Régression PLS sur un processus stochastique

Cristian Preda

CERIM - Département de Statistique
Faculté de Médecine
Université de Lille 2
1, Place de Verdun
59045 Lille Cedex
e-mail : cpreda@univ-lille2.fr

Gilbert Saporta

CNAM - Paris
Chaire de Statistique Appliquée
292, Rue Saint Martin
75141 Paris Cedex 03
e-mail : saporta@cnam.fr

Résumé

Après avoir montré le principe de la régression PLS dans le cas fini, nous allons développer ensuite la régression PLS sur un processus $(X_t)_{t\in[0,T]}$ L_2-continu. On montre l'équivalence avec la régression PLS sur les composantes principales de $(X_t)_{t\in[0,T]}$ ainsi que des propriétés de convergence des approximations données par cette méthode. Les résultats d'une application sur des données boursières seront comparés avec ceux fournis par d'autres méthodes.

Mots clés : régression PLS, opérateur d'Escoufier, analyse en composantes principales.

Abstract

We give an extension of PLS regression to the case where the set of predictor variables forms a L_2-continuous stochastic process and the response is a random vector of finite or infinite dimension. We prove the existence of PLS components as eigenvectors of some operator and also some convergence properties of the PLS approximation. The results of an application to stock-exchange data will be compared with those obtained by others methods.

Key words : PLS regression, Escoufier's operator, principal component analysis.

1 Introduction

Il ne semble pas usuel d'effectuer une régression linéaire sur une infinité de variables explicatives. Cela correspond pourtant à la situation suivante souvent rencontrée (Figure 1) : on observe n courbes (ou trajectoires) en continu sur l'intervalle de temps [0,T] – que nous allons considérer comme réalisations d'un processus stochastique $(X_t)_{t\in[0,T]}$ – et on veut utiliser cette information pour prédire une réponse Y qui peut être X_{T+h} – on parle alors de prédiction à l'horizon h, $h > 0$ – ou une variable aléatoire réelle externe quelconque (par exemple, $(X_t)_{t\in[0,T]}$ peut représenter courbes de températures observées en n lieux et Y le montant de récoltes). Théoriquement, cela s'exprime par la régression de la variable Y sur le processus $(X_t)_{t\in[0,T]}$.

Le but de cet article est d'adapter la régression PLS lorsque l'ensemble de variables explicatives est un processus du second ordre et L_2-continu, $(X_t)_{t\in[0,T]}$, $T \in \mathbf{R}_+$.

Les problèmes posés par la régression linéaire classique sur un processus – l'indétermination des coefficients de régression (Ramsay [10], [11], Saporta [12]) ou encore le choix des composantes principales de $(X_t)_{t\in[0,T]}$ comme variables explicatives (Deville [4], Saporta [12], Aguilera [1]) – trouvent dans ce cadre des solutions satisfaisantes dont les principales propriétés découlent de celles de l'opérateur d'Escoufier associé au processus $(X_t)_{t\in[0,T]}$ (Saporta [12]).

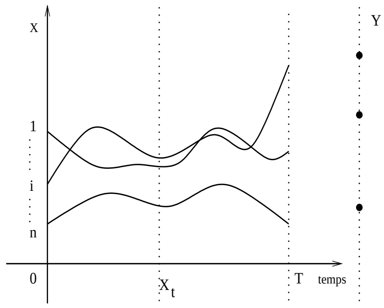

Figure 1: Régression sur un processus

Dans cette note nous introduisons la régression PLS d'un vecteur aléatoire $\mathbf{Y}$ sur un processus stochastique $(X_t)_{t\in[0,T]}$ L_2-conitinu à valeurs dans $\mathbf{R}$. On montre l'existence des composantes PLS ainsi que quelques propriétés de convergence vers la régression linéaire classique. Le cas $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$, offre une alternative aux méthodes de prévision proposées par Aguilera ([1]) et Deville ([4]). Les résultats d'une application sur des données boursières sont comparées avec ceux fournis par d'autres méthodes.

