

# PLS regression with functional predictor and missing data

Cristian PREDA<sup>1</sup>, Gilbert SAPORTA<sup>2</sup>, Mohamed MBAREK<sup>3</sup>

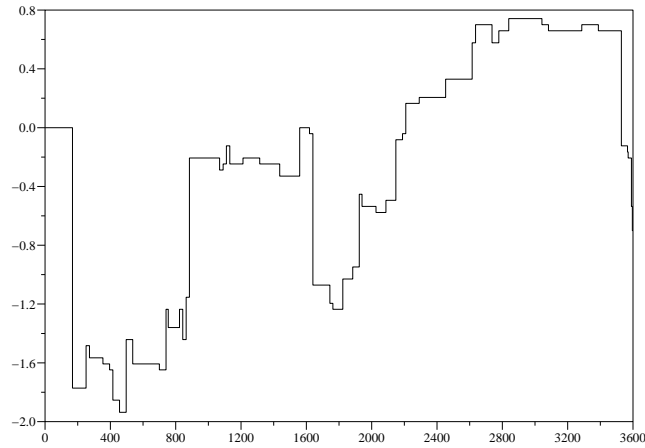
<sup>1</sup> Université des Sciences et Technologies de Lille, France

<sup>2</sup> CNAM Paris, France,

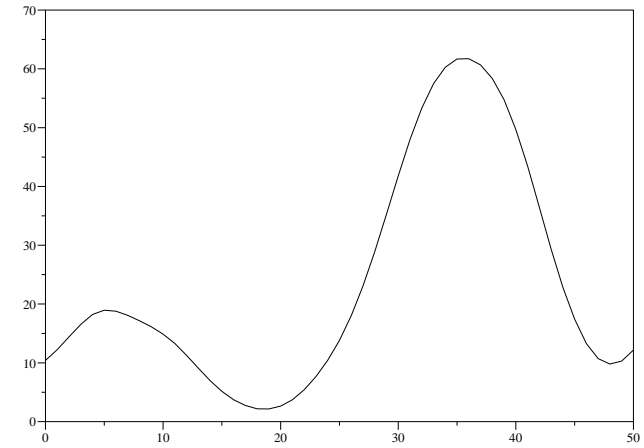
<sup>3</sup> Institut Supérieur de Gestion, Sousse, Tunisia

# Functional data

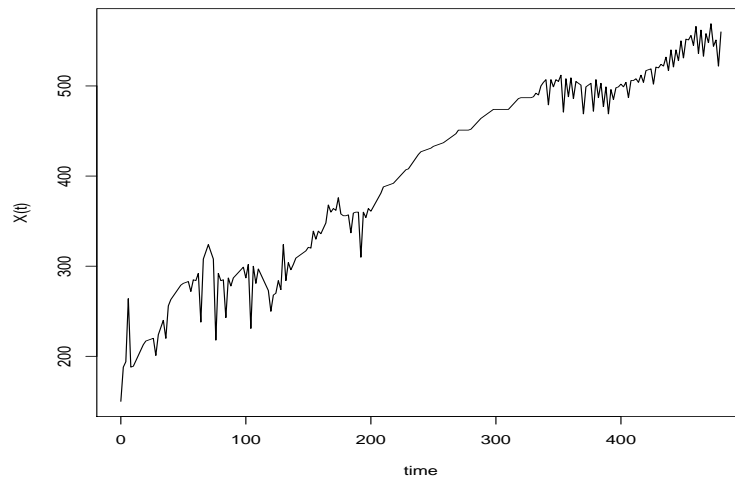
## Examples:



Stock exchange evolution



Gait curve : angular knee



Kneading process: resistance dough

## Model :

$$\mathbf{X} = \{X_t : t \in [0, T]\},$$

$$X_t : \Omega \rightarrow \mathbb{R}$$

## **Models for functional data**

Deville (1974)

Saporta (1981)

