

Supervised and Unsupervised Linear Methods for Functional Data



Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
Conservatoire National des Arts et Métiers
292 rue Saint Martin
F 75141 Paris Cedex 03
Gilbert.saporta@cnam.fr
<http://cedric.cnam.fr/~saporta>

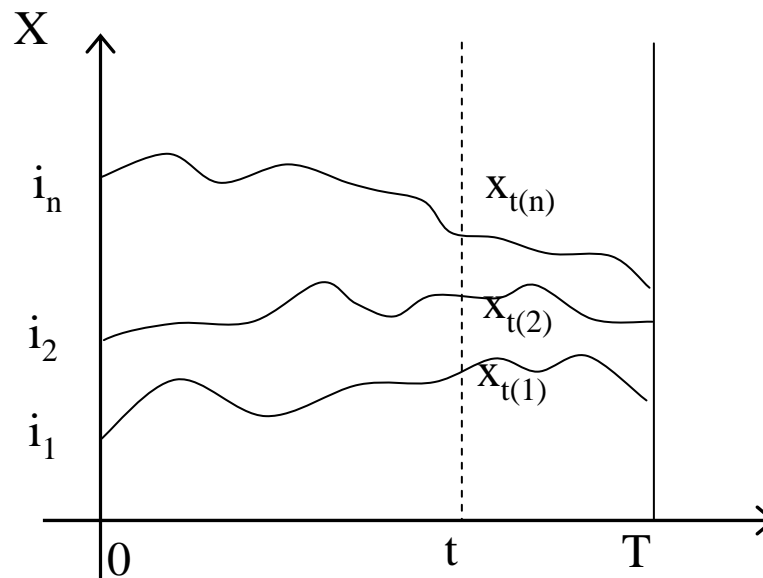
Outline




- 1. Introduction**
- 2. PCA for functional data**
- 3. Regression on functional data**
 - OLS regression
 - PCR and PLS regression
 - Clusterwise regression
- 4. Classification with functional data**
 - Fisher's LDA
 - Anticipated and adaptive prediction
- 5. Conclusion and perspectives**

1. Introduction

- Functional data: curves or paths from a stochastic process X_t



- 
- No response variable
 - unsupervised or exploratory analysis
 - Single response variable Y
 - Y numerical: regression
 - Y categorical: supervised classification, discrimination
 - Common time interval $[0;T]$, zero mean variables

- Pioneering works:

- R.A. Fisher – 1924
- J. C. Deville – 1974
- P. Besse – 1979
- G. Saporta – 1981

- Later:

- Aguilera, Valderrama – 1993, 1995, 1998
- Ramsay, Silverman – 1995, 1997
- Van der Heijden – 1997
- Preda, Cohen – 1999
-

2. *PCA for functional data*

- Unique decomposition where f_i orthonormal functions, and ξ_i uncorrelated variables

$$X_t = \sum_{i=1}^{\infty} f_i(t) \xi_i$$

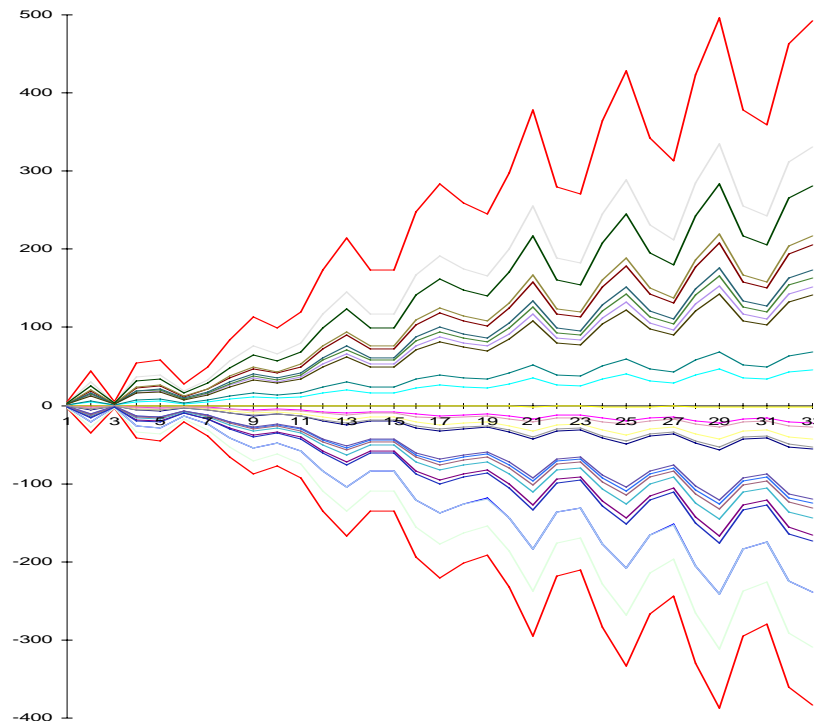
- factor loadings: $\int_0^T C(t, s) f_i(s) ds = \lambda_i f_i(t)$

