Classification supervisée et données fonctionnelles

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

Plan

1. Introduction
2. Régression MCO sur données fonctionnelles
3. Régression PLS fonctionnelle
4. Méthodes linéaires de discrimination
5. Prédiction anticipée
6. Conclusion et perspectives

Travaux réalisés en collaboration avec C.Preda(Univ. Lille2) et D.Costanzo (Univ.Calabria)

1. Introduction

- Données fonctionnelles: courbes ou trajectoires d'un processus stochastique X_t
- Réponse Y
  - Y numérique: régression
  - Y catégorielle: classification supervisée, discrimination
- Intervalle de temps commun $[0;T]$, variables centrées

- Précurseurs:
  - R.A. Fisher – 1924
  - J. C. Deville – 1974
  - P. Besse – 1979
  - G. Saporta – 1981
- Plus récemment:
  - Aguilera, Valderrama – 1993, 1995, 1998
  - Ramsay, Silverman – 1995, 1997
  - Van der Heijden – 1997
  - Preda, Cohen – 1999
  - Cardot, Ferraty, Vieu - 1999, 2005

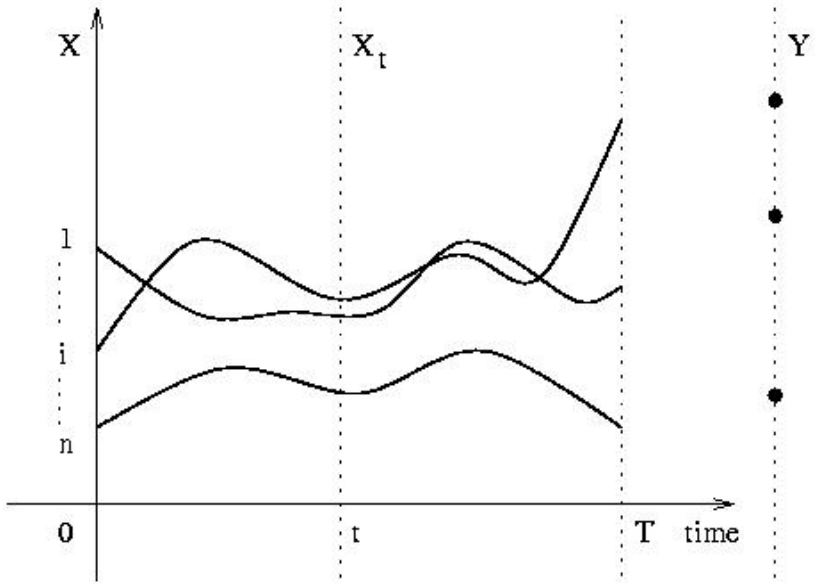
Régression sur données fonctionnelles

- Exemple 1: Y= récolte

  X_t = température

  p= ∞

R.A.Fisher (1924)



- Données de très grande dimension: infinité non dénombrable (en principe..) de prédicteurs
- Combinaison linéaire
  - « Integral regression »

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

    - Au lieu d'une somme finie

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

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

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

where r_1, r_2, ..., r_q are the quantities of rain in the several intervals of time, and a_1, ... 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

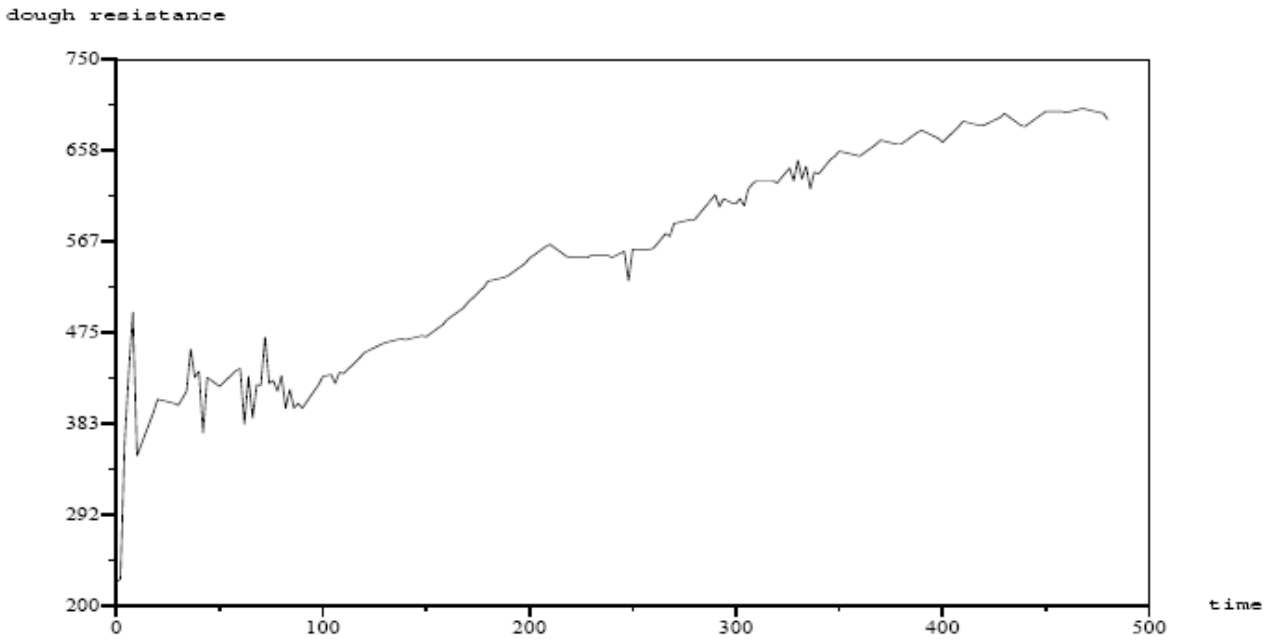
$$\bar{w} = c + \int_0^T ar\,dt, \qquad \text{(III)}$$