Ramsay and Silverman (1997, 2002, 2005)

Ferraty and Vieu (2006)

- Factorial analysis
- Regression models
- Classification
- Clustering

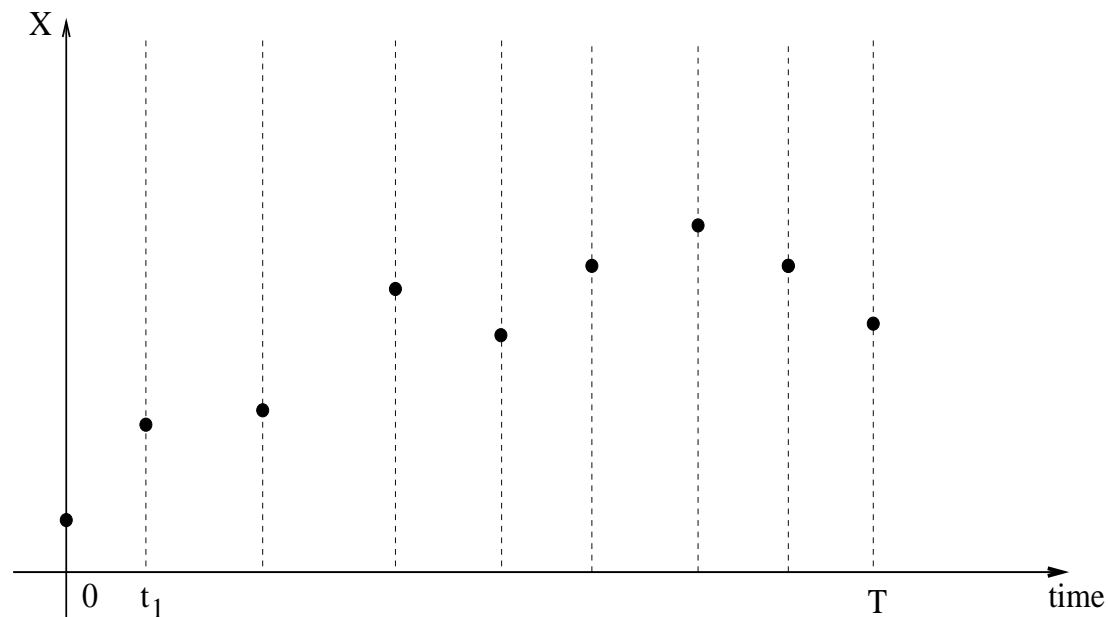
## Missing data for functional predictor

Pleonasm ?

In practice, a curve is observed in a finite number of instants :

