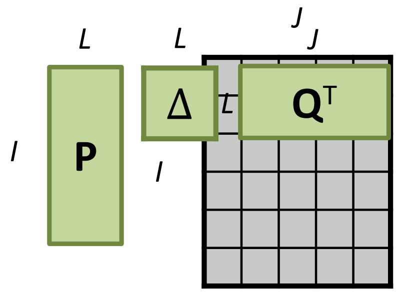
Analyse intégrative avec RGCCA

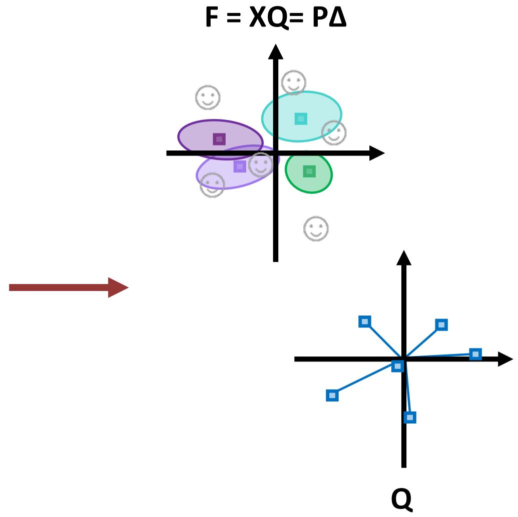
ACTIVITE DE GROUPE

La méthode RGCCA

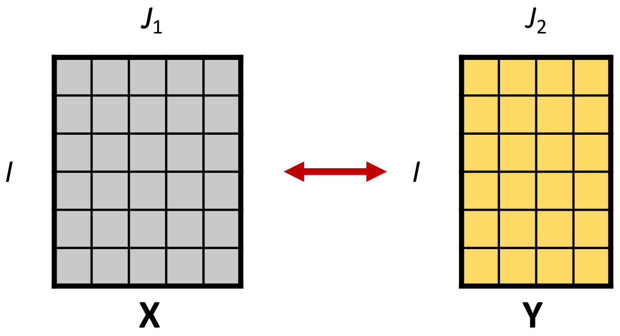
ACP

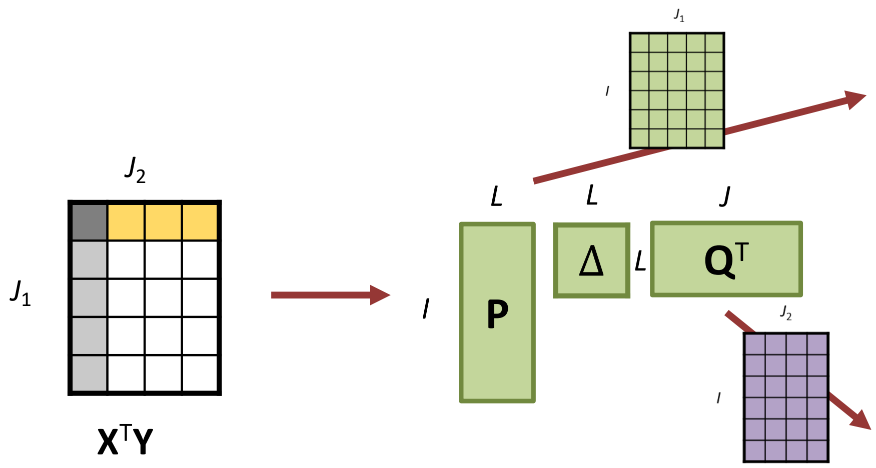


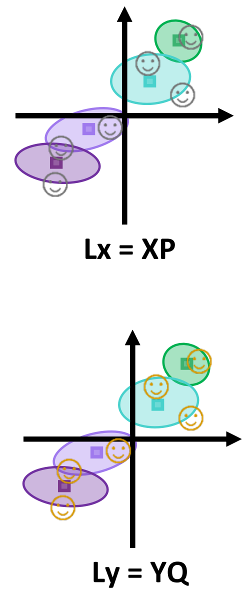
$$P^TP = Q^TQ = I$$



RGCCA







PLS: $P^TP = Q^TQ = I$

CCA: $P^T(X^TX)^{-1}P = Q^T(Y^TY)^{-1}Q = I$

RA: $P^T(X^TX)^{-1}P = Q^TQ = I$

$$P^TP = I$$
$$P^T(X^TX)^{-1}P = I$$

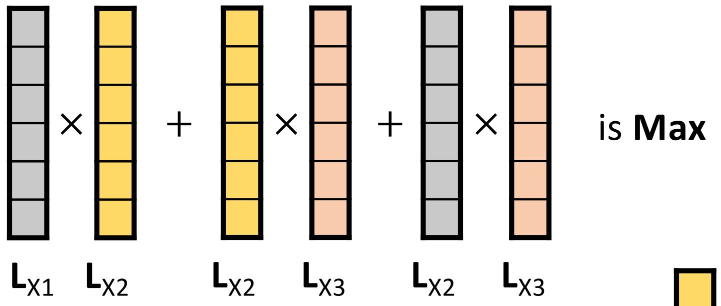
$$\tau \cdot P^TP + (1-\tau) \cdot P^T(X^TX)^{-1}P = I$$

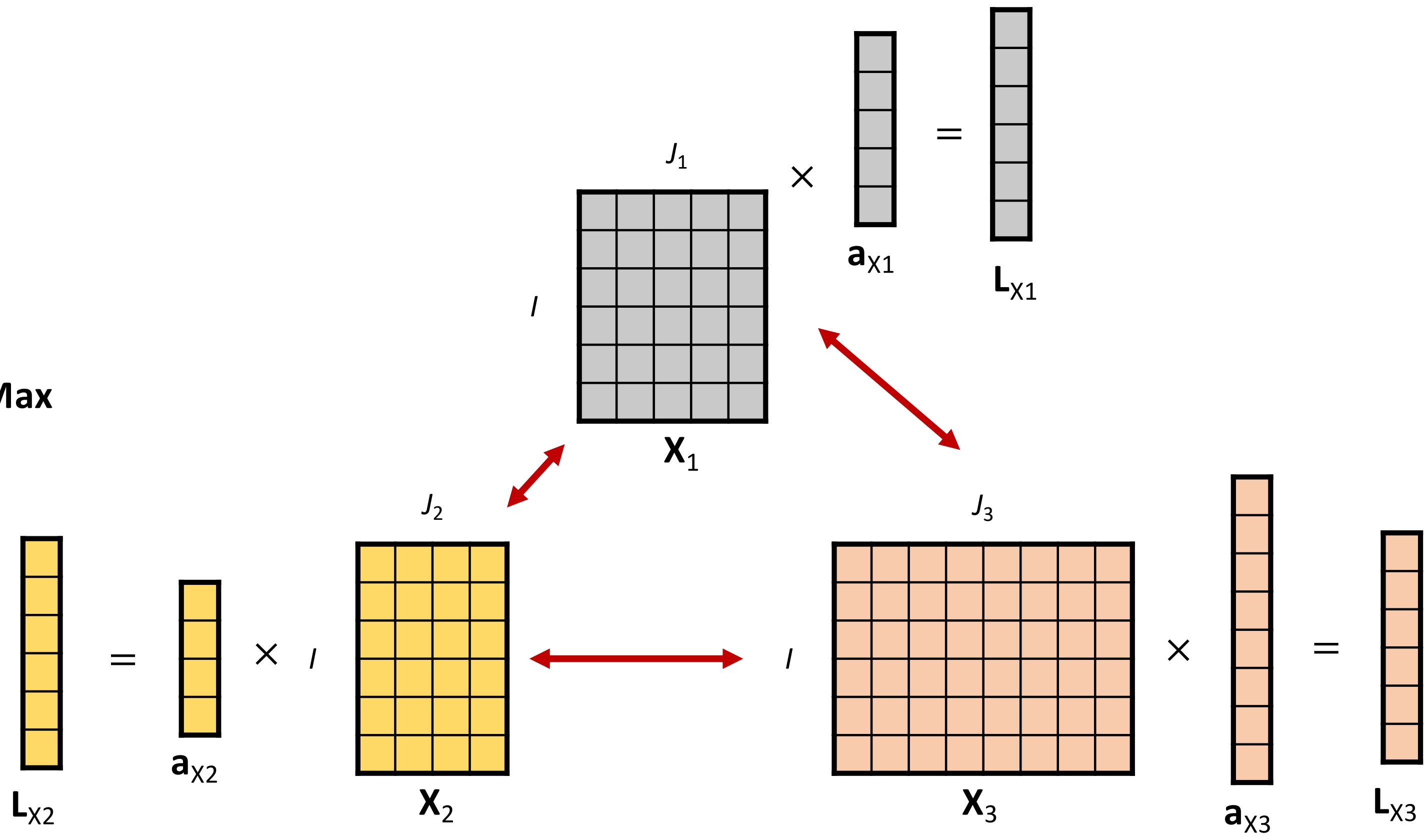
When $\tau = 1$, $\quad P^TP = I$

When $\tau = 0$, $\quad P^T(X^TX)^{-1}P = I$

RGCCA

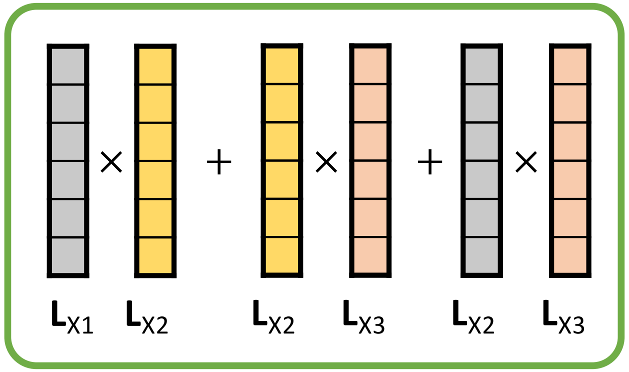
- Start from Component 1
    - Loadings $\mathbf{a}_X$
    - Latent variables





