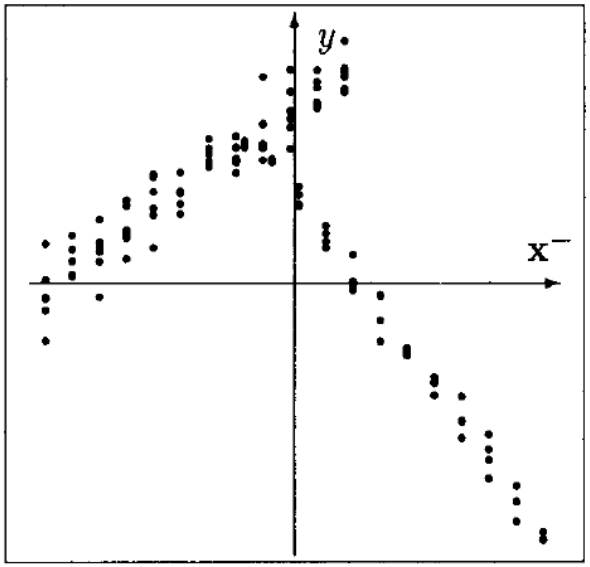
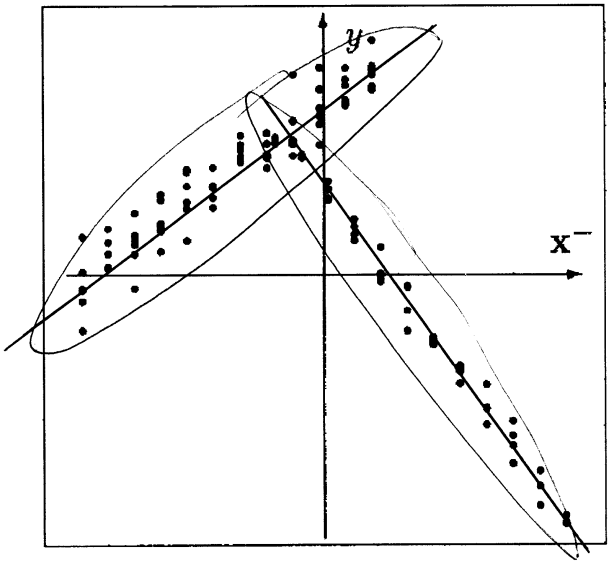
Les méthodes « clusterwise »

Gilbert Saporta (CNAM)
Stéphanie Bougeard (ANSES)
Ndeye Niang-Keita (CNAM)

1.Introduction

- Modèles locaux *versus* modèle global





Adapté de Hennig, 2000

- Classes inconnues *a priori:*
  - *Hétérogenéité non observée ; variable latente catégorielle*
  - *Partitionnement explicatif: recherche simultanée des classes et des modèles par classe*

- Premiers travaux: méthodes typologiques (Diday, 1974) ou clusterwise (Späth, 1979)

- Deux écoles de pensée
  - Méthodes géométriques *versus* mélanges finis
  - K-means *versus* maximum de vraisemblance
  - Analyse des données *versus* statistique mathématique

- Deux objectifs
  - Modéliser *versus* prédire
  - Critères : goodness of fit *versus* capacité prédictive

2. Algorithmes de type k-means

2.1 L'analyse factorielle typologique

E. DIDAY

Introduction à l'analyse factorielle typologique

Revue de statistique appliquée, tome 22, n° 4 (1974), p. 29-38.

<http://www.numdam.org/item?id=RSA_1974__22_4_29_0>

« Recherche simultanée de k variétés d'inertie minimum, ou plans factoriels locaux. On utilise l'algorithme des nuées dynamiques avec des variétés pour noyaux »

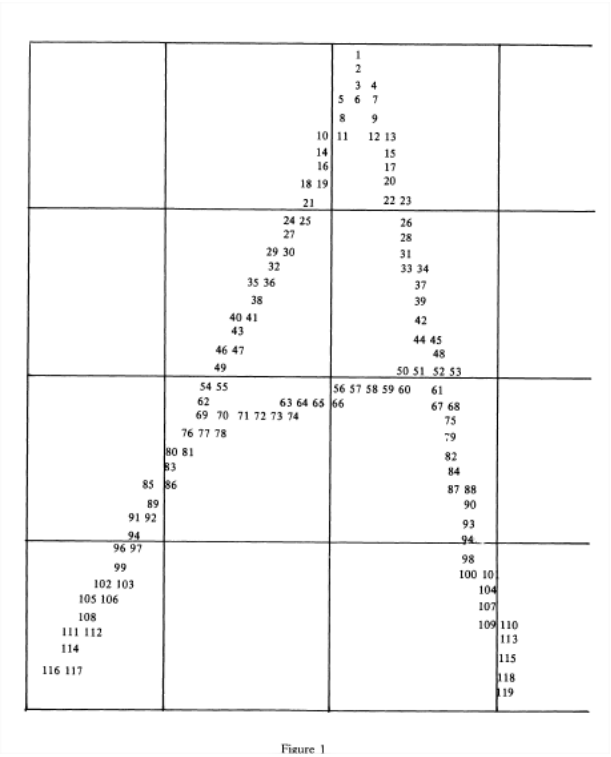


Figure 1

- Algorithme des nuées dynamiques avec distances aux plans factoriels à la place des distances aux centres de gravité
- Rappel: deux variantes principales
  - Mise à jour du critère et de la partition après chaque réaffectation (vrai k-means) algo « stochastique »
  - Mise à jour du critère et de la partition après une passe complète sur les n observations algo « batch »

2.2 Régression typologique ou clusterwise

– Pour chaque classe on ajuste un modèle linéaire, de manière à minimiser la somme des carrés des résidus.

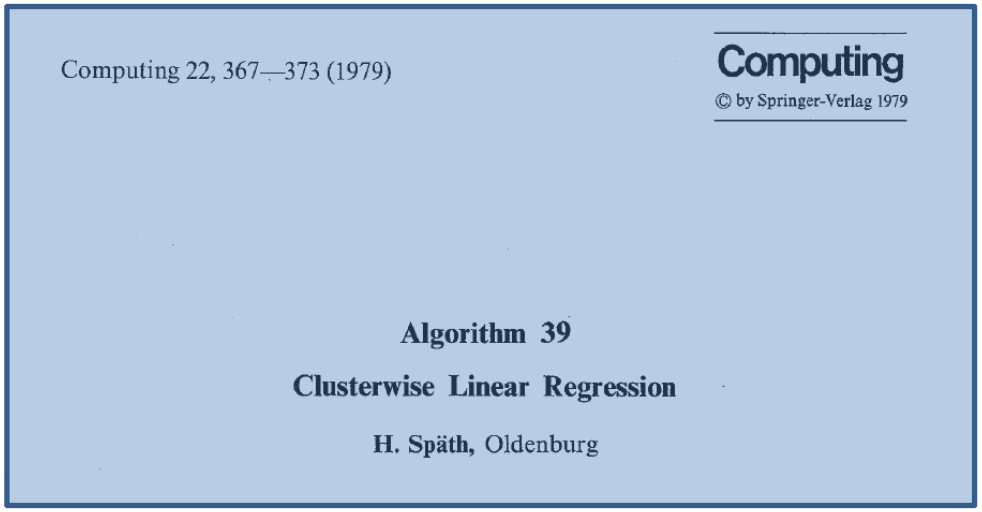
$$\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbf{1}_k(i)\left(y_i-(\alpha_k+\beta_k x_i)\right)^2$$

$$\begin{aligned}\mathrm{V}(Y-\hat{Y}) &= \mathrm{V}(Y-\hat{Y}^L)+\mathrm{V}(\hat{Y}^L-\hat{Y})\\ &= \sum_{i=1}^{s}\mathbf{P}(\{\mathcal{G}=i\})\mathrm{V}(Y-\hat{Y}^i|\mathcal{G}=i)+\mathrm{V}(\hat{Y}^L-\hat{Y}).\end{aligned}$$

Variance résiduelle de la régression globale= variance résiduelle intra cluster + variance due à la différence entre régression locale (clusterwise) et régression globale (OLS)

– Les pionniers:
– thèse de Christian Charles, 1977:
  - Application de l'algorithme des nuées dynamiques
    – Etape 1: à partir d'une partition initiale, on estime séparément k modèles de régression.
    – Etape 2: chaque observation est affectée au cluster (ou modèle) donnant le plus petit résidu carré, ie la meilleure prédiction. Une fois toutes les observations reclassées, on a une nouvelle partition
    – Itération
  - Utilisation de régression ridge et de régression sur composantes principales locales pour les groupes de trop faible effectif

– Späth introduit en 1979 le vocable « clusterwise regression » avec un programme Fortran
  - Même critère
  - Algorithme d'échange (en fait des k-means)



Computing 22, 367—373 (1979)

Computing

Algorithm 39

Clusterwise Linear Regression

H. Späth, Oldenburg

– L’analyse discriminante typologique (Lemoine, 1979) poursuit un but différent:

« Située dans le contexte de la discrimination, l’analyse discriminante typologique permet de remédier à une définition subjective, arbitraire ou imprécise des types a priori. La partition définie a priori sur la population échantillon est remise en cause par optimisation d’un critère qui ne tient compte que de la projection des individus sur les premiers axes de l'analyse factorielle discriminante. »

http://docnum.univ-lorraine.fr/public/UPV-M/Theses/1979/Lemoine.Yves.SMZ79004.pdf

3. **Modèles de mélanges**

Régression sur classes latentes: une extension du modèle de profils latents

	Variables latentes	
Variables manifestes	**qualitative**	**quantitatives**
qualitatives	Classes latentes	Traits latents
quantitatives	**Profils latents**	Analyse factorielle

3.1 L'analyse des profils latents (Lazarsfeld & Henry, 1968)

- Hypothèse: la population consiste en k groupes non observés (*profils latents, classes latentes, composantes du mélange etc.*).
- Modèle classique: mélange multinormal

$$f(\mathrm{x}) = \sum_{j=1}^{k} \pi_j f_j\left(\mathrm{x} / \mu_j, \Sigma_j\right)$$