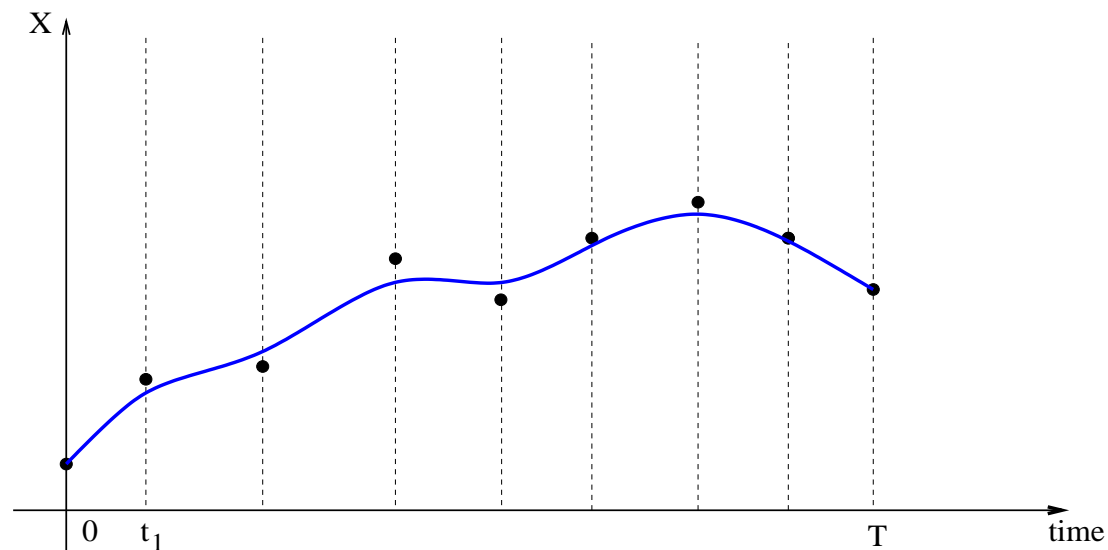
$$0 = t_0 < t_1 < \dots < t_k = T :$$

$$\{(t_0, X_{t_0}), (t_1, X_{t_1}), \dots, (t_k, X_{t_k})\}.$$



## Solution :

- interpolation : linear, spline, ...
- smoothing if data is observed with errors : projection into a finite dimensional function space (B-spline, ...)

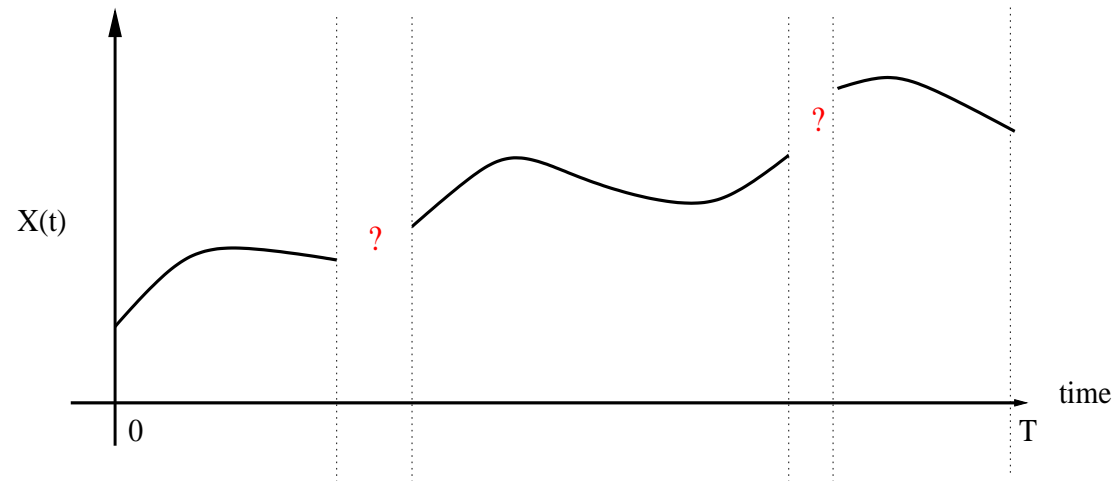


Smoothed curve using cubic B-spline functions

## Missing Completely At Random (MCAR)

Little and Rubin (1987) :

