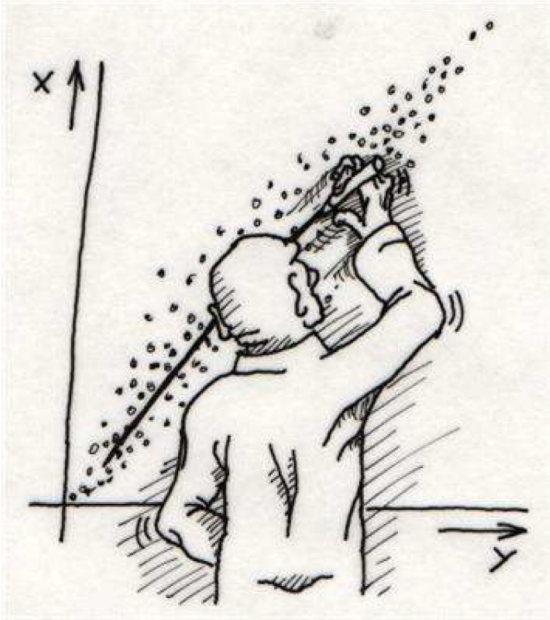
Recherche de composante(s) explicative(s) par pénalisation

UNIVERSITÉ DE MONTPELLIER

Yonatan Carlos Carranza Alarcon

Superviseur: Xavier Bry



Soutenu: 25/06/2016

Plan

I. Introduction
II. Régression
III. Modélisation
IV. Résultats et Conclusions

I. Introduction

- Problématique:
  - L'absence d'échantillon nécessaire pour pouvoir appliquer des méthodes de régression linéaire classique.
  - Les variables explicatives p sont plus grand que l'échantillon n:
    - rang(X) = p >>>n
    - X'X n'est pas inversible

- Motivation:
  - Ce problème est souvent retrouvée dans le domaine de la médecine, la biologie, la chimique et autres, où le processus d'échantillon est très coûteux et parfois des longues périodes.

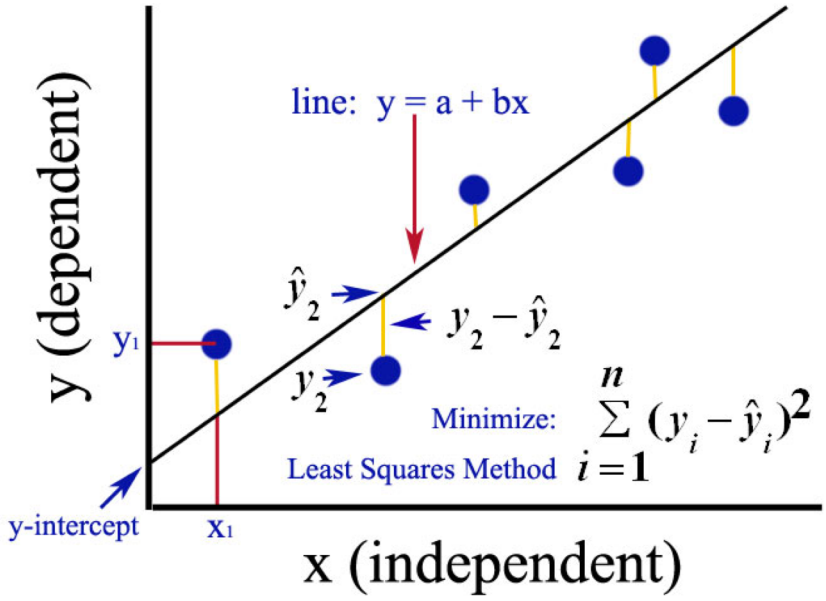
- Objectif
  - Proposer un modèle de régression qui améliore la précision de prévision. C'est-à-dire, ce dernier modèle doit minimiser l'erreur quadratique de prévision.