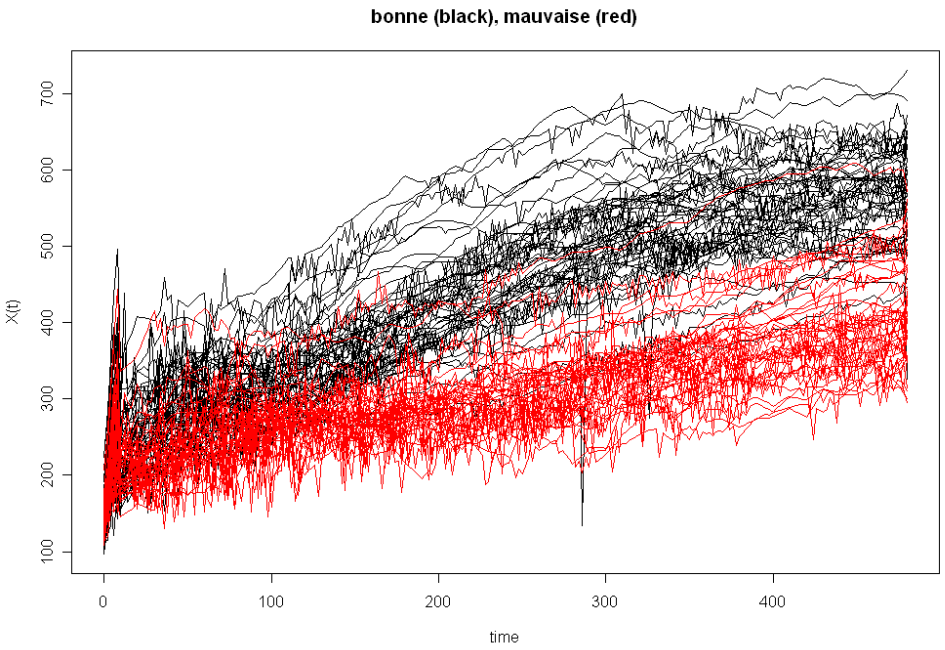
where $r\,dt$ is the rain falling in the element of time dt; the integral is taken over the whole period concerned, and a is a *continuous* function of the time t, which it is our object to evaluate from the statistical data.

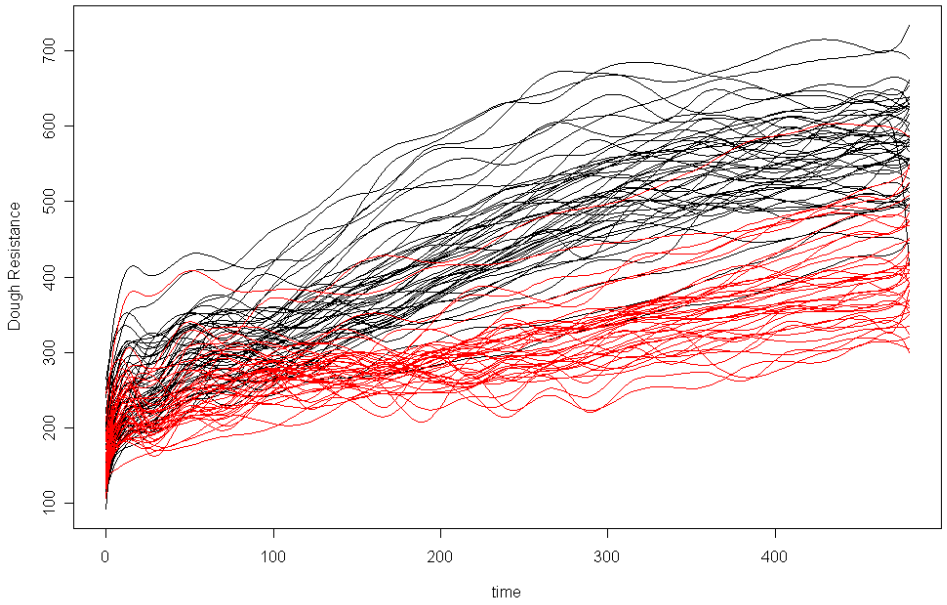
R.A.Fisher « The Influence of Rainfall on the Yield of Wheat at Rothamsted »
Philosophical Transactions of the Royal Society, B, 213, 89-142 (1924)

•Discrimination sur données fonctionnelles

- Exemple 2: courbes de pétrissage pour biscuits (Danone Vitapole)







- Après lissage par B-splines cubiques (Lévéder & al, 2004)

Comment prédire la qualité des biscuits?

