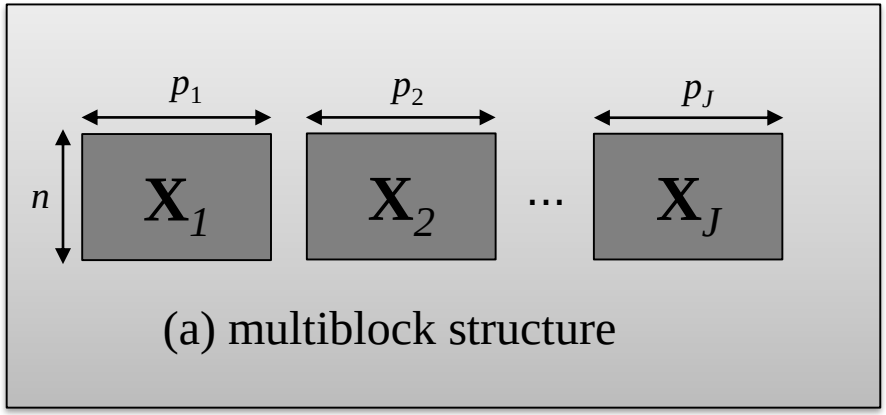
Influence of spatial positions



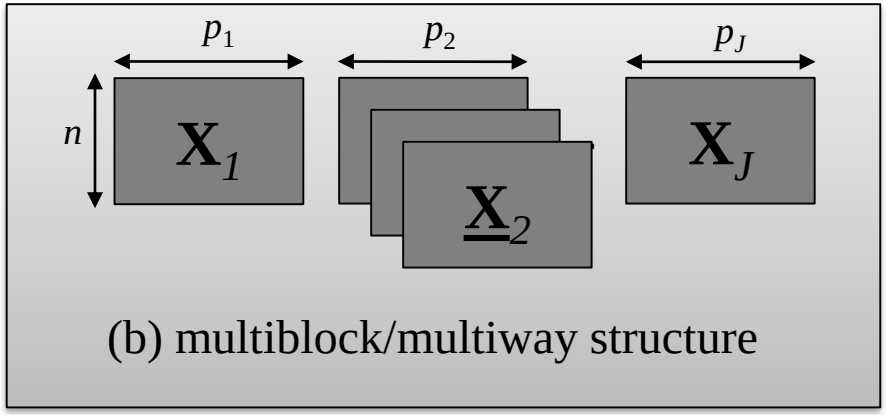
Discriminating voxels within the white matter bundles

Influence of spatial positions

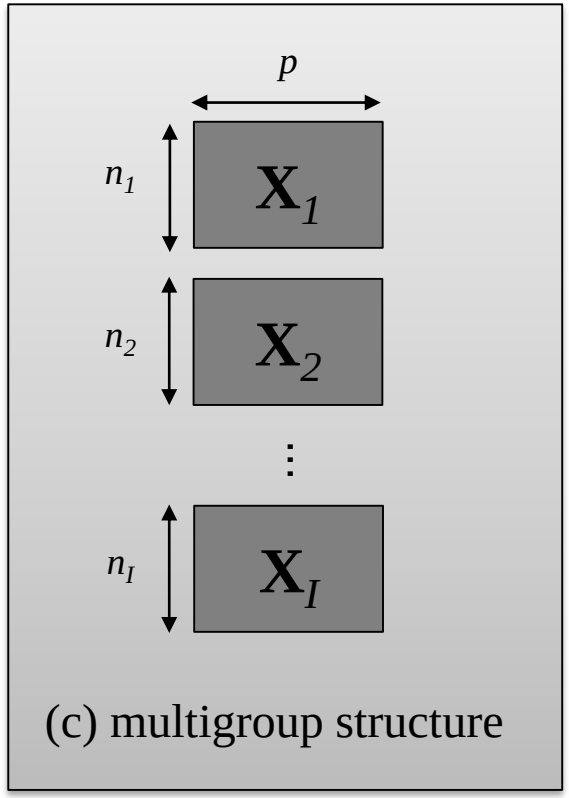
Modality	w_1^K
FA	09887
MD	0,0036
L_1	0,0046
L_t	0,0031



(a) multiblock structure



(b) multiblock/multiway structure



(c) multigroup structure

RGCCA for multiblock data analysis

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

Block component

$$\text{s.t.}\ \left(1-\tau_j\right)\mathrm{var}\left(\mathbf{X}_j\mathbf{w}_j\right) + \tau_j \left\|\mathbf{w}_j\right\|_2^2 = 1\ , j = 1, \ldots, J$$

Block components should verify 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.

$$\mathrm{cov}^2\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right) = \mathrm{var}\left(\mathbf{X}_j\mathbf{w}_j\right)\mathrm{cor}^2\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\mathrm{var}\left(\mathbf{X}_k\mathbf{w}_k\right)$$

RGCCA for multigroup data analysis

$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_I} \sum_{i,l}^{I} c_{il}\, g\left(\left\langle \mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l \right\rangle\right)$$

$$\text{s.t.}\ (1-\tau_i)\|\mathbf{X}_i\mathbf{w}_i\|_2^2 + \tau_i\|\mathbf{w}_i\|_2^2 = 1\ , i = 1, \ldots, I$$