- principal components: $\xi_i = \int_0^T f_i(t) X_t dt$

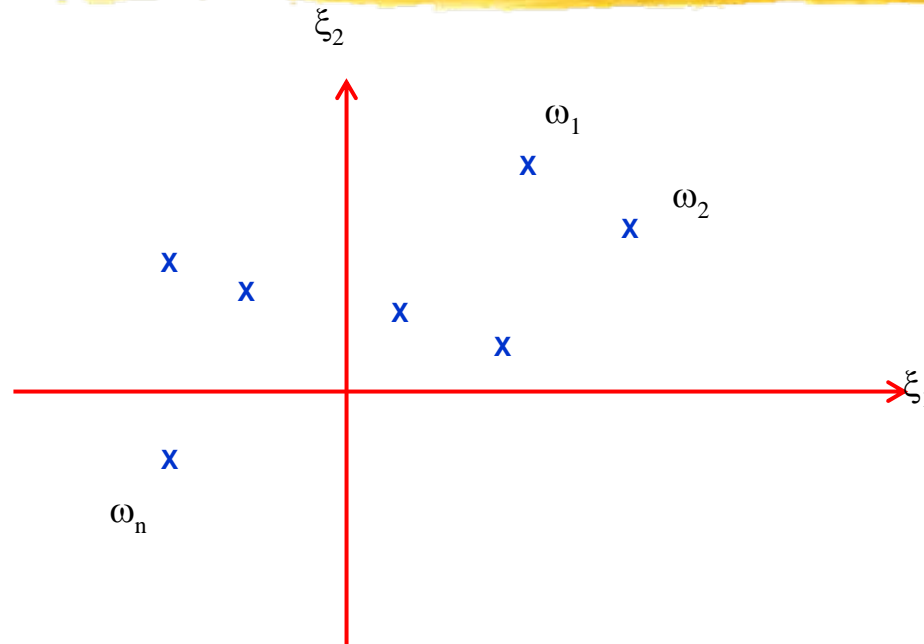
Sum of quasi-deterministic processes

$$X_t(i) = \xi_i f(t)$$

All curves have the same shape except for a constant ξ_i relative to unit i



Principal plane: allows visualisation, clustering



Numerical solution:

- Integral equations cannot be solved in the general case

- For finite n , exact solution :

- W matrix of all inner products between trajectories

$$w_{uv} = \int_0^T x_u(t) x_v(t) dt \quad u, v = 1, 2, \dots, n$$

- principal components are eigenvectors of W
- Factors are

$$f(t) = \frac{1}{n} \frac{1}{\lambda} \sum_{u=1}^n \xi_u X_u(t)$$

Other:

- for step functions: finite number of variables and of units: operators are matrices, but with a very high size
- Approximations by discretisation of time

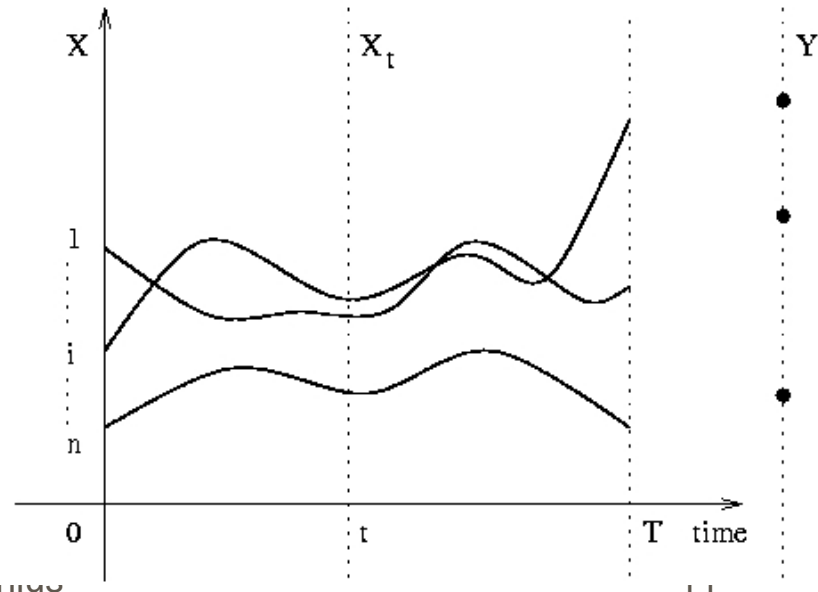
3. Linear Regression on functional data

- Example : R.A.Fisher (1924)

Y = amount of crop

X_t = temperature curves

$p = \infty$



- Infinite number of predictors
- Linear combination
 - « Integral regression »

$$\hat{Y} = \int_0^T \beta(t) X_t dt$$

- instead of a finite sum

$$\hat{Y} = \sum_{j=1}^p \beta_j X_j$$

Disregarding, then, both the arithmetical and the statistical difficulties, which a direct attack on the problem would encounter, we may recognise that whereas with q subdivisions of the year, the linear regression equations of the wheat crop upon the rainfall would be of the form

$$\bar{w} = c + a_1 r_1 + a_2 r_2 + \dots + a_q r_q$$

where r_1, r_2, \dots, r_q are the quantities of rain in the several intervals of time, and a_1, \dots, a_q are the regression coefficients, so if infinitely small subdivisions of time were taken, we should replace the linear regression function by a *regression integral* of the form

$$\bar{w} = c + \int_0^T ar dt, \quad \quad (III)$$

where $r dt$ is the rain falling in the element of time dt ; the integral is taken over the whole period concerned, and a is a *continuous* function of the time t , which it is our object to evaluate from the statistical data.

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)

3.1 The OLS problem

- Minimizing $E\left(Y - \int_0^T \beta(t) X_t dt\right)^2$ leads to normal, or Wiener-Hopf, equations:

$$\text{cov}(X_t, Y) = \int_0^T C(t, s) \beta(s) ds$$

where $C(t, s) = \text{cov}(X_t, X_s) = E(X_t X_s)$

Generalization of $X'y = X'X\beta$


- Picard's theorem: β is unique if and only if:

$$\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i^2} < \infty$$

$$c_i = \text{cov}(Y, \xi_i) = \text{cov}\left(Y, \int_0^T f_i(t) X_t dt\right) = \int_0^T E(X_t Y) f_i(t) dt$$

- Generally not true...especially when n is finite since $p > n$. Perfect fit when minimizing:

$$\frac{1}{n} \sum_{i=1}^n \left(y_i - \int_0^T \beta(t) x_i(t) dt \right)^2$$

- 
- Even if β is unique, Wiener-Hopf equation is not an ordinary integral equation: the solution could be a distribution and not a function
 - Constrained solutions are needed. (cf Green & Silverman 1994, Ramsay & Silverman 1997): “roughness penalty” bounds on the integral of $(\beta'')^2$