II. Régression

- Analyser le relation entre une variable par rapport à une autre ou plusieurs.
- Le plus connu est le modèle de régression linéaire.

$$Y = X\beta + \varepsilon$$

Y : variable à expliquer
X : variables explicatives
ε : erreur de mesure / bruit

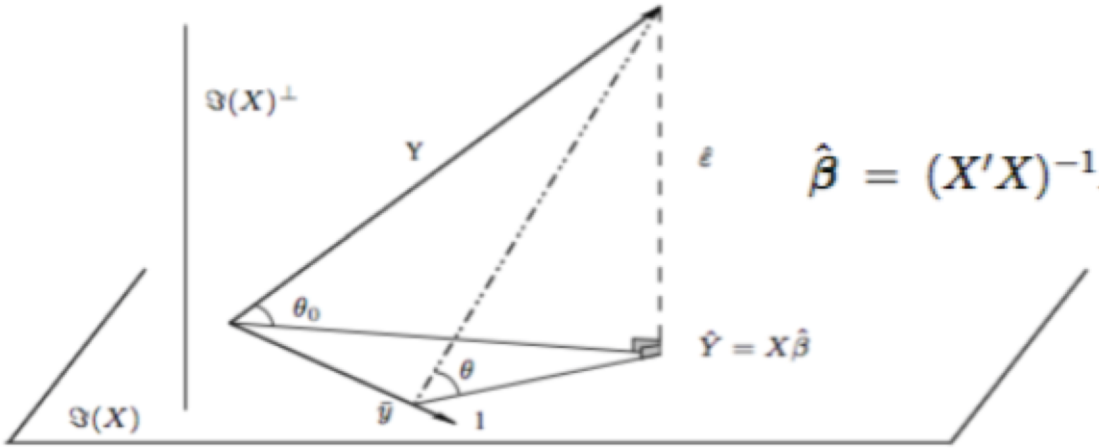


Régression linéaire multiple gaussienne

- Pour de cas où rang(X) <<< n
  - $\mathcal{H}_1 : rang(X) = p$
  - $\mathcal{H}_2 : \mathbb{E}[\varepsilon] = 0, \Sigma_\varepsilon = \sigma^2 \mathbb{I}$
  - $\mathcal{H}_3 : \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$, ils sont i.i.d.

Modèle: $Y = X\beta + \varepsilon$

$$\mathcal{L}(Y; \beta, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{i,j}\beta_i)^2\right]$$



$$\hat{\beta} = (X'X)^{-1}X'Y$$

Régression Ridge

- Pour de cas où rang(X) >>> n
- Réduction de la valeurs des coefficients moins importantes et augmente plus importantes.
- Augmente les valeurs propres de la matrice X'X

$$\hat{\beta}_{ridge} = \arg\min \left[\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + k \sum_{j=1}^{p} \beta_j^2 \right]$$

$$\hat{\beta}_{ridge}(k) = (X'X + k\mathbb{I})^{-1} X'Y$$

K : valeur entre [0,1] et elle est stochastique.

(X'X + κII) est dont inversible.

Régression Composantes Principales

- Le but consiste à ne conserver qu'une partie des composantes principales(variabilité ou inertie global).
- Supprime l'information négligeable (ou bruit) qui n'apporte pas d'information important ou modèle.

Le modèle avec les k premières composantes : $Y = X_1^* \beta_1^* + ... + X_k^* \beta_k^* + \epsilon$

$$\hat{\beta}_{pcr}(k) = (X^{*'} X^*)^{-1} X^{*'} Y$$
$$= (X_{[1:k}^{*'} X_{[1:k]}^*)^{-1} X_{[1:k]}^{*'} Y$$

- Quand s'arrête ?
- Quel est le k optimum ?

Régression de moindre carré partiels

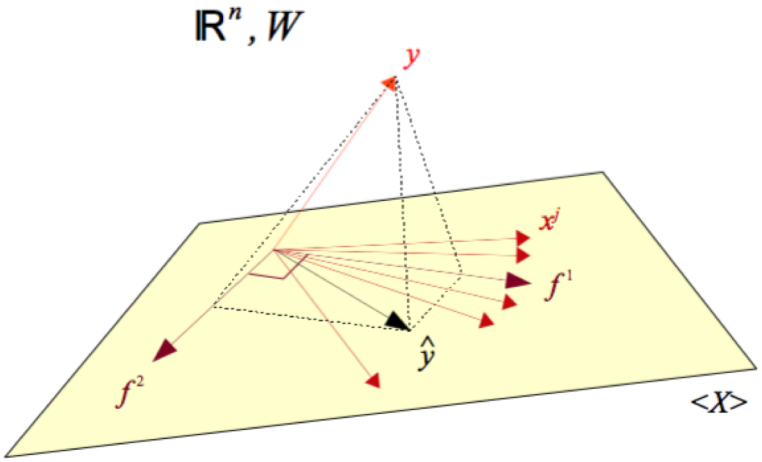
- Le but consiste à trouver des nouvelles variables explicatives T à partir des combinaisons linéaires de la variable explicative départ X.