- Discrimination sur données fonctionnelles
  - Cas particulier de la régression sur données fonctionnelles pour deux classes
- Anticipation
  - déterminer t*<T tel que l'analyse sur [0;t*] donne des prédictions semblables à l'analyse sur [0;T]

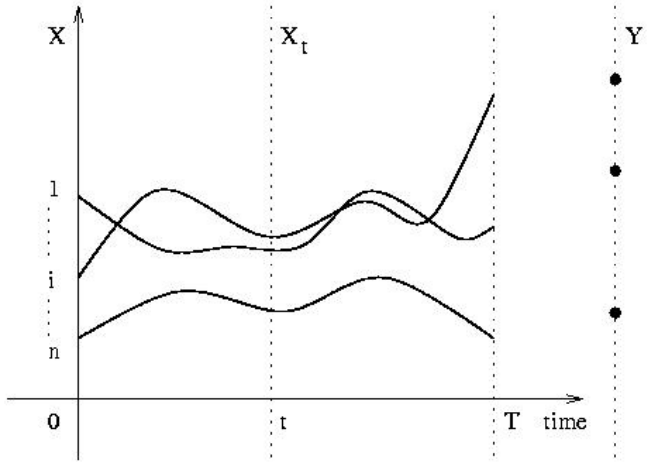
2. Régression MCO sur données fonctionnelles

Y ; X_t ($E(Y)=E(X_t)=0$)

- **2.1 Les mco**
  - Equations normales ou de Wiener-Hopf:

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

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



  - C(t,s)= cov(X_t, X_s)=E($X_t X_s$)

2.2 décomposition de Karhunen-Loeve

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

- facteurs: $\int_0^T C(t,s) f_i(s) ds = \lambda_i f_i(t)$
- Composantes principales: $\xi_i = \int_0^T f_i(t) X_t dt$
- Covariance avec une composante principale:

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

Résolution numérique:

- Equations intégrales non explicites dans le cas général: C(t,s) connu point par point
- Fonctions en escalier: nombre fini de variables et d'individus: opérateurs matriciels mais de grande taille
- Approximations par discrétisation du temps

Une solution exacte:

- **W** matrice des produits scalaires entre trajectoires $w_{uv} = \int_0^T x_u(t)\, x_v(t) dt \quad u, v = 1, 2, .., n$
- Composantes principales: vecteur propres de **W**
- Facteurs principaux

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

- Theorème de Picard: β unique si et seulement si:

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

- Géneralement faux ... Surtout quand n est fini car p >n. Ajustement parfait en minimisant:

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

- Même quand β est unique, « L'équation de Wiener-Hopf n'est pas une équation intégrale ordinaire mais un accouplement entre fonction et distribution dont la solution est plus souvent une distribution qu'une fonction » Paul Kree, 1972

- Nécessité de contraintes. (cf Green & Silverman 1994, Ramsay & Silverman 1997).

2.3 Régression sur composantes principales

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

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

- Approximation de rang q:

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

- Quelles composantes?
  - Les q premières?
  - Les q plus corrélées?

- Les composantes principales sont calculées sans tenir compte de la réponse Y

3. Régression PLS fonctionnelle

- Utiliser les composantes PLS au lieu des composantes principales
- Première composante PLS :

$$\max_w \operatorname{cov}^2\left(Y, \int_0^T w(t) X_t dt\right) \qquad \|w\|^2 = 1$$

$$w(t) = \frac{\operatorname{cov}(X_t, Y)}{\sqrt{\int_0^T \operatorname{cov}^2(X_t, Y) dt}} \qquad t_1 = \int_0^T w(t) X_t dt$$

- Puis itération sur les résidus

- Approximation de Y par X_t d'ordre q:

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

- Convergence :

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

- Mais q doit être fini pour avoir une formule!
- q déterminé par validation croisée
  (Preda & Saporta, 2005)

- Première composante PLS facilement interprétable: coefficients du même signe que $r(y;x_t)$
- Pas d'équation intégrale
- Meilleur ajustement par PLS que par ACP:

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

(De Jong 1993)

4. Discrimination linéaire

4.1 ADL fonctionnelle

- ADL : combinaison linéaire $\int_0^T \beta(t) X_t dt$ maximisant le rapport

  variance inter/variance intra

- Pour 2 groupes la FLD de Fisher s'obtient en régressant Y codé sur X_t

  eg $\sqrt{\frac{p_1}{p_0}}$ and $-\sqrt{\frac{p_0}{p_1}}$

(Preda & Saporta, 2005a)

- La régression PLS avec q composantes donne une approximation de β(t) et du score:

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

- Pour plus de 2 groupes: régression PLS2 entre k-1 indicatrices de Y et X_t

Régression PLS2

- Y multiple: (Y_1, Y_2, …, Y_p)
- Critère de Tucker:

$$\max \ \mathrm{cov}^2\left(\int_0^{\infty} w(t) X_t dt; \sum_{i=1}^{p} c_i Y_i\right)$$

- Composantes PLS :

$$t = \int_0^{\infty} w(t) X_t dt \qquad s = \sum_{i=1}^{p} c_i Y_i$$

Première composante PLS: premier vecteur propre du produit des opérateurs d'Escoufier $\mathbf{W}^X\mathbf{W}^Y$