3.2 Regression on principal components

$$\hat{Y} = \sum_{i=1}^{\infty} \frac{\text{cov}(Y, \xi_i)}{\lambda_i} \xi_i = \sum_{i=1}^{\infty} \frac{c_i}{\lambda_i} \xi_i$$

$$R^2 = r^2(Y, \hat{Y}) = \sum_{i=1}^{\infty} r^2(Y, \xi_i) = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}$$

- Rank q approximation:

$$\hat{Y}^{(q)} = \sum_{i=1}^q \frac{\text{cov}(Y; \xi_i)}{\lambda_i} \xi_i \quad \hat{\beta}^{(q)}(t) = \sum_{i=1}^q \frac{\text{cov}(Y; \xi_i)}{\lambda_i} f_i(t)$$

- 
- Which principal components?
 - First q ?
 - q best correlated with Y ?
 - Principal components are computed irrespective of the response...

3.3 *PLS regression*



- Proposed by H. & S.Wold to solve multicollinearity problems while keeping all variables
- Close to Principal Components Regression:
 - projection on orthogonal linear combinations of the predictors
- Difference: PLS components $t=Xw$ are optimised to be predictive of Y , while principal components explain the variability of X , not of Y

- Tucker's criterium:

$$\max \text{cov}^2(y ; Xw)$$

$$\text{cov}^2(y ; Xw) = r^2(y ; Xw) V(Xw) V(y)$$

- Trade-off between the maximisation of $r(Xw; Y)$ or OLS and maximisation of $V(Xw)$ (PCA)
- First PLS component proportional to:

$$\sum_{j=1}^p \text{cov}(Y; X_j) X_j$$

- Further components by iteration on residuals

3.4 Functional PLS regression

- Use PLS components instead of principal components.
- first PLS component :

$$\max_w \text{cov}^2(Y, \int_0^\infty w(t) X_t dt) \quad \|w\|^2 = 1$$

$$w(t) = \frac{\text{cov}(X_t, Y)}{\sqrt{\int_0^\infty \text{cov}^2(X_t, Y) dt}} \quad t_1 = \int_0^\infty w(t) X_t dt$$

- Higher order PLS components as usual

- order q approximation of Y by X_t :

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \dots + c_q t_q = \int_0^T \hat{\beta}_{PLS(q)}(t) X_t dt$$

- Convergence theorem:

$$\lim_{q \rightarrow \infty} E\left(\left\|\hat{Y}_{PLS(q)} - \hat{Y}\right\|^2\right) = 0$$

- q have to be finite in order to get a formula!
- Usually q is selected by cross-validation
(Preda & Saporta, 2005a)

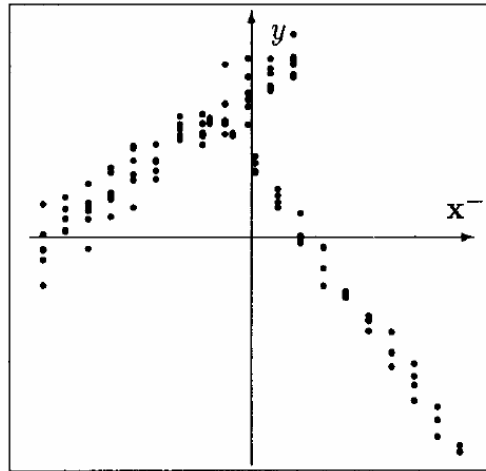
- First PLS component easily interpretable: coefficients with the same sign as $r(y; x_t)$
- No integral equation
- PLS fits better than PCR:

$$R^2(Y; \hat{Y}_{PLS(q)}) \geq R^2(Y; \hat{Y}_{PCR(q)})$$

Same proof as in De Jong, 1993

3.5 Clusterwise regression

A mix between regression and classification

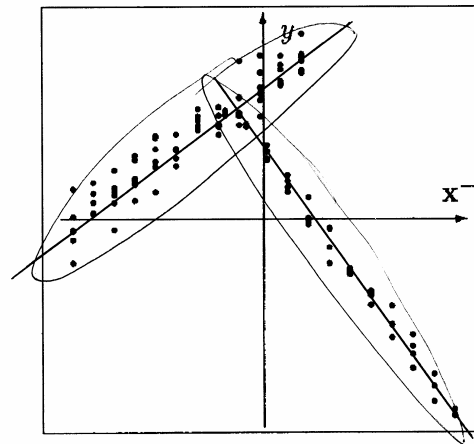


Model

- G , variable with K categories (sub-populations)

$$E(Y | \mathbf{X} = x, G = i) = \alpha^i + \beta^i x$$

$$V(Y | \mathbf{X} = x, G = i) = \sigma_i^2$$



■ OLS and clusterwise regression

\hat{Y} OLS global estimate versus \hat{Y}^L clusterwise "local" estimate

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

- Residual variance of global regression = within cluster residual variance + variance due to the difference between local (clusterwise) and global regression (OLS)

- **Estimation** (Charles, 1977)
 - number of clusters k needs to be known
 - Alternated least squares (k-means)
 - For a given partition: estimate linear regressions for each cluster
 - Reallocate each point to the closest regression model

$$\hat{G}(j) = \arg \min_{j \in \{1, \dots, K\}} (y_j - (\hat{\alpha}^i + \hat{\beta}^i x_j))^2.$$

- Equivalent to ML for fixed regressors, fixed partition model (Hennig, 2000)
- **Optimal k**
 - AIC, BIC, crossvalidation

Clusterwise functional PLS regression

- OLS functional regression not adequate to give estimations in each cluster
- Our proposal: estimate local models with functional PLS regression
- Is the clusterwise algorithm still consistent?