" It should be noted that the model structures resembles quadratic discriminant analysis, with the important difference, of course, that the classes (groups) are unknown." J. Vermunt

- Estimateurs du maximum de vraisemblance obtenus par l'algorithme EM
- Le logiciel Mixmod permet d'estimer 14 modèles de covariance différents

3.2 Régression avec classes latentes

Proposée par DeSarbo et Cron en 1988

We assume y_i is distributed as a finite sum or mixture of conditional univariate normal densities:

$$y_i \sim \sum_{k=1}^{K} \lambda_k f_{ik}(y_i \mid X_{ij}, \sigma_k^2, b_{jk}) \tag{9}$$

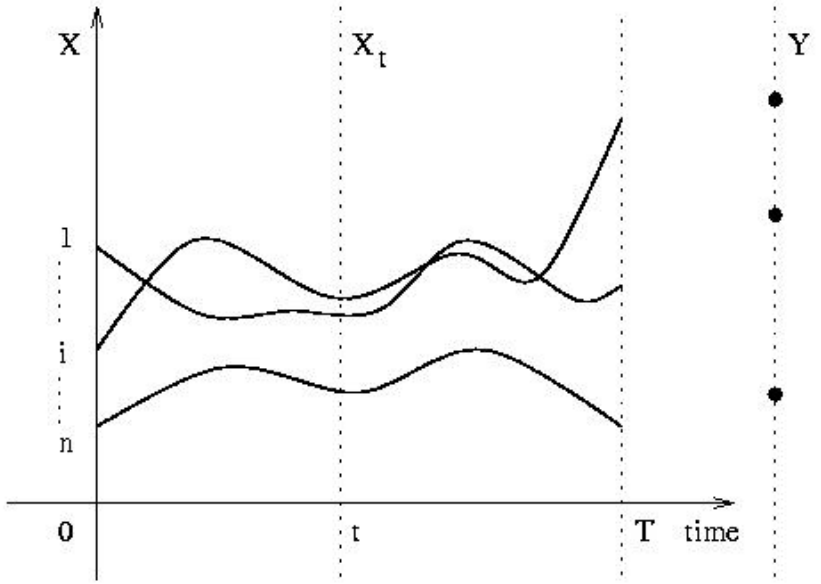
$$= \sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \mathbf{x}_i \mathbf{b}_k)^2}{2\sigma_k^2}\right], \tag{10}$$

- Extension à la régression de Poisson:
  - (Wedel & al. 1993), puis à des modèles linéaires généralisés locaux (Wedel & DeSarbo, 1995)
- Hennig, 2000 propose quelques variantes
- Méthode dominante: maximum de vraisemblance et algorithme EM
  - autre algorithme : IRLS (Lipovetsky & Conklin, 2005)

4. Deux extensions

4.1 Données fonctionnelles

Prédire une variable Y par une variable fonctionnelle:



Y= récolte

X_t = température

p= ∞

R.A.Fisher « The Influence of Rainfall on the Yield of Wheat at Rothamsted »
Philosophical Transactions of the Royal Society, B, 213, 89-142 (1924)

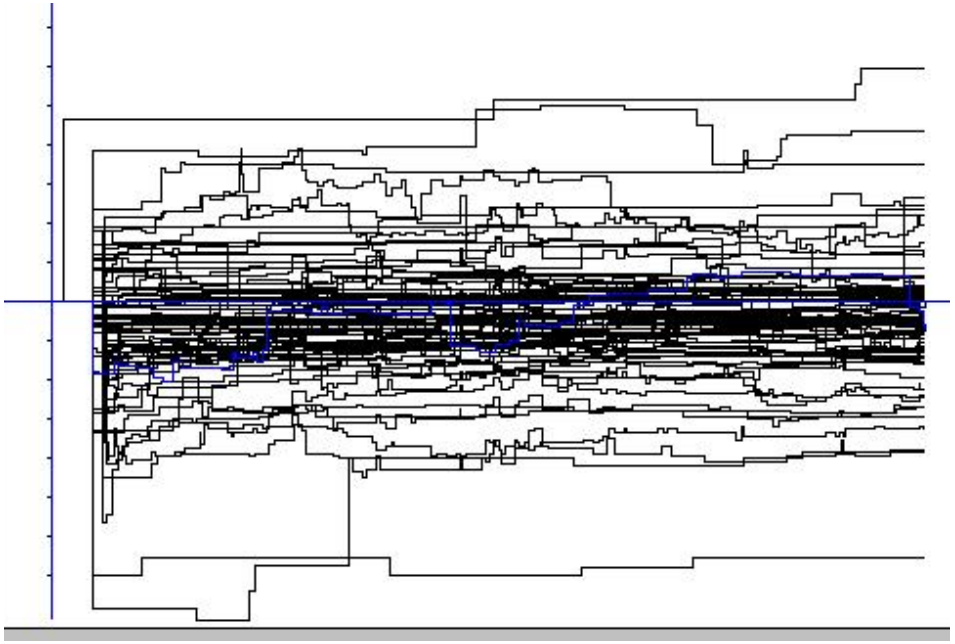
- « Integral regression » $\hat{Y} = \int_0^T \beta(t) X_t dt$

au lieu d'une somme finie $\hat{Y} = \sum_{j=1}^{p} \beta_j X_j$

Problème mal posé, résolu par une régression PLS fonctionnelle (Preda & Saporta, 2005a)

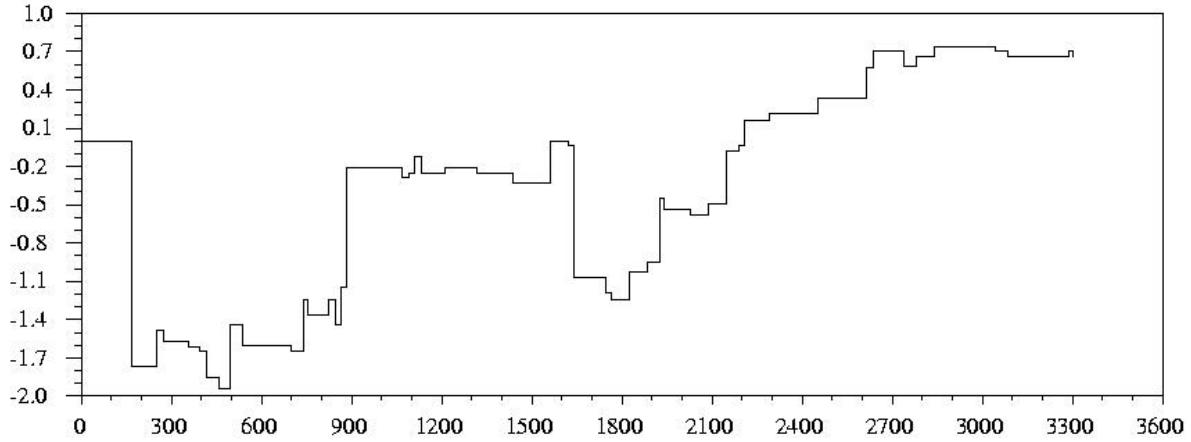
- Etendu dans (Preda & Saporta, 2005b) à la régression clusterwise. Preuve de la décroissance monotone pour un certain nombre de composantes.

- Application à des données financières



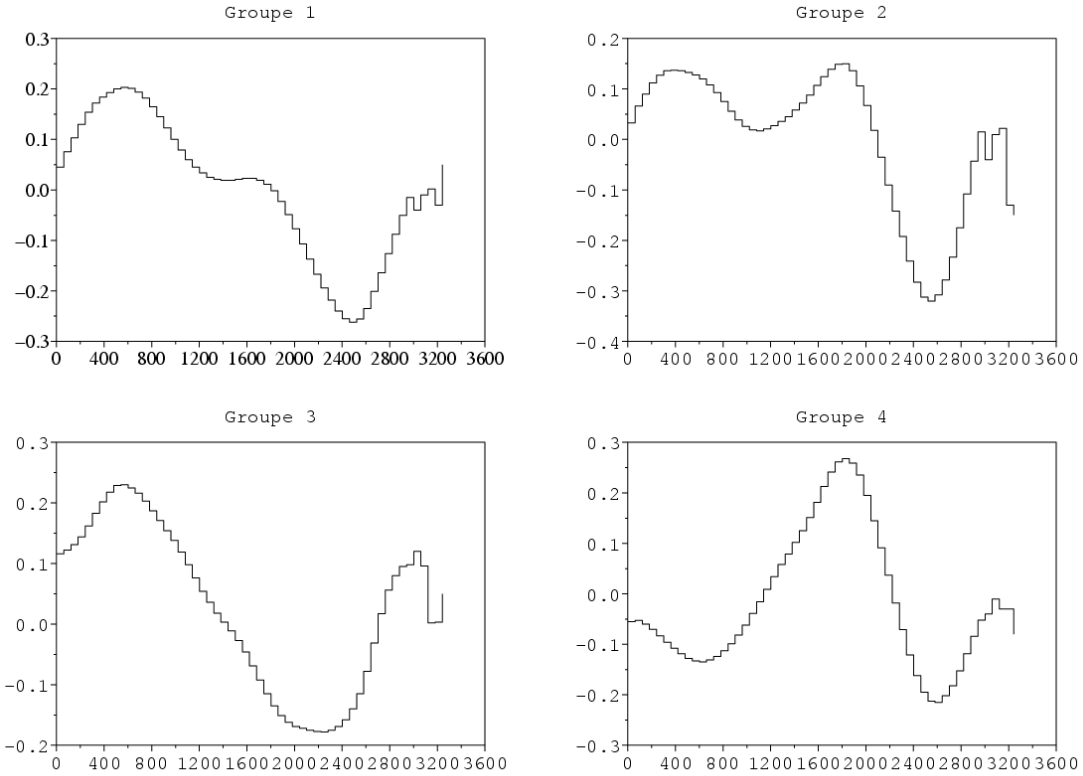
Indices de croissance de 84 actions pendant 1 h

- Prévoir les 5 dernières minutes d'une nouvelle action, connaissant les 55 premieres minutes?