2 Principaux résultats

Soient $(X_t)_{t\in[0,T]}$, $X_t : \mathbf{\Omega} \to \mathbf{R}$, $\forall t \in [0,T]$, un processus stochastique du second ordre, L_2-continu et $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, $p > 1$, un vecteur aléatoire réel défini sur le même espace de probabilité. Supposons que le processus est centré, $E(X_t) = 0, \forall t \in [0,T]$ et $E(Y_i) = 0$, $\forall i = 1, ..., p$.

On définit les opérateurs suivants :

$\mathbf{C}_{YX} : L_2([0,T]) \to \mathbf{R}^p$,

$$f \xrightarrow{\mathbf{C}_{YX}} x, \quad x_i = \int_0^T E(X_t Y_i) f(t) dt, \quad \forall i = 1, ..., p,$$

$\mathbf{C}_{XY} : \mathbf{R}^p \to L_2([0,T])$,

$$x \xrightarrow{\mathbf{C}_{XY}} f, \quad f(t) = \sum_{i=1}^{p} E(X_t Y_i) x_i, \quad \forall t \in [0,T]$$

Notons par $\mathbf{U}_X = \mathbf{C}_{XY} \circ \mathbf{C}_{YX}$ et $\mathbf{U}_Y = \mathbf{C}_{YX} \circ \mathbf{C}_{XY}$.

Soient également $\mathbf{W}^X$, respectivement $\mathbf{W}^Y$, les opérateurs d'Escoufier de $L_2(\mathbf{\Omega})$ associés aux vecteurs $\mathbf{X} = (X_t)_{t\in[0,T]}$, respectivement $\mathbf{Y} = (Y_i)_{i=1,\ldots,p}$ et définis par :

$$\begin{aligned} \mathbf{W}^X Z &= \int_0^T E(X_t Z) X_t dt, && \forall Z \in L_2(\Omega), \\ \mathbf{W}^Y Z &= \sum_{i=1}^{p} E(Y_i Z) Y_i, && \forall Z \in L_2(\Omega). \end{aligned}$$

La composante PLS recherchée par le critère de Tucker est donnée par la proposition suivante :

Proposition 1 *Soit $w \in L_2([0,T])$ et $c \in \mathbf{R}^p$. Alors,*

$$\max_{\substack{w,c \\ \|w\|=1 \\ \|c\|=1}} \quad Cov^2\left(\int_0^T X_t w(t)dt, \sum_{i=1}^{p} c_i Y_i\right)$$

est réalisé pour w, respectivement c, les vecteurs propres correspondants aux plus grandes valeurs propres des opérateurs $\mathbf{U}_X$, respectivement $\mathbf{U}_Y$.

Soit $w_1 \in L_2([0,T])$ la fonction propre correspondante à la plus grande valeur propre de l'opérateur $\mathbf{U}_X$. On définit la première composante PLS de la régression du vecteur $\mathbf{Y}$ sur le processus $(X_t)_{t\in[0,T]}$ par la variable aléatoire :

$$t_1 = \int_0^T X_t w_1(t)dt \tag{1}$$

Théorème 2 *Soient $\mathbf{W}^X$, respectivement $\mathbf{W}^Y$, les opérateurs d'Escoufier associés aux vecteurs $\mathbf{X}$ =$(X_t)_{t\in[0,T]}$, respectivement $\mathbf{Y}$. Alors t_1 est vecteur propre de l'opérateur $\mathbf{W}^X \circ \mathbf{W}^Y$ correspondant à la plus grande valeur propre.*

Soit $X_{0,t} = X_t$, $\forall t \in [0,T]$ et $Y_{0,i} = Y_i$, $\forall i = 1,...,p$. Au pas h de la régression PLS de $\mathbf{Y}$ sur $(X_t)_{t\in[0,T]}$, $h \geq 1$, on calcule la composante t_h comme étant le vecteur propre associé à la plus grande valeur propre de l'opérateur $\mathbf{W}_{h-1}^X \mathbf{W}_{h-1}^Y$,

