# PLS regression on a stochastic process

Cristian Preda[1] and Gilbert Saporta[2]

[1] CERIM - Département de Statistique,
Faculté de Médecine, Université de Lille 2,
1, Place de Verdun, 59045 Lille Cedex, France
(e-mail : cpreda@univ-lille2.fr)
[2] Chaire de Statistique Appliquée
CNAM - Paris
292, Rue Saint Martin, 75141 Paris Cedex 03, France
(e-mail : saporta@cnam.fr)

**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.
**Keywords.** PLS regression, Escoufier's operator, principal component analysis.

## 1 Introduction

It doesn't seem usual to make a linear regression on an infinity of explicative variables. Still that matches the much encountered following situation (Figure 1) : we notice n curves (or trajectories) in continuity on the time lapse $[0, T]$ – which we shall consider as sample paths of a stochastic process $(X_t)_{t\in[0,T]}$ – and we want to use this information to predict a $Y$ answer which can be $X_{T+h}$ or any real extern random variable (for example, $(X_t)_{t\in[0,T]}$ can represent temperature curves observed in $n$ places and $Y$ the amount of crops).

Theoretically, this can be expressed by the regression of the $Y$ variable on the $(X_t)_{t\in[0,T]}$ process.

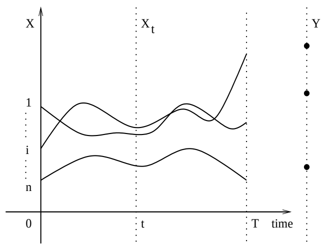


**Fig. 1.** Regression on a stochastic process

The aim of this article is to adapt the PLS regression when all the explicative variables form a stochastic process. The problems brought about by the classical linear regression on a process – the indetermination of the regression coefficients (Ramsay and Dalzell (1991), Ramsay and Silverman (1997), Saporta (1981)) or the choice of the principal components of $(X_t)_{t\in[0,T]}$ as explicative variables (Deville (1978), Saporta (1981), Aguilera et al. (1998))

– get within this framework satisfactory solutions the main characteristics of which derive from those of the Escoufier operator associated with the process $(X_t)_{t\in[0,T]}$ (Saporta (1981)).

We introduce the PLS regression of a random vector $\mathbf{Y}$ on a $L_2$-continuous stochastic process $(X_t)_{t\in[0,T]}$ with values in $\mathbf{R}$. It is an extension of the finite case (a finite set of predictors) developed by Tenenhaus et al. (1995), (1998) and Cazes (1997). We show the existence of PLS components as well as a few convergence properties towards the classical linear regression. The case $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$, presents an alternative to prevision methods proposed by Aguilera et al. (1998) and Deville (1978). The results of an application on stock exchange data are compared with those obtained by other methods.

## 2 Main results

Let $(X_t)_{t\in[0,T]}$ be a random process and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, $p \geq 1$, a random vector defined on the same probability space $(\Omega, \mathcal{A}, P)$. We assume that $(X_t)_{t\in[0,T]}$ and $\mathbf{Y}$ are of second order, $(X_t)_{t\in[0,T]}$ is $L_2$-continuous and for any $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is an element of $L_2([0,T])$. Without loss of generality we assume also that $E(X_t) = 0$, $\forall t \in [0,T]$ and $E(Y_i) = 0$, $\forall i = 1, \ldots, p$.

Under this hypothesis, let $\mathbf{C}_{YX}$ and $\mathbf{C}_{XY}$ be the operators defined as :

$$\mathbf{C}_{YX} : L_2([0,T]) \to \mathbf{R}^p,\ f \xrightarrow{\mathbf{C}_{YX}} x,\quad x_i = \int_0^T E(X_t Y_i) f(t)dt,\quad \forall i = 1, ..., p,$$

$$\mathbf{C}_{XY} : \mathbf{R}^p \to L_2([0,T]),\ x \xrightarrow{\mathbf{C}_{XY}} f,\quad f(t) = \sum_{i=1}^{p} E(X_t Y_i) x_i,\quad \forall t \in [0,T].$$

and denote by $\mathbf{U}_X = \mathbf{C}_{XY} \circ \mathbf{C}_{YX}$ and by $\mathbf{U}_Y = \mathbf{C}_{YX} \circ \mathbf{C}_{XY}$.