- 1366 variables sont nécessaires (nombre d'intervalles où les données restent constantes)

- Quatre clusters d'effectifs (17;32;10;25)



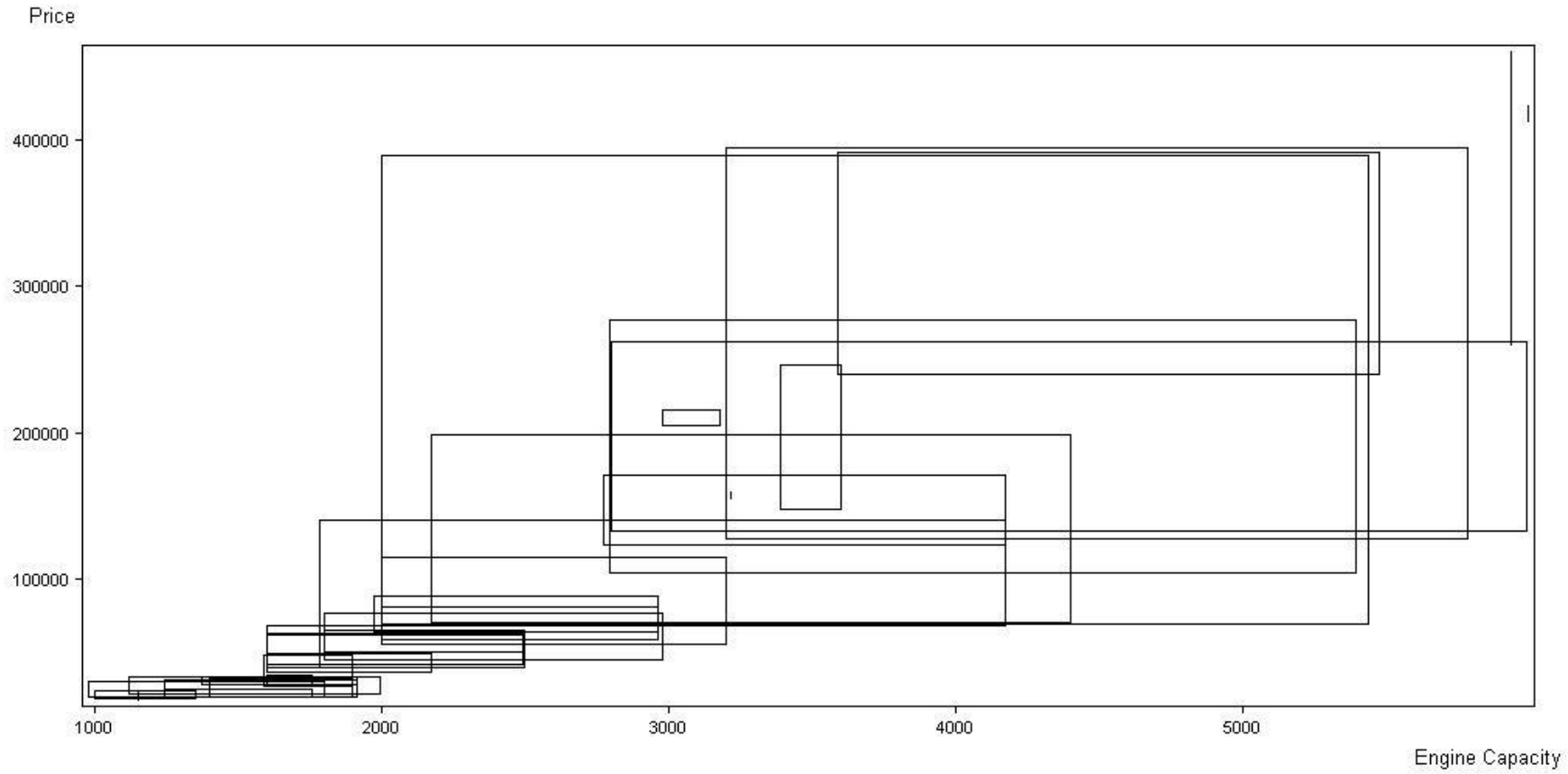
- Affectation de la nouvelle action au cluster le plus proche (ici le n°1) puis application du modèle local (PLS2)

	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	SSE
Observed	**0.700**	**0.678**	**0.659**	**0.516**	**-0.233**	-
PLS(2)	0.312	0.355	0.377	0.456	0.534	0.911
PLS(3)	0.620	0.637	0.677	0.781	0.880	1.295
PCR(3)	0.613	0.638	0.669	0.825	0.963	1.511
CW-PLS(3)	0.643	0.667	0.675	0.482	0.235	0.215
CW-PLS(4)	0.653	0.723	0.554	0.652	-0.324	0.044
CW-PLS(5)	0.723	0.685	0.687	0.431	-0.438	0.055

4.2 Données intervalles

- Carvalho et al. , 2010 : une version clusterwise de la régression « centre et étendue » , pour des données symboliques de type intervalles
- Exemple : 33 véhicules, deux variables

modèle	Y= Prix	X= Cylindrée
Alfa 145	(27806 : 33596)	(1370 : 1910)
Alfa 156	(41593 : 62291)	(1598 : 2492)
Alfa 166	(64499 : 88760)	(1970 : 2959)
Aston Martin	(260500 : 460000)	(5935 : 5935)
Audi A3	(40230 : 68838)	(1595 : 1781)


Price
400000
300000
200000
100000
1000
2000
3000
4000
5000
Engine Capacity

- La régression centre et étendue: deux régressions linéaires, une pour les centres des intervalles, l'autre pour les demi-étendues, empilées en une seule
- On prédit ensuite le min et le max de y
- Version clusterwise:
  - Critère

$$J = \sum_{k=1}^{K} \sum_{i \in P_k} (\epsilon_{i(k)})^T \epsilon_{ik} = \sum_{k=1}^{K} \sum_{i \in P_k} \left[(\epsilon^c_{i(k)})^2 + (\epsilon^r_{i(k)})^2 \right] \qquad (2)$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \left\{ \left[y^c_i - \left(\beta^c_{0(k)} + \sum_{j=1}^{p} \beta^c_{j(k)} \, x^c_{ij} \right) \right]^2 + \left[y^r_i - \left(\beta^r_{0(k)} + \sum_{j=1}^{p} \beta^r_{j(k)} \, x^r_{ij} \right) \right]^2 \right\}$$

  - Optimisation par technique de type nuées dynamiques

Table 1. Fitted regression equations over the whole car interval-valued data set

$K-partition$	cluster k	"Center Model"	"Range Model"
1	1	$\hat{y}^c_{(1)} = -98840.9 + 79.2\, x^c_1$	$\hat{y}^r_{(1)} = -341.4 + 60.9\, x^r_1$
2	1	$\hat{y}^c_{(1)} = -63462.2 + 59.6\, x^c_1$	$\hat{y}^r_{(1)} = -4560.1 + 47.1\, x^r_1$
	2	$\hat{y}^c_{(2)} = -22836.5 + 68.8\, x^c_1$	$\hat{y}^r_{(2)} = 34563.6 + 68.6\, x^r_1$
3	1	$\hat{y}^c_{(1)} = -77422.1 + 82.0\, x^c_1$	$\hat{y}^r_{(1)} = 2229.7 + 92.2\, x^r_1$
	2	$\hat{y}^c_{(2)} = -58484.1 + 71.1\, x^c_1$	$\hat{y}^r_{(2)} = 101952.9 - 546.7\, x^r_1$
	3	$\hat{y}^c_{(3)} = -73362.1 + 62.0\, x^c_1$	$\hat{y}^r_{(3)} = -9755.9 + 53.2\, x^r_1$

Table 2. Determination coefficients for the fitted regression equations over the whole car interval-valued data set

K-partition	1	2		3		
cluster k	1	1	2	1	2	3
$R^2_{c(k)}$	0.93	0.95	0.91	0.97	0.99	0.98
$R^2_{r(k)}$	0.53	0.79	0.66	0.98	0.98	0.83

- Cross validation « 10 fold »
- Agrégation par stacking (Breiman, 1996) des prévisions des K modèles

Table 3. Average Root-mean-square error for the combined estimates of the K regression models

K-partition	1	2	3
$RMSE_L$	96649.28 (13812.49)	90417.42 (13538.22)	94993.75 (11376.24)
$RMSE_U$	143416.6 (17294.02)	135471.4 (17027.49)	137825.9 (14243.29)

- Le modèle à deux classes semble le plus judicieux

5.Problèmes d'implémentation

5.1 Classes d'effectifs inférieurs à p

- L'existence de classes de taille inférieure au nombre de variables interdit d'estimer par les mco des modèles locaux
- Nécessité d'utiliser des régressions régularisées (ridge, rcp, pls …)

5.2 Identifiabilité des modèles de mélange

- Reste un problème complexe:
- Hennig, 2000 examine différents modèles, donne des contre-exemples et indique que

"Mixtures of linear regression models with Gaussian noise are identifiable, if the number of components K is smaller than the minimal number of (feasible) hyperplanes necessary to cover all covariate points (without intercept)."