- Ces dernières variables ne seront pas ajustées par la part de la variabilité qui représente las variables explicatives originales (comme PCR) sinon par leur lien avec la variable à expliquer Y.

$$t^{(k)} = \underset{\substack{t=X^{(k-1)}w,\ w\in\mathbb{R}^p \\ w'w=1}}{\arg\max} < t, Y^{(k)} >$$

$$Y^{(k)} = r_k t^{(k)} + \hat{\epsilon}_k$$

- Quand s'arrête ?
- Quel est le k optimum ?



III. Modélisation

Les régresseurs (ou variables explicatives) sont partitionnés en deux groupes : $X = \{x^1, ..., x^p\}$, dans lequel nous cherchons à faire de la réduction dimensionnelles sous la forme d'une combinaison linéaire x_j, et $T = \{f^1, ..., f^K\}$, qui rassemblera les K composantes principales trouvées de X.

$$Y = X\beta + T\delta + \epsilon \; ; \quad \epsilon \sim N(0; \sigma^2 \mathbb{I})$$
$$\text{où} \quad \beta = \gamma u \quad \text{et} \quad u'u = 1$$
$$\text{donc} \quad \gamma = \|\beta\| \quad \text{et} \; u = \frac{\beta}{\|\beta\|}$$

$$\mathcal{G} = \underset{\substack{\delta, \gamma, \sigma^2, u \\ u'u=1}}{\arg\max} \left[(1-s)\ell + sS(u) \right]$$

Où, ℓ est la log-vraisemblance de $f_Y \sim N(X\beta + T\delta; \sigma^2 \mathbb{I})$

Modèle univarié – T vide

La modélisation: $Y \sim N(X\beta + T\delta; \sigma^2\mathbb{I})$

$$\mathcal{L}(Y; u, \gamma, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right)$$

$$\mathcal{G}' = \underset{\substack{\gamma,\sigma^2,u \\ u'u=1}}{\arg\max} \left[(s-1)\frac{n\ln(\sigma^2)}{2} + (s-1)\frac{\|Y - X\beta\|^2}{2\sigma^2} + su'Nu\right]$$

$$\text{On pose : } \Omega = \left[(s-1)\gamma^2 X'X + 2\sigma^2(sN - \lambda\mathbb{I})\right]$$

$$\iff \hat{u} = \gamma(s-1)\Omega^{-1}X'Y$$

- Le valeur de réglage s est stochastique, elle s'adapte aux données.
- Cependant la matrice Ω pendant l'optimisation n'est pas inversible !!!!!

Modèle univarié – T vide

Théorème 2.1 *Soit X une matrice $m \times p$ dont les coefficients appartiennent au corps K, où $K = \mathbb{R}$ ou $K = \mathbb{C}$. Alors il existe une factorisation, nommé* Décomposition en Valeurs Singulières *de X, de la forme :*

$$X = U\Sigma V'$$

Où si $K = \mathbb{R}$, les matrices unitaires sont de matrices orthogonales et :

- *U une matrice unitaire $m \times m$ sur K.*
- *Σ une matrice diagonale rectangulaire $m \times n$ dont les coefficients diagonaux sont des réels positifs ou nuls et tous les autres sont nuls.*
- *V est la matrice unitaire $n \times n$ sur K.*

En applicant le théorème 2.1 dans notre estimateur u, nous avons donc :

$$\begin{aligned} X = U\Sigma V' &\iff X'X = V\Sigma^2 V' \text{ et } N = V\Sigma W^*\Sigma V', \text{ où } W^* = U'WU \\ &\iff \Omega^* = \left[(s-1)\hat{\gamma}^2\Sigma^2 + 2\hat{\sigma}^2(s\Sigma W^*\Sigma - \hat{\lambda}\mathbb{I})\right] \\ &\iff \hat{u} = \hat{\gamma}(s-1)V{\Omega^*}^{-1}\Sigma U'Y \end{aligned} \tag{2.13}$$

D'où, la matrice Ω^* est une matrice carrée diagonale et elle est donc pseudo-inversible.

