

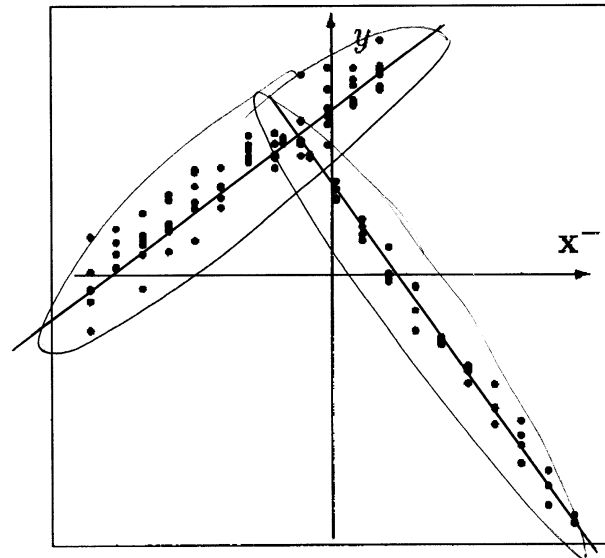
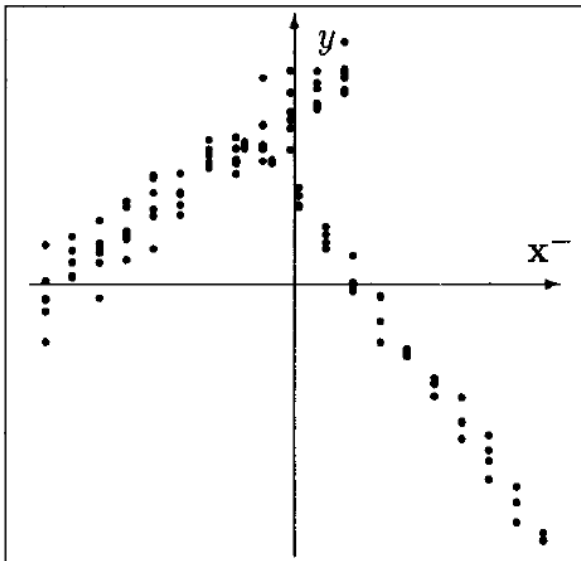
Clusterwise Methods: Past and Present

Gilbert Saporta
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
<http://cedric.cnam.fr/~saporta>

1. Introduction

- Local models *versus* global models



Hennig, 2000

- Unknown partition :
 - *Unobserved heterogeneity ; latent classes*
 - *Simultaneous search for classes and models for each class*
- Pioneering works: **typological** methods (Diday, 1974) or **clusterwise** (Späth, 1979)

2. k-means like methods

2.1 Typological PCA

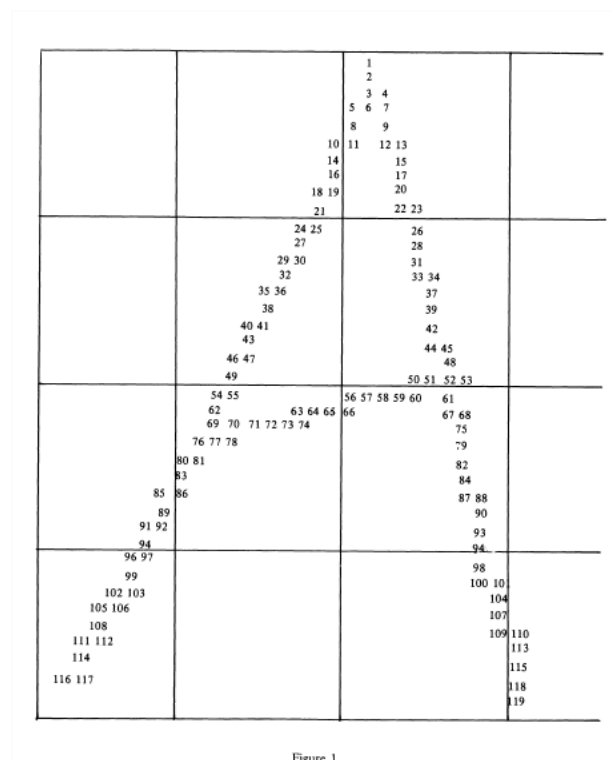
E. DIDAY

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

« Simultaneous search of k subspaces with minimal inertia , ie local factorial planes.
 Dynamic clustering algorithm with factorial planes as kernels »



- Distance to the factorial plane instead of distances to the barycenters
- Two types of updating
 - Update criterium and partition after each reallocation (true k-means) « stochastic » algorithm
 - Update criterium and partition after a « pass » or complete run of all observations « batch » algorithm

2.2 Clusterwise regression

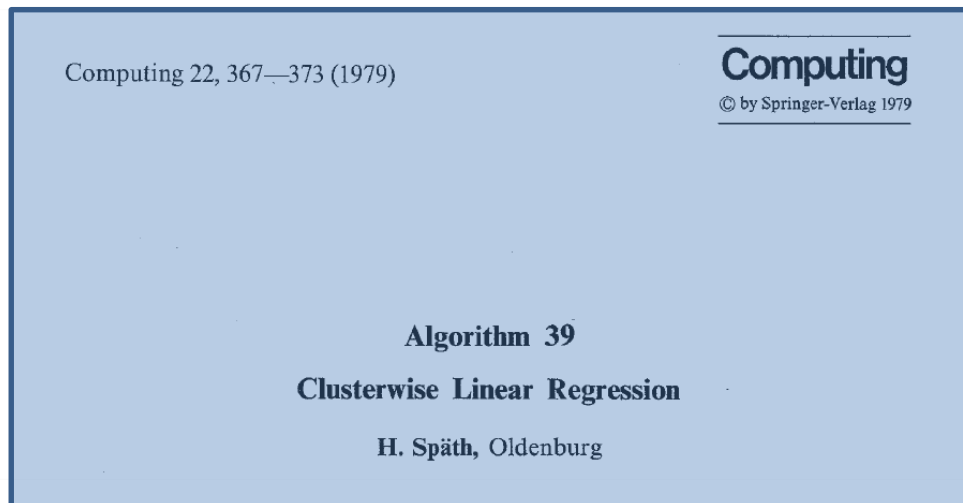
- For each class, fit a linear model by least squares.