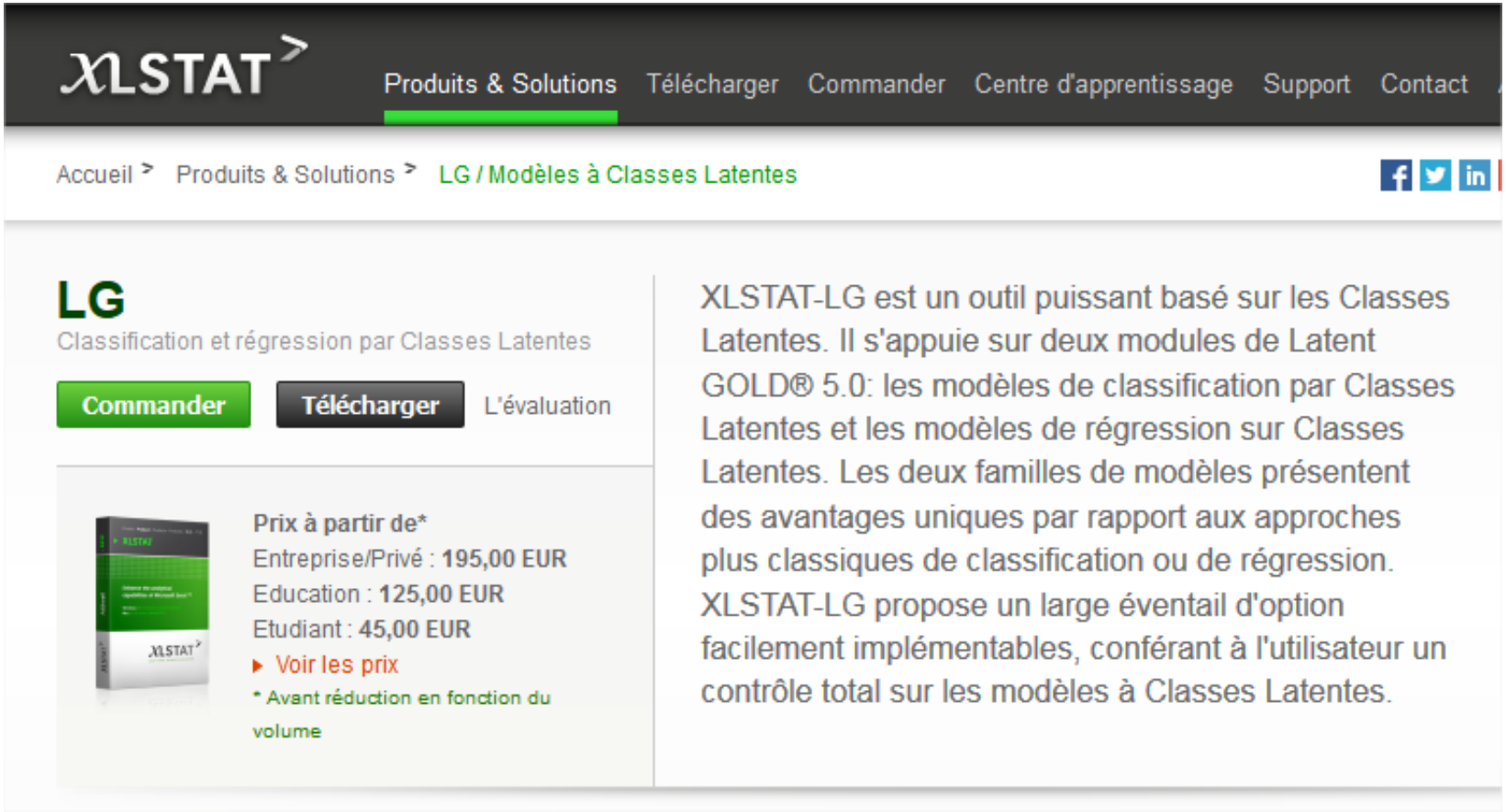
5.3 Logiciels

- Logiciels de latent class regression
  - Commerciaux:
    - GLIMMIX * http://www.scienceplus.com/glimmix
    - LatentGold 5.0 et XLSTAT-LG
  - Libres :
    - Flexmix 2.3-13 , janvier 2015 (Leisch, 2004)

---

* Ne pas confondre avec les procédures homonymes de SAS et R

Pub…


XLSTAT
Produits & Solutions
Télécharger
Commander
Centre d'apprentissage
Support
Contact
Accueil > Produits & Solutions > LG / Modèles à Classes Latentes
LG
Classification et régression par Classes Latentes
Commander
Télécharger
L'évaluation
Prix à partir de*
Entreprise/Privé : 195,00 EUR
Education : 125,00 EUR
Etudiant : 45,00 EUR
Voir les prix
* Avant réduction en fonction du volume
XLSTAT-LG est un outil puissant basé sur les Classes Latentes. Il s'appuie sur deux modules de Latent GOLD® 5.0: les modèles de classification par Classes Latentes et les modèles de régression sur Classes Latentes. Les deux familles de modèles présentent des avantages uniques par rapport aux approches plus classiques de classification ou de régression. XLSTAT-LG propose un large éventail d'option facilement implémentables, conférant à l'utilisateur un contrôle total sur les modèles à Classes Latentes.

- Logiciels régression typologique
  - ?

- Autres algorithmes de régression typologique
  - Recuit simulé (DeSarbo et al., 1989)
  - Amélioration par la metaheuristique VNS « Recherche à Voisinage Variable » (Caporossi & Hansen, 2005)
  - Optimisation globale (Carbonneau et al., 2014)
- Mais pas de comparaison systématique, ni de packages

5.4 Choix du nombre de classes

- Toujours un problème! K souvent supposé fixé.
- **Approche modèle** : AIC, BIC et leurs variantes (CAIC, AIC3 etc.)

$$AIC = -2\ln(L) + 2\left(K(p+3)-1\right)$$

$$BIC = -2\ln(L) + \ln(n)\left(K(p+3)-1\right)$$

- Nécessite des hypothèses de distribution

- AIC et BIC ne sont semblables qu'en apparence
- **Théories différentes**
  - AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée
  - BIC : choix bayesien de modèles
    - m modèles M_i paramétrés de probabilités *a priori* $P(M_i)$ égales
- **Illogique d'utiliser les deux simultanément**

(Saporta, 2008)

- A la recherche du « vrai » modèle
  - On suppose que le vrai modèle fait partie des modèles en compétition
  - Si *n* tend vers l'infini la probabilité que le *BIC* choisisse le vrai modèle tend vers 1, ce qui est faux pour l'*AIC*.
  - *AIC* va choisir le modèle qui maximisera la vraisemblance de futures données et réalisera le meilleur compromis biais-variance
- "Essentially, all models are wrong, but some are useful " G.Box (1987)
- "The Truth Is Out There" (X-Files, 1993)

5.4 Choix du nombre de classes, suite

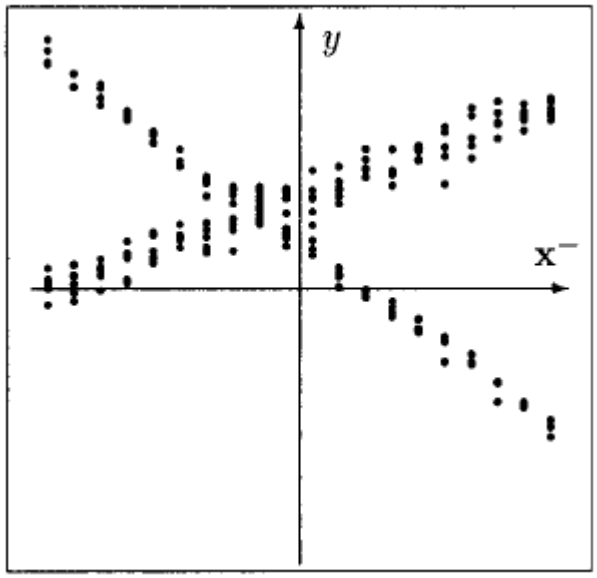
- **Approche apprentissage**: validation croisée, ce qui suppose de savoir prédire de nouvelles observations

5.5 Prévision de nouvelles observations

On connait seulement les x

- Règle « dure »: affecter au cluster le plus proche et appliquer le modèle correspondant
  - Comment affecter? Distance au centre, ppv etc.
- Règle « douce »: moyenne pondérée des prévisions de chaque modèle
  - Bayesian Model Averaging : poids = probabilités *a posteriori* de chaque modèle *ie* d'appartenance aux clusters
- Règle aléatoire: tirage selon les probas *a posteriori*

- Optique analyse des données (k-means)
  - Pas de difficulté technique particulière à combiner discrimination et régression, sauf pour des petits clusters.
  - Quelques incohérences à mêler probabilités et géomètrie…
  - Un cas difficile:



Hennig, 2000

- Optique classes latentes
  - ça se complique!
  - Dans flexmix, on ne peut calculer les probas *a posteriori* que si y est connu

$$\mathrm{P}(j|x,y,\psi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_k \pi_k f(y|x,\theta_k)}$$

"We have no solution for this problem: Without y you cannot determine the likelihood and hence not into which cluster the observation belongs. You could calculate predictions for each cluster, but then you have K answers, not one. " (F.Leisch , communication personnelle)

6. Propositions et expérimentations

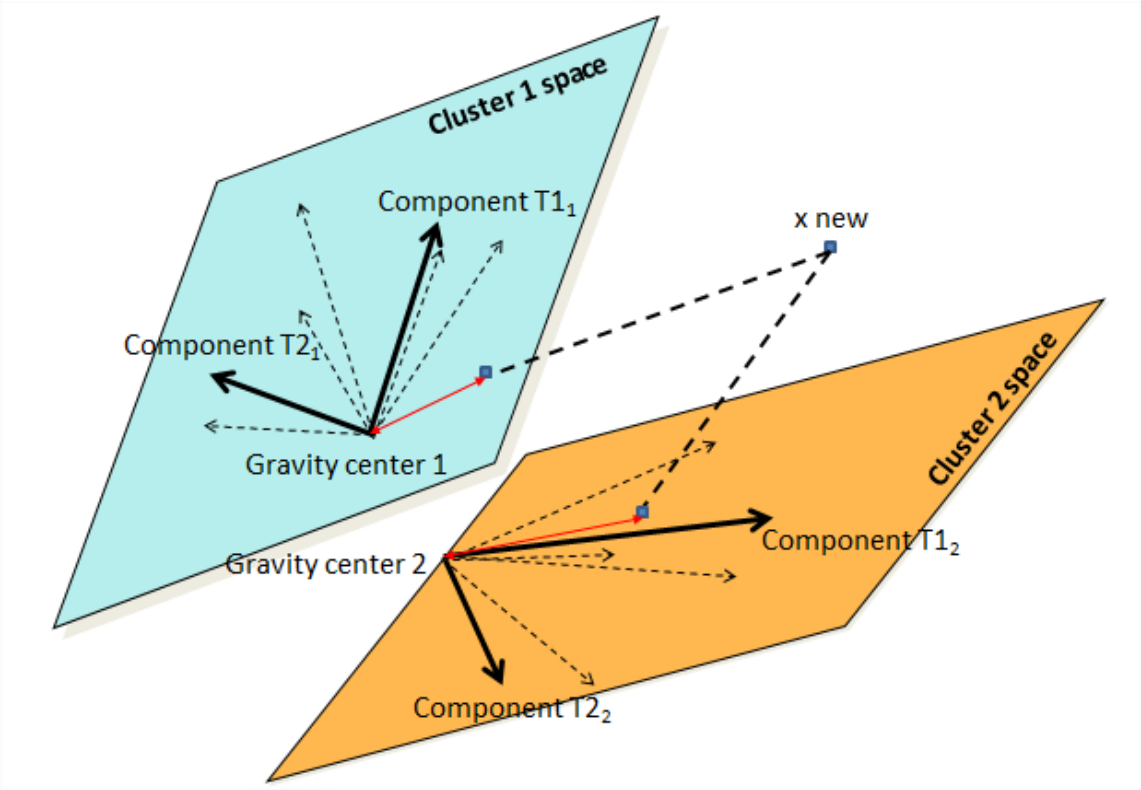
6.1 critère et algorithme

- Régression clusterwise (un seul Y) , critère du RMSE
- Algorithme stochastique avec plusieurs passes (arrêt quand moins de 1% de variation du RMSE) et plusieurs initialisations aléatoires, puis choix du meilleur modèle par VC-10