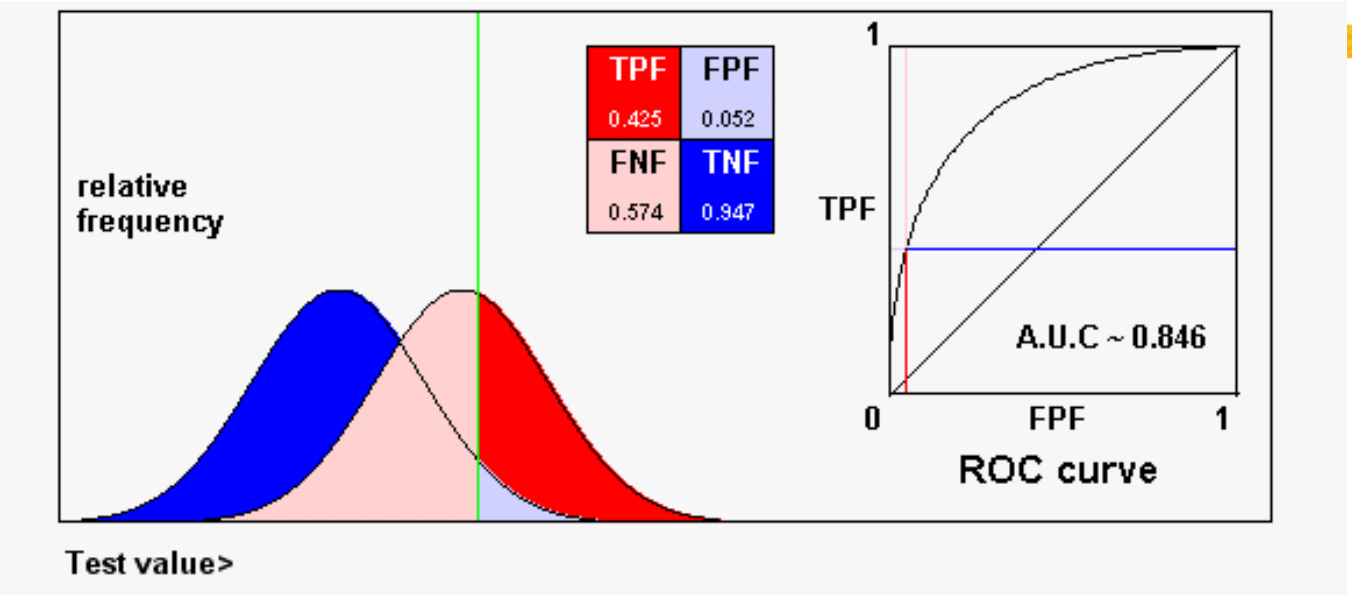
$$\mathbf{W}^X Z = \int_0^T E(X_t Z) X_t \, \mathrm{d}t, \quad \mathbf{W}^Y Z = \sum_{i=1}^{p} E(Y_i Z) Y_i, \quad \forall Z \in L_2(\Omega).$$

Preda & Saporta, 2002 & 2005a ; Barker & Rayens , 2003

4.2 Mesures de qualité

- Pour k=2 : courbe ROC et AUC
  - Pour un seuil s , x est classé en 1 si $d_T(x)>s$
  - Sensibilité ou taux de vrais positifs: $P(d_T(x)>s/Y=1)=1-\beta$
  - 1- Spécificité ou 1-taux de vrais négatifs: $P(d_T(x)>s/Y=0)=\alpha$

Courbe ROC



- En cas de discrimination parfaite :
  courbe confondue avec les côtés du carré
- Si distribution conditionnelles identiques, courbe confondue avec la diagonale

- Courbe ROC invariante pour toute transformation monotone croissante
- Surface sous la courbe: mesure de performance permettant de comparer (partiellement) des modèles
- $$AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s) = P(X_1 > X_2)$$

On tire une obs de G_1 et une de G_2

- AUC estimée par la proportion de paires concordantes

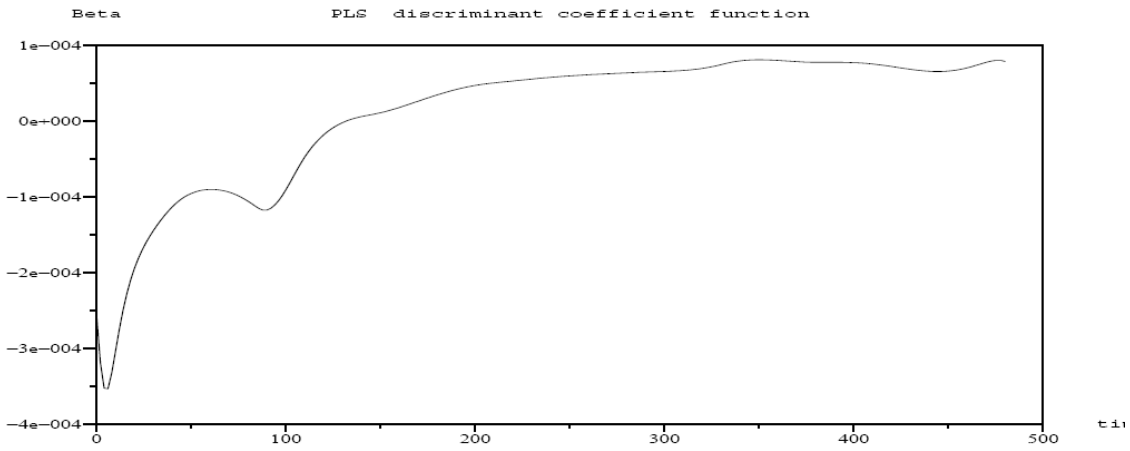
$$c = n_c / n_1 n_2$$

- n_c statistique de Wilcoxon-Mann-Whitney

$U+W= n_1n_2+0.5n_1(n_1+1)$ $\quad AUC=U/n_1n_2$

courbes de pétrissage

- Après T= 480s de pétrissage, on obtient des biscuits de qualités Y
- 115 observations: 50 « bon », 40 «mauvais » and 25 «indéterminés »
- 241 mesures à pas constant
- Lissage avec B-splines cubiques , 16 nœuds