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

- The fit is always better than with the global model

- Ph.D. Christian Charles (1977):
 - An application of dynamic clustering
 - Step 1: define an initial partition and estimate k local models.
 - Step 2: Each observation is reallocated to the cluster giving the smallest regression residual ie the best prediction. Once all observations being reallocated (or not) , replace the initial partition
 - Iterate step 1 and 2 until convergence
 - Ridge or principal components regression for small clusters .

- Späth (1979) coined the expression « clusterwise regression » together with a Fortran package
 - Same criterium
 - Exchange algorithm (k-means)



3. Mixture models

Latent class regression

Proposed by DeSarbo & Cron (1988)

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

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

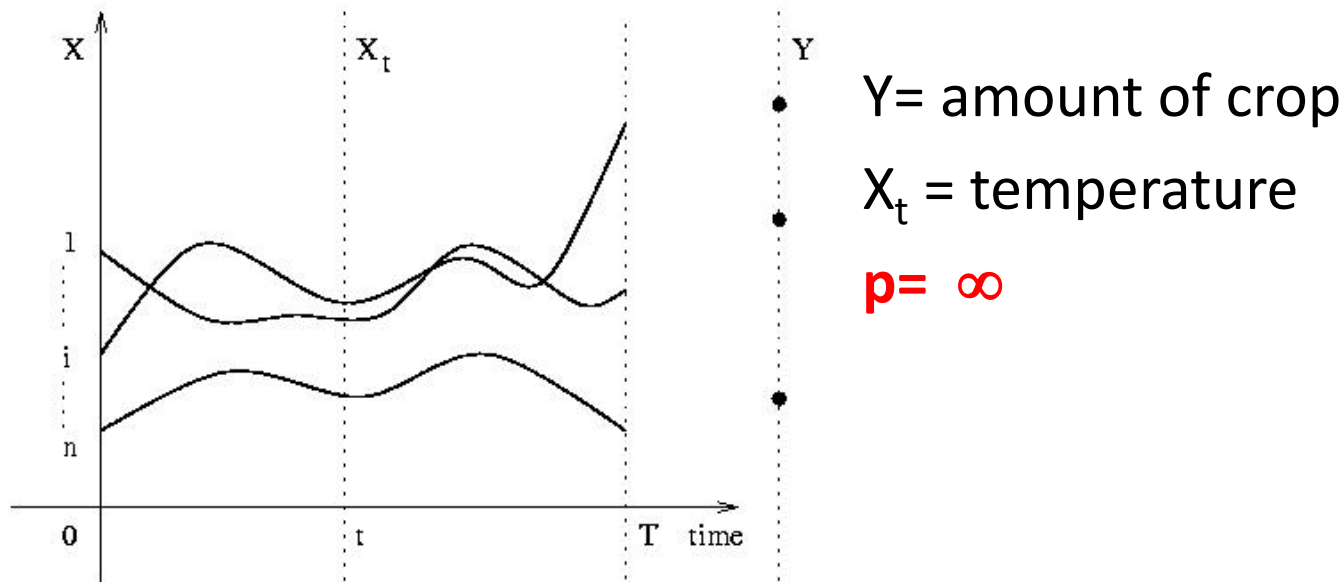
$$= \sum_{k=1}^K \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp \left[\frac{-(y_i - \mathbf{x}_i \mathbf{b}_k)^2}{2\sigma_k^2} \right], \quad (10)$$

- Extension to Poisson regression :
 - (Wedel & al. 1993), generalized local linear models (Wedel & DeSarbo, 1995)
- Mainstream methodology: maximum likelihood through EM algorithm

4. Two generalizations

4.1 Functional data

Prediction of Y with a functional variable :



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)

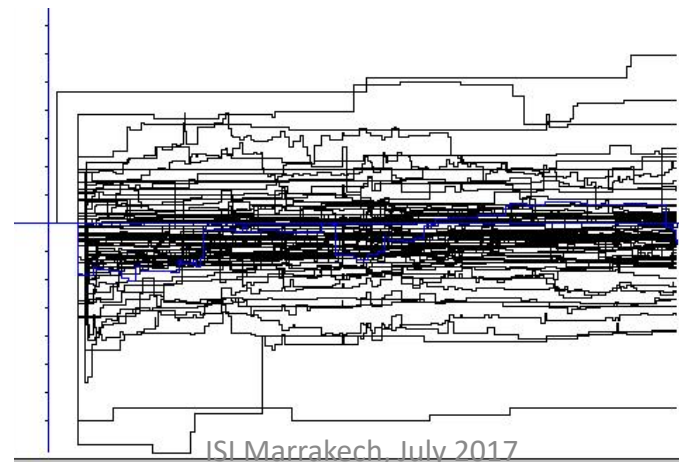
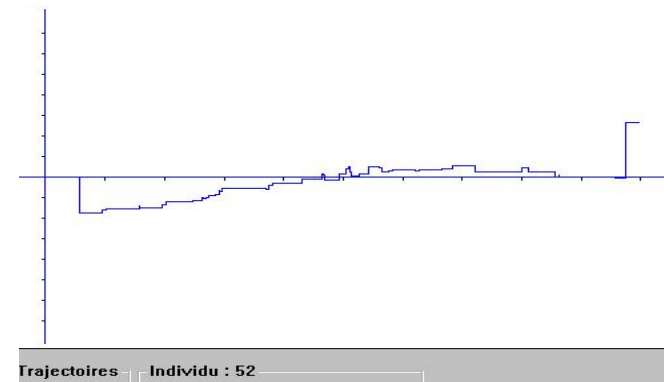
- « Integral regression » $\hat{Y} = \int_0^T \beta(t) X_t dt$