Group loadings

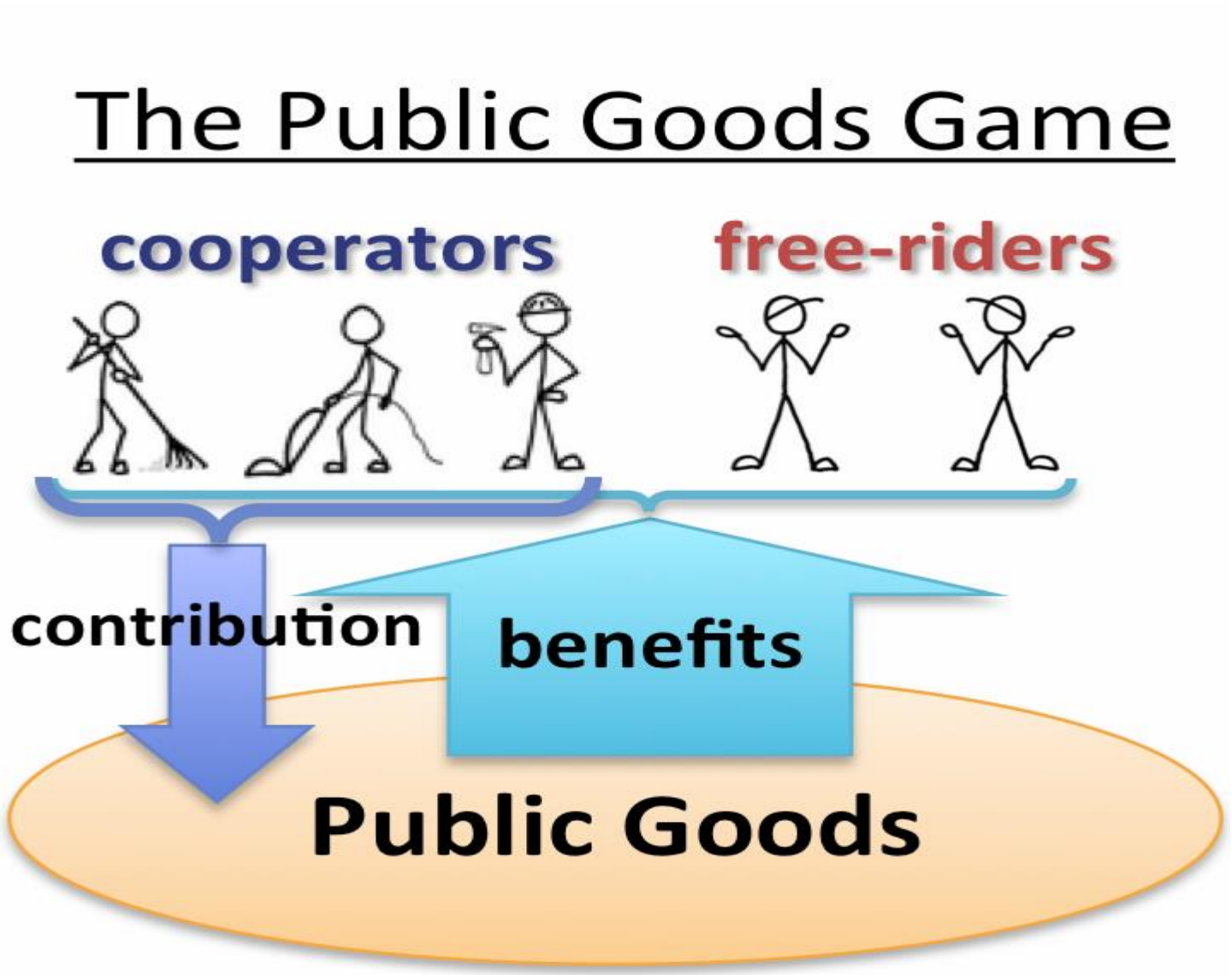
Group components

Group loadings and group components should verify the following properties at the same time:

- Group component $\mathbf{X}_i\mathbf{w}_i$ well explains their own block.
- Small angle between group loadings if groups are connected.

$$\langle \mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l \rangle = \cos\left(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i, \mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l\right) \times \left\|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{w}_i\right\| \times \left\|\mathbf{X}_l^\top \mathbf{X}_l \mathbf{w}_l\right\|$$

Multi-group data analysis
The public good dilemma game

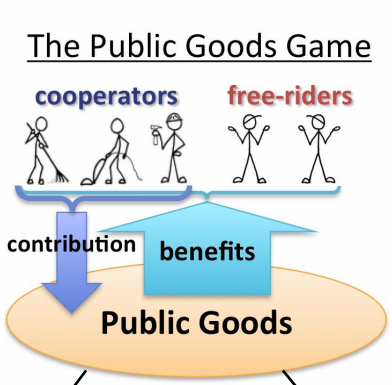


Example from Kim De Roover et al.(2013) for illustrating Clusterwise SCA-ECP

Hypothesis

1. When both players contribute equally, they want to be rewarded equally.
2. Violation of equality is perceived as unjust and elicits anger.
3. In case of inequality, the subject's negative emotional reactions may be modified if the advantaged person is in personal need, because feelings of sympathy and empathy can be aroused.

An experimental study on 282 cooperators: Two factors "Reward" and "Empathy" are fully crossed



Reward is equal
(Equal condition)

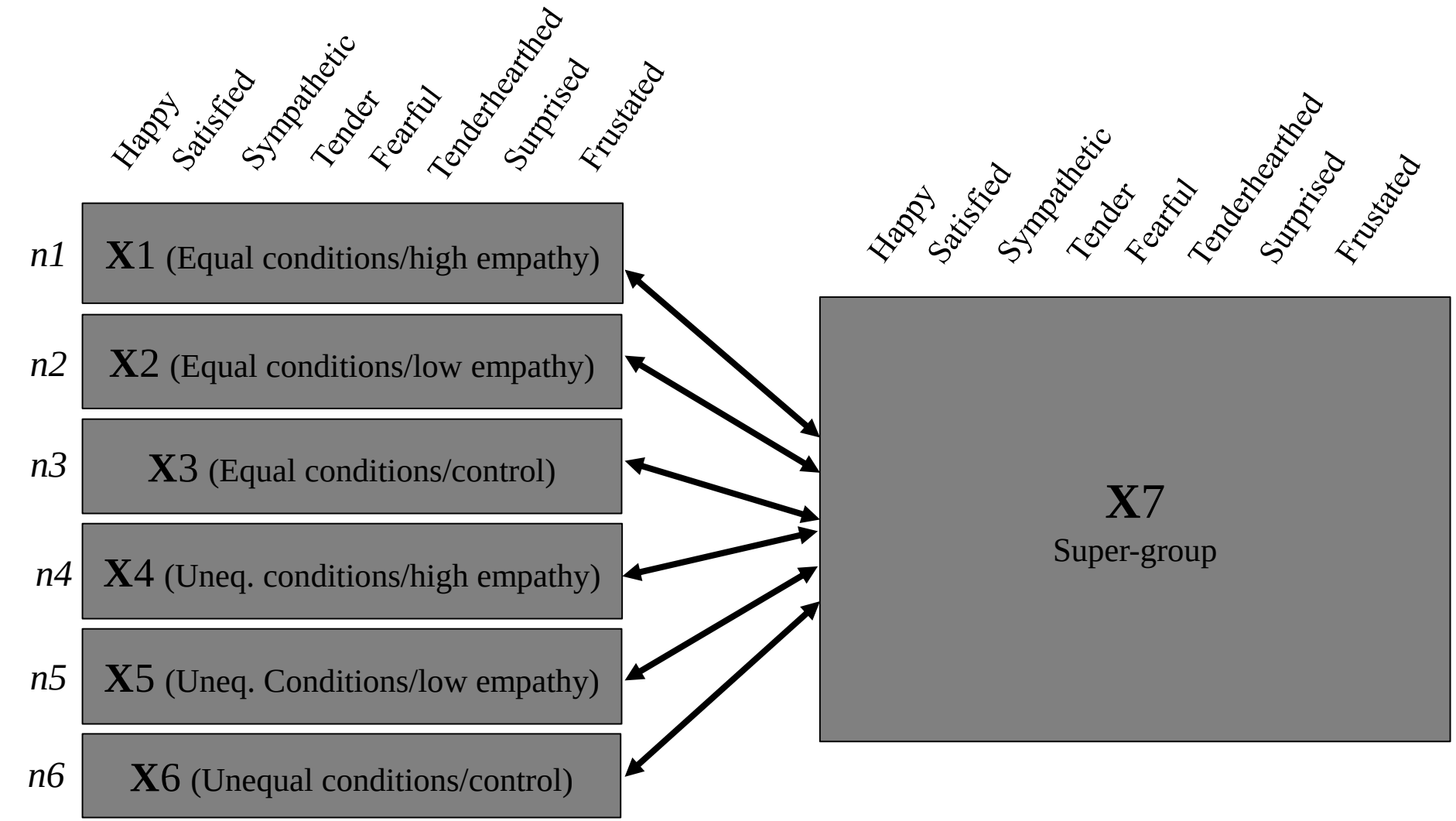
Reward is unequal
(unequal condition)