RGCCA

- Start from Component 1
    - Loadings $\mathbf{a}_X$
    - Latent variables

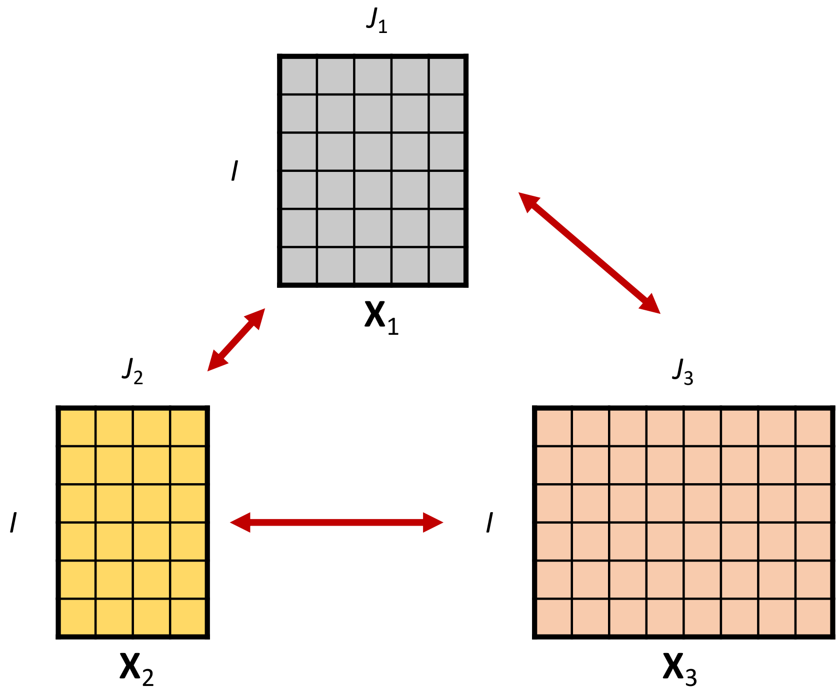


is **Max**

What is this?

Sum of Cross-Product (SCP) or Covariance

When SS = 1, covariance = correlation



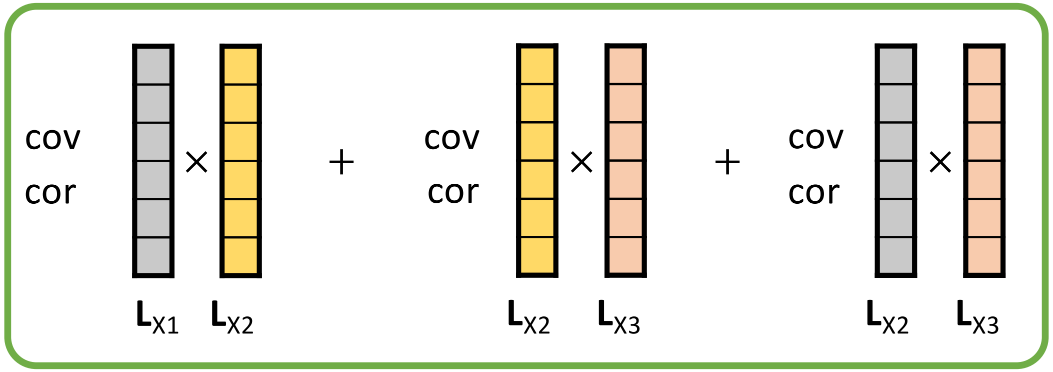
$$\boxed{\tau \cdot \mathbf{a}^T\mathbf{a} + (1-\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}}$$

When $\tau = 1$, **covariance** is max

When $\tau = 0$, **correlation** is max

RGCCA

- Start from Component 1
    - Loadings $\mathbf{a}_X$
    - Latent variables



is **Max**

What is this?

$\boxed{\tau \cdot \mathbf{a}^T\mathbf{a} + (1-\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}}$

When $\tau = 1$, **covariance** is max → PLS

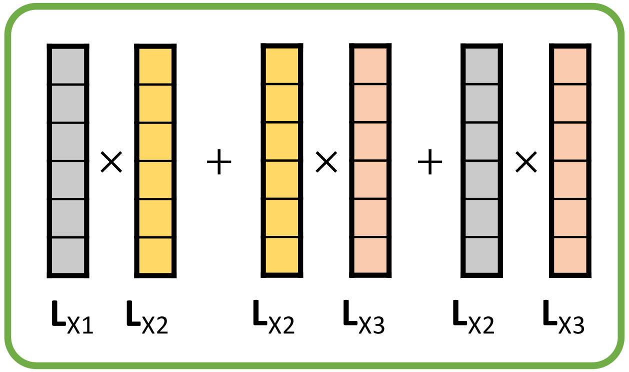
When $\tau = 0$, **correlation** is max → CCA

When $\tau_1 = 0$, but $\tau_2 = 1$, **coefficient of regression** is max → RA

$$\frac{\mathbf{L}_{X1} \cdot \mathbf{L}_{X2}}{\sqrt{\mathbf{L}_{X1}^2}}$$

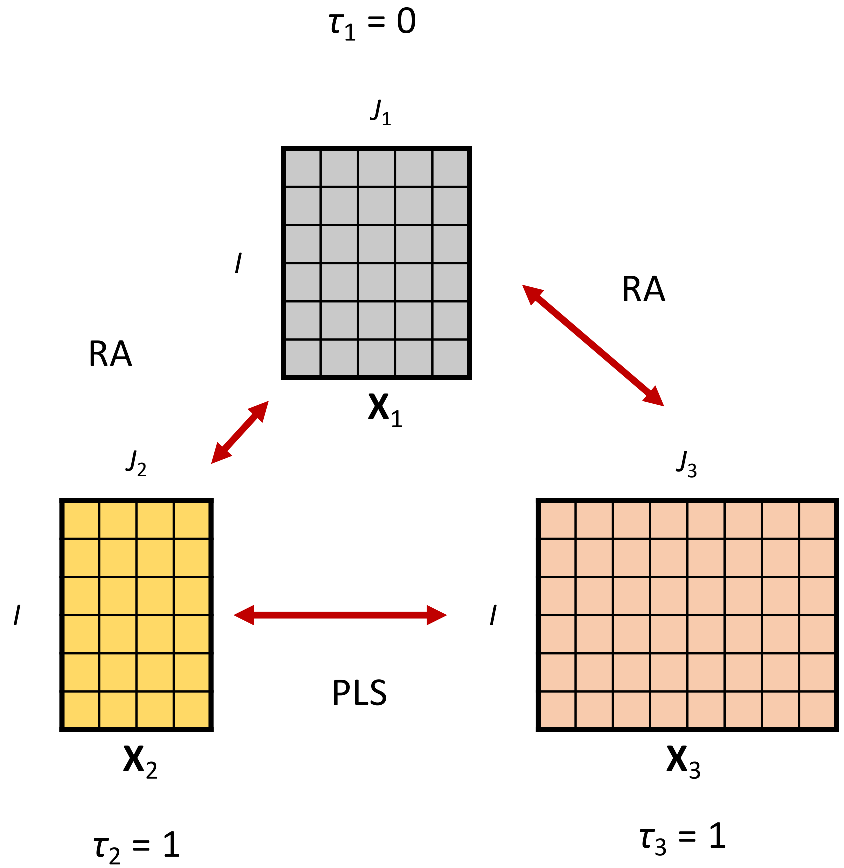
RGCCA

- Start from Component 1
    - Loadings $\mathbf{a}_X$
    - Latent variables



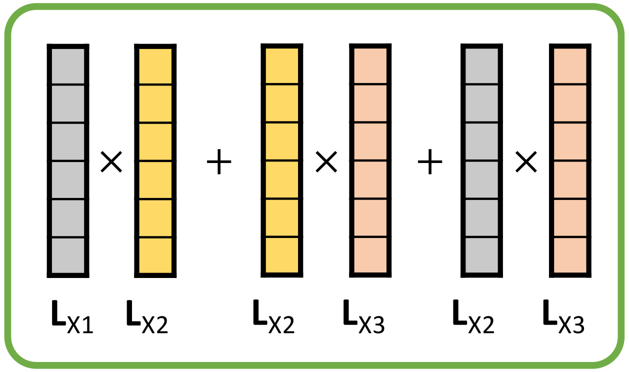
is **Max**

$$\tau \cdot \mathbf{a}^T\mathbf{a} + (1-\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}$$