instead of a finite sum $\hat{Y} = \sum_{j=1}^p \beta_j X_j$

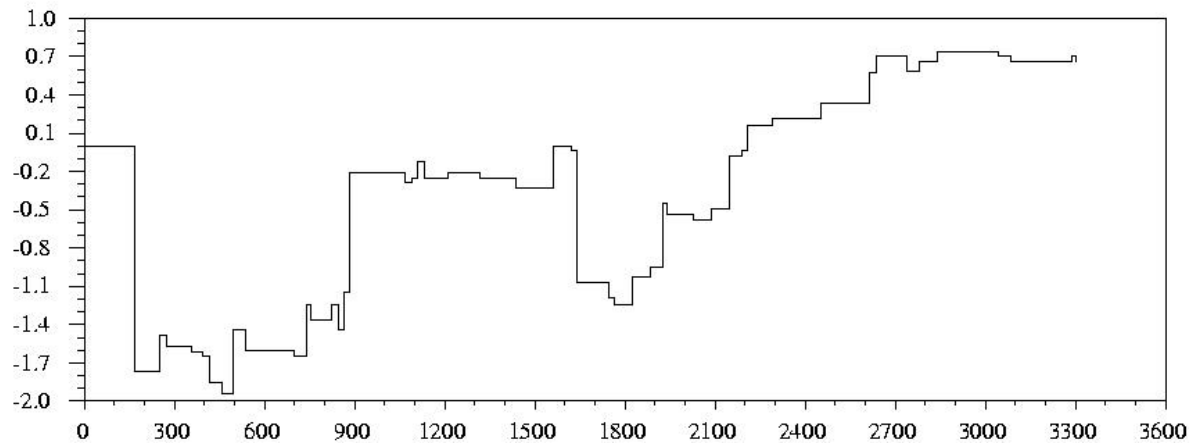
- Ill-posed problem (more variables than observations)
- Preda & Saporta, 2005a: functional PLS regression
- Preda & Saporta, 2005b: clusterwise functional regression.

Application to stock market data

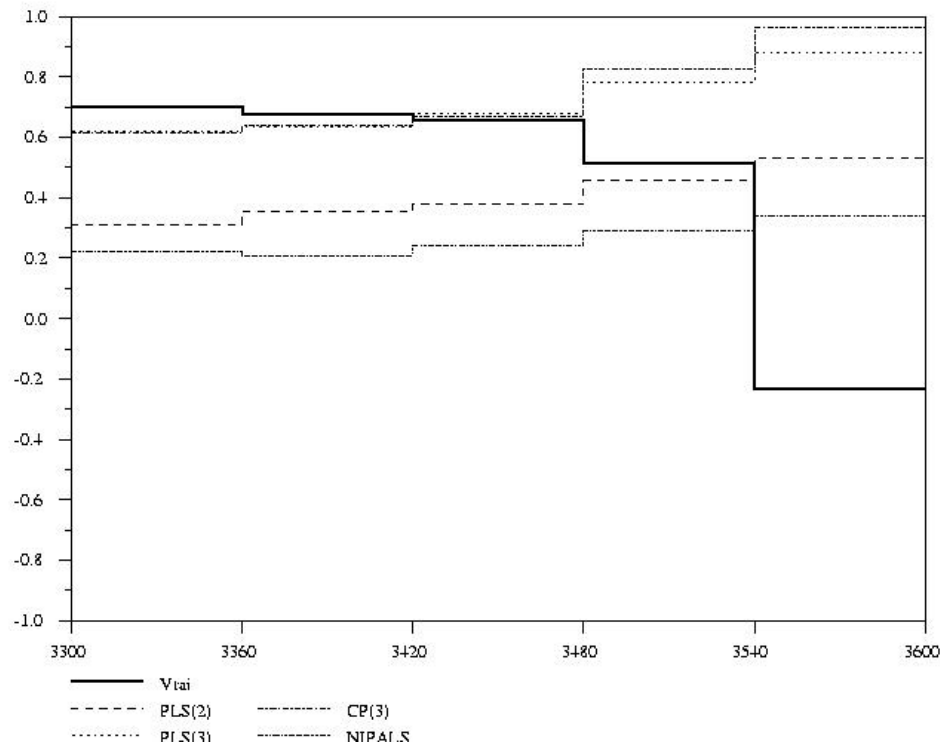
- Growth index during 1 hour (between 10h and 11h) of 84 shares at Paris Stock Exchange



- How to predict a new share between 10h55 and 11h using data between 10h and 10h55?

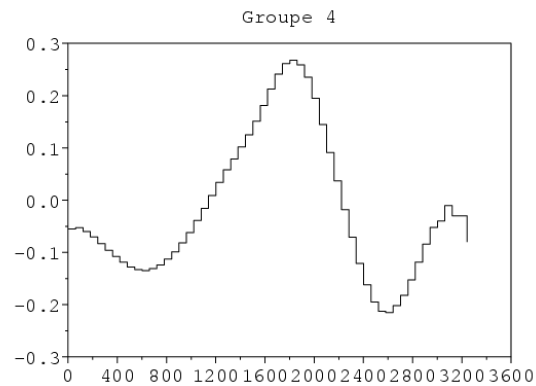
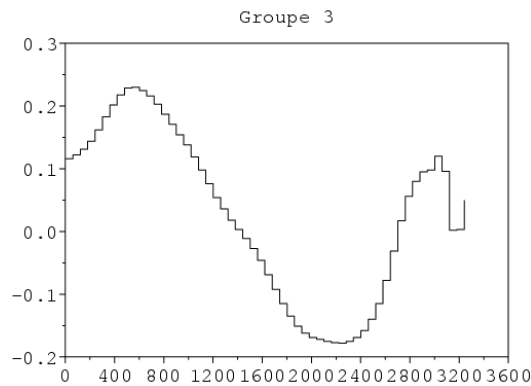
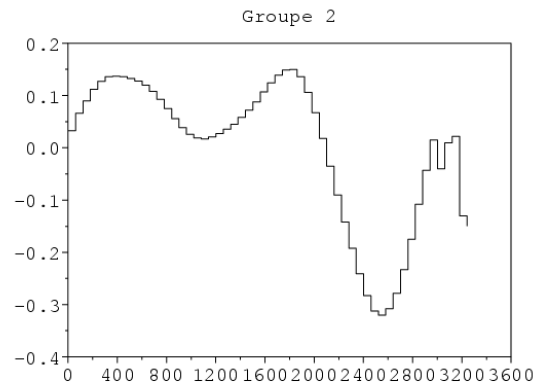
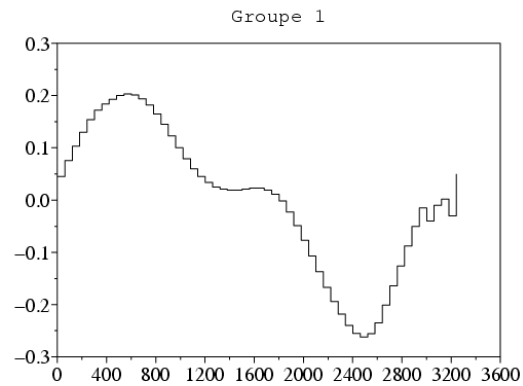