*The occurrence of missing data is independent of the data structure.*

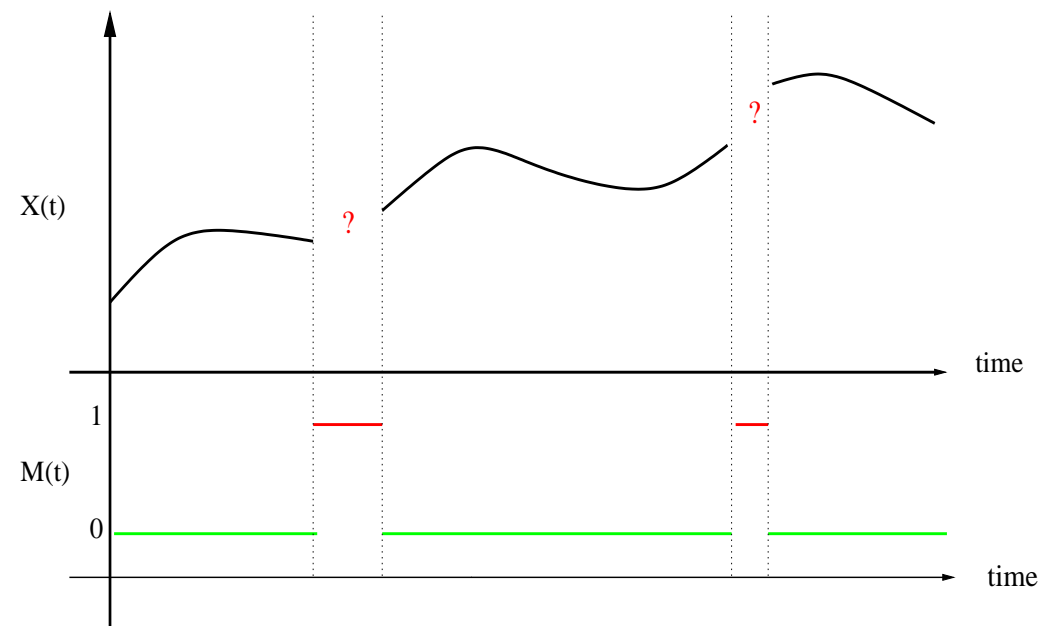


What model for missing data occurrences ?

## The ‘Missing’ process $M(t)$

Two-state process :

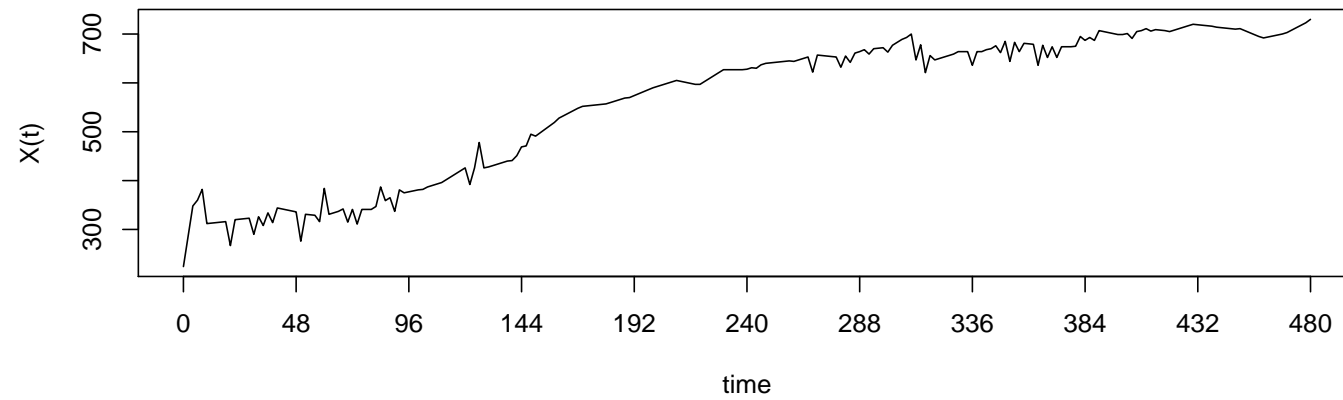
$$M(t) = \begin{cases} 1, & \text{if data is missing at time } t \\ 0, & \text{otherwise.} \end{cases}$$