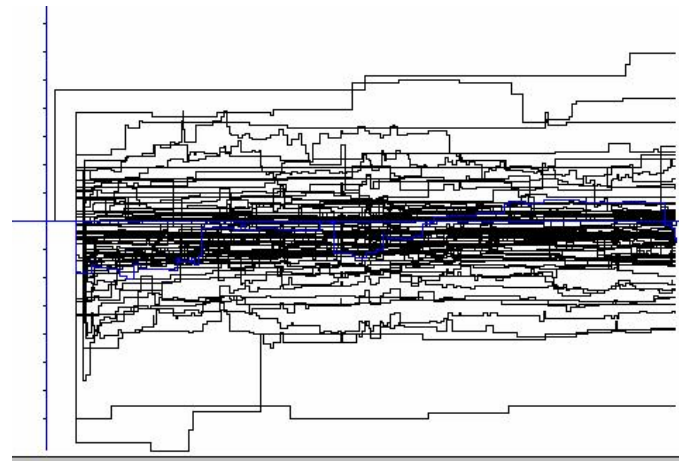
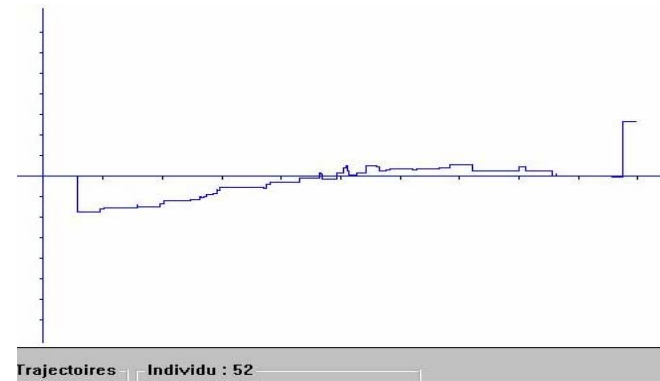
Proposition 1. *For each step s of the clusterwise PLS regression algorithm there exists a positive integer $q(s)$ such that $\hat{G}_{PLS,s}$ and $\{\hat{\alpha}_{PLS,s}^i, \hat{\beta}_{PLS,s}^i\}_{i=1}^K$ given by the PLS regressions using $q(s)$ PLS components preserve the decreasing monotonicity of the sequence $\{\mathcal{V}(\hat{G}_{PLS,s}, \{\hat{\alpha}_{PLS,s}^i, \hat{\beta}_{PLS,s}^i\}_{i=1}^K)\}_{s \geq 1}$*

- Proof in Preda & Saporta, 2005b

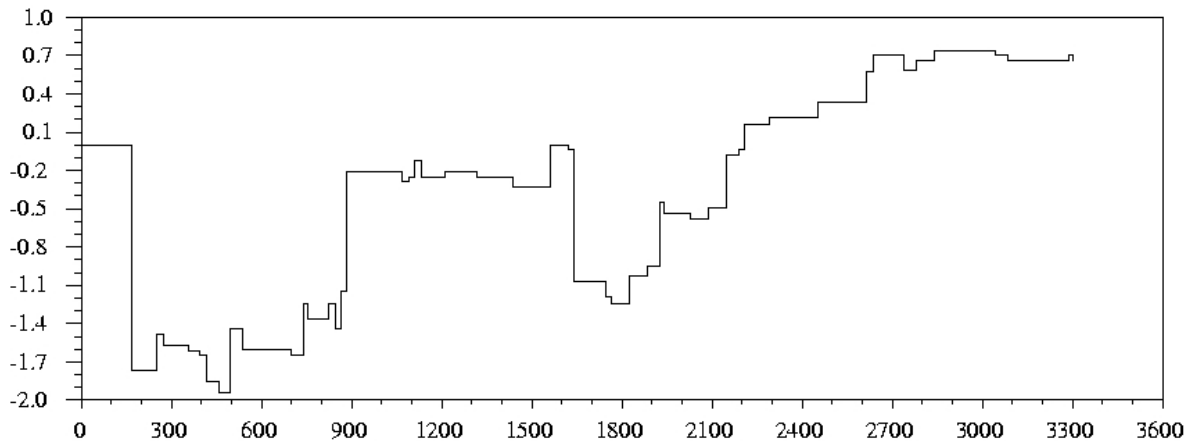
- Prediction:
 - Allocate a new observation to a cluster (nearest neighbor or other classification technique)
 - Use the corresponding local model

Application to stock market data

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



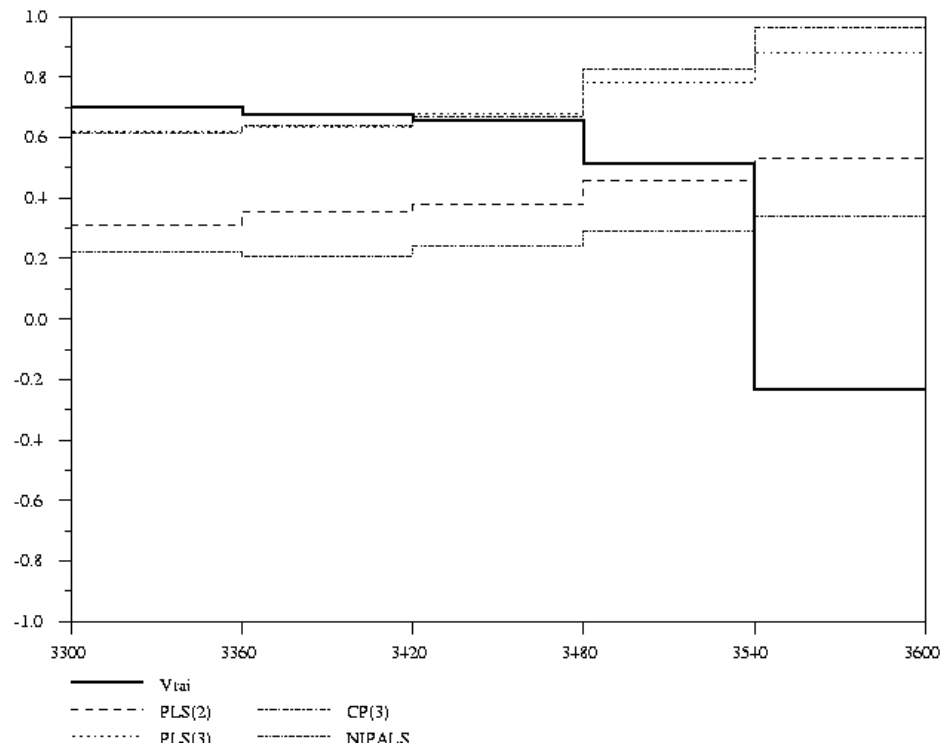
- How to predict a new share between 10h55 and 11h using data between 10h and 10h55?