- **Performance pour Y={bon,mauvais}**
  - On divise 100 fois les données en apprentissage et test (60, 30)
    - Taux d'erreur moyen
      - 0.142 avec 3 composantes principales
      - 0.112 avec 3 composantes PLS
    - AUC moyen= 0.746

4.3 Régression logistique fonctionnelle

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

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

Hypothèse: β(t) et les trajectoires sont dans le même espace de dimension fini (Ramsay et al., 1997)

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

D'où une régression logistique classique:

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

avec $\mathbf{C} = (c_{iq})$ $\Phi = (\phi_{kq} = \int_T \psi_k(t)\psi_q(t)\mathrm{d}t)$

Aguilera *et al.* (2006) utilisent les composantes principales de X_t comme base

5. Prédiction anticipée

- Chercher t*<T tel que l'analyse sur [0;t*]donne des prédictions semblables à l'analyse sur [0;T]
- Solution:
  - En augmentant s depuis 0 , chercher la première valeur telle que AUC(s) ne diffère pas significativement de AUC(T)

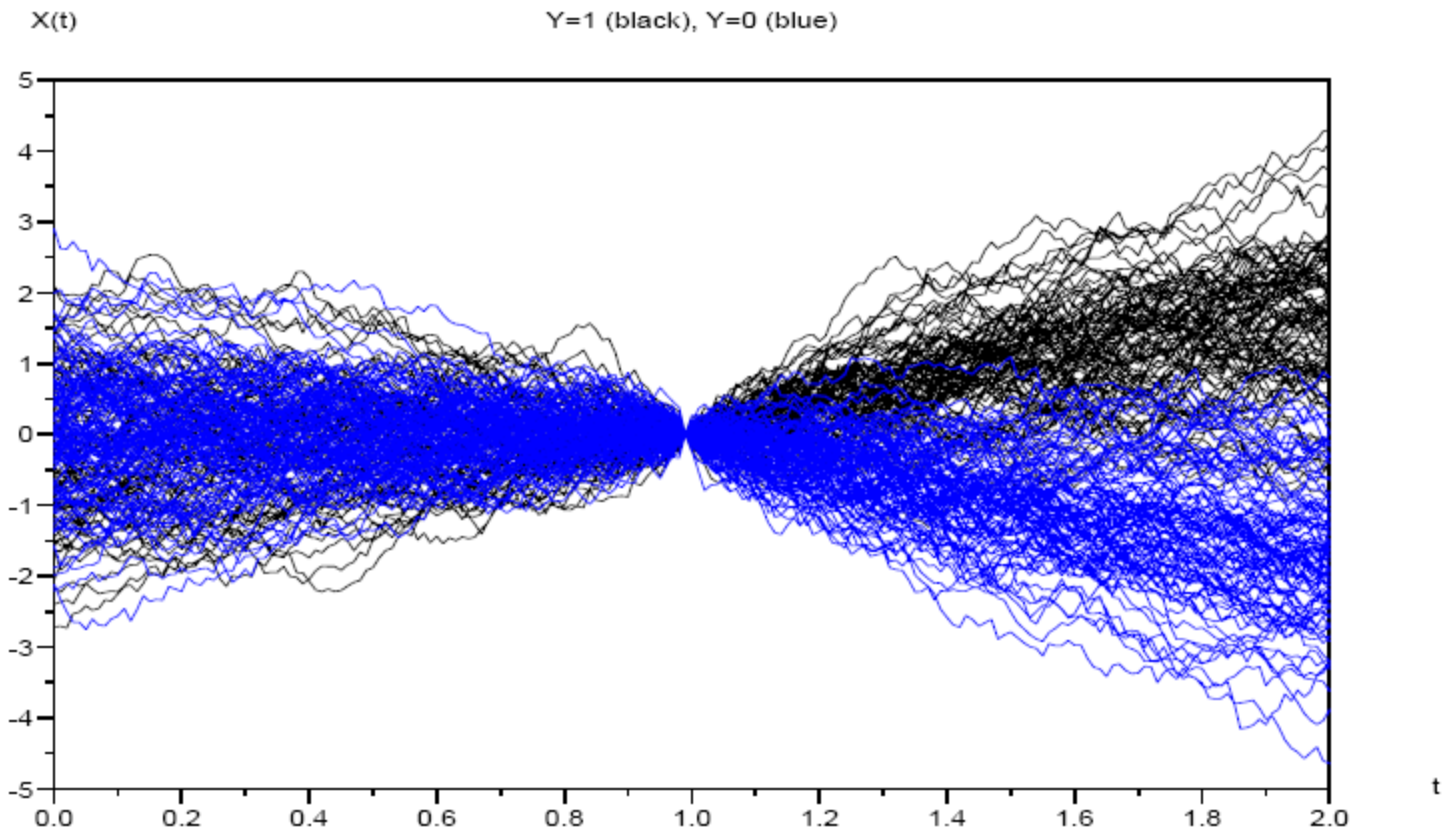
- Test d'égalité via une procédure bootstrap
  - Rééchantillonnage des données, stratifié pour conserver les proportions des classes
  - A chaque réplication b on calcule $AUC_b(s)$ et $AUC_b(T)$
  - Test basé sur les différences (Student ou Wilcoxon pour données appariées)
    $\delta_b = AUC_b(s) - AUC_b(T)$