Modèle univarié – T non vide

La modélisation: $Y \sim N(X\beta + T\delta; \sigma^2\mathbb{I})$

$$\mathcal{L}_Y(X,T;u,\delta,\gamma,\sigma^2) = \prod_{i=1}^{n} f_Y(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta - T\delta\|^2\right)$$

$$\mathcal{Q} = \underset{\substack{\delta,\gamma,\sigma^2,u \\ u'u=1 \\ u_k'X'WXu=0}}{\arg\max} \left[(s-1)\frac{n\ln(\sigma^2)}{2} + (s-1)\frac{\|Y - \gamma Xu - T\delta\|^2}{2\sigma^2} + su'Nu\right]$$

Définition 2.1 *Nous soumettrons à notre programme à optimiser la contrainte d'orthogonalité aux composantes principales précédents. C'est-à-dire :* $\forall k \in \{1,...,K\}$, $f_k = Xu_k$ *et* $F^{k-1} = [f^1,...,f^{k-1}]$;

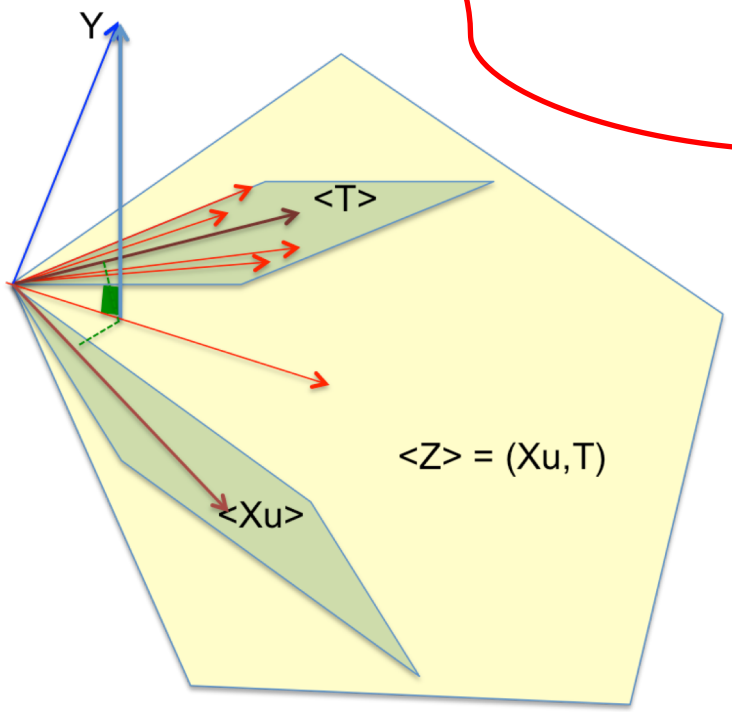
$$\forall k : f^1,...,f^{k-1} \perp f^k \iff F^{k-1}Wf^k = 0 \iff F^{k-1}WXu_k = 0$$

D'où, la covariance des composantes principales est :$cov(f_i,f_j) = 0, \forall i,j \in \{1,...,K\}$ *et* $j \neq i$.

Modèle univarié – T non vide

- Étape (1):

$$Q' = \arg\max_{\delta,\gamma,\sigma^2} \left[n\ln(\sigma^2) + \frac{\left\| Y - \gamma X \hat{u}_{[t]} - T\delta \right\|^2}{\sigma^2} \right]$$



$$Q'' = \arg\max_{\delta,\gamma,\sigma^2} \left[n\ln(\sigma^2) + \frac{\|Y - Z\Lambda\|^2}{\sigma^2} \right]$$

$$\hat{\Lambda} = (Z'Z)^{-1}Z'Y \qquad \hat{\Lambda} = (\hat{\gamma}, \hat{\delta})'.$$

Modèle univarié – T non vide

- Étape (2):

$$Q''' = \underset{\substack{u,\ u'u=1 \\ A_k'u=0 \\ \text{où } A_k'=u_kX'WX}}{\arg\max} \left[(s-1)\frac{\left\| Y - \hat{\gamma}_{[t]} X u - T\hat{\delta}_{[t]} \right\|^2}{2\hat{\sigma}^2_{[t]}} + su'Nu \right]$$

où $A_k' = u_k' X'WX \iff A_k' u = 0, \forall k \in \{1, 2, ..., K\}$

Nous posons : $\Omega = \left[(s-1)\hat{\gamma}^2 X'X + 2\hat{\sigma}^2(sN - \hat{\lambda}\mathbb{I}) \right]$

$$\iff \hat{u}_{[t+1]} = \Omega^{-1}\left[\hat{\gamma}(s-1)X'\left(Y - T\hat{\delta}\right) + \hat{\sigma}^2 A_k \hat{\tau} \right]$$

Ω Pas inversible, encore SVD !!!!