6.2 Prédiction sur composantes

- Réduction de dimension pour éviter les problèmes dus aux classes d'effectifs trop faibles
- Composantes dépendant des classes pour une meilleure prévision, calculables pour les nouvelles observations
  - Composantes globales (Esposito-Vinzi et al., 2003, 2005) par PLS-DA

- Composantes locales
  - K régressions PLS, puis calcul de distances
  - Probas *a posteriori* proportionnelles à $\exp(-d_j^2)$ pondérées par les poids des classes



6.3 choix du nombre de composantes et de groupes

- CV-10
- Prévision «soft » de nouvelles observations par règle bayesienne

$$\hat{y}(i) = \sum_{k=1}^{K} \pi_k(i)\hat{y}_k(i)$$

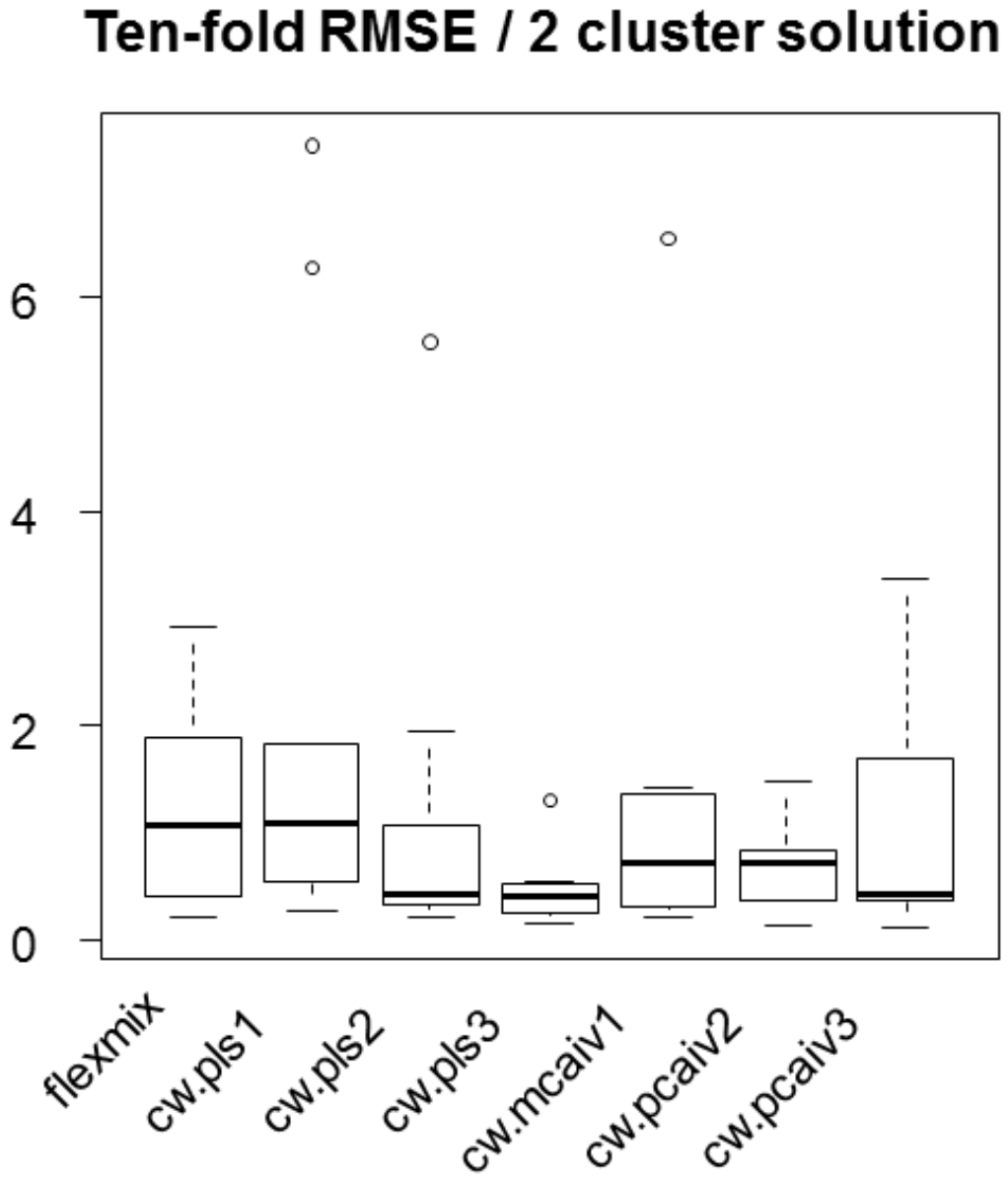
- Si les classes sont bien séparées prévision « soft » peu différente de prévision « hard »

6.4 Exemples

- Exemple 1 : données de Späth (1978) un peu bruitées

Table 1

i	y_i	a_{i1}	a_{i2}	a_{i3}
1	960.0	60.0	18.0	1.0
2	830.0	220.0	0.0	1.0
3	1260.0	180.0	14.0	1.0
4	610.0	80.0	6.0	1.0
5	590.0	120.0	1.0	1.0
6	900.0	100.0	9.0	1.0
7	820.0	170.0	6.0	1.0
8	880.0	110.0	12.0	1.0
9	860.0	160.0	7.0	1.0
10	760.0	230.0	2.0	1.0
11	1020.0	70.0	17.0	1.0
12	1080.0	120.0	15.0	1.0
13	960.0	240.0	7.0	1.0
14	700.0	160.0	0.0	1.0
15	800.0	90.0	12.0	1.0
16	1130.0	110.0	16.0	1.0
17	760.0	220.0	2.0	1.0
18	740.0	110.0	6.0	1.0
19	980.0	160.0	12.0	1.0
20	800.0	80.0	15.0	1.0


Ten-fold RMSE / 2 cluster solution
6
4
2
0
flexmix
cw.pls1
cw.pls2
cw.pls3
cw.mcaiv1
cw.pcaiv2
cw.pcaiv3

	Expected	Flexmix	cw.pls(1)	cw.pls(2)	cw.pls(3)	cw.pcaiv(1)	cw.pcaiv(2)	cw.pcaiv(3)
Adjusted Rand index	-	0,46	-0,04	-0,05	-0,04	-0,04	-0,04	-0,04
Intercept - Cluster 1	471,58	-4452,76	666,44	258,02	237,47	260,70	260,70	237,47
X1 - Cluster 1	1,24	4,24	0,72	2,70	2,78	2,06	2,06	2,78
X2 - Cluster 1	33,4	33,18	0,06	35,20	35,98	30,60	30,60	35,98
X3 - Cluster 1	0	4453,06	0,00	0,00	0,00	0,00	0,00	0,00
Intercept - Cluster 2	84,56	3185,31	860,18	306,74	260,70	237,47	237,47	260,70
X1 - Cluster 2	3,6	1,66	1,63	1,90	2,06	2,78	2,78	2,06
X2 - Cluster 2	34,21	31,53	0,06	28,28	30,60	35,98	35,98	30,60
X3 - Cluster 2	0	-2814,58	0,00	0,00	0,00	0,00	0,00	0,00

- **Flexmix retrouve correctement les classes mais retrouve difficilement l'intercept et le coefficient de X3 (bruit ajouté pour la réalisation des calculs)**
- **Cw.pls retrouve mal les classes et les intercepts, mais retrouve correctement les coefficients de régression des variables, notamment pour les modèles à 2 et 3 composantes.**
- **Cw.pcaiv retrouve mal les classes mais correctement les coefficients de régression des variables (stable quelque soit le nombre de composantes du modèle).**

- Exemple 2 : DeSarbo 1988

TABLE 1

Synthetic Regression Data

i	X_1	X_2	y	
1	1	-3	-5	
2	1	-2	-3	
3	1	-1	-1	
4	1	0	1	GROUP 1
5	1	1	3	$y_i = 2X_{2i}+1$
6	1	2	5	
7	1	3	7	
8	1	-3	5	
9	1	-2	3	
10	1	-1	1	
11	1	0	-1	GROUP 2
12	1	1	-3	$y_i = -2X_{2i}-1$
13	1	2	-5	
14	1	3	-7	

	Expected	Flexmix	cwpls(1)	cwpls(2)	cwpcaiv(1)	cwpcaiv(2)
Adjusted Rand index	-	1	1	1	0,713	1
Intercept - Cluster 1	1	1,290	1,006	1,006	-41,053	1,290
X1 - Cluster 1	0	-0,287	-0,003	-0,003	42,018	-0,287
X2 - Cluster 1	2	2,004	2,004	2,004	2,062	2,004
Intercept - Cluster 2	-1	-0,704	-1,006	-1,006	1,501	-0,704
X1 - Cluster 2	0	-0,296	0,007	0,007	-2,497	-0,296
X2 - Cluster 2	-2	-2,000	-1,999	-1,999	-2,008	-2,000