Obviously, $\mathbf{U}_X$ and $\mathbf{U}_Y$ are self-adjoint, positive and compact operators with the same spectrum. Therefore, the spectral analyses of $\mathbf{U_X}$ and $\mathbf{U_Y}$ lead to a countable set of positive eigenvalues.

The following proposition justify this analysis and gives the solution to the PLS problem :

**Proposition 1 (Tucker criterion)**

$$\max_{\substack{w, c \\ w \in L_2([0,T]), \|w\| = 1 \\ c \in \mathbf{R}^p, \|c\| = 1}} Cov^2\left(\int_0^T X_t w(t)dt, \sum_{i=1}^{p} c_i Y_i\right)$$

*is reached for $w$, respectively $c$, the eigenvectors associated to the largest eigenvalue of $\mathbf{U}_X$, respectively of $\mathbf{U}_Y$.*

Let $w_1 \in L_2([0,T])$ be the eigenfunction of $\mathbf{U_X}$ associated to the largest eigenvalue. Then, the first PLS component (Tenenhaus et al. (1995)) of the regression of $\mathbf{Y}$ on the process $(X_t)_{t\in[0,T]}$ is the random variable defined as :

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

Denote by $\mathbf{W}^X$, respectively $\mathbf{W}^Y$, the Escoufier's operators[1] (Escoufier (1970)) associated to $(X_t)_{t\in[0,T]}$, respectively to $\mathbf{Y}$, defined by :

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

Our main result is the following theorem.
**Theorem 2** *$t_1$ is the eigenvector of the $\mathbf{W}^X\mathbf{W}^Y$ associated to the largest eigenvalue.*

The PLS regression is an iterative method. Let $X_{0,t} = X_t$, $\forall t \in [0,T]$ and $Y_{0,i} = Y_i$, $\forall i = 1,...,p$. At the step $h$, $h \geq 1$, of the PLS regression of $\mathbf{Y}$ on $(X_t)_{t\in[0,T]}$, we define the $h^{th}$ PLS component, $t_h$, by the eigenvector associated to the largest eigenvalue of the operator $\mathbf{W}_{h-1}^X \mathbf{W}_{h-1}^Y$,

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

where $\mathbf{W}_{h-1}^X$, respectively $\mathbf{W}_{h-1}^Y$, are the Escoufier's operators associated to $(X_{h-1,t})_{t\in[0,T]}$, respectively to $\mathbf{Y}_{h-1} = (Y_{h-1,i})_{i=1,\ldots,p}$ and $\lambda_{\max}$ the largest eigenvalue of $\mathbf{W}_{h-1}^X \mathbf{W}_{h-1}^Y$. Finally, the PLS step is completed by the ordinary linear regression of $X_{h-1,t}$ and $Y_{h-1,i}$ on $t_h$. Let $X_{h,t}$, $t \in [0,T]$ and $Y_{h,i}$, $i = 1,...,p$ be the random variables which represent the error of these regressions :

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

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

As in the finite case (Tenenhaus et al. (1995)), the next statements hold :
**Proposition 3** *For each $h \geq 1$ :*

*a)* $\{t_h\}_{h\geq 1}$ *forms an orthogonal system in* $L_2(X)$,
*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)* $X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_h(t)t_h + X_{h,t}, \quad t \in [0,T],$
*d)* $E(Y_{h,i}t_j) = 0, \quad \forall i = 1,...,p, \forall j = 1,...,h,$
*e)* $E(X_{h,t}t_j) = 0, \quad \forall t \in [0,T], \forall j = 1,...,h.$

From the Proposition 3-b), the PLS approximation of $\mathbf{Y}$ by $(X_t)_{t\in[0,T]}$ at step $h$, $h \geq 1$, is given by :

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

[1] The spectral analysis of this operator leads to the principal component analysis of the associated variable. See Deville (1974) and Escoufier (1970) for details.