En posant : $\Omega^* = \left[(s-1)\hat{\gamma}^2\Sigma^2 + 2\hat{\sigma}^2(s\Sigma W^*\Sigma - \hat{\lambda}\mathbb{I}) \right]$, où : $W^* = U'WU$

$$X = U\Sigma V' \iff \hat{u}_{[t+1]} = V\Omega^{*-1}\left[\hat{\gamma}(s-1)\Sigma U'\left(Y - T\hat{\delta}\right) + \hat{\sigma}^2 \Sigma W^* \Sigma V' u_k \hat{\tau} \right] \quad (2.28)$$

- Condition d'arrêt:

$$\left\| \hat{u}_{[t]} - \hat{u}_{[t+1]} \right\|^2 > 10^{-6} \text{ ou } \langle\, \hat{u}_{[t]} \mid \hat{u}_{[t+1]} \,\rangle^2 < 1 - 10^{-6}.$$

Validation du modèle de régression

- Centrage et réduction

$$\hat{\sigma}_{X_j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 / n$$

$$\tilde{X} = (X_j - \bar{x}_j \mathbb{I}) / \hat{\sigma}_{X_j} \qquad \bar{x}_j = \sum_{i=1}^{n} x_{ij} / n \qquad \tilde{Y} = Y - \bar{y}$$

- Prévision d'un nouveau échantillon

$$\hat{y}^{p}_{xbry,n+1} = \left[\sum_{j=1}^{p} \left(\frac{x_{n+1,j} - \bar{x}_j}{\hat{\sigma}_{X_j}} \left[\hat{\beta}_{xbry}(\bar{s}) \right]_j \right) \right] + \left[\sum_{j=1}^{p} \left(\frac{x_{n+1,j} - \bar{x}_j}{\hat{\sigma}_{X_j}} \left[\hat{u}_k(\bar{s}) \hat{\delta}(\bar{s}) \right]_j \right) \right] + \bar{y}$$

- Apprentissage et Validation Croisé

$$PRESS(s) = \left\| \hat{Y}^{p}_{xbry,v}(s) - Y_v \right\|^2$$

$$(\tilde{X}_a, \tilde{Y}_a) \text{ et } (\tilde{X}_v, \tilde{Y}_v)$$

Choisir minimum

$$\hat{s} = \underset{s \in]0,1[}{\arg\min} \left\| \hat{Y}^{p}_{xbry,v}(s) - Y_v \right\|^2$$

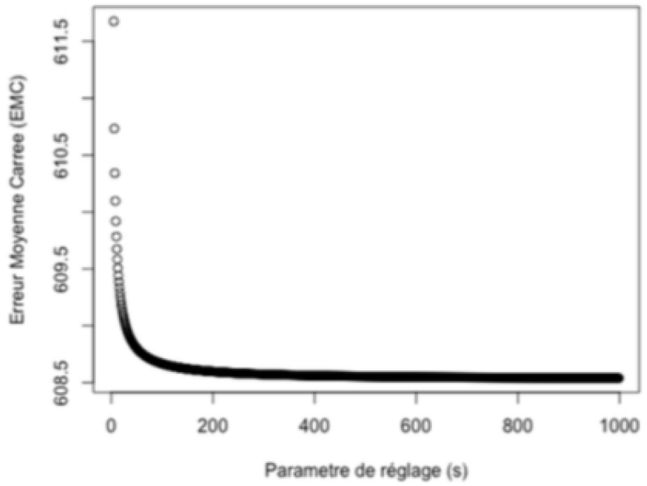
IV. Résultats et Conclusions
Jeu de données

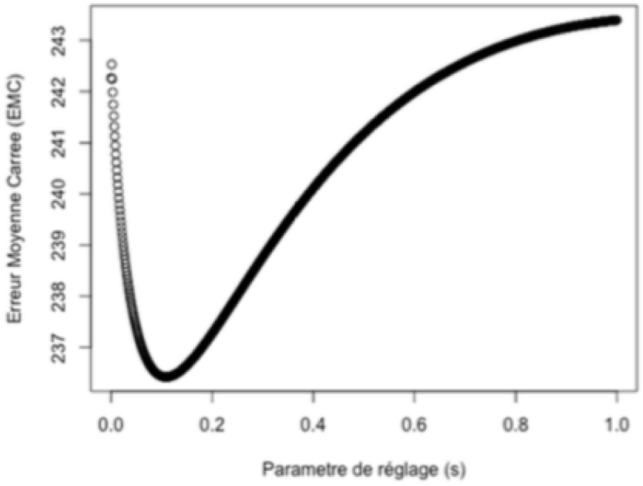
- Nous sommes en présence de biscuits non cuits pour lesquels nous souhaitons connaître rapidement et à moindre coût, la composition en quatre ingrédients : les lipides, les sucres, la farine et l'eau

- Processus alternatif est la mesure d'un spectre d'absorbance dans le domaine proche infrarouge. Ce processus nous donne 700 variables explicatives pour trouver la composition du biscuit non cuit.