- Exact computations need 1366 variables
(number of intervals where the 85 curves are constant)
- Discretisation in 60 intervals.
- Global regressions: crash of share # 85 not detected!



- Clusterwise PLS regression

- Four clusters (17;32;10;25)
- Number of PLS component for each cluster: 1; 3; 2 ; 2 (cross-validation)



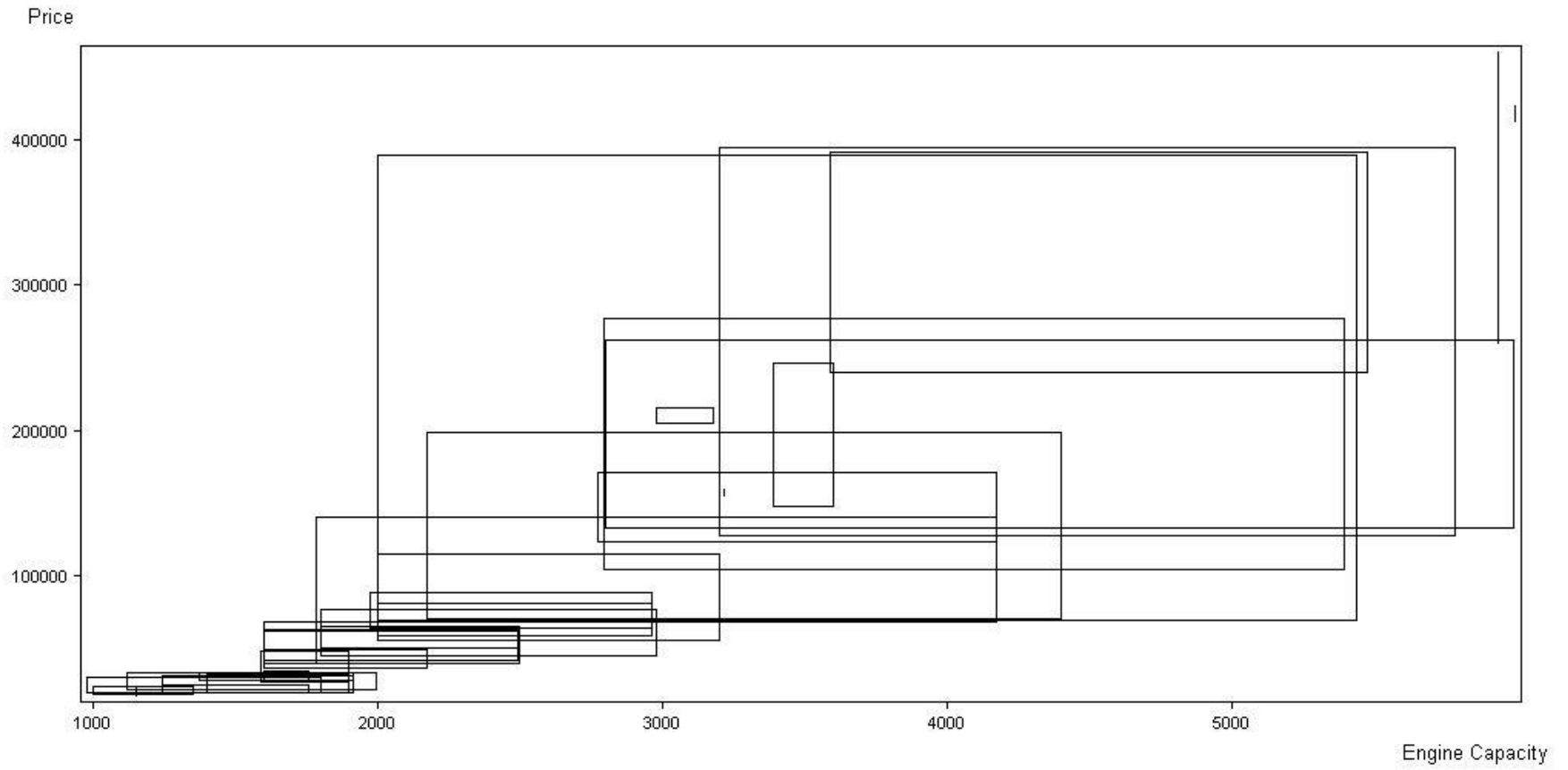
- Allocation to nearest cluster (n°1) and application of local model (PLS2)

	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	<i>SSE</i>
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 Interval data

- Carvalho et al. , 2010 : a clusterwise generalization of « center and range » regression for symbolic interval data
- Example : 33 cars, two variables

car	Y= Price	X= Displacement
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)