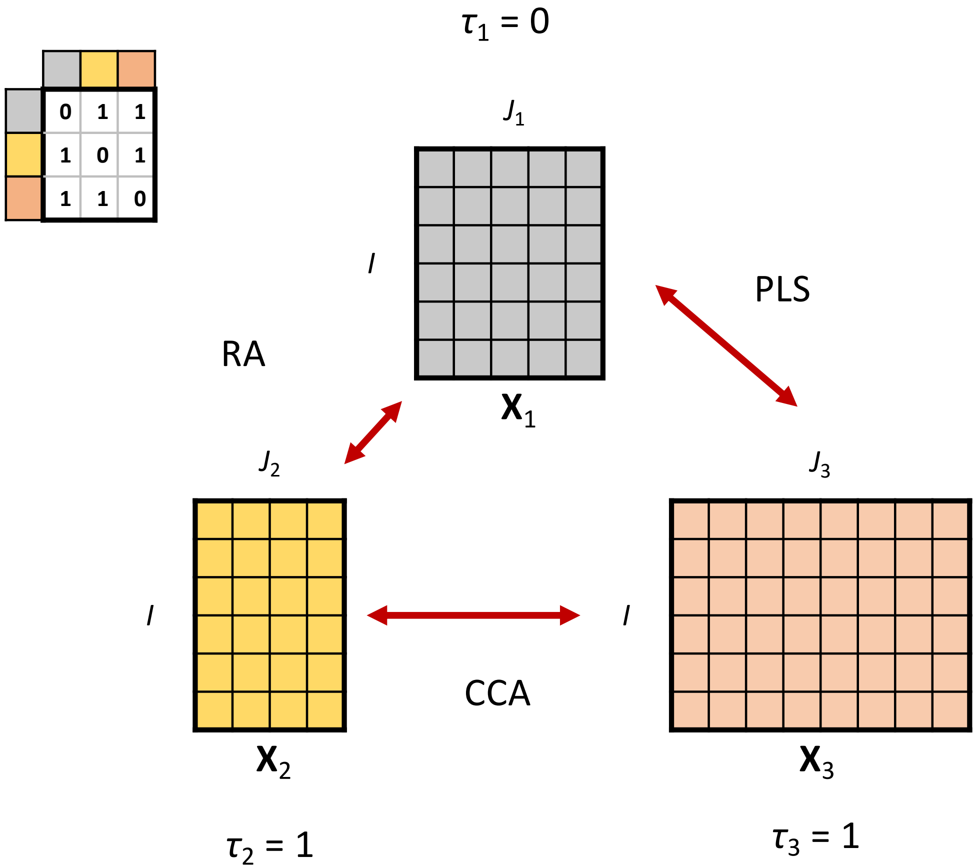
RGCCA

- Start from Component 1
  - Loadings $\mathbf{a}_X$
  - Latent variables
  - Connection matrix **C**



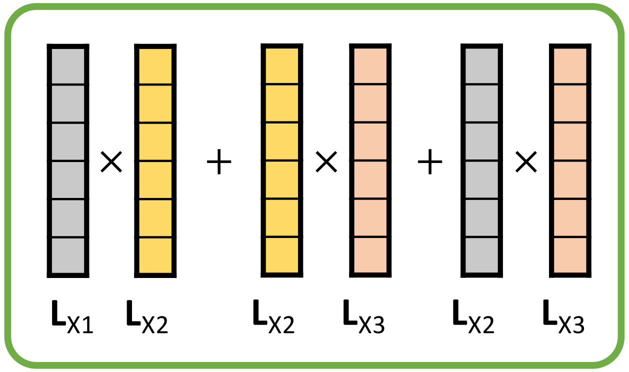
is **Max**

$$\tau \cdot \mathbf{a}^T\mathbf{a} + (1-\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}$$



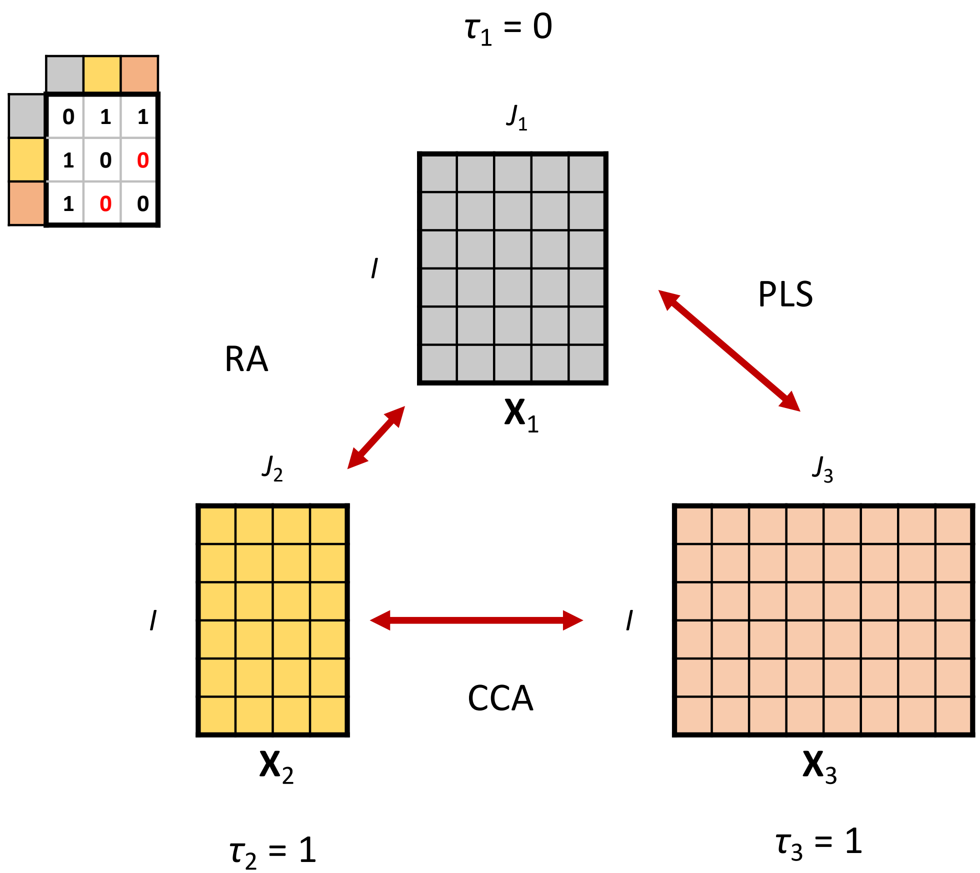
RGCCA

- Start from Component 1
  - Loadings $\mathbf{a}_X$
  - Latent variables
  - Connection matrix **C**



is **Max**

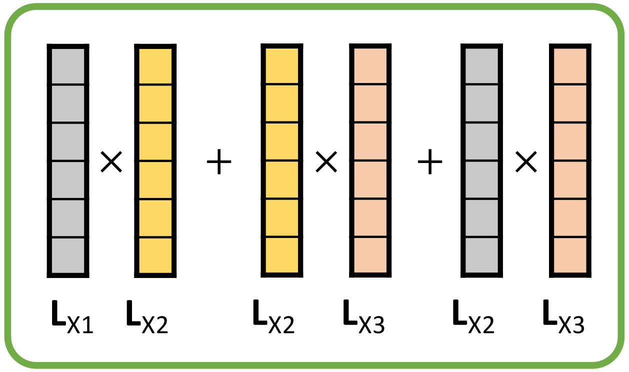
$$\boxed{\tau \cdot \mathbf{a}^T\mathbf{a} + (1\text{-}\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}}$$