5.2 Données simulées

- Deux classes équiprobables
- W(t) brownien standard

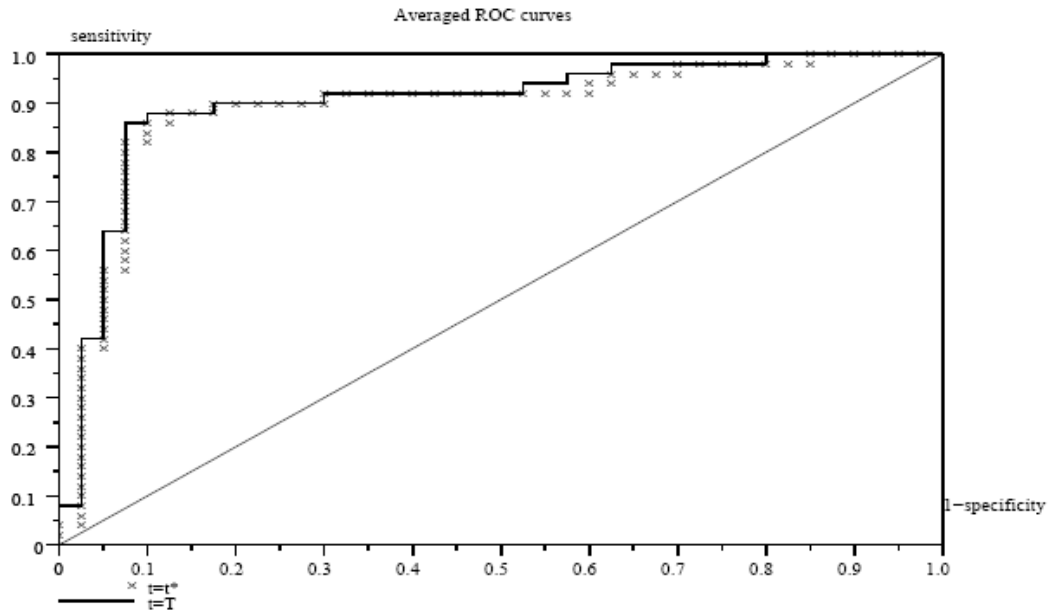
$$\text{Class } \{Y=0\} : X_t = \begin{cases} W(1-t), & 0 \le t \le 1 \\ -2\sin(t-1) + W(t-1), & 1 < t \le 2 \end{cases}$$

$$\text{Class } \{Y=1\} : X_t = \begin{cases} W(1-t), & 0 \le t \le 1 \\ 2\sin(t-1) + W(t-1), & 1 < t \le 2 \end{cases}$$



Sample of size $n = 100$ for each class of Y.

- Avec B=50



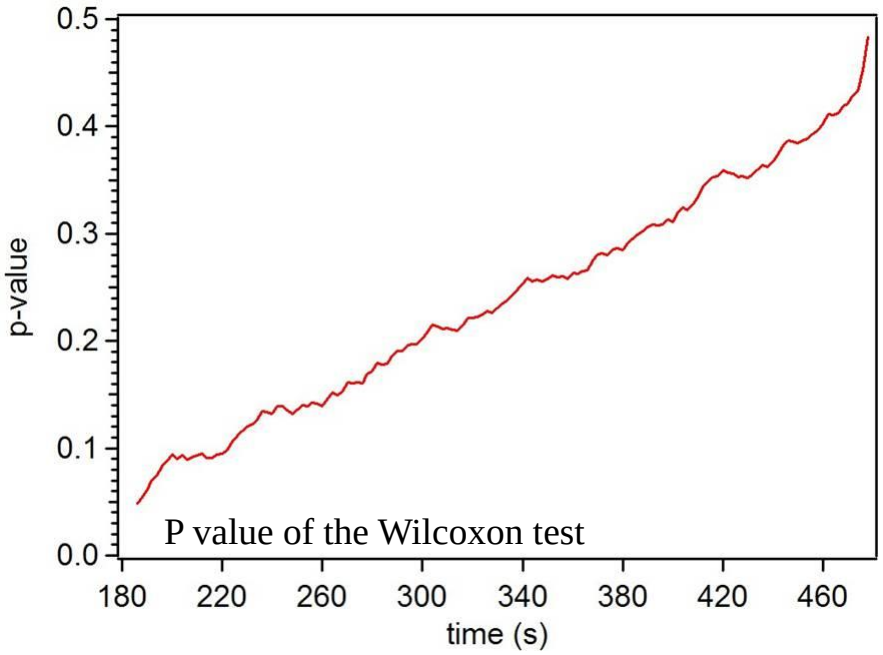
For the Wilcoxon test (one-tailed) with the first error type fixed to 0.05, the minimum t^* for which the test is not significant is $t^* = 1.46$. The Wilcoxon statistic is 1.582 and the two averaged AUC corresponding to $t^* = 1.46$ and $T = 2$ are $\overline{AUC}(t^*) = 0.866$ and $\overline{AUC}(T) = 0.872$.