High empathy condition | **Control condition** | **Low empathy condition** | **High empathy condition** | **Control condition** | **Low empathy condition**

Reward is equal for both players (Equal condition).
Reward is higher for the free-rider (Unequal condition).

High empathy condition: the cooperator receives a message about a negative personal event that happened to the free-rider.
Low empathy condition: the same message is received, but the cooperator is asked to take an objective perspective.
Control condition: the cooperator does not receive any message.

The public good dilemma game data set



$$\max_{\mathbf{a}_1,\ldots,\mathbf{a}_6} \sum_{i=1}^{6} \cos\left(\mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}_i, \mathbf{X}_7^\top \mathbf{X}_7 \mathbf{a}_7\right) \times \left\|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}_i\right\| \times \left\|\mathbf{X}_7^\top \mathbf{X}_7 \mathbf{a}_7\right\| \text{ s.t. } \quad \|\mathbf{X}_i \mathbf{a}_i\| = 1,\ i = 1, \ldots, 7$$

Similar Loadings

$\sqrt{p \times \mathrm{AVE}_j}$

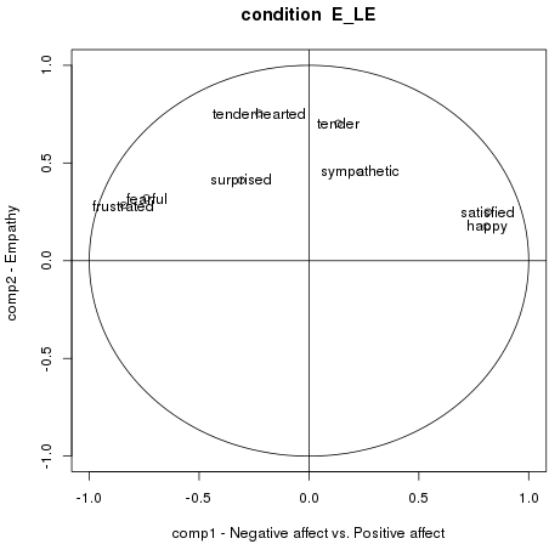
$\sqrt{p \times \mathrm{AVE}_7}$

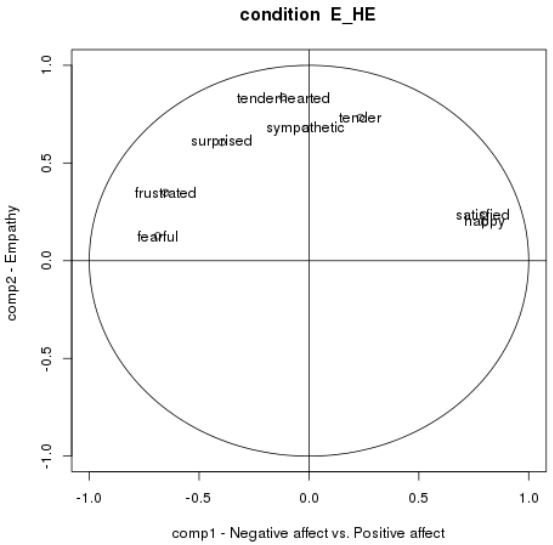
Low Empathy

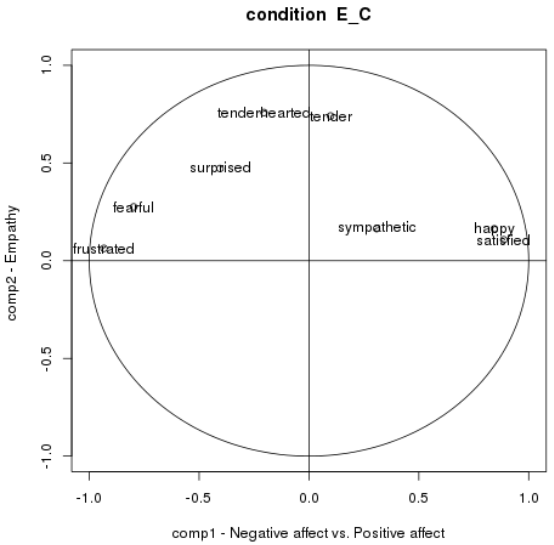
High Empathy

Control

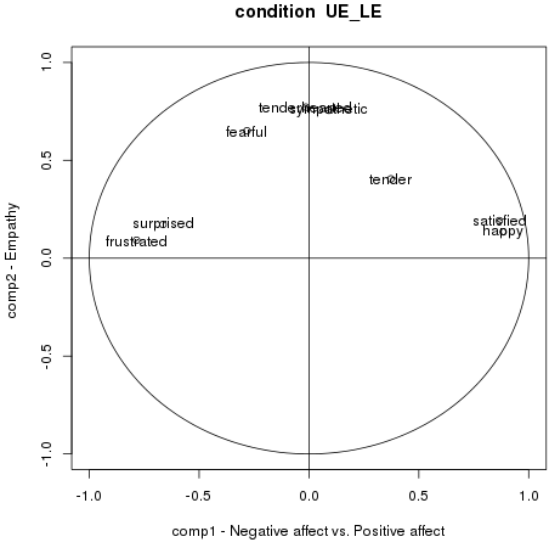
Equal Condition

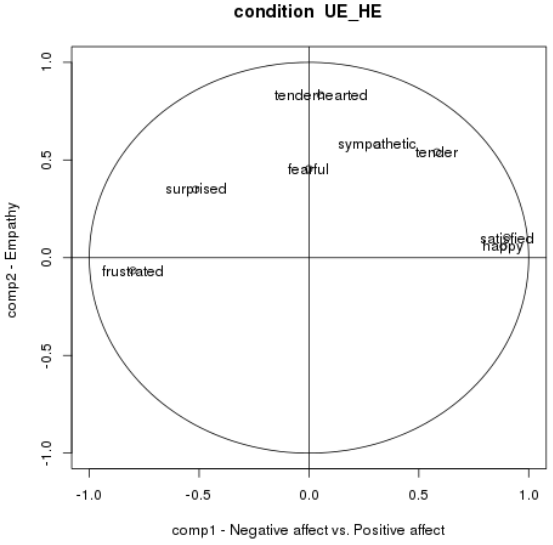


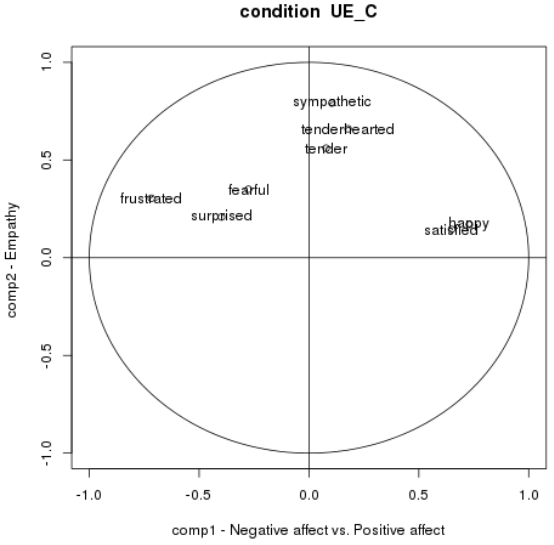




Unequal Condition



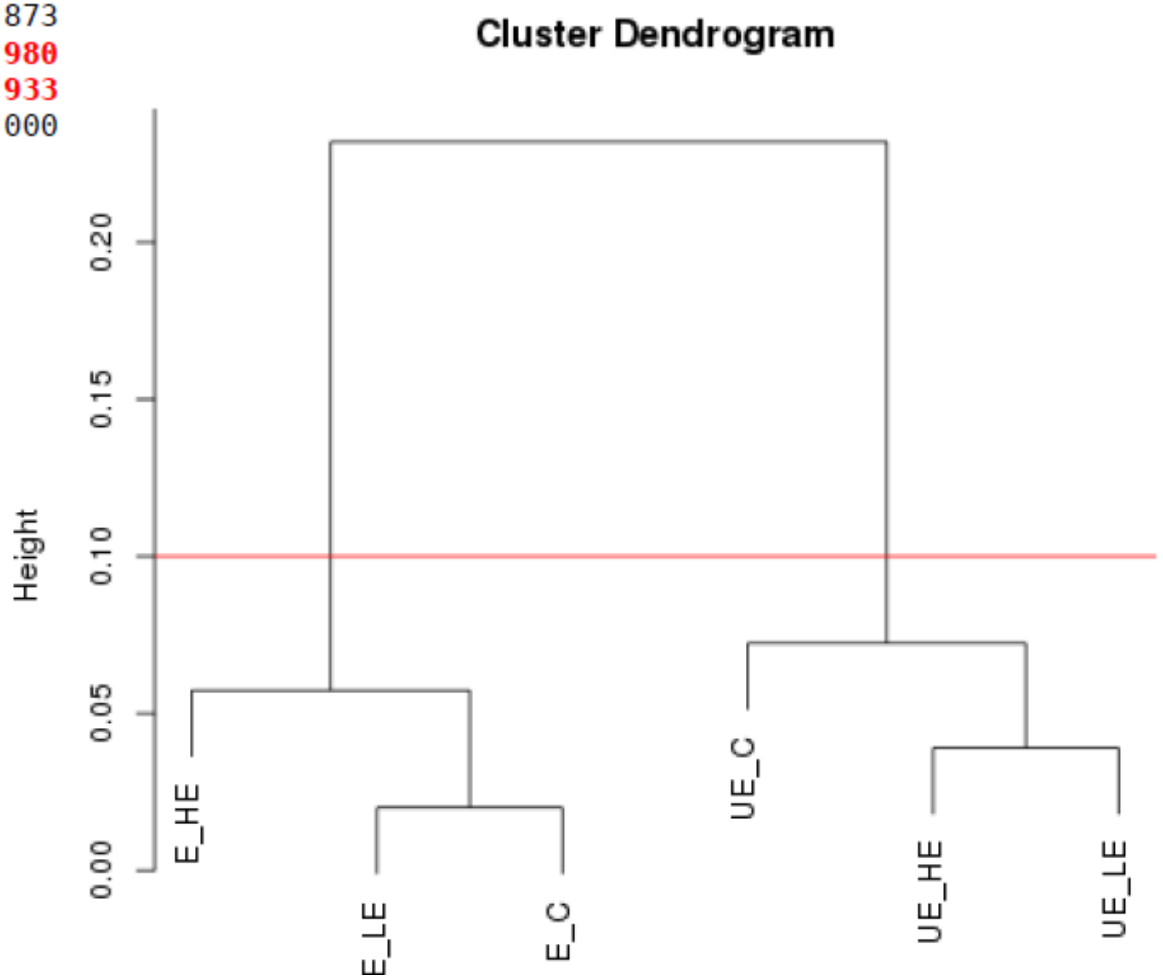