- Exact computations need 1366 variables
(number of intervals where the 85 curves are constant)
- Discretisation in 60 intervals.
- Comparison between PCR and PLS:

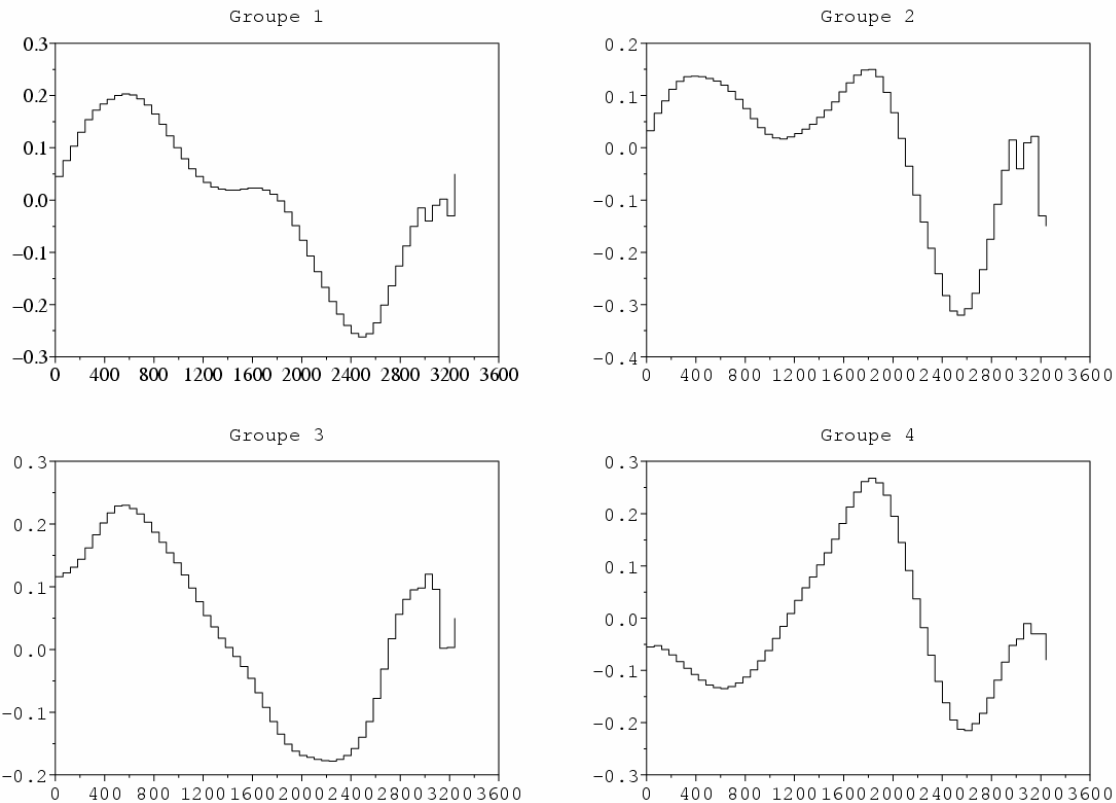
	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	$SSE = \sum_{i=56}^{60} (\hat{m}_i - m_i)^2$
	0.700	0.678	0.659	0.516	-0.233	-
PLS(1)	-0.327	-0.335	-0.338	-0.325	-0.302	3.789
PLS(2)	0.312	0.355	0.377	0.456	0.534	0.928
PLS(3)	0.620	0.637	0.677	0.781	0.880	1.318
PCR(1)	-0.356	-0.365	-0.368	-0.355	-0.331	4.026
PCR(2)	-0.332	-0.333	-0.335	-0.332	-0.298	3.786
PCR(3)	0.613	0.638	0.669	0.825	0.963	1.538

■ Crash of share 85 not detected!