- Center and range regression: two linear regressions centers and mid-ranges stacked in a unique data matrix, hence the predicted interval
- Clusterwise generalization:
 - Minimize over partition and coefficients

$$\begin{aligned}
 J &= \sum_{k=1}^K \sum_{i \in P_k} (\epsilon_{i(k)})^T \epsilon_{ik} = \sum_{k=1}^K \sum_{i \in P_k} [(\epsilon_{i(k)}^c)^2 + (\epsilon_{i(k)}^r)^2] \quad (2) \\
 &= \sum_{k=1}^K \sum_{i \in P_k} \left\{ \left[y_i^c - \left(\beta_{0(k)}^c + \sum_{j=1}^p \beta_{j(k)}^c x_{ij}^c \right) \right]^2 + \left[y_i^r - \left(\beta_{0(k)}^r + \sum_{j=1}^p \beta_{j(k)}^r x_{ij}^r \right) \right]^2 \right\}
 \end{aligned}$$

- Dynamic clustering algorithm

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

K - partition	cluster k	“Center Model”	“Range Model”
1	1	$\hat{y}_{(1)}^c = -98840.9 + 79.2 x_1^c$	$\hat{y}_{(1)}^r = -341.4 + 60.9 x_1^r$
2	1	$\hat{y}_{(1)}^c = -63462.2 + 59.6 x_1^c$	$\hat{y}_{(1)}^r = -4560.1 + 47.1 x_1^r$
	2	$\hat{y}_{(2)}^c = -22836.5 + 68.8 x_1^c$	$\hat{y}_{(2)}^r = 34563.6 + 68.6 x_1^r$
3	1	$\hat{y}_{(1)}^c = -77422.1 + 82.0 x_1^c$	$\hat{y}_{(1)}^r = 2229.7 + 92.2 x_1^r$
	2	$\hat{y}_{(2)}^c = -58484.1 + 71.1 x_1^c$	$\hat{y}_{(2)}^r = 101952.9 - 546.7 x_1^r$
	3	$\hat{y}_{(3)}^c = -73362.1 + 62.0 x_1^c$	$\hat{y}_{(3)}^r = -9755.9 + 53.2 x_1^r$

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
$R_{c(k)}^2$	0.93	0.95	0.91	0.97	0.99	0.98
$R_{r(k)}^2$	0.53	0.79	0.66	0.98	0.98	0.83

- Cross validation « 10 fold »
- Model aggregation: stacking (Breiman, 1996)
predictions of the K models

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)

- The two-classes model seems the best

5. Implementation issues

5.1 High dimensional data

- If the cluster size is less than p : no OLS solution for local
- Regularized regression (ridge, pcr, pls ...)

5.2 Identifiability of mixture models

- Cf Hennig (2000)

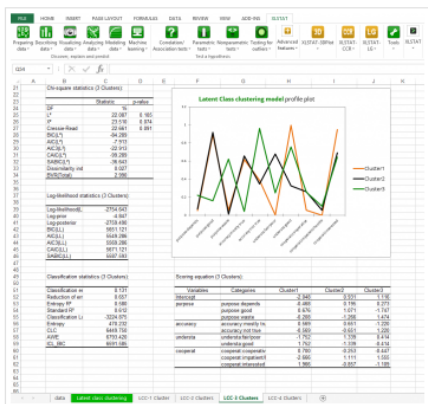
"Mixtures of linear regression models with Gaussian noise are identifiable, if the number of components K is smaller than the minimal number of (feasible) hyperplanes necessary to cover all covariate points (without intercept)."

5.3 Software

- Latent class regression
 - Commercial:
 - GLIMMIX * <http://www.scienceplus.com/glimmix>
 - LatentGold 5.0 and XLSTAT-LG
 - Free :
 - Flexmix 2.3-13 , january 2015 (Leisch, 2004)
 - Mixmod

* Not to be confused with homonymous SAS or R procedures

XLSTAT-LatentClass



€235.00 Company/Private
1 user | Annual
license

PRICES

€145.00 Academic

PRICES

DESCRIPTION

FEATURES

TUTORIALS

SPECIFICATIONS

The XLSTAT-LatentClass Option (only for Windows OS)

XLSTAT-LatentClass is a reduced version of Statistical Innovations' highly acclaimed LatentGOLD® package, which can help you discover the segments hidden in your data.

Latent Class (LC) Cluster models and LC Regression models both offer unique features compared to traditional clustering approaches. Unlike the ad-hoc clustering algorithms, LC is based on a formal statistical model and provides probability-based classification, formal model selection criteria and optimal handling of missing data.

XLSTAT-LatentClass offers a wide variety of easily implementable options that let users gain full control over the Latent Class models. For example, it is possible to include variables of different scales (continuous, ordinal or nominal) within the same model and to use scoring equations to be able to classify new cases into their most likely latent class as a function of the observed variables.

Latent GOLD is a registered trademark of Statistical Innovations.

SHOW FEATURES

VIEW ALL TUTORIALS

- Some clusterwise regression algorithms
 - Simulated annealing (DeSarbo et al., 1989)
 - VNS metaheuristics « Variable Neighbour Search » (Caporossi & Hansen, 2005)
 - Global optimization (Carbonneau et al., 2014)
- No systematic comparison, no simple package available
- Multiblock clusterwise
 - Package R `mbclusterwise` (Bougeard, 2016)

5.4 How to predict new observations knowing only the predictors

- « Hard » rule: allocate to the nearest cluster and apply the relevant model
 - How? Distance, knn etc.
- « Soft » rule: weighted average of the k predictions
 - Bayesian Model Averaging : weight = posterior probabilities of each model given the observation
- Random rule: according to posterior probabilities

- k-means clusterwise regression
 - No technical difficulty if some kind of regularized regression is used for small size clusters.
- Latent class regression
 - Becomes tricky!
 - In flexmix, posterior probabilities need the true class

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

“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 , personal communication)