$$\mathbf{W}_{h-1}^X \mathbf{W}_{h-1}^Y t_h = \lambda_h t_h, \tag{2}$$

où $\mathbf{W}_{h-1}^X$, respectivement $\mathbf{W}_{h-1}^Y$ sont les opérateurs d'Escoufier associés au vecteurs $\mathbf{X} = (X_{h-1,t})_{t\in[0,T]}$, respectivement $\mathbf{Y}_{h-1} = (Y_{h-1,i})_{i=1,...,p}$. On calcule ensuite les résidus :

$$X_{h,t} = X_{h-1,t} - p_h(t)t_h, \quad t \in [0,T],$$

$$Y_{h,i} = Y_{h-1,i} - c_{h,i}t_h, \qquad i = 1,...,p,$$

où $p_h(t) = \dfrac{E(X_{h-1}t_h)}{E(t_h^2)}$, $\forall t \in [0,T]$ et $c_{h,i} = \dfrac{E(Y_{h-1,i}, t_h)}{E(t_h^2)}$, $\forall i = 1,...,p$.

Pour tout $h \geq 1$, t_h est combinaison linéaire des résidus de la régression de $X_{h-1,t}$ sur t_{h-1}. On a les propriétés suivantes :

Proposition 3 $\forall h \geq 1$ *:*

a) $\{t_h\}_{h\geq 1}$ *forment un système orthogonal dans* $L_2(X)$,

b) $Y_i = c_{1,i}t_1 + c_{2,i}t_2 + \ldots c_{h,i}t_h + Y_{h,i}, \quad i = 1,...,p,$

c) $X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots p_h(t)t_h + X_{h,t},$

d) $E(Y_{h,i}t_j) = 0, \quad i = 1,...,p, \forall j = 1,...,h,$

e) $E(X_{h,t}t_j) = 0, \quad t \in [0,T], \forall j = 1,...,h,$

L'approximation de $\mathbf{Y}$ donnée par la régression PLS sur $(X_t)_{t\in[0,T]}$ au pas h, $h \geq 1$, est donnée par :

$$\hat{\mathbf{Y}}_h = c_1 t_1 + \ldots + c_h t_h, \quad c_i \in \mathbf{R}^p, i = 1, ..., p. \tag{3}$$

Si $\hat{\mathbf{Y}}$ est l'approximation de $\mathbf{Y}$ donnée par la régression linéaire classique sur $(X_t)_{t\in[0,T]}$, on a la convergence en moyenne quadratique de $\{\hat{\mathbf{Y}}_h\}_{h\geq 1}$ vers $\hat{\mathbf{Y}}$:

Proposition 4

$$\lim_{h\to\infty} E(\hat{\mathbf{Y}}_h - \hat{\mathbf{Y}})^2 = 0 \tag{4}$$

Remarque (Le cas continu) Les résultats précédents restent valables dans le cas où $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$. On obtient alors les formules de décomposition suivantes :

$$X_t = \begin{cases} t_1 p_1(t) + \ldots t_h p_h(t) + X_{h,t}, & \forall t \in [0,T], \\ \\ t_1 c_1(t) + \ldots t_h c_h(t) + X_{h,t}, & \forall t \in [T, T+a], \end{cases} \tag{5}$$

où $p_h(t) = \dfrac{E(X_{h-1,t} t_h)}{E(t_h^2)}$, $\forall t \in [0,T]$ et $c_h(t) = \dfrac{E(X_{h-1,t} t_h)}{E(t_h^2)}$, $\forall t \in [T, T+a]$.

Pour tout $s \in [0,a]$, la prévision de X_{T+s} à l'aide du passé, $(X_t)_{t\in[0,T]}$, est donc donnée par

$$\hat{X}_{T+s} = t_1 c_1(T+s) + \ldots + t_h c_h(T+s). \tag{6}$$

Les propriétés relatives à la convergence de la régression PLS vers la régression linéaire restent valables et dans ce cas.

3 Application sur des données boursières

La régression PLS sur un processus présentée dans les sections précédentes sera utilisée pour prédire le comportement des actions boursières sur une certaine période de temps. De telles données constituent un bon exemple de réalisation d'un processus stochastique à temps continu pour lequel les hypothèses d'existence des moments de second ordre et de continuité en moyenne quadratique sont tout à fait raisonnables.