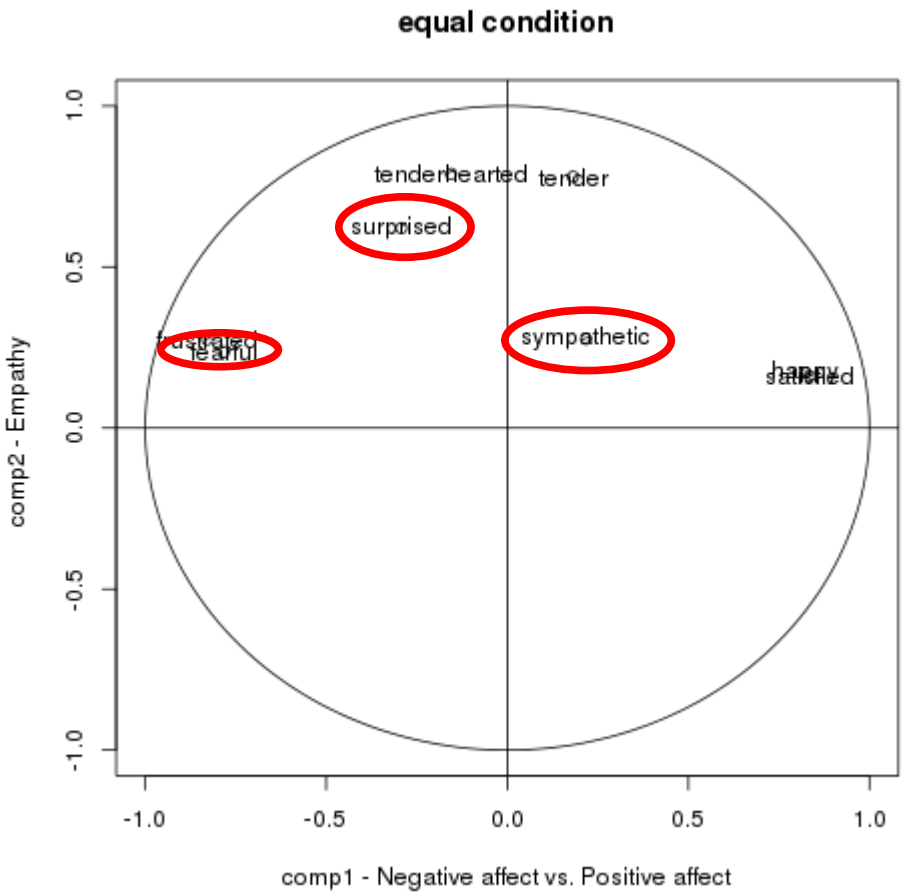
Clustering of groups according to cosine between group-loadings

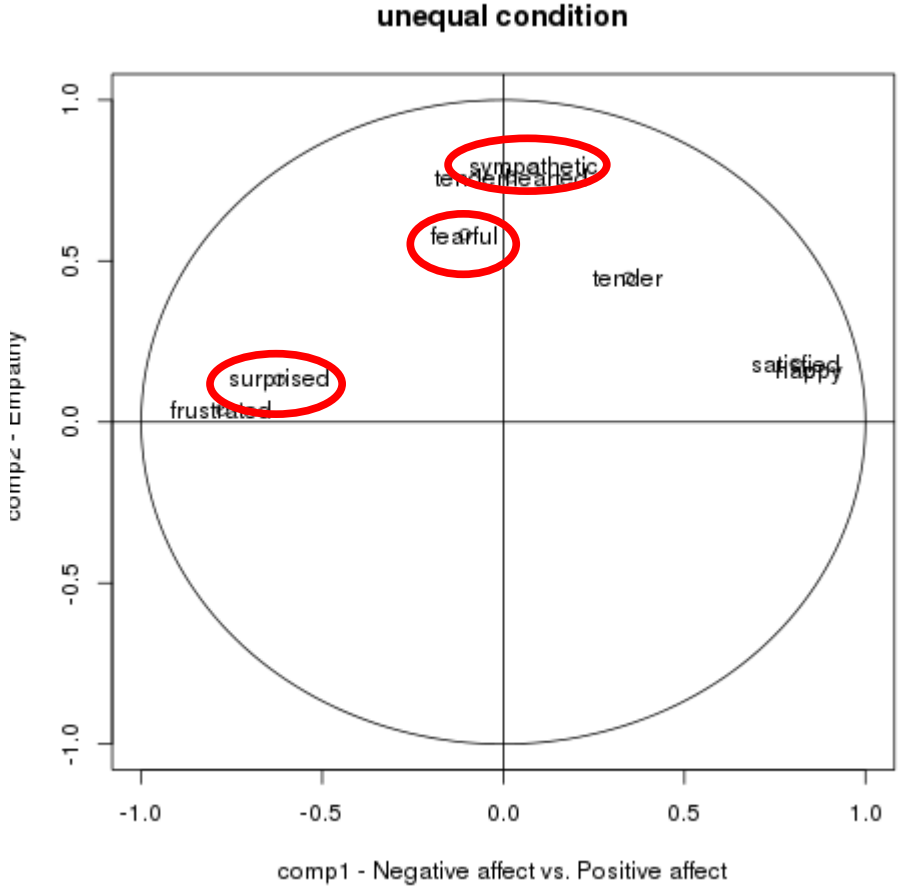
COS

	UE_HE	UE_LE	UE_C	E_LE	E_HE	E_C
UE_HE	1.000	**0.961**	**0.916**	0.877	0.870	0.866
UE_LE	**0.961**	1.000	**0.956**	0.906	0.897	0.871
UE_C	**0.916**	**0.956**	1.000	0.926	0.935	0.873
E_LE	0.877	0.906	0.926	1.000	**0.971**	**0.980**
E_HE	0.870	0.897	0.935	**0.971**	1.000	**0.933**
E_C	0.866	0.871	0.873	**0.980**	**0.933**	1.000



PCA + Varimax on standardized groups





Technical conclusions

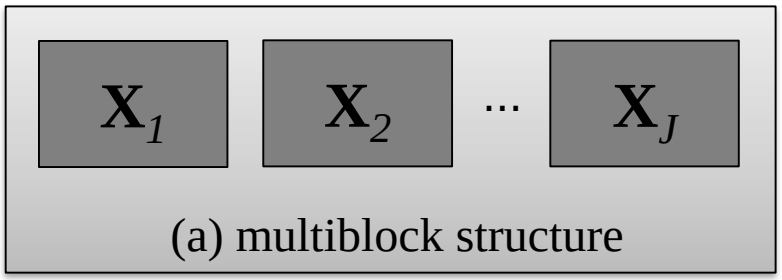
1. Depending on the dimension of the blocks, use either the **primal or the dual algorithm**.

2. 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.