6. Some experiments*

6.1 criterium and algorithm

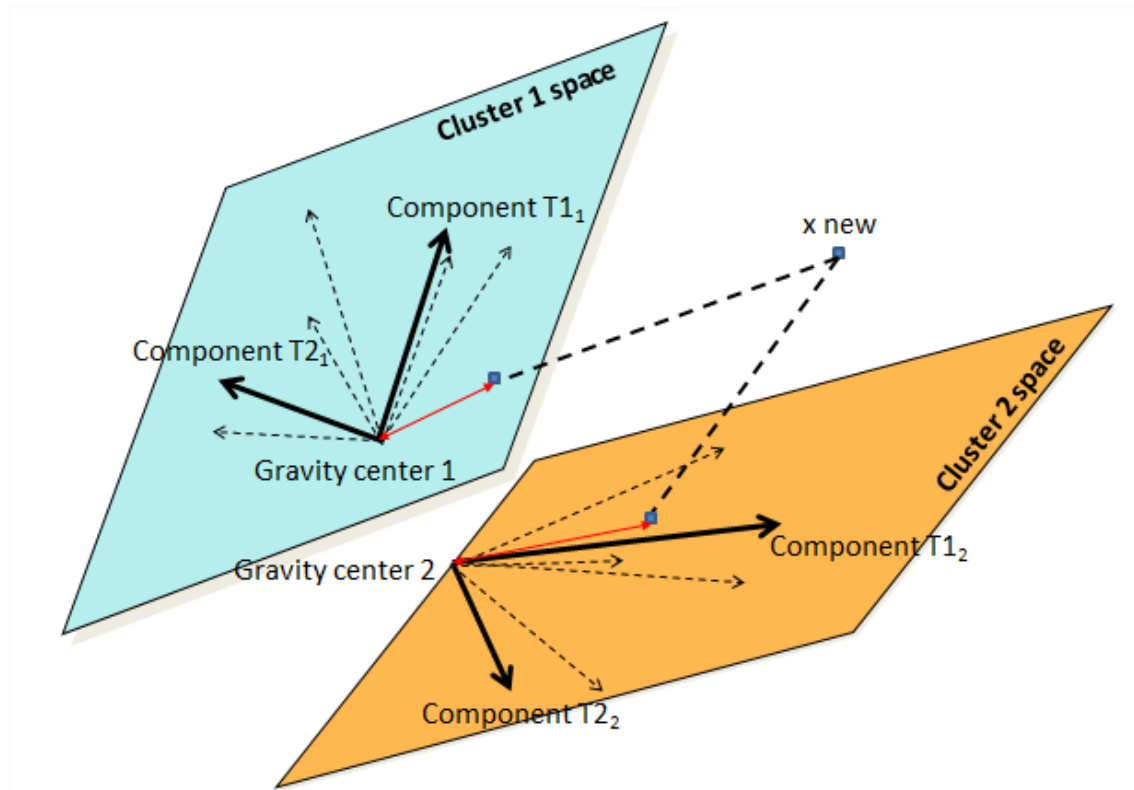
- RMSE
- stochastic algorithm with several passes of all data (stopping rule if RMSE's variation is less than 1%) . Several random initialisations, best one chosen by 10 fold CV.

* With N.Niang & S.Bougeard

6.2 Use of components

- Dimension reduction to avoid issues related to small size clusters.
 - Global components(Esposito-Vinzi et al., 2003, 2005) through PLS-DA
 - Cluster dependent components for a better fit

- Local components
 - K PLS regressions (subspaces)
 - Posterior probabilities proportional to $\exp(-d_j^2)$ weighted by cluster sizes



- CV-10 for choosing both the number of clusters and the numbers of local components
- bayesian rule for «soft » predictions