5.3 Courbes de pétrissage

Prédiction anticipée

- B=50
- t*=186

- Il est donc possible de réduire de plus de moitié la durée d'étude!



5.4 Prévision adaptative

- Au lieu d'un t* commun, adapter t* à chaque trajectoire nouvelle ω , connaissant le début de la trajectoire. Pour certaines trajectoires il pourra être nécessaire d'observer le processus plus longtemps que sur [0, t*] , pour d'autres non.
- t* devient une v.a. t*(ω)

- Procédure proche dans son esprit des tests séquentiels:
  - On discrétise [0, T] avec un pas h
  - Si à t, on arrête d'observer $X(\omega)$ et que l'on prend une décision de classement alors t*=t , sinon on continue jusqu'à t+h etc.
- La décision dépend de la similarité de $X(\omega)$ avec des observations x_i en tenant compte de la prédiction que l'on peut faire de Y

« Taux de conservation »

- d_t score discriminant calculé sur [0,t]
- $\Omega_\omega(t)$ ensemble des observations prédites comme ω au temps t.
- $p_{0|\Omega_\omega(t)}$ proportion classée dans le groupe Y=0 au temps T . Idem pour

$$p_{1|\Omega_\omega(t)} \quad p_{0|\bar{\Omega}_\omega(t)} \quad p_{1|\bar{\Omega}_\omega(t)}$$

- On a $p_{0|\Omega_\omega(t)} + p_{1|\Omega_\omega(t)} = 1$ et $p_{0|\bar{\Omega}_\omega(t)} + p_{1|\bar{\Omega}_\omega(t)} = 1$

- Deux taux de conservation:

$$C_{\Omega_\omega(t)} = \max\{p_{0|\Omega_\omega(t)}, p_{1|\Omega_\omega(t)}\}$$
$$C_{\overline{\Omega}_\omega(t)} = \max\{p_{0|\overline{\Omega_\omega}(t)}, p_{1|\overline{\Omega}_\omega(t)}\}$$

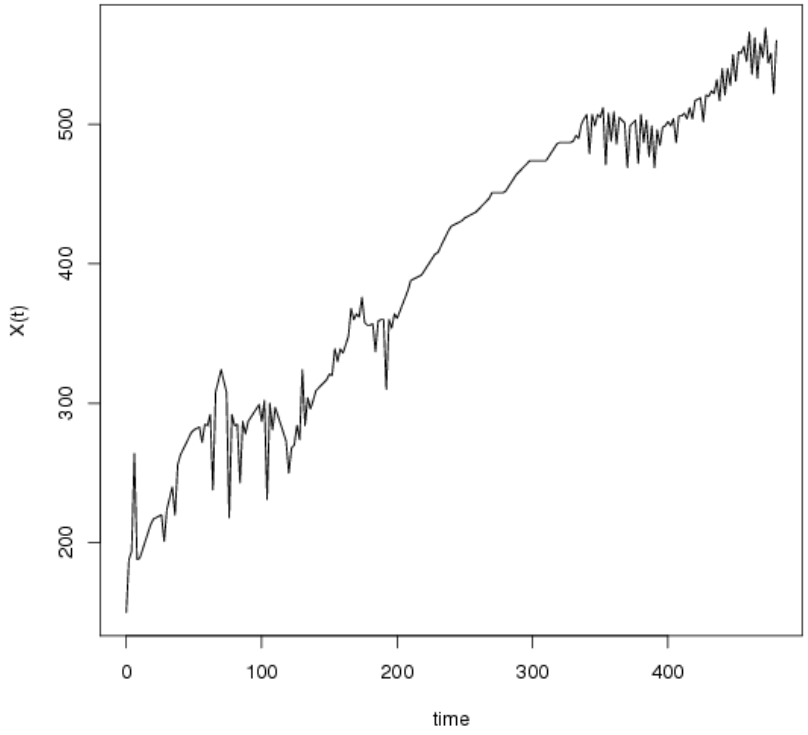
- Taux global de conservation $\min\left(C_{\Omega_\omega(t)}; C_{\bar{\Omega}_\omega(t)}\right)$
- Pour tout t $0.5 \leq C_\Omega(\omega, t) \leq 1$
- $C_\Omega(\omega, T) = 1.$