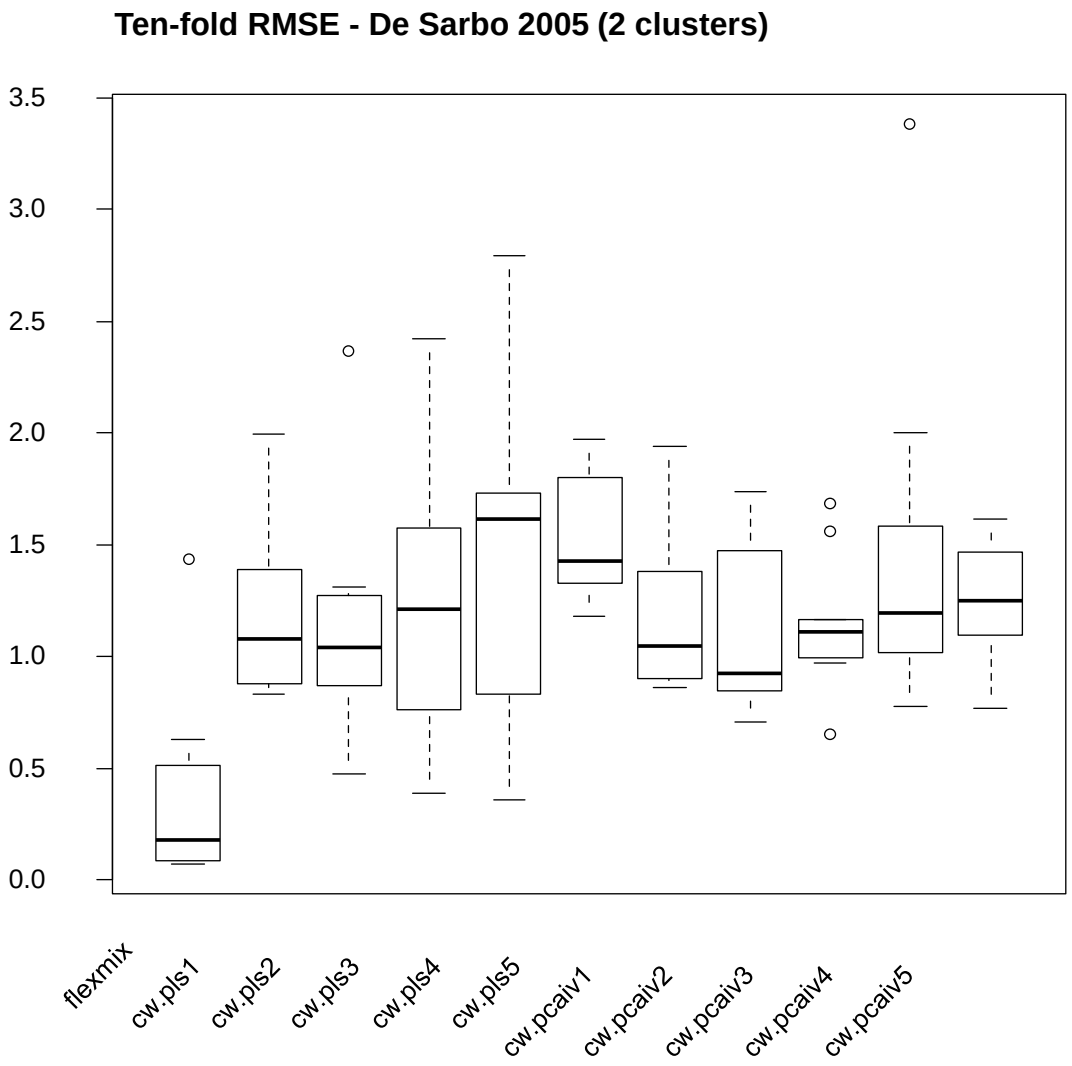
- **Excellente performance de cw.pls quelque soit le nombre de composantes utilisées.**
- **Flexmix présente de bons résultats.**
- **Performance décevante de cw.pcaiv avec une composante.**

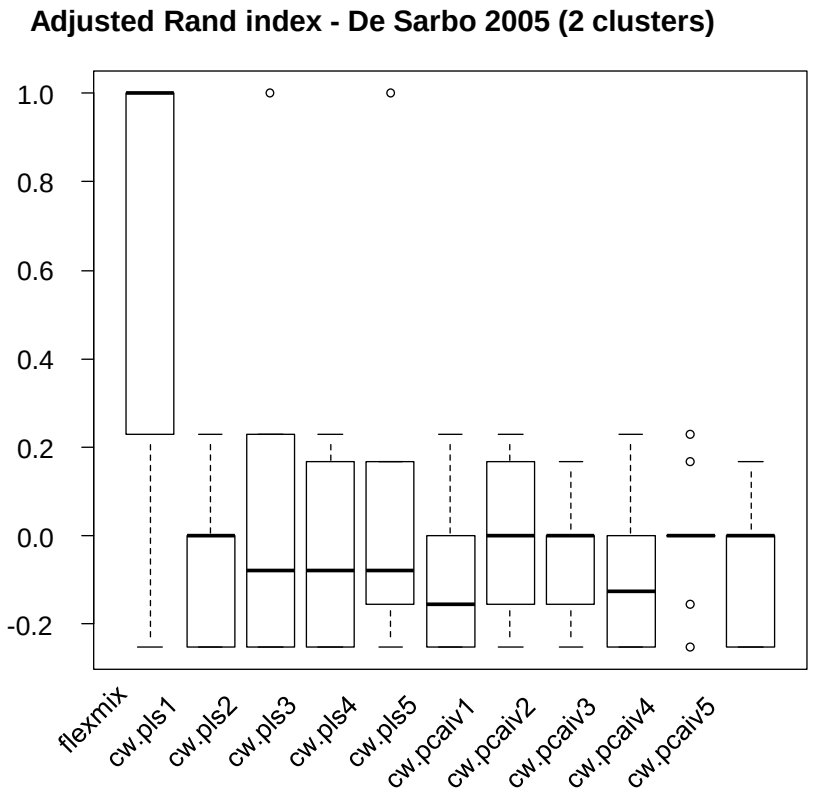
- Exemple 3: DeSarbo et al. 2005

Table 19.1 Synthetic Regression Data

X_1	X_2	X_3	X_4	X_5	Y
0.995	1.893	−1.291	−0.560	0.508	−3.287
−1.268	−0.988	0.244	0.716	−2.261	−8.254
0.135	0.725	−0.511	0.712	−1.551	−11.867
0.687	0.223	1.441	0.098	−0.352	10.859
1.020	−0.355	−0.660	0.090	−1.544	−4.733
0.313	0.999	0.365	−0.991	1.000	12.191
0.584	−1.775	−1.702	1.843	−1.267	−25.508
−1.342	−0.976	−1.381	0.290	1.525	−11.996
−0.159	−0.132	0.217	−1.643	−0.278	9.533
1.089	0.629	1.148	−1.138	−1.299	15.716
0.322	−0.983	−1.673	0.101	−0.287	−14.693
0.082	−0.009	1.068	−1.023	0.565	16.379
−0.923	2.327	−0.989	1.637	1.023	−14.245
−0.138	0.122	−0.967	−0.365	−0.535	−6.939
−0.650	1.494	0.040	−0.086	−0.806	−1.569
1.227	0.475	2.090	−0.294	0.643	21.104
0.987	0.389	−1.791	0.758	1.574	−12.890
−1.786	0.484	0.495	0.545	−0.448	−1.547
−1.330	−1.583	0.606	0.948	−0.137	−3.635
−0.213	−1.400	−0.942	−0.659	0.042	−1.644
−0.305	0.531	0.627	0.068	0.214	3.152
−0.907	−0.296	−0.528	−0.423	−0.603	−3.425
−0.051	−1.204	0.553	−0.044	0.373	4.460
0.247	−0.288	−0.623	0.371	−1.285	−9.167
−0.261	1.024	−2.361	0.188	−1.183	−20.099
−1.072	−0.044	0.486	0.176	0.788	−2.941
−0.179	0.345	−0.655	2.860	2.211	18.596
−0.777	−0.390	−1.587	0.682	1.438	14.039
−1.311	−0.671	0.715	0.440	0.181	−0.625
0.145	0.321	−1.310	−0.085	0.843	7.317
−1.229	−1.133	−0.148	0.865	−0.007	8.799
0.421	0.147	−0.215	0.075	−0.492	2.029
1.088	−1.352	−0.531	−0.222	−0.048	0.930
−1.210	−0.100	−0.935	−0.142	−0.742	10.174
−0.202	1.088	−0.514	1.695	0.146	14.756

2 segments de 25 obs.