Path model

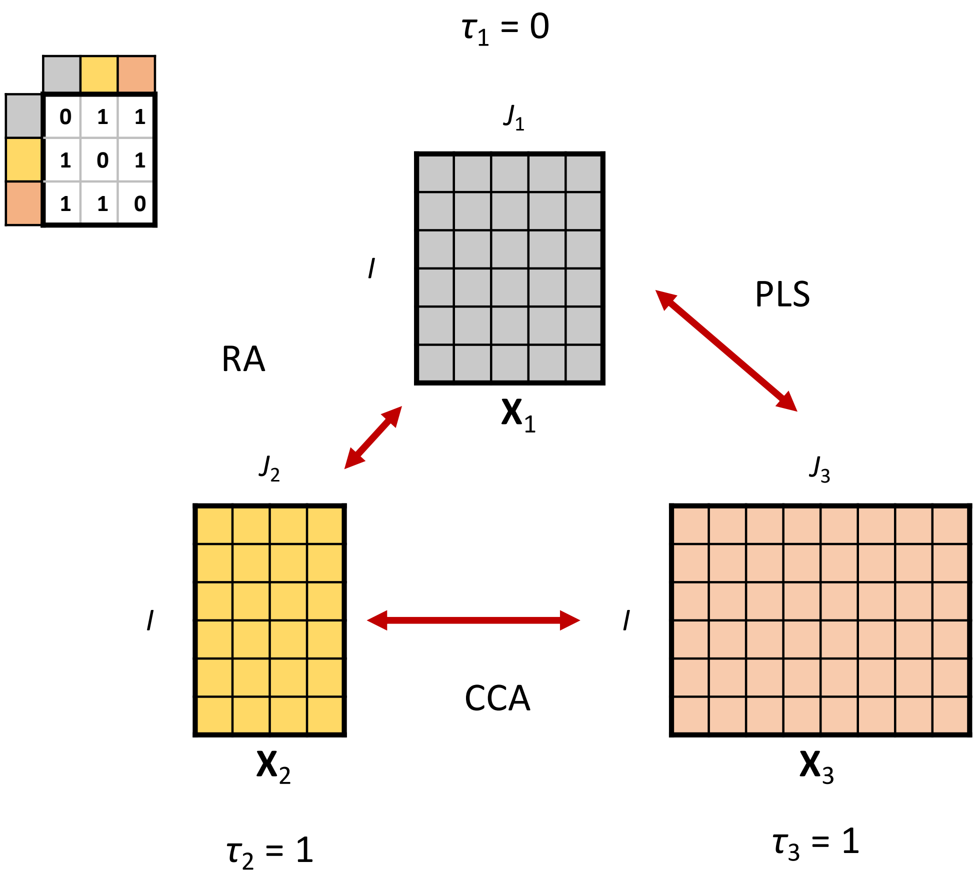
RGCCA

- Start from Component 1
  - Loadings $\mathbf{a}_X$
  - Latent variables
  - Connection matrix **C**



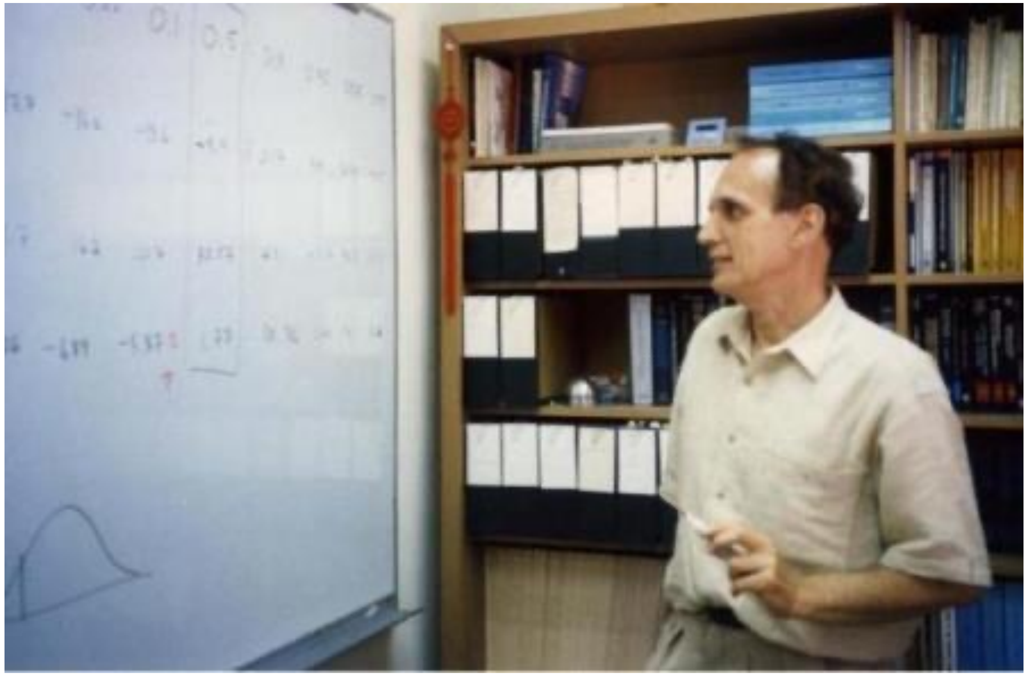
is **Max**

$$\tau \cdot \mathbf{a}^T\mathbf{a} + (1-\tau) \cdot \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} = \mathbf{I}$$

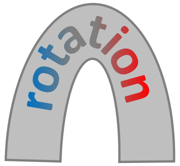


Sélectionner des variables : bootstrap, parcimonie, rotations etc.

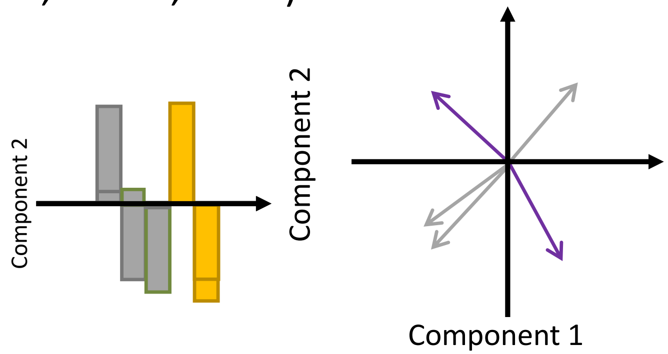
Bootstrap



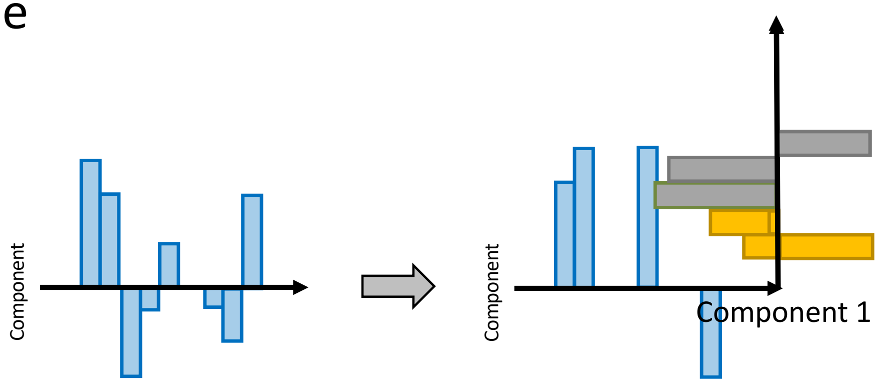
https://www.stat.auckland.ac.nz/~wild/BootAnim/index.html



- In psychometric tradition: rotation (Thurstone, 1935; Kaiser, 1958)
  - Works when data have clear factor structure
  - Does not work when, …
    - Number of dimensions is unknown *a-priori*
    - Current data-driven approach



- Modern statistics: sparsification (Tibshirani, 1996; Zou, Hastie, & Tibshirani, 2006)
  - Simplify the interpretation
  - Explain the largest amount of variance
  - Smallest number of strong variables