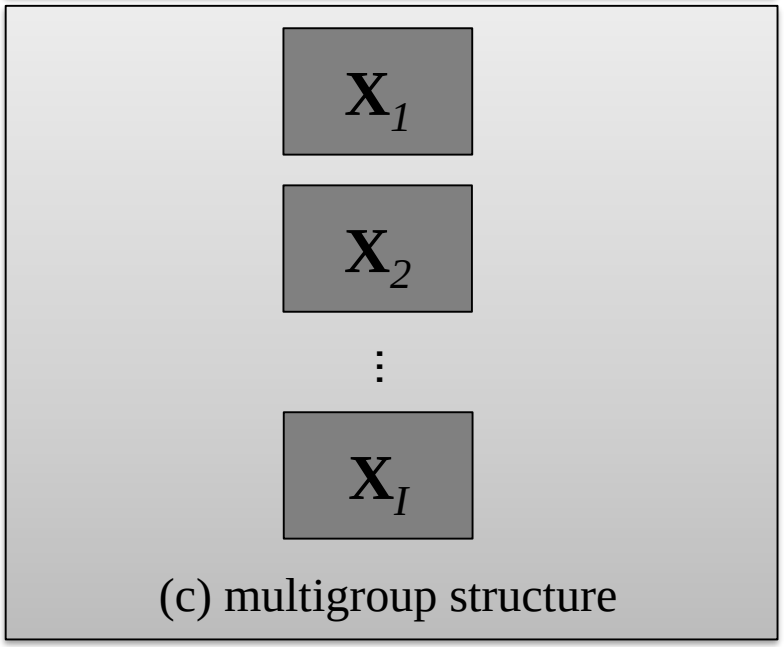
3. **Sparse constraints are useful when the relevant variables are masked by (too many) noisy variables.**

4. **Kronecker constraints are used for threeway data configuration.**

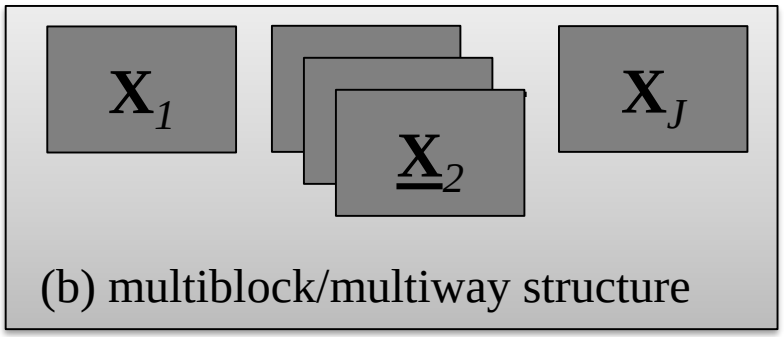
General conclusion



(a) multiblock structure



(c) multigroup structure



(b) multiblock/multiway structure

RGCCA for multiblock, multigroup or multiway data allows analyzing the data in their natural (but complex) structure.

ICM
Institut du Cerveau
et de la Moelle épinière

Economic inequality and political instability Data from Russett (1964)

Economic inequality

Agricultural inequality

GINI : Inequality of land distributions

FARM : % farmers that own half of the land (> 50)

RENT : % farmers that rent all their land

Industrial development

GNPR : Gross national product per capita ($ 1955)

LABO : % of labor force employed in agriculture

Political instability

INST : Instability of executive (45-61)

ECKS : Nb of violent internal war incidents (46-61)

DEAT : Nb of people killed as a result of civic group violence (50-62)

D-STAB : Stable democracy

D-UNST : Unstable democracy

DICT : Dictatorship

Economic inequality and political instability (Data from Russett, 1964)

	X_1			X_2		X_3			
	Gini	**Farm**	**Rent**	**Gnpr**	**Labo**	**Inst**	**Ecks**	**Deat**	**Demo**
Argentine	86.3	98.2	32.9	374	25	13.6	57	217	2
Australie	92.9	99.6	3.27	1215	14	11.3	0	0	1
Autriche	74.0	97.4	10.7	532	32	12.8	4	0	2
⋮									
France	58.3	86.1	26.0	1046	26	16.3	46	1	2
⋮									
Yougoslavie	43.7	79.8	0.0	297	67	0.0	9	0	3

Agricultural inequality

GINI : Inequality of land distributions

FARM : % farmers that own half of the land (> 50)

RENT : % farmers that rent all their land

Industrial development

GNPR : Gross national product per capita ($, 1955)

LABO : % of labor force employed in agriculture

Political instability

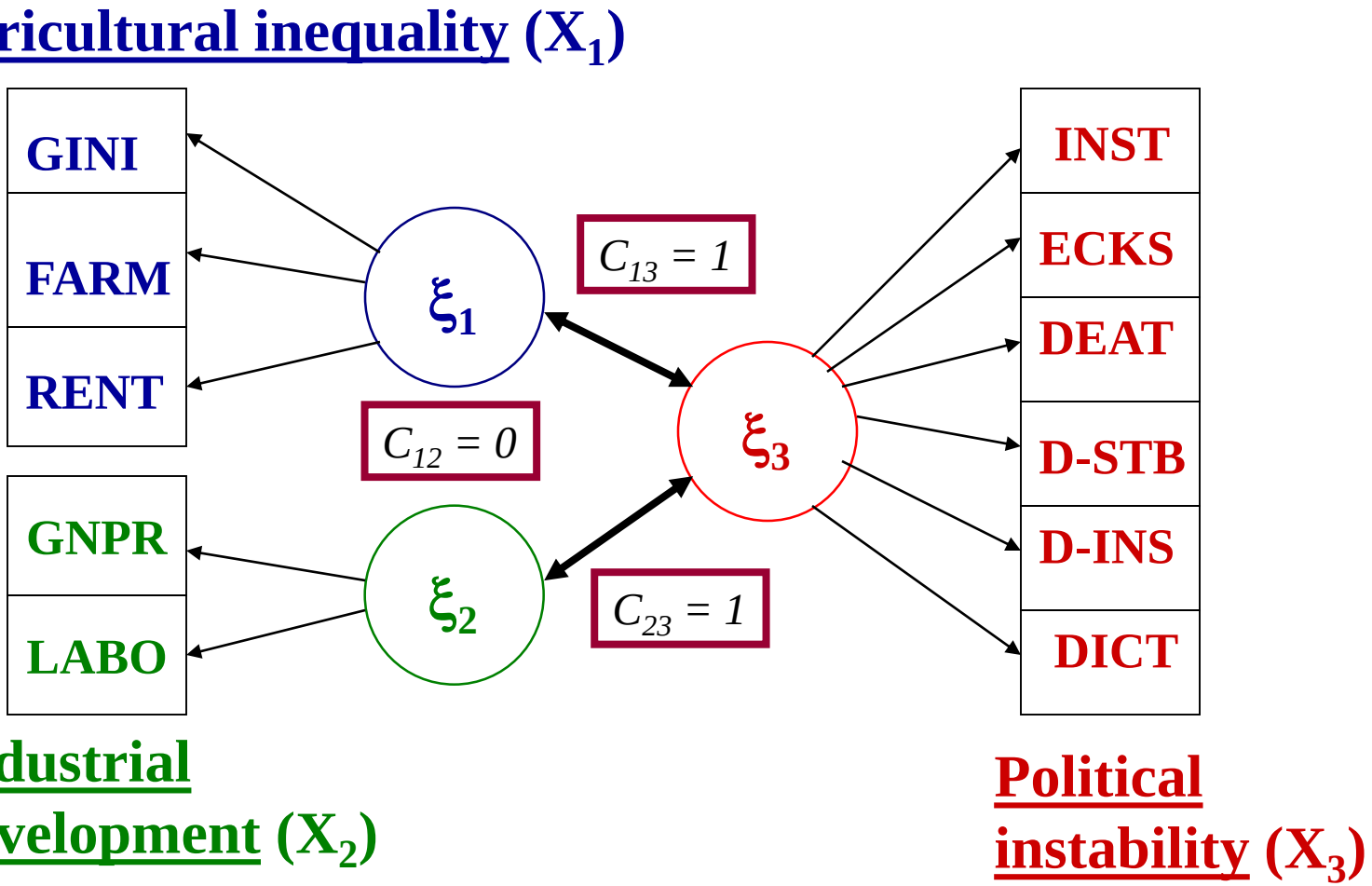
INST : Instability of executive (45-61)