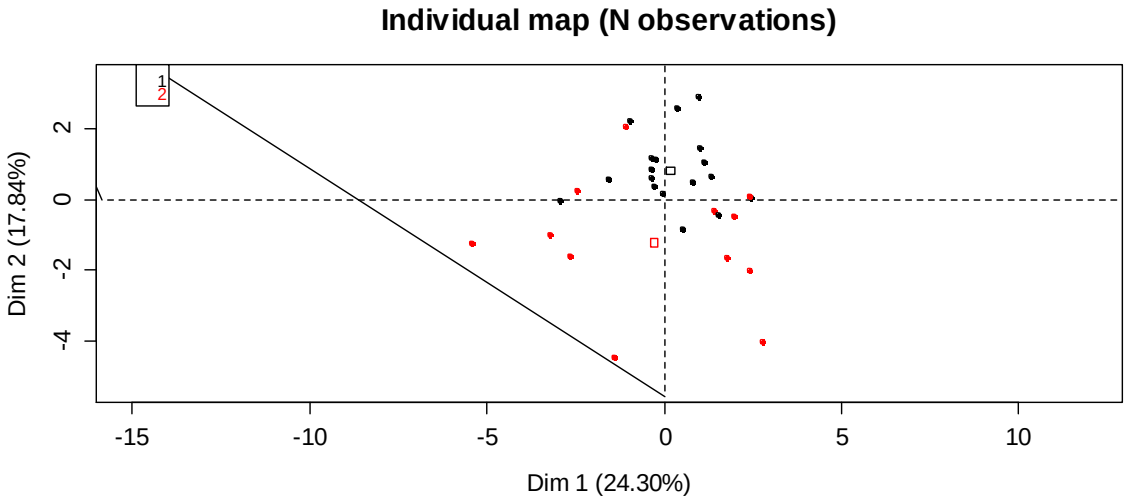

Ten-fold RMSE - De Sarbo 2005 (2 clusters)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
flexmix
cw.pls1
cw.pls2
cw.pls3
cw.pls4
cw.pls5
cw.pcaiv1
cw.pcaiv2
cw.pcaiv3
cw.pcaiv4
cw.pcaiv5

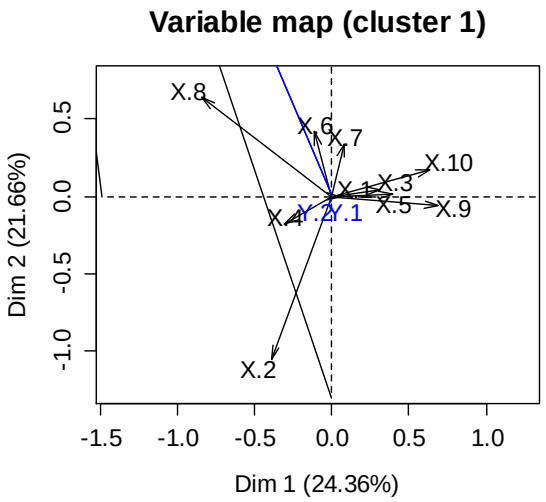


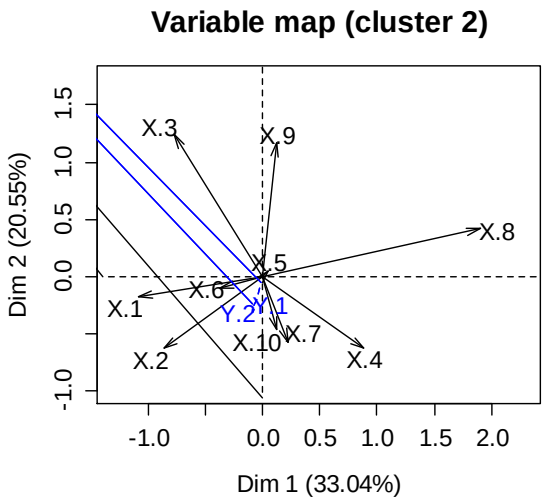
- **Excellente performance de Flexmix qui retrouve très bien les classes et assure du coup une bonne prédiction, mais utilise l'info du numéro de classe…**
- **Cw.pls et cw.pcaiv ont des performances plus moyennes car ces méthodes retrouvent mal les classes.**

6.5 Surapprentissage

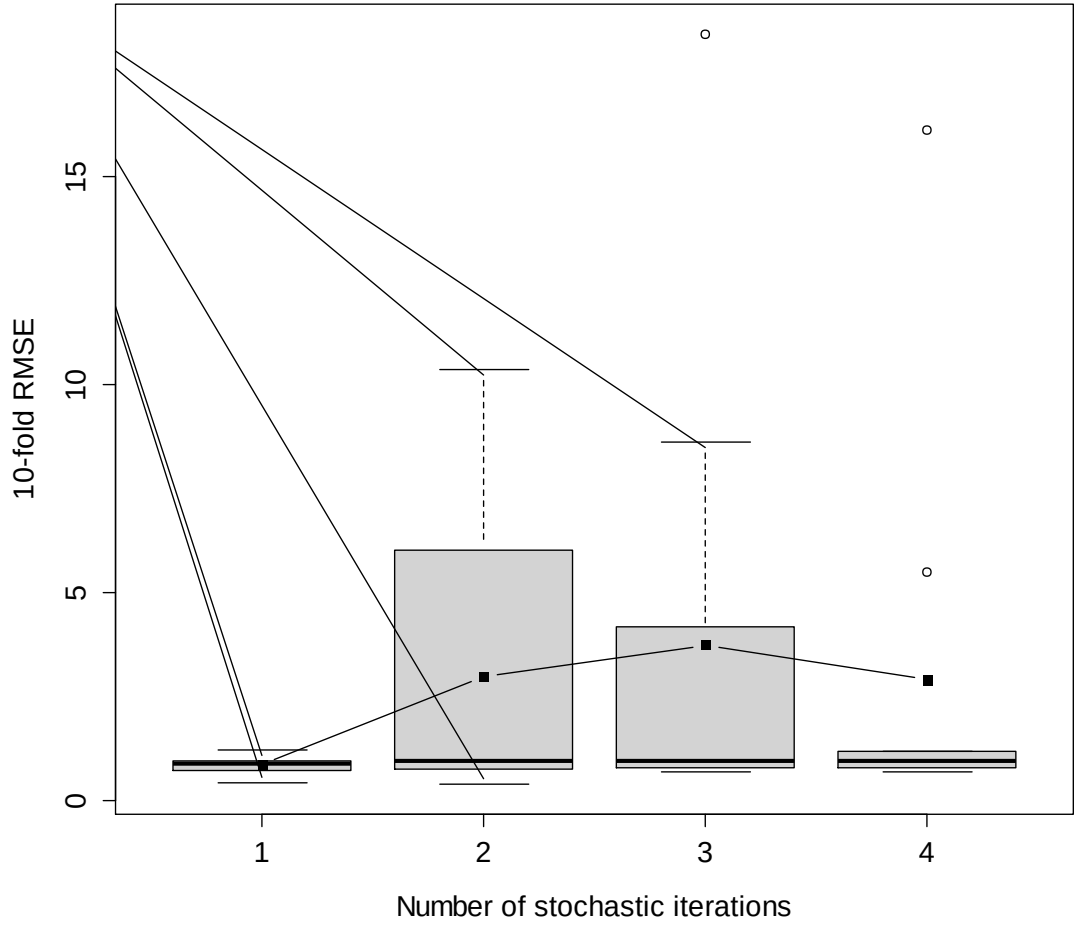
- Augmenter le nombre de passes (lectures complètes du fichier) conduit à du surapprentissage comme dans les réseaux de neurones