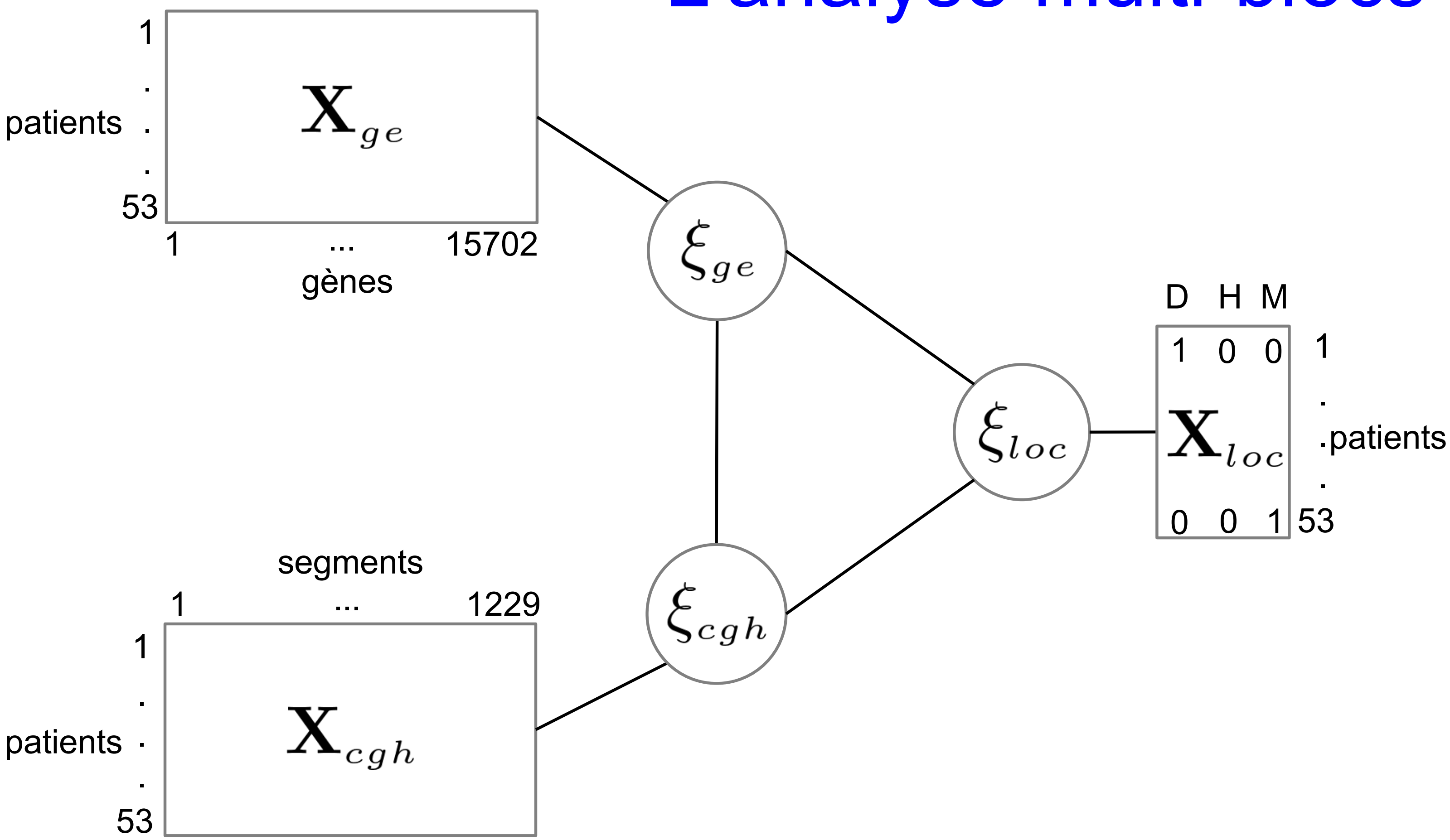


Credits: Ju-Chi Yu

RGCCA en pratique

Sur les données fruits parce que c’est plus facile

L'analyse multi-blocs

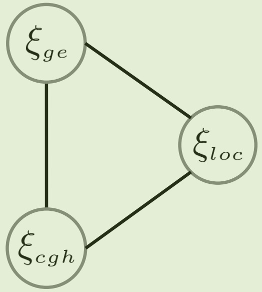


Une méthode adaptative

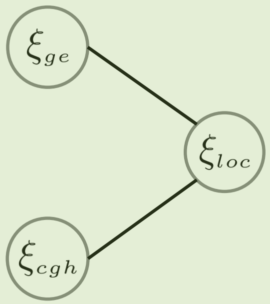
Schémas : g(x)

- $g(x) = x$
  Horst
- $g(x) = |x|$ Centroïde
- $g(x) = x^2$
  Factoriel

Dessins : $\mathbf{C} = (c_{jk})$



$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

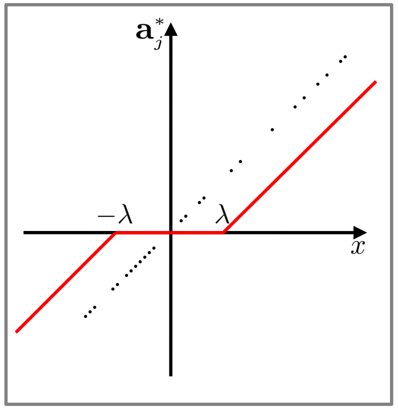


$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Le critère SGCCA

Sparse Generalized Canonical Correlation Analysis

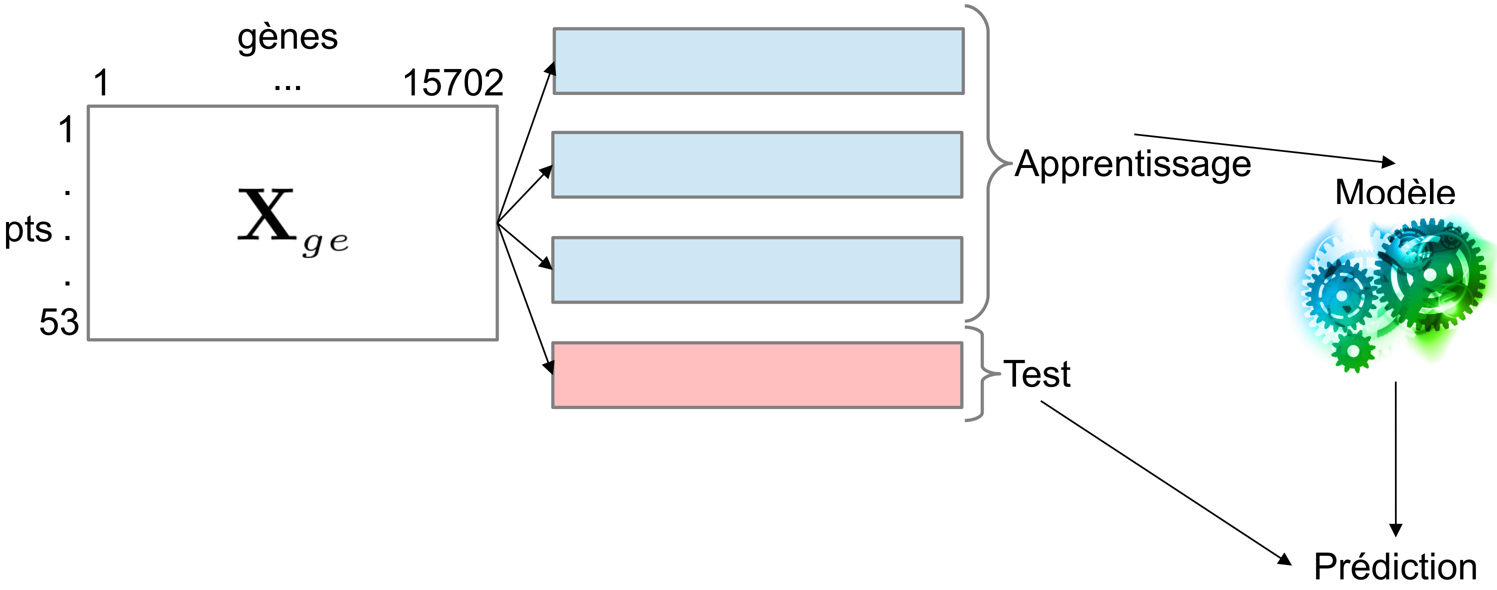
Soient J blocs $\mathbf{X}_j$ de variables centrées, de dimensions n x p_j décrivant n individus.
Soit un réseau de connexions entre les blocs, défini par la matrice $\mathbf{C} = (c_{jk})$: $c_{jk} = 1$ si $\mathbf{X}_j$ et $\mathbf{X}_k$ sont connectés et 0 sinon.
Ici, $\tau = 1$, pour j=1, …, J.
Application d'une pénalité sur la norme l_1 des vecteurs de poids $\mathbf{a}_j$.



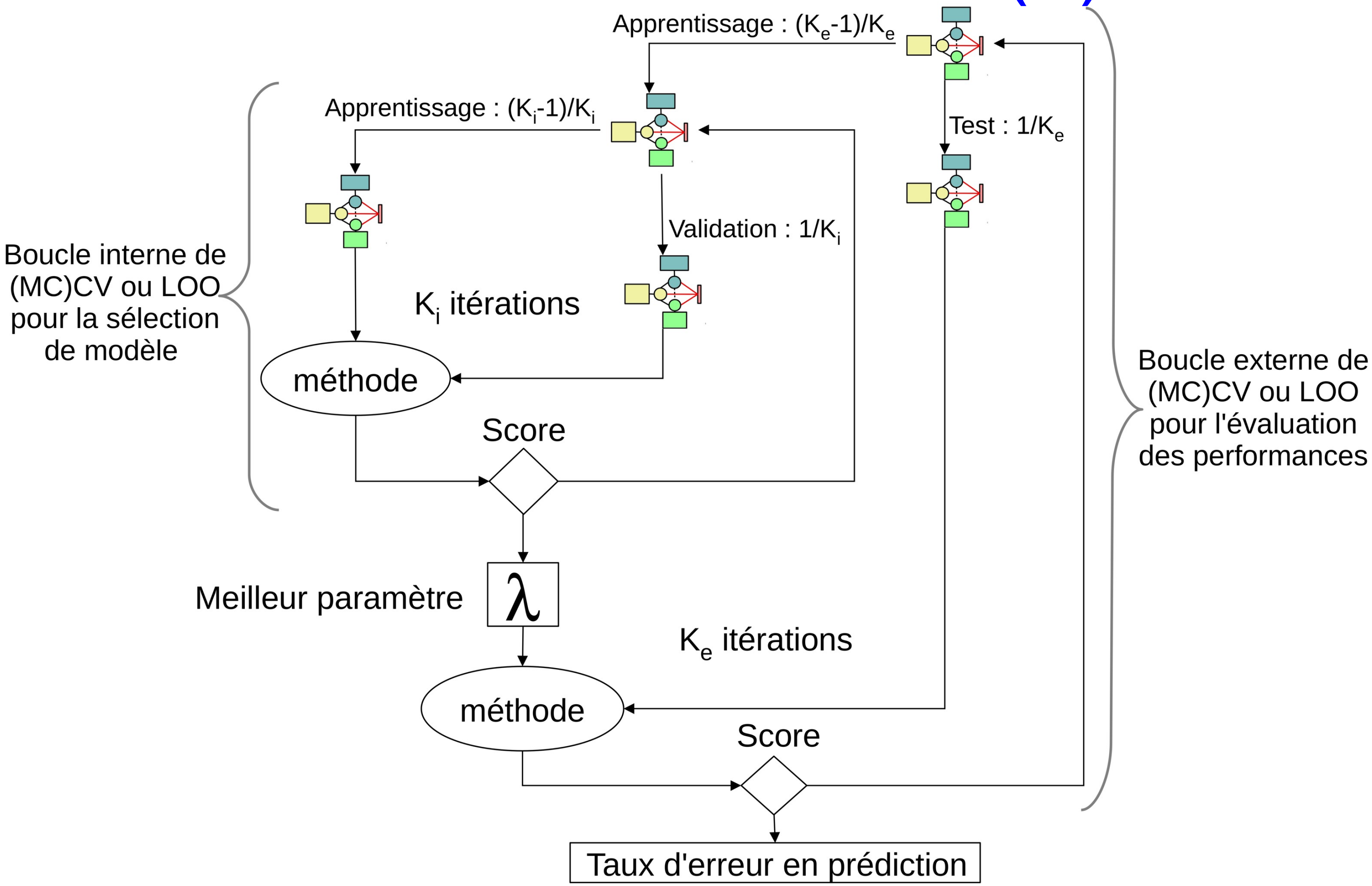
$$\begin{cases} \max\limits_{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_J} & \sum_{j,k=1; j\neq k}^{J} c_{jk}\, g(\mathrm{cov}(\mathbf{X}_j \mathbf{a}_j, \mathbf{X}_k \mathbf{a}_k)) \\ \text{sous contraintes} & \|\mathbf{a}_j\|_2^2 = 1 \text{ et } \|\mathbf{a}_j\|_1 \leq s_j, j = 1, \ldots, J \end{cases}$$