ECKS : Nb of violent internal war incidents (46-61)

DEAT : Nb of people killed as a result of civic group violence (50-62)

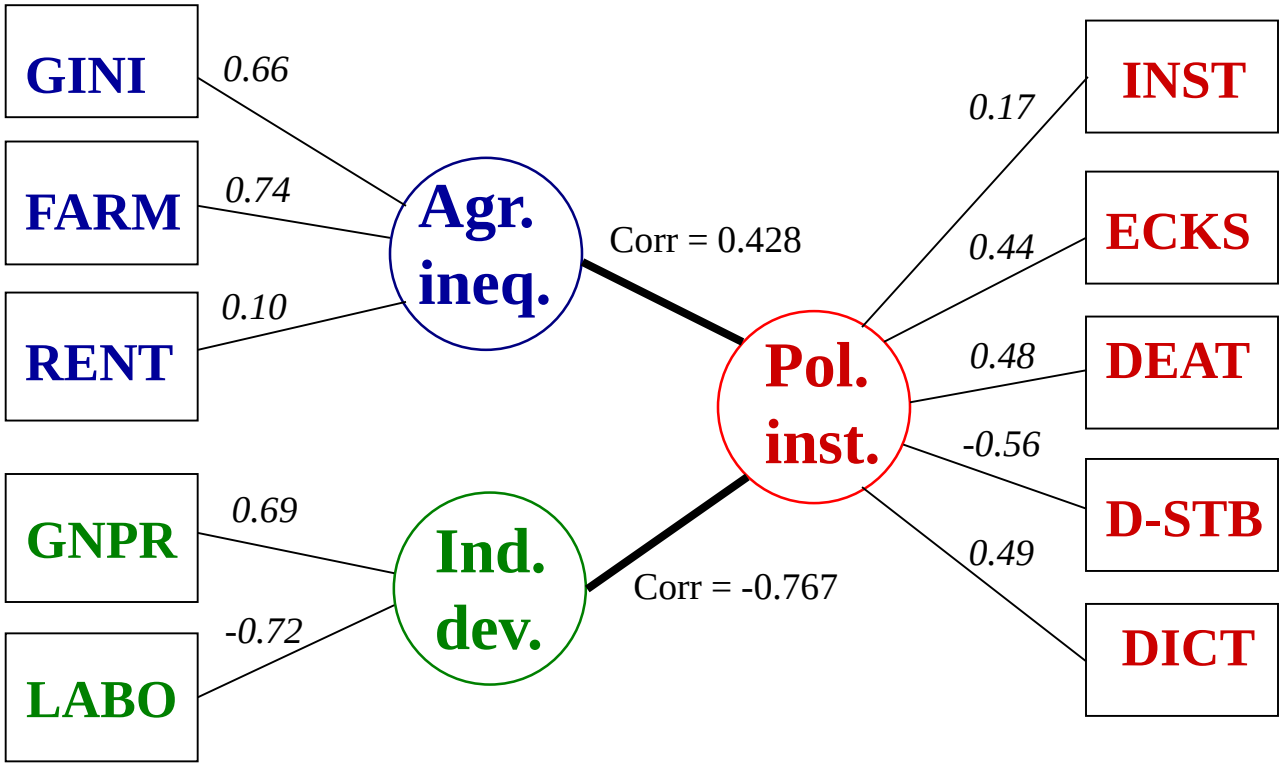
DEMO : Stable democracy (1), Unstable democracy (2) or Dictatorship (3)

Path Diagram



Outer weight vectors

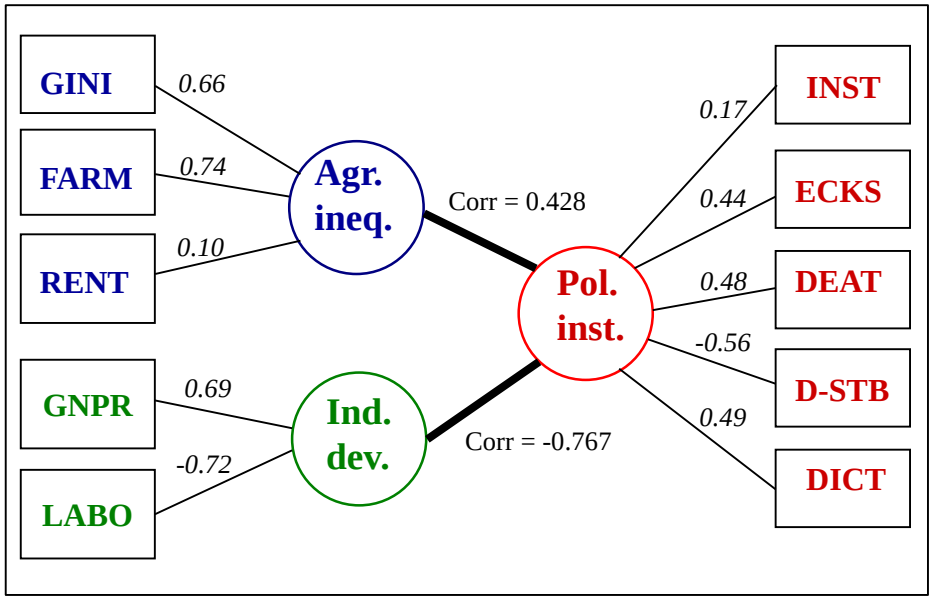
Agricultural inequality (X_1)