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

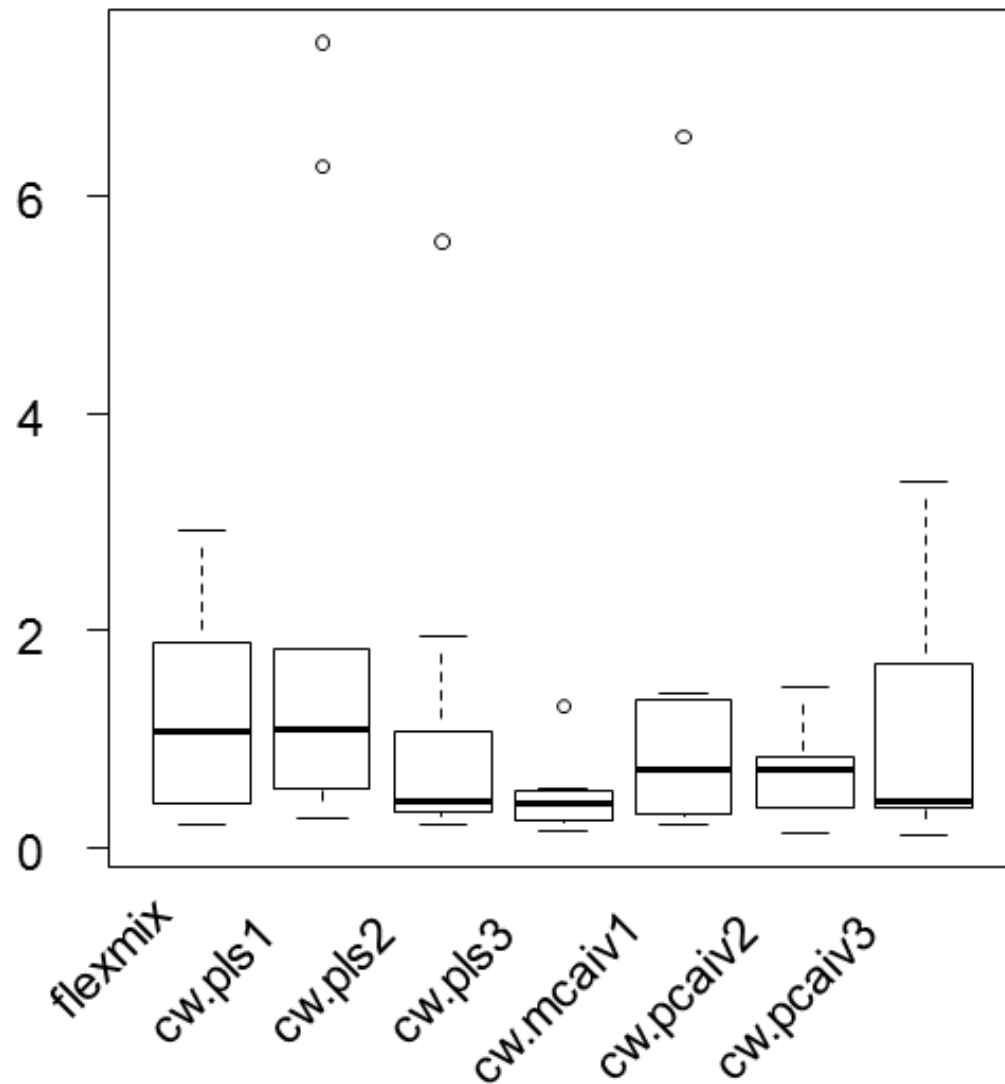
6.3 Examples

- Example 1 : Späth data (1978) with added noise

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



	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

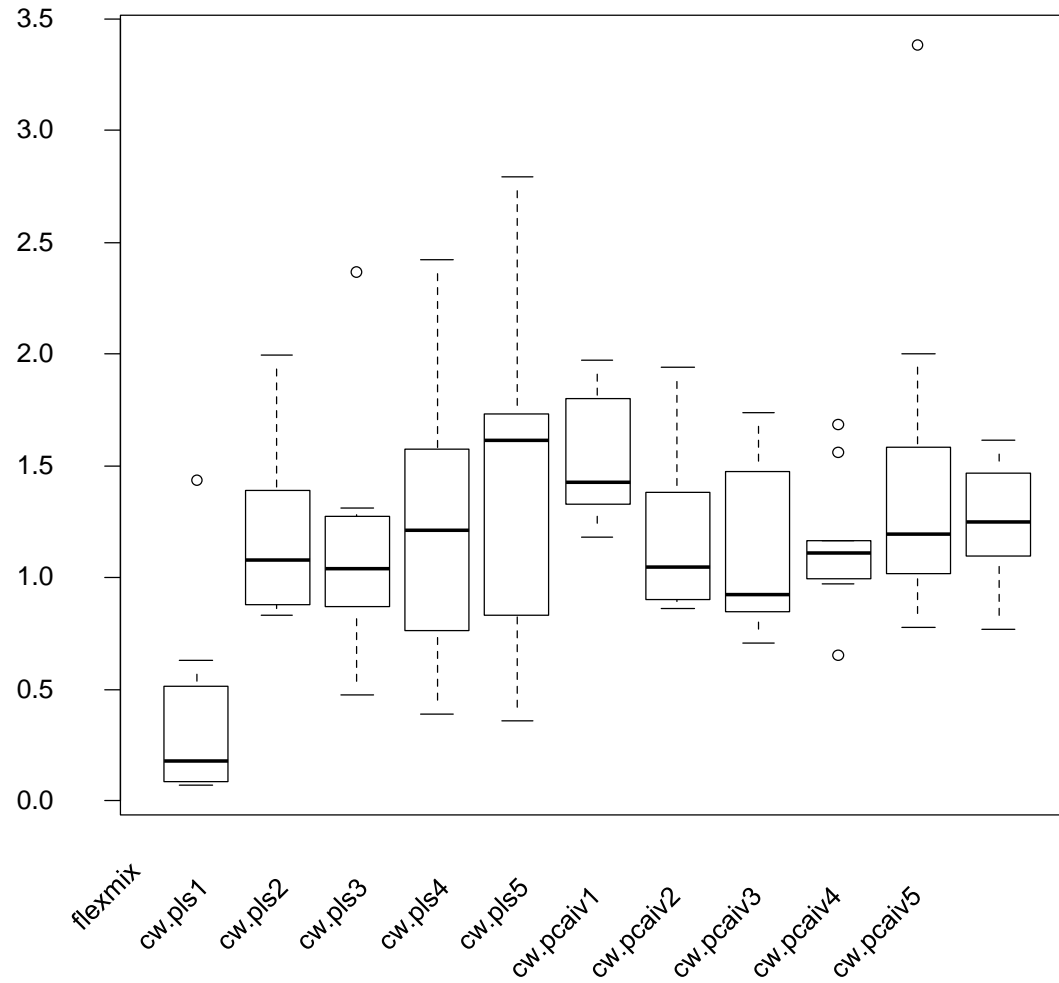
- Exemple 2: DeSarbo et al. 2005

Table 19.1 Synthetic Regression Data

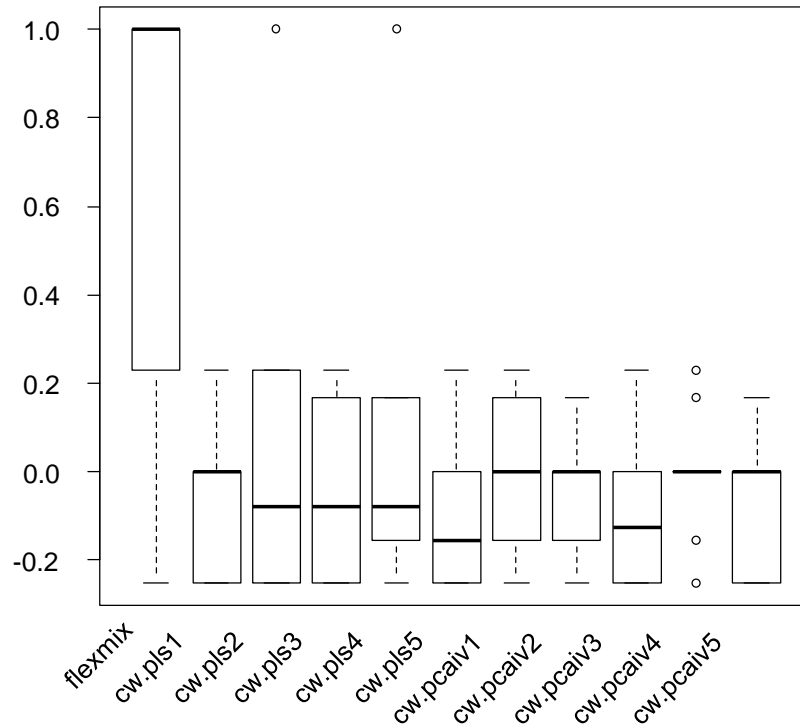
X_1	X_2	X_3	X_4	X_5	Y
0.995	1.893	-1.291	-0.560	0.508	-3.287
-1.268	-0.988	0.244	0.716	-2.261	-8.254
0.135	0.725	-0.511	0.712	-1.551	-11.867
0.687	0.223	1.441	0.098	-0.352	10.859
1.020	-0.355	-0.660	0.090	-1.544	-4.733
0.313	0.999	0.365	-0.991	1.000	12.191
0.584	-1.775	-1.702	1.843	-1.267	-25.508
-1.342	-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
-0.702	1.088	-0.514	1.695	0.146	14.756

2 clusters of 25 obs.