Tenenhaus, Philippe *et al* (2014)

Validation croisée (I)

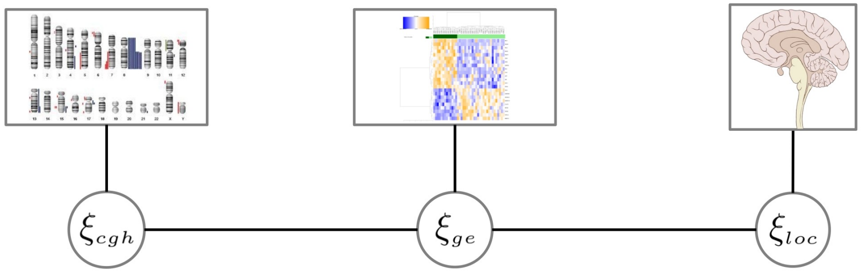


Validation croisée (II)

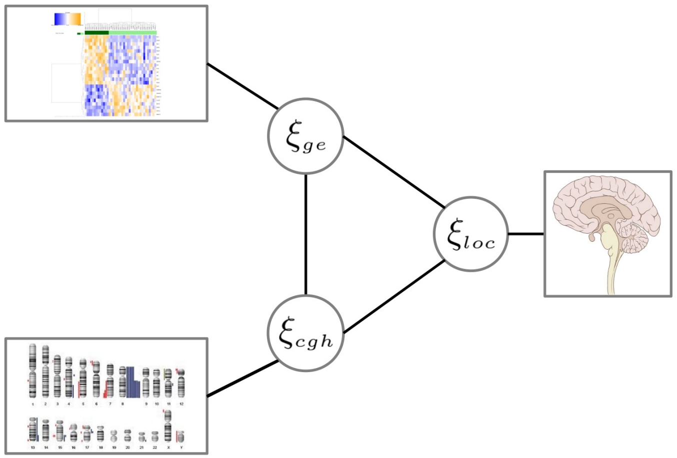


Application aux données pHGG

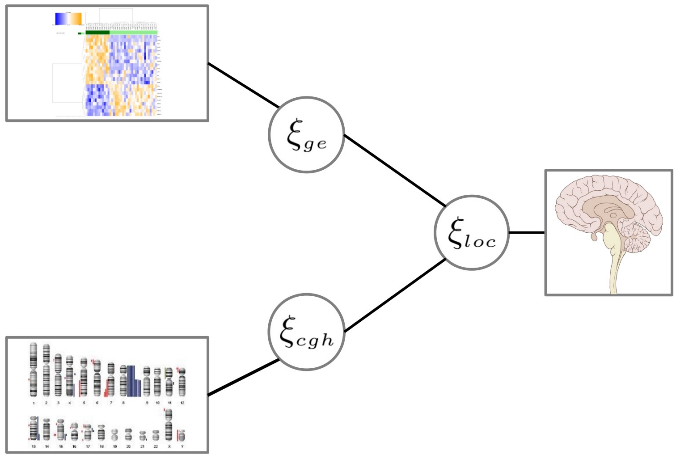
Cascade



Complet

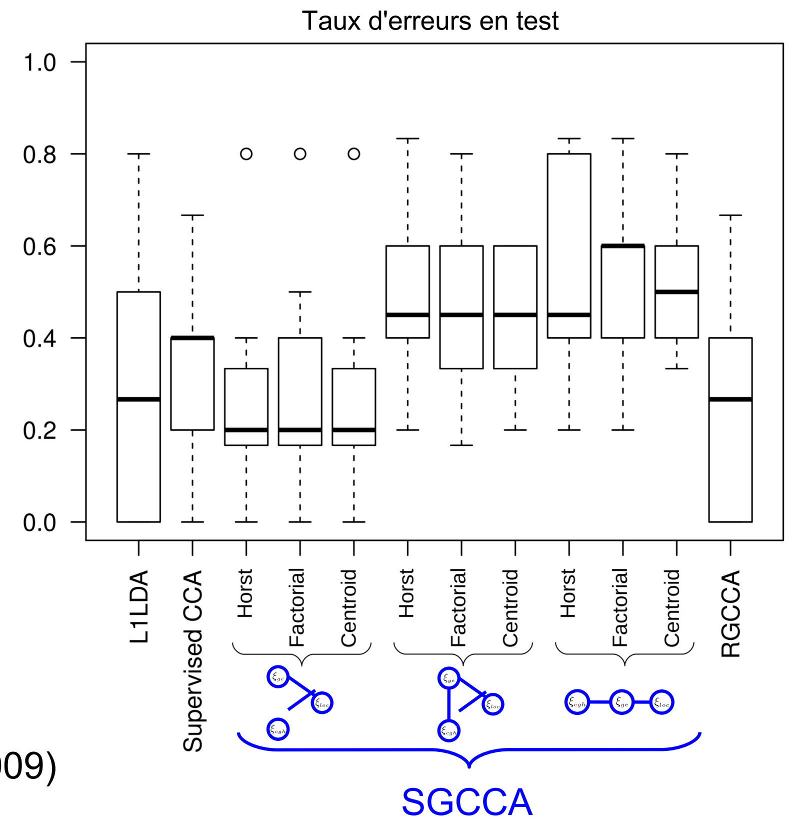


Hiérarchique



Performances en prédiction

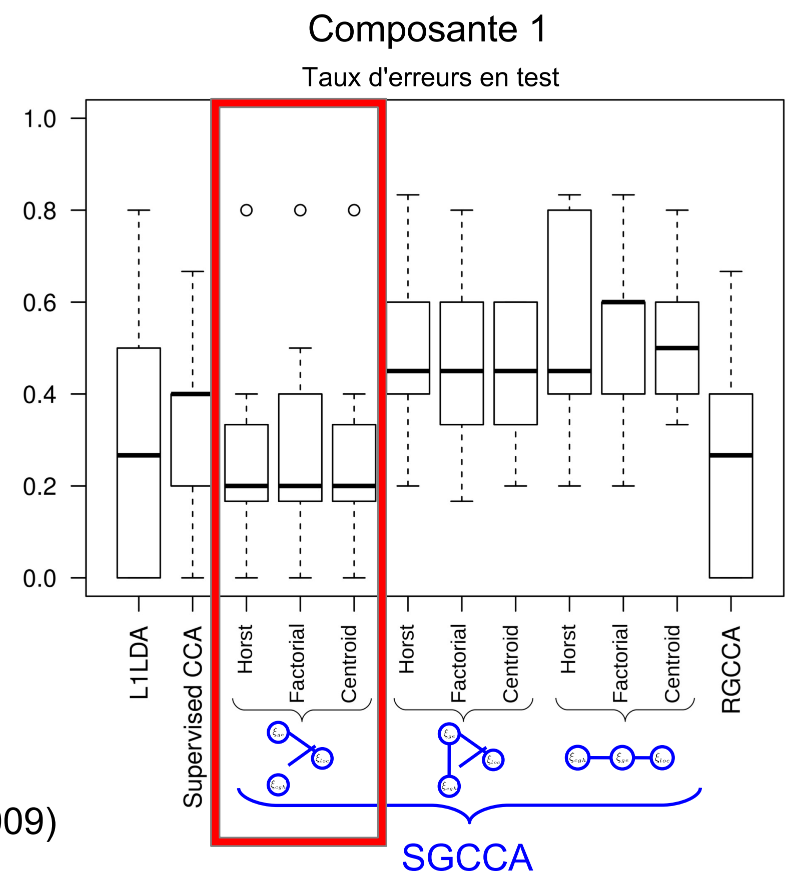
Composante 1



Supervised CCA
Witten et Tibshirani (2009)

L1-LDA
Witten et Tibshirani (2011)