Denote by $\hat{\mathbf{Y}}$ the approximation of $\mathbf{Y}$ given by the ordinary linear regression on $(X_t)_{t\in[0,T]}$. Then, the sequence $\{\hat{\mathbf{Y}}_h\}_{h\geq 1}$ is convergent in $L_2(\Omega)$ and the limit is $\hat{\mathbf{Y}}$ :

**Proposition 4**

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

Finally, the choice of $h$ using the cross-validation criterion (Green and Silverman (1994), Tenenhaus (1998)) remains applicable in this case.

**Remark 5**

a) **(The continuous case)** The previous results are still valid for the particular case $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$. Indeed, because of the $L_2$ continuity of the process $(X_t)_{t\in[0,T+a]}$, $\mathbf{C}_{X,Y}$ and $\mathbf{C}_{Y,X}$ are compacts and therefore, $\mathbf{U}_X$ and $\mathbf{U}_Y$ are compacts. The results of the Proposition 1 and Theorem 2 are preserved.
The decomposition formulas (Proposition 3-b,c) become in this case :

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

For each $s \in [0,a]$, the "forecast" of $X_{T+s}$ by $(X_t)_{t\in[0,T]}$ is given by :

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

b) **(Approximation)** Let $\Delta = \{0 = t_0 < t_1 < \ldots < t_p = T\}$, $p \geq 1$, be a discretization of $[0,T]$ and consider the process $(X_t^{\Delta})_{t\in[0,T]}$ defined as (Preda (1999)) :

$$X_t^{\Delta} = \frac{1}{t_{i+1} - t_i}\int_{t_i}^{t_{i+1}} X_t dt, \quad \forall t \in [t_i, t_{i+1}[, \quad \forall i = 0, ..., p-1.$$

Denote by $m_i$ the random variable $\frac{1}{t_{i+1}-t_i}\int_{t_i}^{t_{i+1}} X_t dt$, $i = 0, \ldots, p-1$. Then, the approximation of the PLS regression on $(X_t)_{t\in[0,T]}$ by those on $(X_t^{\Delta})_{t\in[0,T]}$ is equivalent to the PLS regression on the finite set $\{m_i\sqrt{t_{i+1} - t_i}\}$, $i = 0, \ldots, p-1$.

## 3 Application on stock exchange data

The PLS regression on a process presented in the previous sections will be used to predict the behaviour of shares on a certain lapse of time.

We have 84 shares quoted at the Paris stock exchange, for which we know the whole behavior of the growth index on a gap of one hour (between $10^{00}$ o'clock opening time and $11^{00}$ o'clock) ; a share is likely to change every second. We also know the evolution of the growth index of a new share (noted 85) on the gap $10^{00}$ - $10^{55}$. The aim is to predict the way that share will behave during the five minutes gap (between $10^{55}$ and $11^{00}$) using a PLS model built starting from the 84 shares the evolution of which is thoroughly known in the gap $10^{00}$- $11^{00}$ o'clock.

With the notations introduced in the previous section and considering the time gap expressed in seconds, this is the PLS regression of $(X_t)_{t\in[T,T+a]}$ on $(X_t)_{t\in[0,T]}$, where $T = 3300$ and $a = 300$.

We are going to use the approximation given in Remark 5-b) by taking an equidistant discretization of the gap $[0, 3600]$ in 60 sub gaps. The previsions obtained will then match the average level of the growth index of share 85 considered on each 60 seconds gap $[60 \cdot (i-1), 60 \cdot i)$, $i = 56, \ldots, 60$.