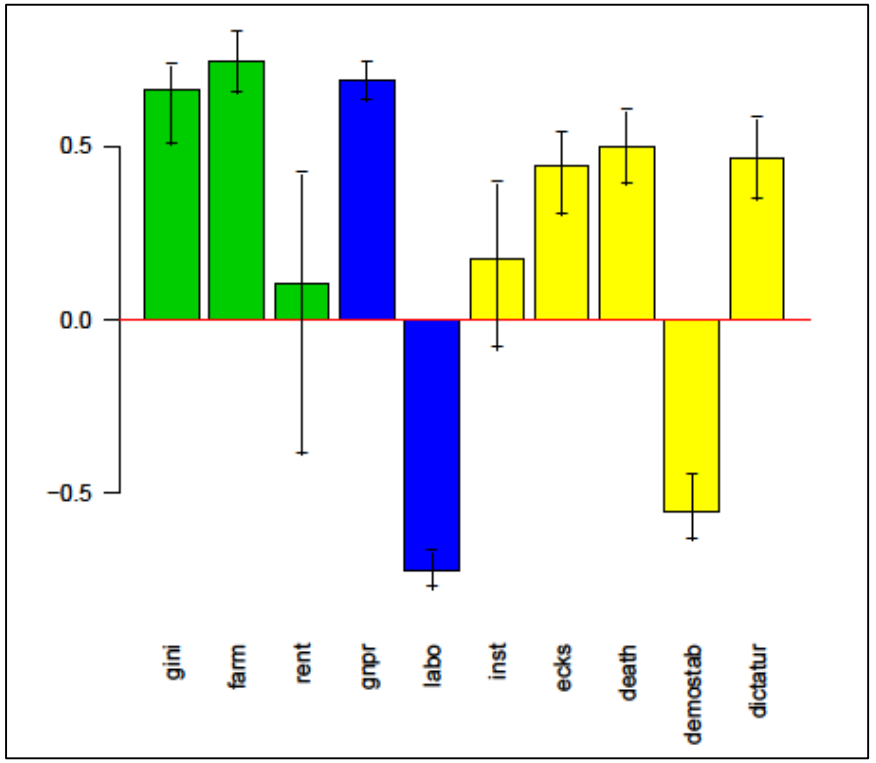


Industrial development (X_2)

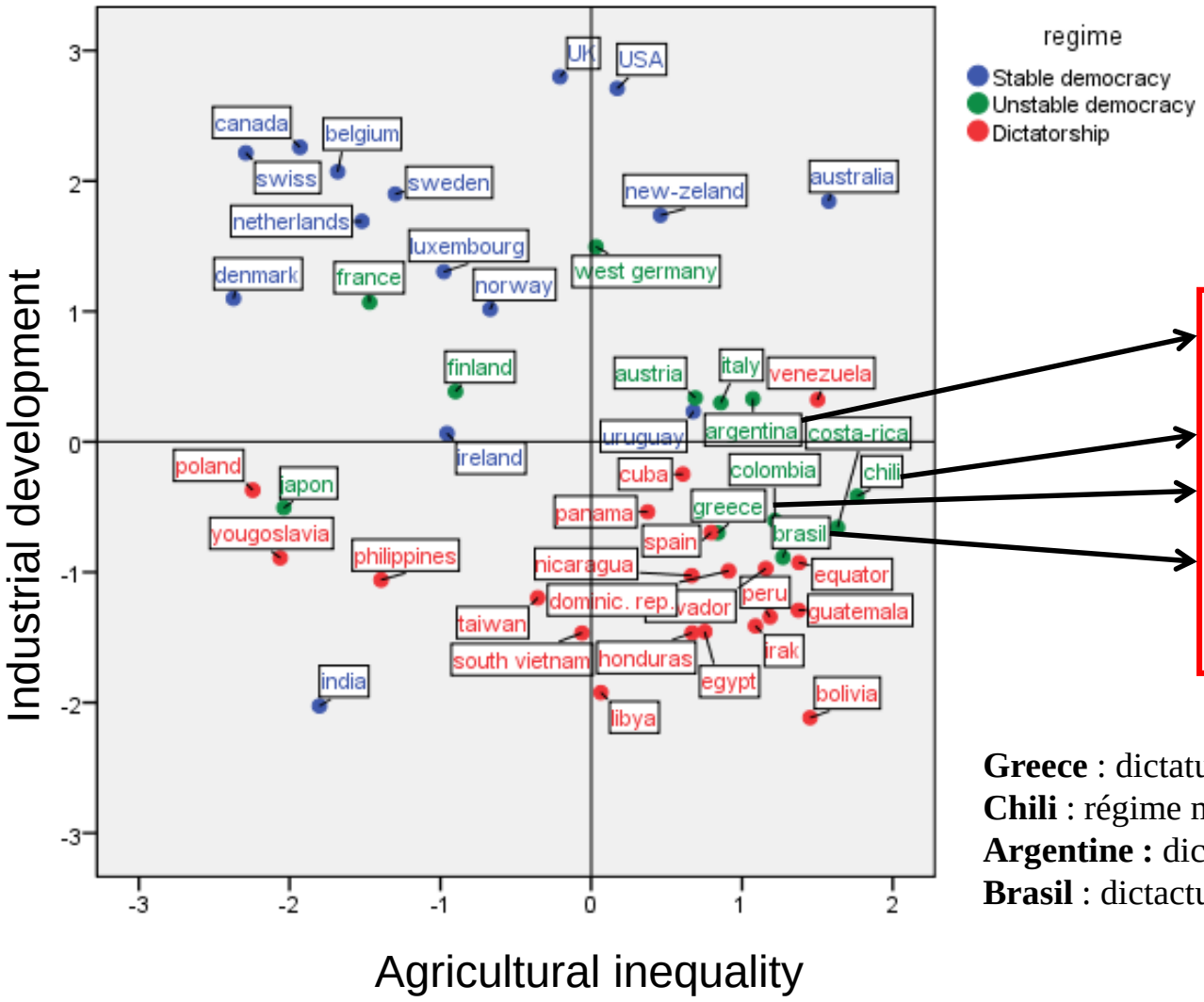
Political instability (X_3)

Bootstrap confidence intervals





Data vizualization



These countries have known a period of dictatorship after 1964.

Greece : dictature des colonnels 1967-1974
Chili : régime militaire de Pinochet 1973-1990
Argentine : dictature militaire 1976-1983
Brasil : dictacture militaire de Branco 1964-1985