Clusterwise PLS

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

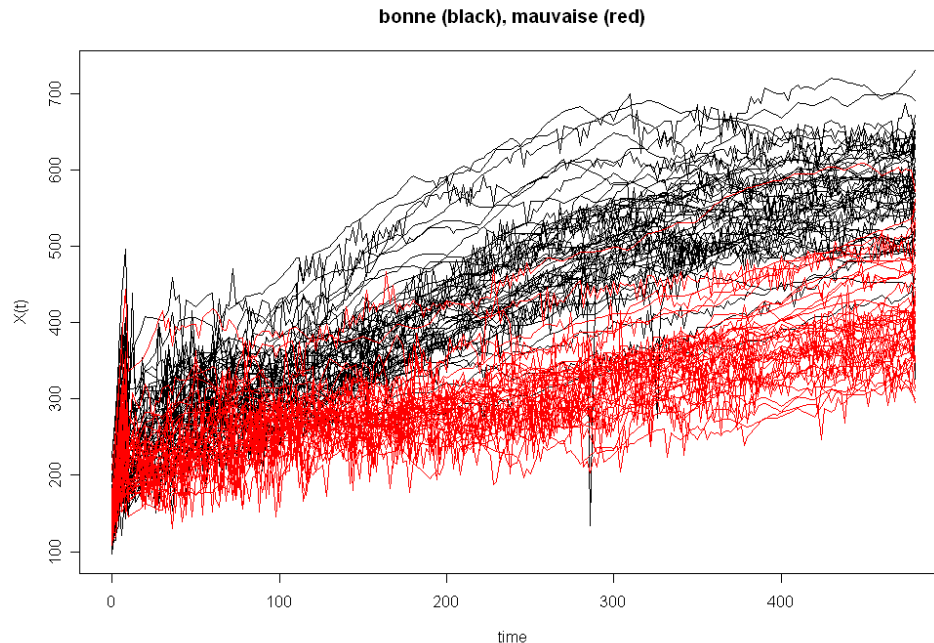


- Share 85 classified into cluster 1

	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	<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. Linear methods for functional discrimination

- Example : Kneading curves for cookies (Danone)



How to predict the quality of the cookies?

4.1 Functional LDA

- LDA : linear combinations $\int_0^T \beta(t) X_t dt$
maximizing the ratio:

Between group variance / Within group variance

- For two groups Fisher's LDF via a regression between coded Y and X_t

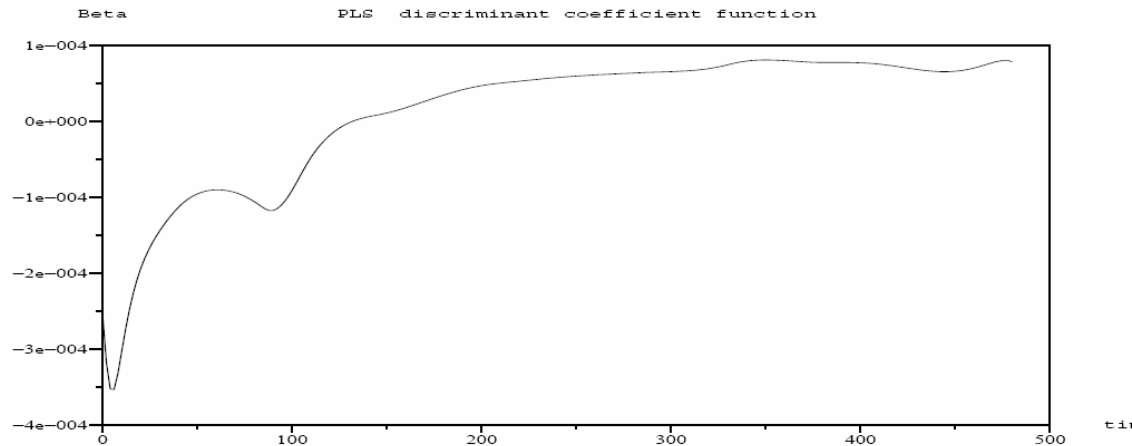
$$\sqrt{\frac{p_1}{p_0}} \quad \text{and} \quad -\sqrt{\frac{p_0}{p_1}}$$

- Same drawbacks as OLS regression
- PLS regression with q components gives an approximation of $\beta(t)$ and of the **score**

$$d_T = \Phi_{PLS}(X) = \int_0^T \hat{\beta}_{PLS}(t) X_t dt$$

Kneading curves

- After $T = 480$ s of kneading, one gets cookies where quality is Y
- 115 observations: 50 « good », 40 « bad » and 25 « unknown »
- 241 equally spaced measurements



- **Performance for $Y = \{\text{good}, \text{bad}\}$**
 - Repeat 100 times the split into learning and test samples of size (60, 30)
 - Average error rate
 - 0.142 with principal components
 - 0.112 with (3) PLS components
 - Average AUC = 0.746

4.3 Functional logistic regression

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \int_0^T x_i(t)\beta(t)dt; \quad i = 1, \dots, n$$

$$\pi_i = P(Y = 1 | X = x_i(t); t \in T)$$

Assumption: parameter function and sample paths are in the same finite space (Ramsay et al., 1997)

$$\beta(t) = \sum_{q=1}^p b_q \psi_q(t) = \mathbf{b}' \boldsymbol{\psi} \quad x_i(t) = \sum_{q=1}^p c_{iq} \psi_q(t) = \mathbf{c}'_i \boldsymbol{\psi}$$

Comes down to standard logistic regression:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha\mathbf{1} + \mathbf{C}\Phi\mathbf{b}$$

where $\mathbf{C} = (c_{iq})$ $\Phi = (\phi_{kq} = \int_T \psi_k(t)\psi_q(t)dt)$

Principal components are used as a basis expansion by Aguilera *et al.* (2006)