Using the SIMCA-P software (see Tenenhaus (1998) for details) we are going to build several PLS models according to the number of components chosen for the regression. So we are going to refer to the model with $k$ PLS components as PLS(k). The previsions obtained with those models will be compared to those given by the regression on the principal components (models quoted with CP(k)) and the algorithm NIPALS (see Tenenhaus (1998) for details). To rate the quality of those models we are going to compare the previsions obtained by each model, quoted with $\{\hat{m}\}_i(85)$, with true values $\{m\}_i(85)$, $i = 56, \ldots, 60$, observed previously.

Table 1 gives the parts of variance associated to the three first PLS components, respectively to the three first principal components of the set $\{m_i\}_{i=1,\ldots,55}$.

| PLS | % | % cum. |
|---|---|---|
| $t_1$ | 91.7 | 91.7 |
| $t_2$ | 2.9 | 94.6 |
| $t_3$ | 2.7 | 97.3 |

| ACP | % | % cum. |
|---|---|---|
| $\xi_1$ | 91.7 | 91.7 |
| $\xi_2$ | 4.0 | 95.7 |
| $\xi_3$ | 2.5 | 98.2 |

Table 1: Parts of variance

The forecasts of the some models are presented in Table 2.

| | $\hat{m}_{56}(85)$ | $\hat{m}_{57}(85)$ | $\hat{m}_{58}(85)$ | $\hat{m}_{59}(85)$ | $\hat{m}_{60}(85)$ | $SSE = \sum_{i=56}^{60} (\hat{m}_i - m_i)^2$ |
|---|---|---|---|---|---|---|
| **Observed** | **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 |
| CP(1) | -0.356 | -0.365 | -0.368 | -0.355 | -0.331 | 4.026 |
| CP(2) | -0.332 | -0.333 | -0.335 | -0.332 | -0.298 | 3.786 |
| CP(3) | 0.613 | 0.638 | 0.669 | 0.825 | 0.963 | 1.538 |
| NIPALS | 0.222 | 0.209 | 0.240 | 0.293 | 0.338 | 1.000 |

Table 2: Forecasts and errors

Considering $SSE$ as global measure of the quality model, the best forecasts are those of the PLS regression in two steps. The models PLS(3) and

CP(3) are particularly good for a small horizon (3300, 3480), but much less for the interval of time (3480-3600). The NIPALS algorithm provides also, on an average, good predictions.

## 4 Conclusions

The PLS regression on a process offers an alternative to the regression on the principal components. It gives a solution to the problems linked with the correlation of the predictors and in case the number of observations is smaller than the number of explicative variables, which is often the case in this context.

## References

AGUILERA, A.M., OCAÑA F. and VALDERAMA, 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.

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

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

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

ESCOUFIER, Y. (1970): *Echantillonnage dans une population de variables aléatoires réelles*, Publications de l'Institut de Statistique de l'Université de Paris, 19, Fasc. 4, p. 1-47.

GREEN, P.J. and SILVERMAN, B. W. (1994): *Nonparametric Regression and generalized linear models. A roughness penalty approach*, Monographs on statistic and applied probability, No. 58, Chapman & Hall.

LEBART, L., MORINEAU, A. and PIRON, M. (1995): *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

PALM, R. and IEMMA, A.F. (1995): *Quelques alternatives à la régression classique dans le cas de colinéarité*, Rev. Statistique Appliquée XLIII (2), p. 5-33.

PREDA, C. (1999): *Analyse factorielle d'un processus : problèmes d'approximation et de régression* , Thèse de doctorat de l'Université de Lille 1, No. 2648.

RAMSAY, J.O. and DALZELL, C.J. (1991): *Some tools for functional data analysis*, Journal of Royal Statistical Society (B), 53, No. 3, p. 539-572.

RAMSAY, J.O. and SILVERMAN, B.W. (1997): *Functional Data Analysis*, Springer Series in Statistics, Springer-Verlag, New York.

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.

TENENHAUS, M., GAUCHI, J.P. and MENARDO, C. (1995): *Régression PLS et applications*, Revue de Statistique Appliquée, XLIII (1), p. 7-63.

TENENHAUS, M. (1998): *La régression PLS. Théorie et pratique*, Editions Technip, Paris.