Nous disposons d'un ensemble de 84 actions cotées à la Bourse de Paris pour lesquelles on connaît complètement le comportement de l'indice de croissance[1] sur un intervalle d'une heure (entre 10^{00}h – l'heure d'ouverture – et 11^{00}h). On connaît également l'évolution de l'indice de croissance d'une nouvelle action (notée 85) sur l'intervalle 10^{00}h - 10^{55}h. Le but est de prédire le comportement de cette action sur l'intervalle de 5 minutes entre 10^{55}h - 11^{00}h utilisant un modèle PLS construit à l'aide des 84 actions dont l'évolution est entièrement connue sur l'intervalle 10^{00}h - 11^{00}h.

Une action est susceptible de changer toutes les secondes : nous allons donc considérer les actions comme étant des réalisations indépendantes d'un processus stochastique $\{X_t : t \in [0, 3600]\}$ (l'intervalle de temps est exprimé ici en secondes). Avec les notations introduites dans la section précédente, il s'agit de la régression PLS de $\{X_t : t \in [T, T+a]\}$ sur $\{X_t : t \in [0,T]\}$ avec $T = 3300$ et $a = 300$.

[1]Au moment t, l'indice de croissance d'une action ω est défini par $X_t(\omega) = \dfrac{v(t) - v(0)}{v(0)}$, où $v(t)$ est la valeur de la cotation de l'action ω à l'instant t et $v(0)$ sa valeur de l'ouverture.

4 Conclusions

Nous avons développé dans cette article la régression PLS sur un processus L_2-continu. Le point clé de cette étude est l'exploitation des propriétés de l'opérateur d'Escoufier associé au processus.

La régression PLS sur un processus offre une alternative à la régression sur les composantes principales Elle donne une solution aux problèmes liés à la corrélation des prédicteurs et au cas où le nombre d'observations est inférieur au nombre de variables explicatives, comme il arrive souvent dans ce contexte.

Bibliographie

[1] Aguilera A.M., Ocaña F., Valderrama M.J. (1998) *An approximated principal component prediction model for continous-time stochastic process*, Applied Stochastic Models and Data Analysis, Vol. 13, p. 61-72.

[2] Cazes P. (1997)*Adaptation de la reǵression PLS au cas de la régression après Analyse des Correspondances Multiples*, Revue de Statistique Appliquée, XLIV (4), p. 35-60.

[3] Deville J.C. (1974) *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE, No. 15, p 3-101.

[4] Deville J. C. (1978) *Analyse et prévision des séries chronologiques multiples non stationnaires*, Statistique et Analyse des Données, No. 3, p. 19-29.

[5] Escoufier Y. (1970) *Echantillonage dans une population de variables aléatoires réelles*, Publications de l'Institut de Statistique de l'Université de Paris, 19, Fasc. 4, p. 1-47.

[6] Green P.J., Silverman B. W. (1994) *Nonparametric Regression and generalized linear models. A roughness penalty approach*, Monographs on statistic and applied probability, No. 58, Chapman & Hall.

[7] L. Lebart, A. Morineau, M. Piron (1995) *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

[8] Palm R., Iemma A.F. (1995) *Quelques alternatives à la régression classique dans le cas de colinéarité*, Rev. Statistique Appliquée XLIII (2), p. 5-33.

[9] Preda C. (1999) *Analyse factorielle d'un processus : problèmes d'approximation et de régression* , Thèse de doctorat de l'Université de Lille 1.

[10] Ramsay J.O., Dalzell C.J. (1991) *Some tools for functional data analysis*, Journal of Royal Statistical Society (B), 53, No. 3, p. 539-572.

[11] Ramsay J.O., Silverman B.W. (1997) *Functional Data Analysis*, Springer Series in Statistics, Springer-Verlag, New York.

[12] Saporta G. (1981) *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O., No. 37-38, Université Pierre et Marie Curie, Paris.

[13] Tenenhaus M., Gauchi J.P., Ménardo C. (1995) *Régression PLS et applications*, Revue de Statistique Appliquée, XLIII (1), p. 7-63.

[14] Tenenhaus M. (1998) *La régression PLS. Théorie et pratique*, Editions Technip, Paris.