4.4 Anticipated and adaptive prediction

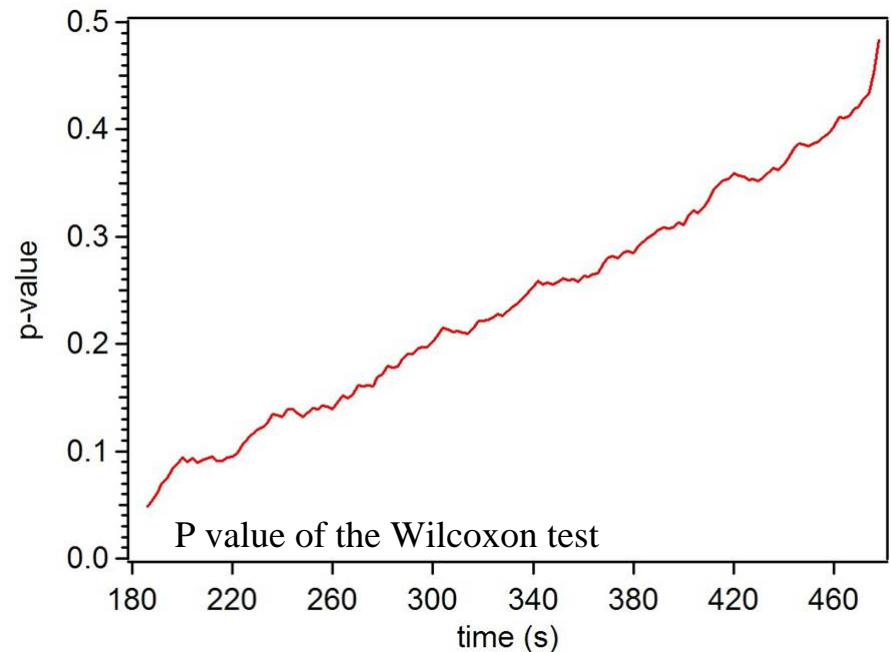
Anticipated prediction (Costanzo et al, 2006)

- $t^* < T$ such that the analysis on $[0; t^*]$ give predictions almost as good as with $[0; T]$
- Solution:
 - When increasing s from 0 to T , look for the first value such that $AUC(s)$ does not differ significantly from $AUC(T)$

- A bootstrap procedure
 - Stratified resampling of the data
 - For each replication b , $AUC_b(s)$ and $AUC_b(T)$ are computed
 - Student's T test or Wilcoxon on the B paired differences $\delta_b = AUC_b(s) - AUC_b(T)$

Application to kneading curves

- Anticipated prediction
 - $B=50$
 - $t^*=186$ minimum value for which the difference between AUC is not significant at .05
- The recording period of the resistance dough can be reduced to **less than half** of the current one!



Adaptive forecasting



- Instead of a common optimal time t^* , adapt t^* to each new trajectory ω given its incoming measurements: $t^*(\omega)$
- For some cases it could be necessary to observe the process during a longer period than $[0, t^*]$, while for others a shorter period could be enough.
- t^* becomes a random variable

- Procedure close in spirit to sequential tests:
 - Discretization of $[0, T]$
 - At t , decide if we stop the observation of $X(\omega)$ (classification decision) then $t^*=t$, or if we continue till $t+h$
- Decision depends on the similarity of $X(\omega)$ with some observations x_i with respect to the prediction of Y

Conservation rate

- d_t discriminant score using only $[0,t]$
- $\Omega_\omega(t)$ set of observations having the same prediction as ω at time t .
- $p_0|\Omega_\omega(t)$ proportion classified in state 0 at time T . Same for $p_1|\Omega_\omega(t)$
- $\text{Max} \{p_0|\Omega_\omega(t) ; p_1|\Omega_\omega(t)\} = \text{Conservation rate} = C_{\Omega_\omega(t)}$ for Ω . Same for $\bar{\Omega}$
- Global conservation rate $\min\left(C_{\Omega_\omega(t)}; C_{\bar{\Omega}_\omega(t)}\right)$

Adaptive rule

Given a confidence conservation threshold $\gamma \in (0, 1)$, e.g. $\gamma = 0.90$, we define the following rule :

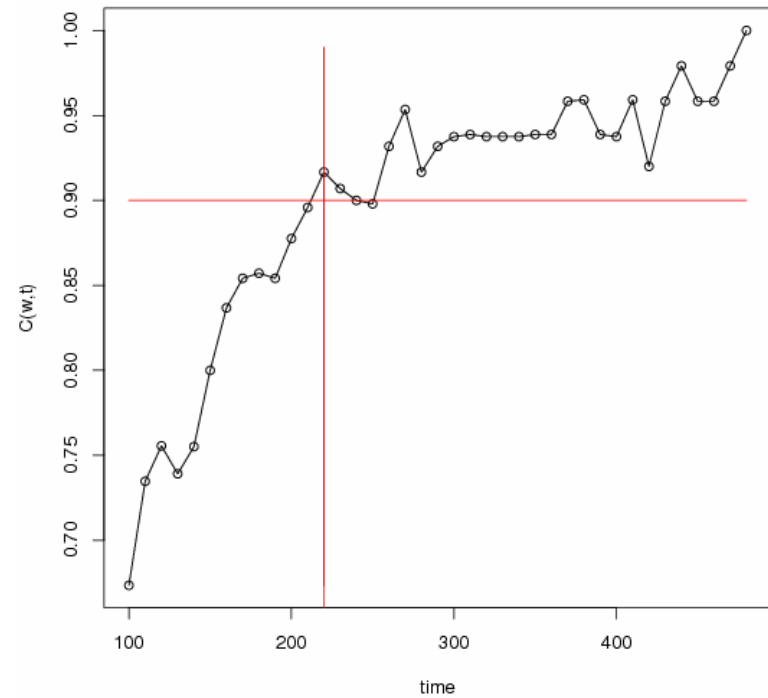
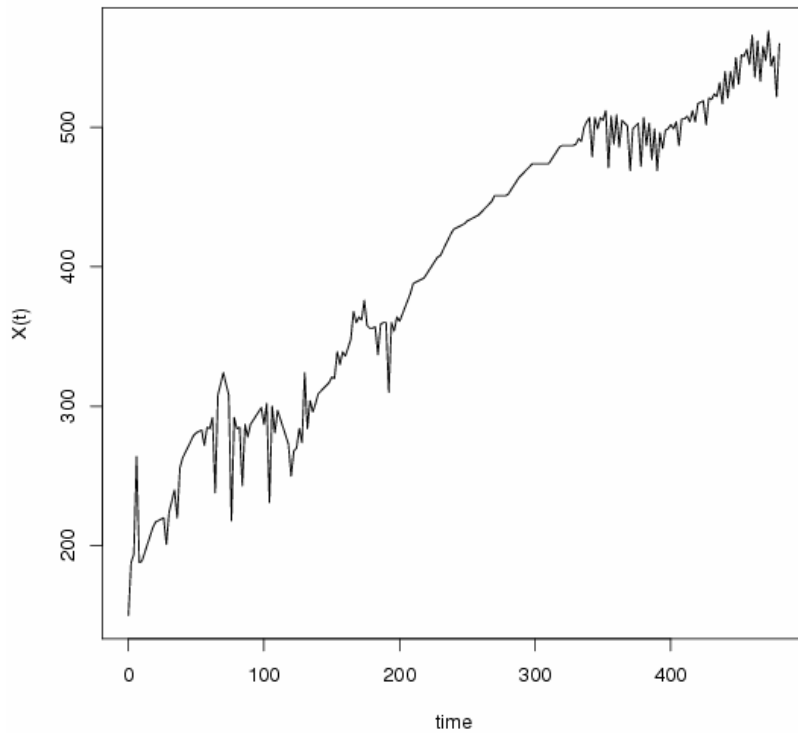
Adaptive prediction rule for ω and t :