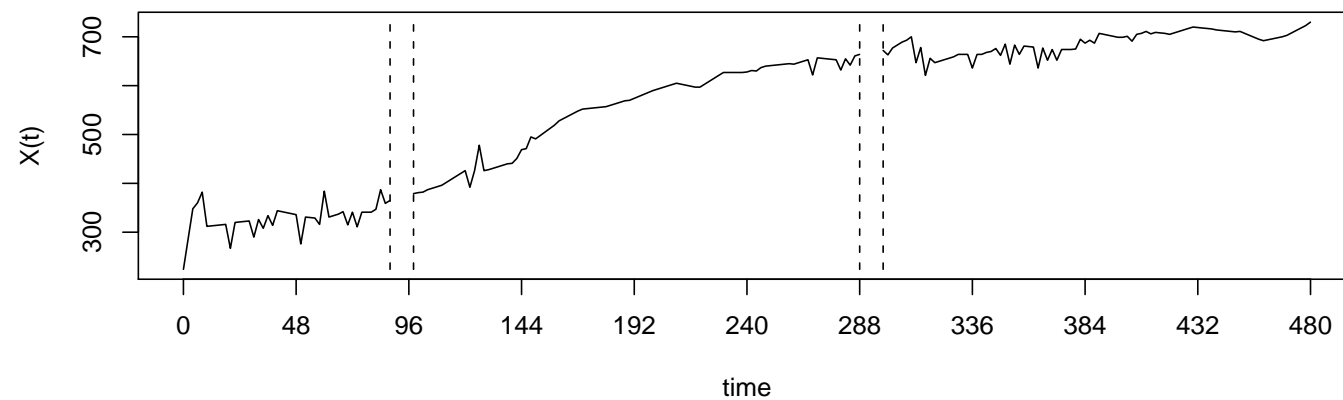
$\text{MCAR} = \{X_t, t \in [0, T]\}$  and  $\{M_t, t \in [0, T]\}$  are independent.

## Example : kneading data

**Complete curve**



**Curve with missing data**



## Measure of missing information :

*Mean Time of Missing Observation* (MTMO) :

$$\text{MTMO} = \frac{1}{T} \int_0^T U(t) dt,$$

where  $U(t) = \mathbb{P}(M(t) = 1)$ .

### Example :

Two state markovian jump process with continuous time

- State 0 : Exp( $\lambda$ )

- State 1 : Exp( $\mu$ )

$$\text{MTMO} = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{(\lambda + \mu)^2 T} \left( 1 - e^{-\frac{\lambda + \mu}{T}} \right)$$

$T = 1, \lambda_1, \mu = 100$  : MTMO = 0.009802.

The process is unobservable about of 1% of time.

# Imputation of missing data.

## Time-average approximation and NIPALS algorithm

Let  $X_1, \dots, X_n$  be  $n$  independent observations (curves) of  $X$ .

**Time-average approximation** (Preda, 2000) :

Consider  $\Delta = \{t_0 < t_1 < \dots < t_K = T\}$  an equidistant discretization of  $[0, T]$  such that

-  $t_i = i \times \delta, \delta = \frac{T}{K}, K \geq 1$ .

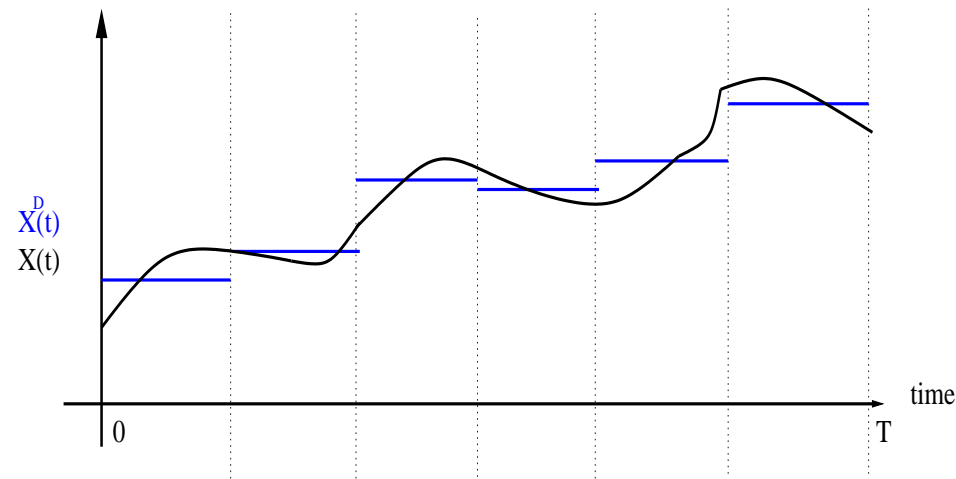
- for all curves, any missing data interval is of the form  $[t_i, t_j]$