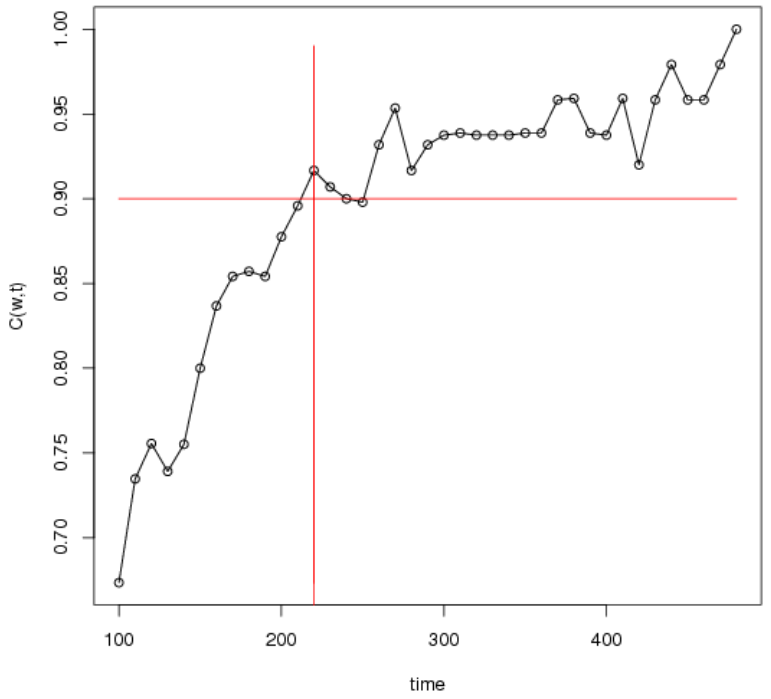
Règle de décision

- En prenant un niveau de confiance γ, ici 0.9, on définit la règle adaptative pour ω à l'instant t:
  - (1) Si $C_\Omega(\omega,t) > \gamma$ alors l'observation de ω sur [0, t] suffit pour prédire $Y(\omega)$ car la prédiction à t est la même que la prédiction à T du sous-groupe $\Omega_\omega(t)$
  - (2) Si $C_\Omega(\omega,t) < \gamma$, on continue d'observer jusqu'à t+h et on recommence jusqu'à satisfaction de (1)

Application

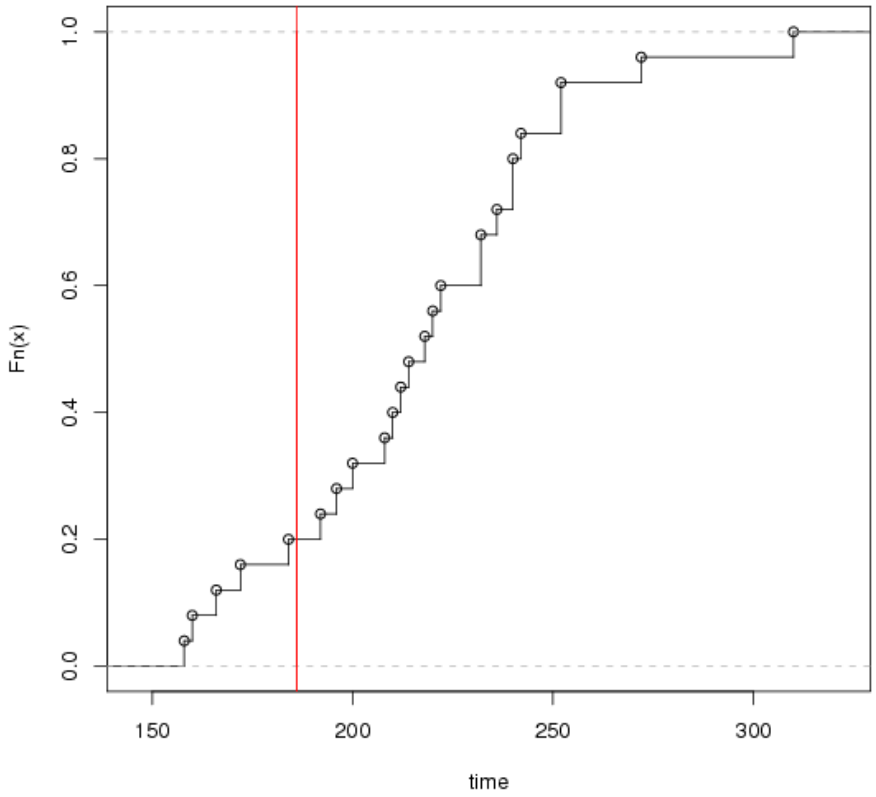
une nouvelle farine





taux de conservation

25 « nouvelles » farines



- Distribution cumulée empirique de t*(ω) . 5 points plus en avance que t= 186, temps optimal pour la prédiction anticipée. 10 prédites comme «mauvaise».

6.Conclusions et perspectives

- La régression PLS permet d'effectuer une prédiction linéaire de manière simple et efficace
- Nécessité de prétraitements pour données bruitées
- Prédiction anticipée via une procédure bootstrap, possibilité de prédiction « on-line »: adapter t* pour chaque nouvelle courbe
- En cours:
  - Comparaison avec régression logistique PLS fonctionnelle et autres approches

Références

- 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. of Chemometrics* 17:166–173
- Dabo-Niang S., Ferraty F. (2008): *Functional and Operatorial Statistics*, Springer-Verlag
- 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
- Lévéder C., Abraham C., Cornillon P. A., Matzner-Lober E., Molinari N. (2004) Discrimination de courbes de pétrissage. *Chimiometrie 2004*, 37–43.
- Preda C. , Saporta G. (2005) PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, 48, 149-158.
- Preda C., Saporta G., Lévéder C., (2007) PLS classification of functional data, *Computational Statistics,* 22(2), 223-235
- Ramsay J.O. , Silverman (2005) *Functional data analysis*, 2nd edition, Springer