- Echantillons de validation et apprentissage:
  - 700 Variables explicatives
  - 40 biscuits non cuits pour l'apprentissage
  - 32 biscuits non cuits pour la validation

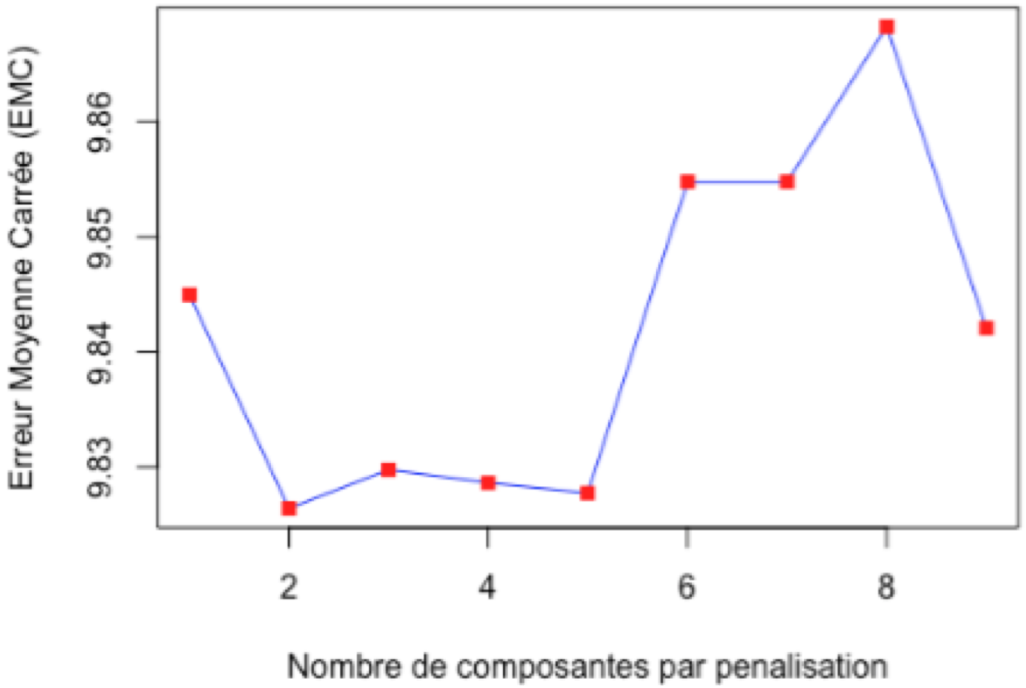
Première Composante Principale





Modèle de Regression	Paramètre de réglage/nb. Composantes	EMQ*
Moindre carrés multiple	—	4304
Ridge	0.1081081	4.95
Composantes principales (PCR)	3 Comps.	1.03
Moindres carrés partiels (PLS)	5 Comps.	0.78
Nouvelle approche	T vide et 0.998	14.79

K Composante principale permises



Modèle de Regression	Paramètre de réglage/nb. Composantes	EMQ*
Moindre carrés multiple	—	4304
Ridge	0.1081081	4.95
Composantes principales (PCR)	3 Comps.	1.03
Moindres carrés partiels (PLS)	5 Comps.	0.78
Nouvelle approche	T avec 2 comps. et 0.9391	9.8264
Nouvelle approche	T avec 5 comps. et 0.9451	9.8277

Avec 2 composantes à peu près 98% Information ACP

Conclusions

- La généralisation du modèle de régression univariée à un modèle multivariée (i.e. la variable explicative a plusieurs colonnes à expliquer), aurait enté intéressante à l'analyser et le comparer avec les méthodes existantes.

- Enfin et le plus intéressant aurait été de modéliser autres variantes de la fonction de pertinence structurelle S(u), par exemple :

$$S(u) = \left(\sum_{j=1}^{J} (u' N_j u)^l\right)^{1/l}.$$

Des questions ?