Ten-fold RMSE - De Sarbo 2005 (2 clusters)



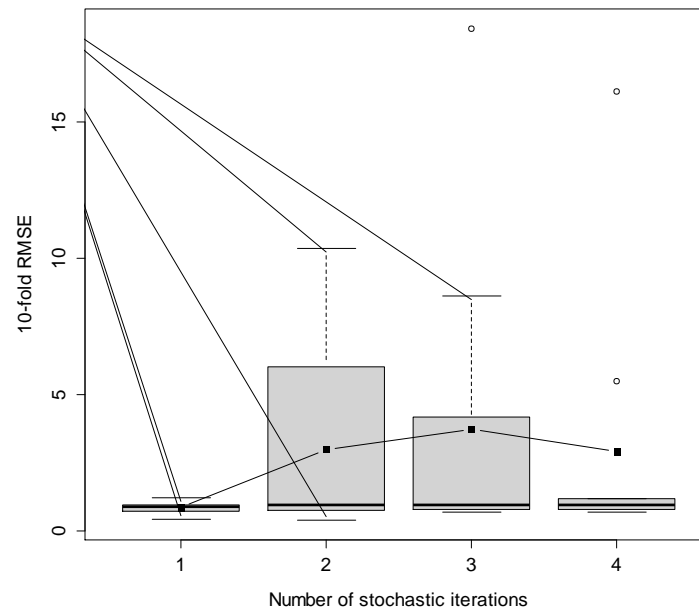
Adjusted Rand index - De Sarbo 2005 (2 clusters)



- Excellent performance of Flexmix which recovers the true clusters and hence give good predictions **but uses cluster information ...**

6.5 Overfitting

- Too many complete runs leads to overfitting, like in NN

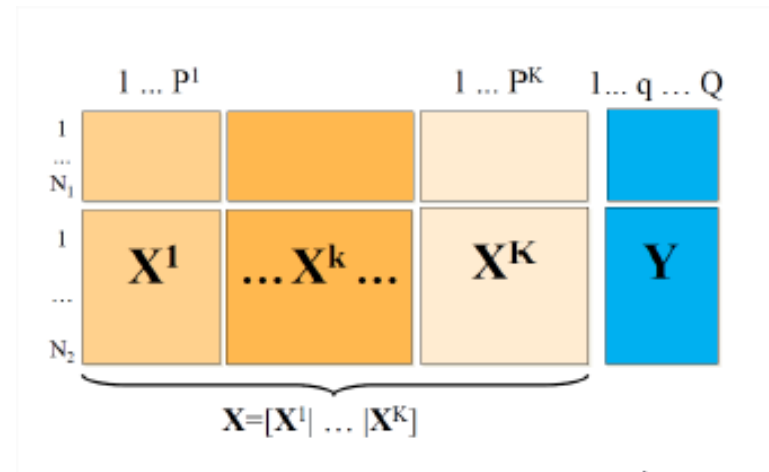
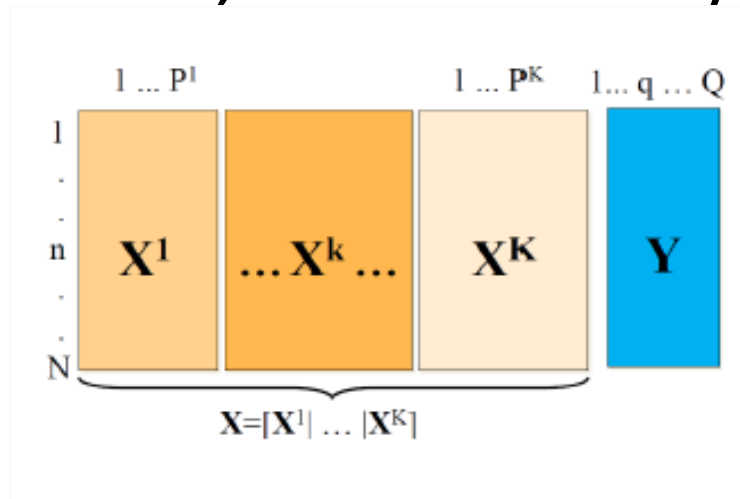


7. Conclusions

- Publications bias in favour of mixture models
- PLS components and cross validation efficient for prediction
- Opposition between generative models and prediction (Breiman, 2001)
 - Two different goals (Saporta, 2008, Shmueli, 2010)
 - **Modelling** *versus* **prediction**
 - Criteria: goodness of fit *versus* prediction accuracy

In progress

- Easy extension to multivariate Y (PLS2 ...)
- Generalisation to multiblock regression (Niang et al. , CARME 2015)



and sparse PLS (Bougeard et al. , PLS 2017)

References

- BOUGEARD, S. (2016) <https://cran.r-project.org/web/packages/mbclusterwise/mbclusterwise.pdf>
- BREIMAN, L. (1996): Stacked Regressions. *Machine Learning* 24, pp.49-64.
- BREIMAN, L. (2001) : Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231
- 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*. Ph.D. 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.

- 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, N., BOUGEARD, S., SAPORTA, G., ABDI, H. (2015) Clusterwise multiblock PLS, *CARME 2015*, pp.58, Naples
- 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
- SHMUELI G. (2010) To explain or to predict? *Statistical Science*, 25, 289–310
- 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.