- (1) if $C_{\Omega}(\omega, t) \geq \gamma$ then the observation of X for ω on the time interval $[0, t]$ is sufficient for the prediction of $Y(\omega)$. $\hat{Y}(\omega)$ is then the same as the prediction at time T of the subgroup of $\Omega_{\omega}(t)$ corresponding to $C_{\Omega_{\omega}(t)}$.
- (1) if $C_{\Omega}(\omega, t) < \gamma$ then the observation process of X for ω should continue after t . Put $t = t + h$ and repeat the adaptive prediction procedure.

Then, $t^*(\omega)$ is the smallest t such that the condition (1) of the adaptive prediction rule is satisfied.

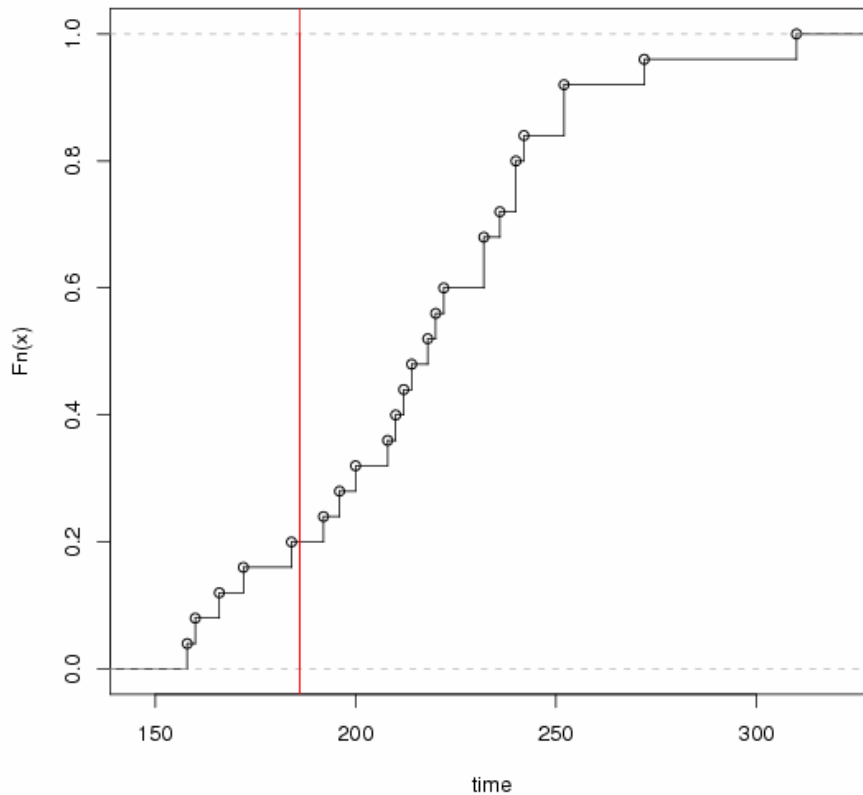
Application

A new flour



Conservation rate

25 « new » flours



- Empirical cumulative distribution function of t^* . 5 time points are earlier than the optimal time for anticipated prediction ($t=186$). 10 flours are predicted in the "good" class.

5. *Conclusions and perspectives*

- Karhunen-Loeve decomposition: a powerful tool for exploratory analysis
- PLS regression: efficient and easy way to get linear prediction for functional data
- Anticipated prediction has been solved by means of a bootstrap procedure
- « on-line » forecasting: adapt t^* to each new trajectory given its incoming measurements.
- Clusterwise discrimination when heterogeneity is present
- Multiple predictors

References

- Aguilera A.M., Escabias, M. & Valderrama M.J. (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data, *Computational Statistics & Data Analysis*, 50, 1905-1924
- Barker M., Rayens W. (2003) Partial least squares for discrimination. *J Chemometrics* 17:166–173
- Costanzo D. , Preda C. , Saporta G. (2006). Anticipated prediction in discriminant analysis on functional data for binary response . In *COMPSTAT2006*, p. 821-828, Physica-Verlag
- Hennig, C., (2000). Identifiability of models for clusterwise linear regression. *J. Classification* 17, 273–296.
- Preda C. , Saporta G. (2005a): PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, 48, 149-158.
- Preda C. & Saporta G. (2005b) Clusterwise PLS regression on a stochastic process . *Computational Statistics & Data Analysis*, 49(1): 99-108, 2005.
- Preda C., Saporta G. & Lévédér C., (2007) PLS classification of functional data, *Computational Statistics*, 22(2), 223-235
- Saporta G., Preda C. Adaptive Forecasting on Functional Data (2008). In *SIS, Univ. Calabria*, 25-27 june
- Ramsay & Silverman (1997) *Functional data analysis*, Springer