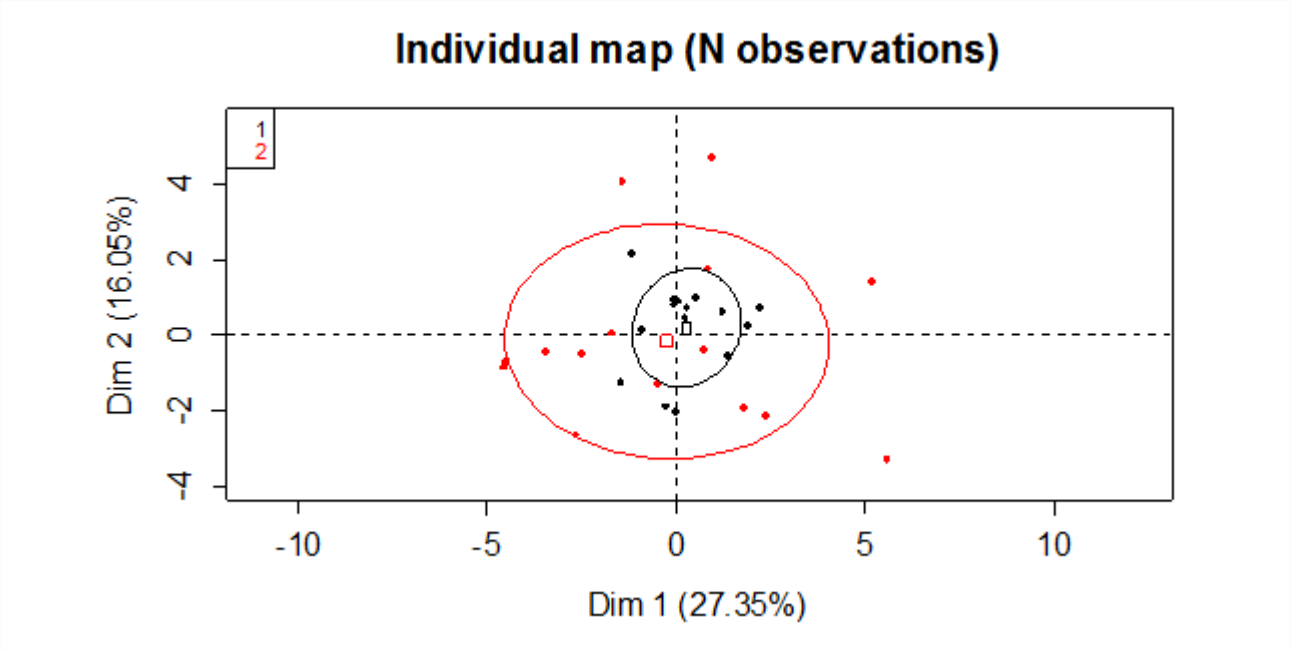


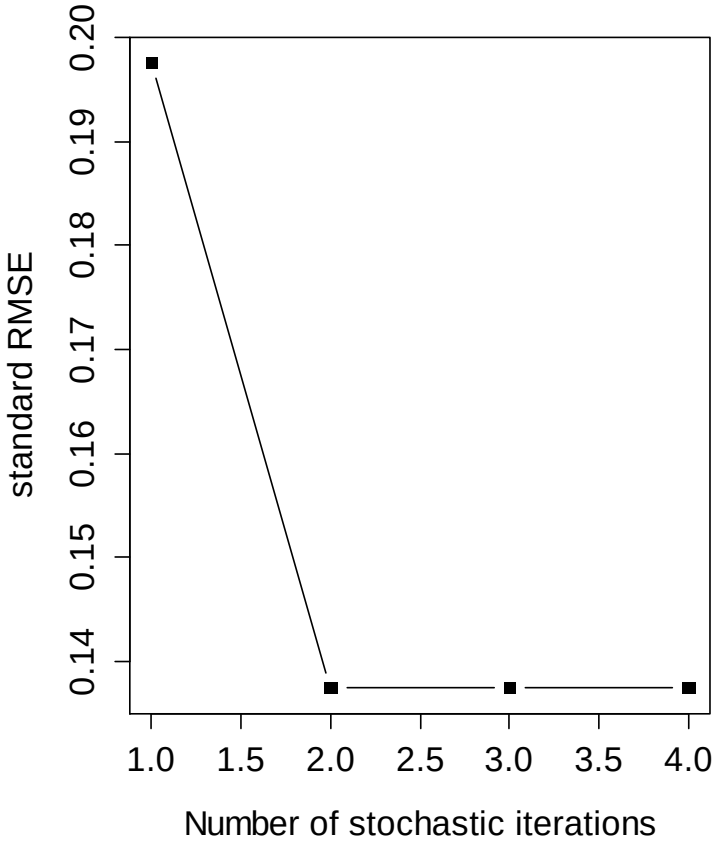


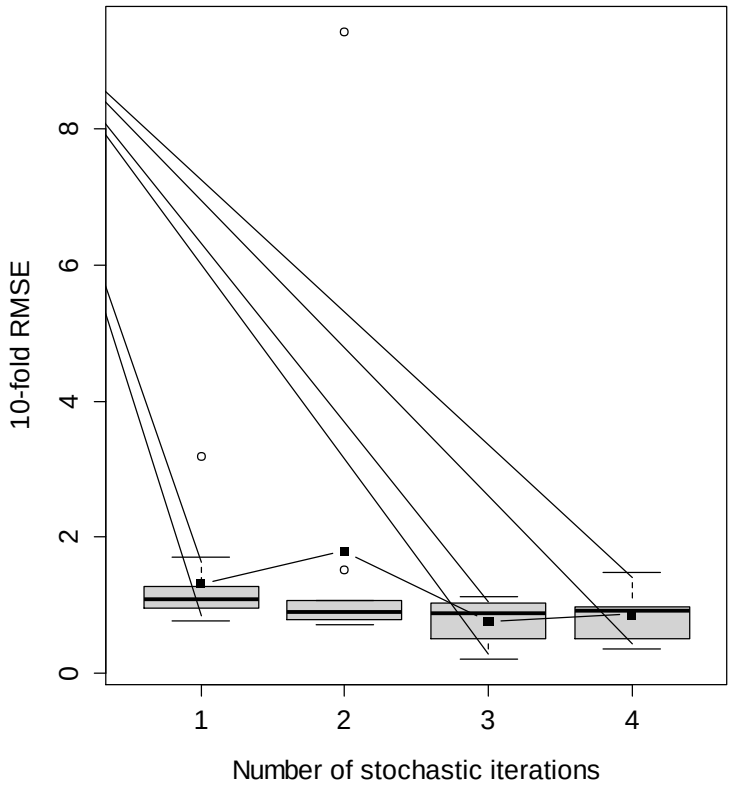
- G=2 clusters with the same size, <u>different volume</u>, same orientation and same shape,
- <u>Gravity centers are not clearly separated</u>,
- Limited number of N=30 individuals compared to P=20 explanatory variables X (max. of 15 individuals within each cluster),
- Q=2 dependent variables Y, <u>weakly</u> positively linked with X for the cluster 1 and <u>weakly</u> negatively for the cluster 2.


10-fold RMSE
0
5
10
15
1
2
3
4
Number of stochastic iterations

- Surapprentissage d'autant plus marqué que les classes sont mal séparées



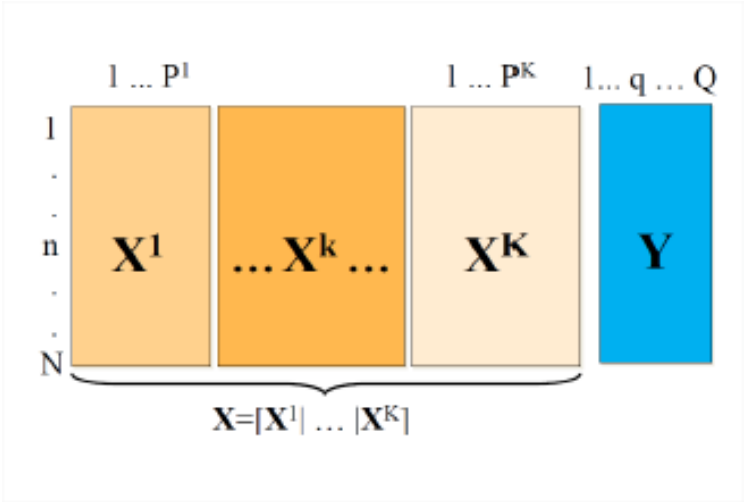

standard RMSE
0.14
0.15
0.16
0.17
0.18
0.19
0.20
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of stochastic iterations

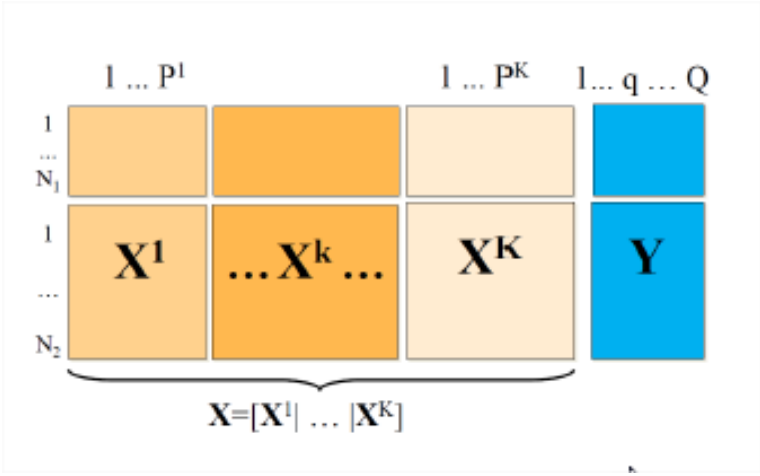

10-fold RMSE
0
2
4
6
8
1
2
3
4
Number of stochastic iterations

7. Conclusions et perspectives

- Opposition modélisation-prévision
- Déséquilibre des publications en faveur des des modèles de mélange
- Proposition de démarche basée sur composantes PLS et validation croisée
- Mise en évidence de phénomènes de surapprentissage

- Extension facile au cas d'une réponse Y vectorielle (PLS2 …)
- Généralisation au cas multibloc (Niang & Saporta, 2014, Niang et al. , congrès CARME 2015)





Merci pour votre attention

Références

- BREIMAN, L. (1996): Stacked Regressions. *Machine Learning* 24, pp.49-64.
- CAPOROSSI, G. , HANSEN, P. (2005): Variable Neighborhood Search for Least Squares Clusterwise Regression, *Cahiers du GERAD,* HEC Montréal
- CARBONNEAU, R.A., CAPOROSSI, G., and HANSEN, P. (2014) : Globally Optimal Clusterwise Regression By Column Generation Enhanced with Heuristics, Sequencing and Ending Subset Optimization, Journal of Classification, 31, pp.219-241
- CHARLES, C. (1977): *Régression Typologique et Reconnaissance des Formes.* Thèse de doctorat, Université Paris IX.
- De CARVALHO, F., SAPORTA, G., QUEIROZ, D. (2010): A Clusterwise Center and Range Regression Model for Interval-Valued , *COMPSTAT'2010, 19th International Conference on Computational Statistics,* pp.461-468,
- DESARBO, W.S. and CRON, W.L. (1988): A maximum likelihood methodology for clusterwise linear regression*. Journal of Classification*, 5, pp.249-282.
- DESARBO, W.S , KAMAKURA W.A. , WEDEL M. (2005): Latent Structure Regression, In: *Handbook of Marketing Research* , R. Grover & M. Vriens, (eds), London, Sage, 394-417.

- DIDAY, E. (1974): Introduction à l'analyse factorielle typologique, *Revue de Statistique Appliquée*, 22, 4, pp.29-38
- ESPOSITO-VINZI, V. LAURO, C., AMATO, S.(2005): PLS Typological Regression: Algorithmic, Classification and Validation Issues, in *New Developments in Classification and Data Analysis,* pp.133-140, Springer
- ESPOSITO-VINZI, V. LAURO, C., (2003): PLS Regression and Classification, In: *Proceedings of the PLS'03 International Symposium*, DECISIA, pp. 45-56
- HENNIG, C. (1999): Models and methods for clusterwise linear regression. *In:Classification in the Information Age*, Springer, pp.179-187.
- HENNIG, C. (2000): Identifiability of models for Clusterwise linear regression. *Journal of Classification*, 17, pp.273-296.
- LAZARSFELD, P.F. and HENRY, N.W. (1968): *Latent structure analysis*. Houghton Mifflin
- LEISCH, F. (2004) : FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(8).

- LEMOINE, Y. (1979): *Classification et discrimination : analyse discriminante typologique et applications*, Thèse de doctorat, Université de Metz
- NIANG- KEITA,N., SAPORTA G. (2014): Régression typologique pour données multi-blocs, *46 èmes journées de statistique*, Rennes
- PREDA, C. and SAPORTA, G. (2005a) : PLS regression on a stochastic process. *Computational Statistics and Data Analysis,* 48, pp.149-158
- PREDA, C. and SAPORTA, G. (2005b): Clusterwise PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 49, pp.99–108.
- SAPORTA, G. (2008): Models for Understanding versus Models for Prediction, in *Proceedings COMPSTAT'08*, Brito, P. (ed.), Springer, pp.315-322,
- SPÄTH, H. (1979): Clusterwise linear regression, *Computing*, 22, pp.367-373
- WEDEL M., DESARBO W.S. (1995): A Mixture Likelihood Approach for Generalized Linear Models, *Journal of Classification*, 12, 21–55.