Then

$$X_t^\Delta = m_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} X_t dt, i = 1, \dots, K.$$

$$\{X_t, t \in [0, T]\} \approx \{X_t^\Delta, t \in [0, T]\} = \{m_i, i = 1, \dots, K\}.$$

Example :



Time-average approximation

Missing data correspond to missing values for the random variables  $m_i, i = 1, \dots, K$ .

NIPALS algorithm (Tenenhaus, 1998) is used for imputation of missing data.

**Missing data for a curve is approximated by the imputation of the corresponding time-average approximations,  $\hat{m}_i$ 's.**

## PLS regression and missing data

$$Y = f(X) + \varepsilon$$

- $Y$  :
- scalar ( $\mathbb{R}$ )
  - vectorial ( $\mathbb{R}^p$ )
  - fonctional ( $L_2([0, T])$ )
  - categorical

Model :

- parametrical (PCR, PLS)
- non-parametrical (RKHS)

## The linear model

$$f(X) = \langle X, \beta \rangle_{L_2([0, T])} = \int_0^T X_t \beta(t) dt, \quad \beta \in L_2([0, T]).$$

Least squares criterion :

$$\mathbb{E}(Y X_t) = \int_0^T \mathbb{E}(X_t X_s) \beta(s) ds \quad (\text{W-H})$$

Penalized Least Squares criterion :

- principal component regression
- Projection
- Partial Least Squares (PLS)

# The PLS approach

Tucker Criterion :

$$\max_{w, c} \text{Cov}^2 \left( \underbrace{\int_0^T X_s w(s) ds}_{\mathbf{t}}, \sum_{i=1}^p c_i Y_i \right)$$
$$w \in L_2([0, T]), \|w\| = 1$$
$$c \in \mathbf{R}^p, \|c\| = 1$$

Solution (Preda and Saporta (2005)) :

$$\mathbf{W}^X \mathbf{W}^Y \mathbf{t} = \lambda_{\max} \mathbf{t}$$

**Theorem [PLS expansion]** For all  $h \geq 1$  :

a)  $\{t_i\}_{1 \leq i \leq h}$  - orthogonal system in  $L_2(X)$ ,

$$b) \mathbf{Y} = \underbrace{c_1 t_1 + c_2 t_2 + \dots + c_h t_h}_{\hat{\mathbf{Y}}_{PLS(h)}} + \mathbf{Y}_h = \underbrace{\int_0^T X_t \beta_{PLS(h)}(t) dt}_{\hat{\mathbf{Y}}_{PLS(h)}} + \mathbf{Y}_h,$$

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

PLS approximation :