Performances en prédiction

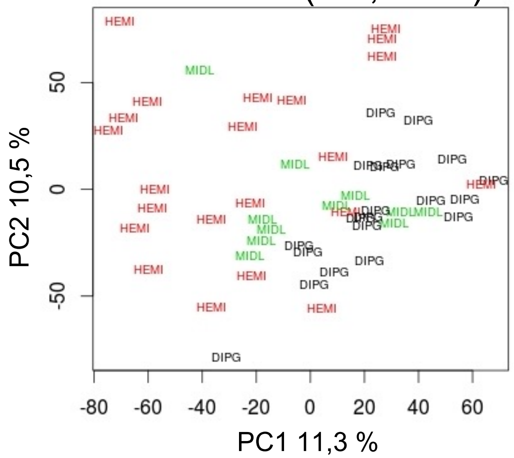


Supervised CCA
Witten et Tibshirani (2009)

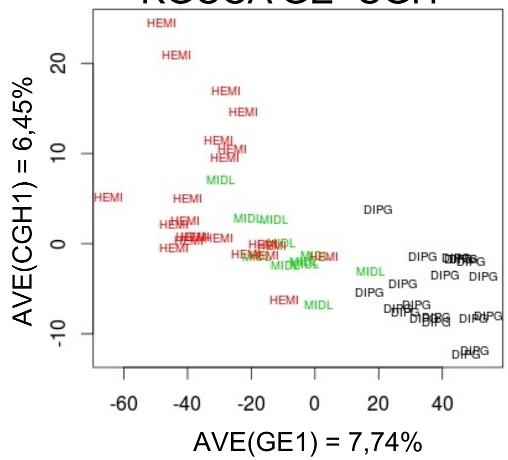
L1-LDA
Witten et Tibshirani (2011)

Visualisation des données pHGG

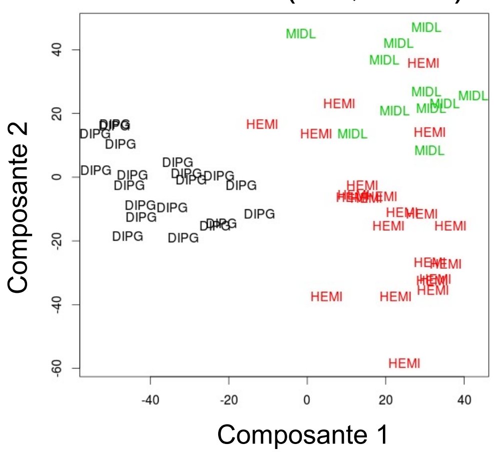
ACP concat(GE, CGH)



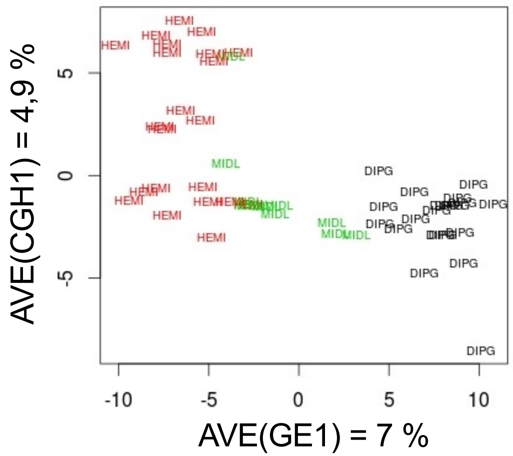
KGCCA GE+CGH



PLS concat(GE, CGH)



SGCCA GE+CGH



Stabilité des signatures

$$\kappa_{Fleiss} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Method	Fleiss' κ (GE)	Length of the GE signature	Fleiss' κ (CGH)	Length of the CGH signature
Supervised CCA	0.130	455.3	0.116	36.5
ℓ_1-LDA	**0.476**	**9790.0**	0.322	480.9
Horst SGCCA (Design 1)	0.103	132.2	0.071	35.9
Factorial SGCCA (Design 1)	0.071	79.0	0.014	59.6
Centroid SGCCA (Design 1)	0.137	73.6	0.105	22.6
Horst SGCCA (Design 2)	**0.468**	**61.1**	**0.296**	**33.6**
Factorial SGCCA (Design 2)	**0.439**	**42.0**	**0.343**	**37.6**
Centroid SGCCA (Design 2)	**0.478**	**40.6**	**0.317**	**34.8**
Horst SGCCA (Design 3)	0.071	83.6	0.074	40.7
Factorial SGCCA (Design 3)	0.061	118.3	0.026	49.2
Centroid SGCCA (Design 3)	0.040	75.5	0.035	40.2

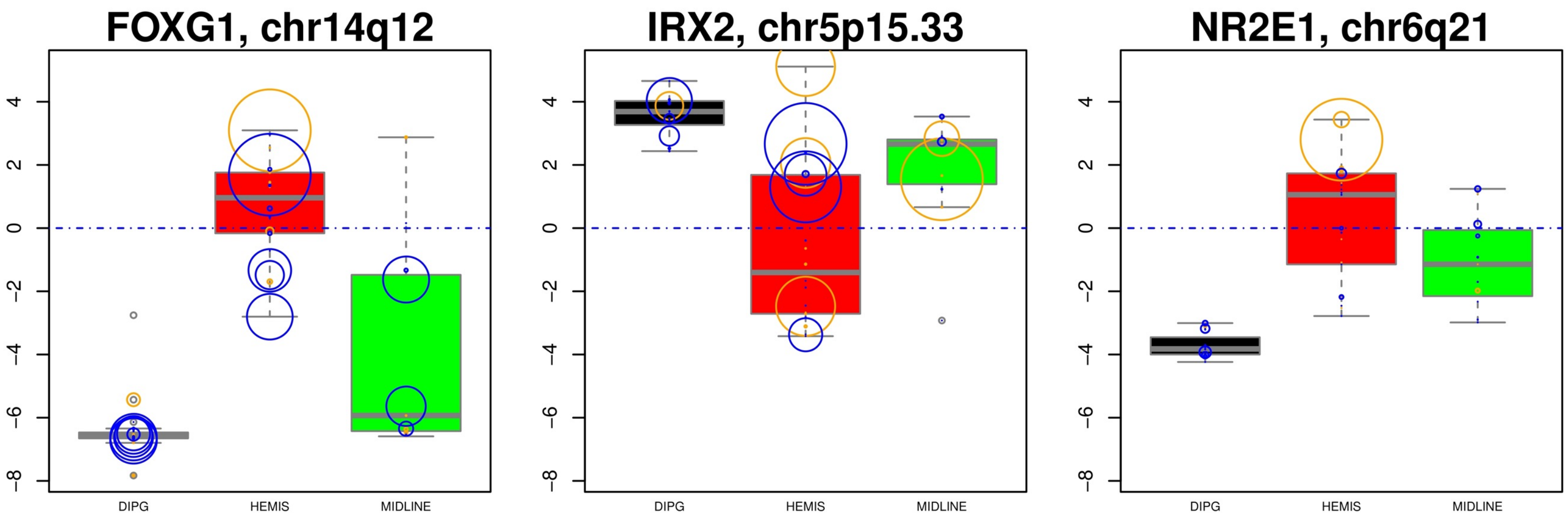
Stabilité des signatures

$$\kappa_{Fleiss} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Method	Fleiss' κ (GE)	Length of the GE signature	Fleiss' κ (CGH)	Length of the CGH signature
Supervised CCA	0.130	455.3	0.116	36.5
ℓ_1-LDA	**0.476**	**9790.0**	0.322	480.9
Horst SGCCA (Design 1)	0.103	132.2	0.071	35.9
Factorial SGCCA (Design 1)	0.071	79.0	0.014	59.6
Centroid SGCCA (Design 1)	0.137	73.6	0.105	22.6
Horst SGCCA (Design 2)	**0.468**	**61.1**	**0.296**	**33.6**
Factorial SGCCA (Design 2)	**0.439**	**42.0**	**0.343**	**37.6**
Centroid SGCCA (Design 2)	**0.478**	**40.6**	**0.317**	**34.8**
Horst SGCCA (Design 3)	0.071	83.6	0.074	40.7
Factorial SGCCA (Design 3)	0.061	118.3	0.026	49.2
Centroid SGCCA (Design 3)	0.040	75.5	0.035	40.2

Interprétation biologique

- SGCCA : signature de 82 gènes dont 23 impliqués dans le développement et l'organisation spatiale du cerveau.



- Identification de plusieurs gènes de la voie Wnt : SFRP2, WNT5A, DAAM2, FZD7, VAX2.

Zhang *et al* (2011)

Conclusion (I)

- RGCCA : cadre statistique général
  - Vaste exploration des données
  - Faibles temps de calcul
- KGCCA :
  - Gestion des données de grandes dimensions
- SGCCA :
  - listes de variables très courtes
- Multiblog :
  - Prédiction d'une variable binaire
  - Optimisation du calcul
- Multiblox :
  - Modélisation du risque instantané

RGCCA

Tenenhaus et Guillemot (2013)

Grande couverture des questions biologiques fréquemment rencontrées

Conclusion (II)

- Identification de gènes liés à la localisation et donc peut-être liés à la tumorigenèse des gliomes malins pédiatriques → validation en humide
- Outils pour modéliser le risques instantané de décès → mise en œuvre de la version parcimonieuse sur la cohorte afin de sélectionner les gènes liés au pronostic
- Rôle modeste des données de CGH

La liste des méthodes

Methods	$g(x)$	τ_j	$\mathbf{C}$
Canonical Correlation Analysis (Hotelling 1936)	x	$\tau_1 = \tau_2 = 0$	$\mathbf{C}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
Inter-battery Factor Analysis (Tucker 1958) or **PLS Regression** (Wold, Martens, and Wold 1983)	x	$\tau_1 = \tau_2 = 1$	$\mathbf{C}_1$
Redundancy Analysis (Van den Wollenberg 1977)	x	$\tau_1 = 1\,;\, \tau_2 = 0$	$\mathbf{C}_1$
Regularized Redundancy Analysis (Takane and Hwang 2007; Bougeard, Hanafi, and Qannari 2008; Qannari and Hanafi 2005)	x	$0 \le \tau_1 \le 1\,;\, \tau_2 = 0$	$\mathbf{C}_1$
Regularized Canonical Correlation Analysis (Vinod 1976; Leurgans *et al.* 1993; Shawe-Taylor and Cristianini 2004)	x	$0 \le \tau_1 \le 1\,;$ $0 \le \tau_2 \le 1$	$\mathbf{C}_1$

Methods	$g(x)$	τ_j	$\mathbf{C}$
SUMCOR (Horst 1961)	x	$\tau_j = 0, j = 1, \ldots, J$	$\mathbf{C}_2 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 1 \end{pmatrix}$
SSQCOR (Kettenring 1971)	x^2	$\tau_j = 0, j = 1, \ldots, J$	$\mathbf{C}_2$
SABSCOR (Hanafi 2007)	$\lvert x\rvert$	$\tau_j = 0, j = 1, \ldots, J$	$\mathbf{C}_2$
SUMCOV-1 (Van de Geer 1984)	x	$\tau_j = 1, j = 1, \ldots, J$	$\mathbf{C}_2$
SSQCOV-1 (Hanafi and Kiers 2006)	x^2	$\tau_j = 1, j = 1, \ldots, J$	$\mathbf{C}_2$
SABSCOV-1 (Tenenhaus and Tenenhaus 2011; Kramer 2007)	$\lvert x\rvert$	$\tau_j = 1, j = 1, \ldots, J$	$\mathbf{C}_2$
SUMCOV-2 (Van de Geer 1984)	x	$\tau_j = 1, j = 1, \ldots, J$	$\mathbf{C}_3 = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}$
SSQCOV-2 (Hanafi and Kiers 2006)	x^2	$\tau_j = 1, j = 1, \ldots, J$	$\mathbf{C}_3$
PLS path modeling - mode B (Wold 1982; Tenenhaus, Vinzi, Chatelin, and Lauro 2005)	$\lvert x\rvert$	$\tau_j = 0, j = 1, \ldots, J$	$c_{jk} = 1$ for two connected block and $c_{jk} = 0$ otherwise

Methods	$g(x)$	τ_j	$\mathbf{C}$
Generalized CCA (Carroll 1968a)	x^2	$\tau_j = 0, j = 1, \dots, J+1$	$\mathbf{C}_4 = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}$
Generalized CCA (Carroll 1968b)	x^2	$\tau_j = 0, j = 1, \dots, J_1$; $\tau_j = 1, j = J_1 + 1, \dots, J$	$\mathbf{C}_4$
Hierarchical PCA (Wold, S. and Kettaneh, N. and Tjessem, K. 1996)	x^4	$\tau_j = 1, j = 1, \dots, J$; $\tau_{J+1} = 0$	$\mathbf{C}_4$
Multiple Co-Inertia Analysis (Chessel and Hanafi 1996; Westerhuis *et al.* 1998; Smilde *et al.* 2003)	x^2	$\tau_j = 1, j = 1, \dots, J$; $\tau_{J+1} = 0$	$\mathbf{C}_4$
Multiple Factor Analysis (Escofier and Pages 1994)	x^2	$\tau_j = 1, j = 1, \dots, J+1$	$\mathbf{C}_4$

Les équations

Analyse en Composantes Principales

L'Analyse en Composantes Principales de la matrice des données centrées $\mathbf{X}$*, de dimension* $n \times p$*, avec éventuellement* $p > n$*, est la recherche d'une combinaison linéaire des variables de* $\mathbf{X}$*, notée* $\mathbf{y} = \mathbf{X}\mathbf{w}$ *(première composante principale de* $\mathbf{X}$*) avec* $\mathbf{w} \in \mathbb{R}^p$*, telle que la variance de* $\mathbf{y}$ *soit maximale. Cela revient à résoudre le problème d'optimisation suivant :*

$$\begin{cases} \mathbf{w}_1 = \underset{\mathbf{w}}{argmax} \;\; var(\mathbf{X}\mathbf{w}) \\ sous\ contrainte \quad \|\mathbf{w}\| = 1 \end{cases} \tag{3.1}$$

Traduction : on cherche une variable latente qui ait une variance maximale

Analyse des Corrélations Canoniques

*Posons $\mathbf{X}_1$ et $\mathbf{X}_2$, les blocs correspondants aux deux ensembles de variables mesurées sur un ensemble de n individus. L'**Analyse Canonique des Corrélations** de $\mathbf{X}_1$ et $\mathbf{X}_2$ est la recherche des vecteurs canoniques $\mathbf{a}_1 \in \mathbb{R}^{p_1}$ et $\mathbf{a}_2 \in \mathbb{R}^{p_2}$ tels que les variables canoniques $\mathbf{y}_1 = \mathbf{X}_1\mathbf{a}_1$ et $\mathbf{y}_2 = \mathbf{X}_2\mathbf{a}_2$ soient de corrélation maximale.*

$$\begin{cases} \underset{\mathbf{a}_1,\mathbf{a}_2}{max} \quad cor(\mathbf{X}_1\mathbf{a}_1, \mathbf{X}_2\mathbf{a}_2) \\ sous\ contraintes \quad var(\mathbf{X}_1\mathbf{a}_1) = var(\mathbf{X}_2\mathbf{a}_2) = 1 \end{cases} \tag{3.3}$$

Traduction : on cherche une variable latente par bloc qui soient les plus corrélées

Analyse Inter-Batteries

L'$\boxed{AFIB}$ de $\mathbf{X}_1$ et $\mathbf{X}_2$ est la recherche de combinaisons linéaires des variables de $\mathbf{X}_1$, notée $\mathbf{y}_1 = \mathbf{X}_1\mathbf{a}_1$ avec $\mathbf{a}_1 \in \mathbb{R}^{p_1}$ et une combinaison linéaire des variables de $\mathbf{X}_2$, notée $\mathbf{y}_2 = \mathbf{X}_2\mathbf{a}_2$ avec $\mathbf{a}_2 \in \mathbb{R}^{p_2}$ telle que la covariance entre ces deux variables soit maximale.

$$\begin{cases} \underset{\mathbf{y}_1 \text{ et } \mathbf{y}_2}{max}\ cov(y_1, y_2) = cor(\mathbf{X}_1\mathbf{a}_1, \mathbf{X}_2\mathbf{a}_2)\sqrt{var(\mathbf{X}_1\mathbf{a}_1)}\sqrt{var(\mathbf{X}_2\mathbf{a}_2)} \\ sous\ contraintes\ \ \|\mathbf{a}_1\| = \|\mathbf{a}_2\| = 1 \end{cases} \tag{3.6}$$

Traduction : on cherche une variable latente par bloc dont la covariance soit maximale

RGCCA

Regularized **G**eneralized **C**anonical **C**orrelation **A**nalysis
Regularized = compatible grandes dimensions (# variables >> # samples)

On considère J blocs $\mathbf{X}_j$ de variables centrées, de dimension $n \times p_j$ décrivant n individus. On considère également un réseau de connexions entre les blocs, en définissant la matrice $\mathbf{C} = (c_{jk})$: $c_{jk} = 1$ si $\mathbf{X}_j$ et $\mathbf{X}_k$ sont connectés et 0 sinon. On recherche les combinaisons linéaires standardisées $\mathbf{y}_j = \mathbf{X}_j\mathbf{a}_j$ solution du problème d'optimisation suivant :

$$\begin{cases} \underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{max} \sum_{j,k=1;j\neq k}^{J} c_{jk}\, g\left(cov(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)\right) \\ sous\ contraintes\ \ Var(\mathbf{X}_j\mathbf{a}_j) = \mathbf{a}_j^T \Sigma_{jj}\mathbf{a}_j = 1, \quad j = 1, \ldots, J \end{cases} \tag{3.25}$$

SGCCA

Sparse **G**eneralized **C**anonical **C**orrelation **A**nalysis
Sparse = sélection de variables

$$\begin{cases} \underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{\max} \quad \sum_{j,k=1;j\neq k}^{J} c_{jk} g(\operatorname{cov}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)) \\ \\ \text{sous contraintes} \quad \|\mathbf{a}_j\|_2 = 1 \text{ et } \|\mathbf{a}_j\|_1 \leq s_j, j = 1, \ldots, J \end{cases}$$