$$\hat{\mathbf{Y}}_{PLS(h)} = \int_0^T X_t \beta_{PLS(h)}(t) dt$$

If  $X_t$  is approximated by  $X_t^\Delta$  then

$$PLS(Y/X^\Delta) = PLS(Y/m_1, \dots, m_K).$$

## Simulation study

$$Y = \int_0^1 \beta(t) X_t dt + \varepsilon,$$

where

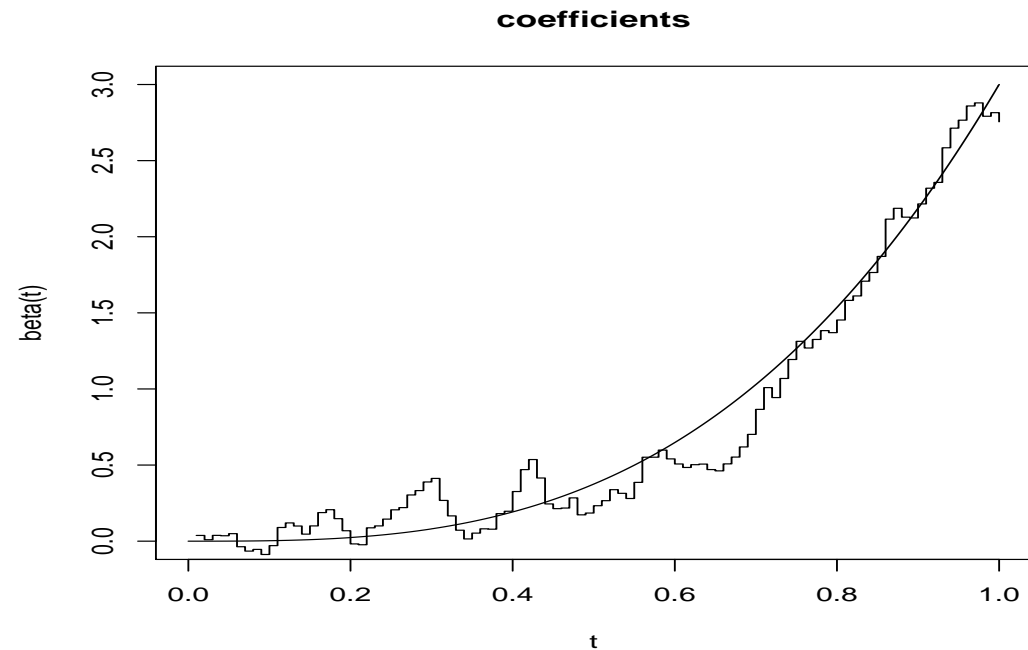
- $\beta(t) = 3t^3$
- $X_t$  : standard Brownian motion on  $[0, 1]$
- $\varepsilon$  :  $\mathbb{E}(\varepsilon) = 0$ ,  $\mathbb{V}(\varepsilon) = 0.1$

$$\mathbb{V}(Y) = 0.5, R^2(X, Y) = 0.8$$

-  $n = 100$  curves observed on  $[0, 1]$  in 1000 equidistant time points.

- Time-average approximation :  $\delta = 1/100$

( $\{m_i : i = 1, \dots, 100\}$ )



Regression coefficient function with complete data ( $R^2 = 0.7645$ )

## The fit with missing data

Exponential r.v's are simulated with a precision of 1/1000.

Parameters	MTMO	R <sup>2</sup>
<i>Complete data</i>	0	<b>0.7645</b>
$\lambda = 1, \mu = 100$	0.00980	0.7263
$\lambda = 1, \mu = 50$	0.01922	0.7288
$\lambda = 1, \mu = 20$	0.04535	0.7144
$\lambda = 2, \mu = 20$	0.08677	0.6625
$\lambda = 2, \mu = 10$	0.15277	0.6218
$\lambda = 2, \mu = 5$	0.24493	0.4872

## Conclusion and Perspectives

- Two state Markovian jump process as model for missing data process in the MCAR case.
- Time-average approximation and NIPALS algorithm for imputation data.
- Influence of the amount of missing data (MTMO) on the quality regression model.
- Other models for the missing data process (dependence models).
- Other methods for imputation missing data.

# References

1. Little R.J.A., Rubin D. B. (1987) *Statistical analysis with missing data*, Wiley.
2. Preda C., Saporta G. (2005) *PLS regression on a stochastic process*, Computational Statistics and Data Analysis, 48, 149-158.
3. Preda C. (2000) *Approximation par moyennage de l'analyse en composantes principales d'un processus stochastique*, Comptes Rendus de l'Académie des Sciences de Paris, T. 330, Série I, 1–6.
4. Ramsay J.O., Silverman B.W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*, Springer.
5. Saporta G. : *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O, Université Pierre et Marie Curie,37-38, Paris, (1981).