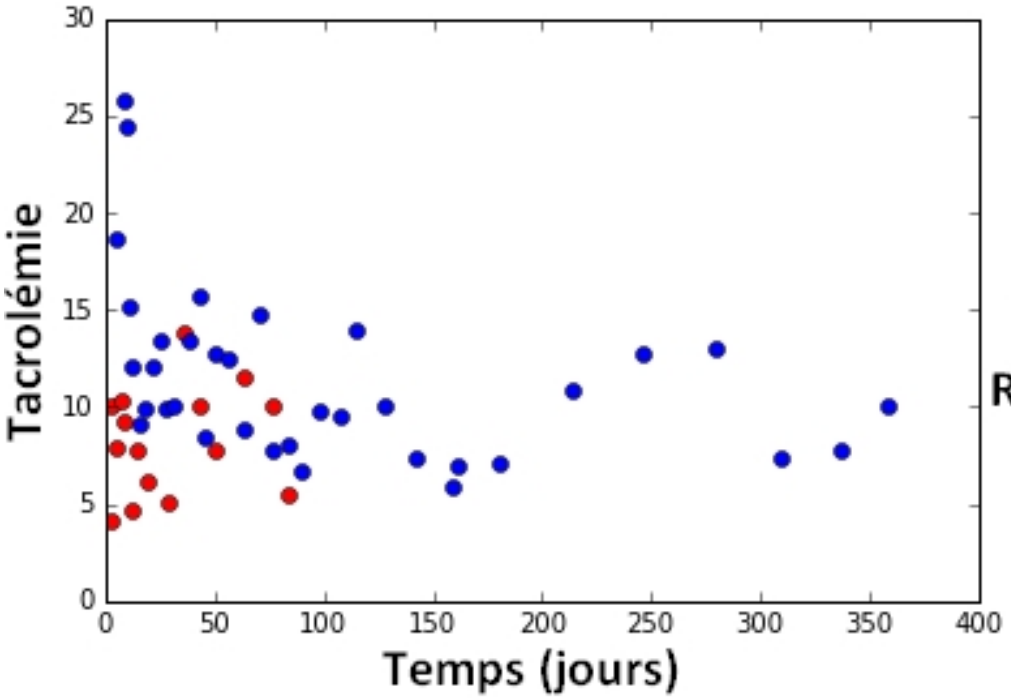
Régression sur données médicales longitudinales, comment aborder la faible densité d'échantillonnage?

Christophe BOTELLA

Journée SFdS Biopharmacie & Santé, 28/11/2016

Données et problème



Taux de créatinine Y :

Y = 234

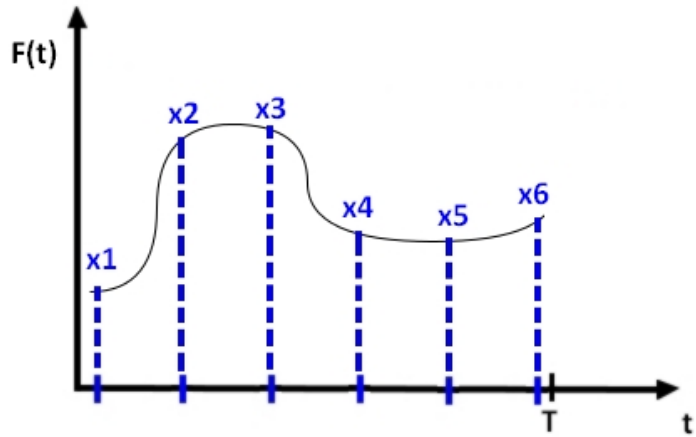
Y = 335

Régression ?

pour tout patient $i \in [|1, N|]$:

- T_i est la durée de mesure (durée d'observation du signal).
- $t^i = (t_1^i, ..., t_{m_i}^i)^T \in \mathbb{R}^{m_i}$ est le vecteur des instants de mesure de la tacrolémie du patient i dans l'intervalle $[0, T_i]$ et m_i le nombre d'instants.
- $X^i = (X_1^i, ..., X_m^i)\mathbb{R}^{m_i}$ est le vecteur des valeurs associées.
- $Z_i \in \mathbb{R}^p$ est le vecteur des variables scalaires du patient.
- Y_i est la variable réelle de sortie, le taux de créatinine.

La régression linéaire fonctionnelle



$x = (x_1, ..., x_6)^T$, $\beta \in \mathbb{R}^6$, Modèle linéaire : $Y = x^T\beta + \epsilon$?

Régression linéaire fonctionnelle (*Ramsay & Silverman*, 2006) :

$$Y = \int_0^T \beta(t)F(t)dt + \epsilon, \quad \beta \in \mathbb{R}^{[0,T]}$$

On modélise $\beta := b^T\phi$, Où $\phi = (\phi_1, ..., \phi_K) \in (\mathbb{R}^{[0,T]})^K$ forme une base fonctionnelle linéairement libre dans $\mathbb{R}^{[0,T]}$.

Extension : Durée de mesure variable

$$Y_i = Z_i^T \alpha + \frac{1}{T_i} \int_0^{T_i} F_i(s)\beta(s, T_i)ds + \epsilon_i$$

F_i obtenue par un prétraitement de la série temporelle (t^i, X^i), en modélisant la courbe par une combinaison de fonctions splines $\Rightarrow$ Paramétrique.

Base de fonctions bidimensionnelles,
$\forall s, T \in \mathbb{R}^2$, $\beta(s, T) = b^T \phi(s, T)$.

Régression sur Processus Gaussien

Mesure à des temps $t = (t_1, ..., t_m)^T \in \mathbb{R}^m$
d'un vecteur aléatoire $X = (X(t_1), ..., X(t_m))^T$.
On peut modéliser X par :

$$X = F(t) + \zeta$$

$\zeta \sim \mathcal{N}(0_m, \gamma^2 I_m)$
$F \sim \mathcal{GP}(0_{\mathbb{R}^{\mathbb{R}}}, k(.,.,\theta)) \Rightarrow F(t) \sim \mathcal{N}(0_m, K)$

Où $K = \begin{pmatrix} k(t_1, t_1, \theta) & \cdots & \\ \vdots & \ddots & \\ k(t_m, t_1, \theta) & \cdots & k(t_m, t_m, \theta) \end{pmatrix}$

Alors

$$X \sim \mathcal{N}(0_m, K + \sigma^2 I_m).$$

Inférence de θ, γ^2 par maximisation de la vraisemblance gaussienne.

Densité prédictive

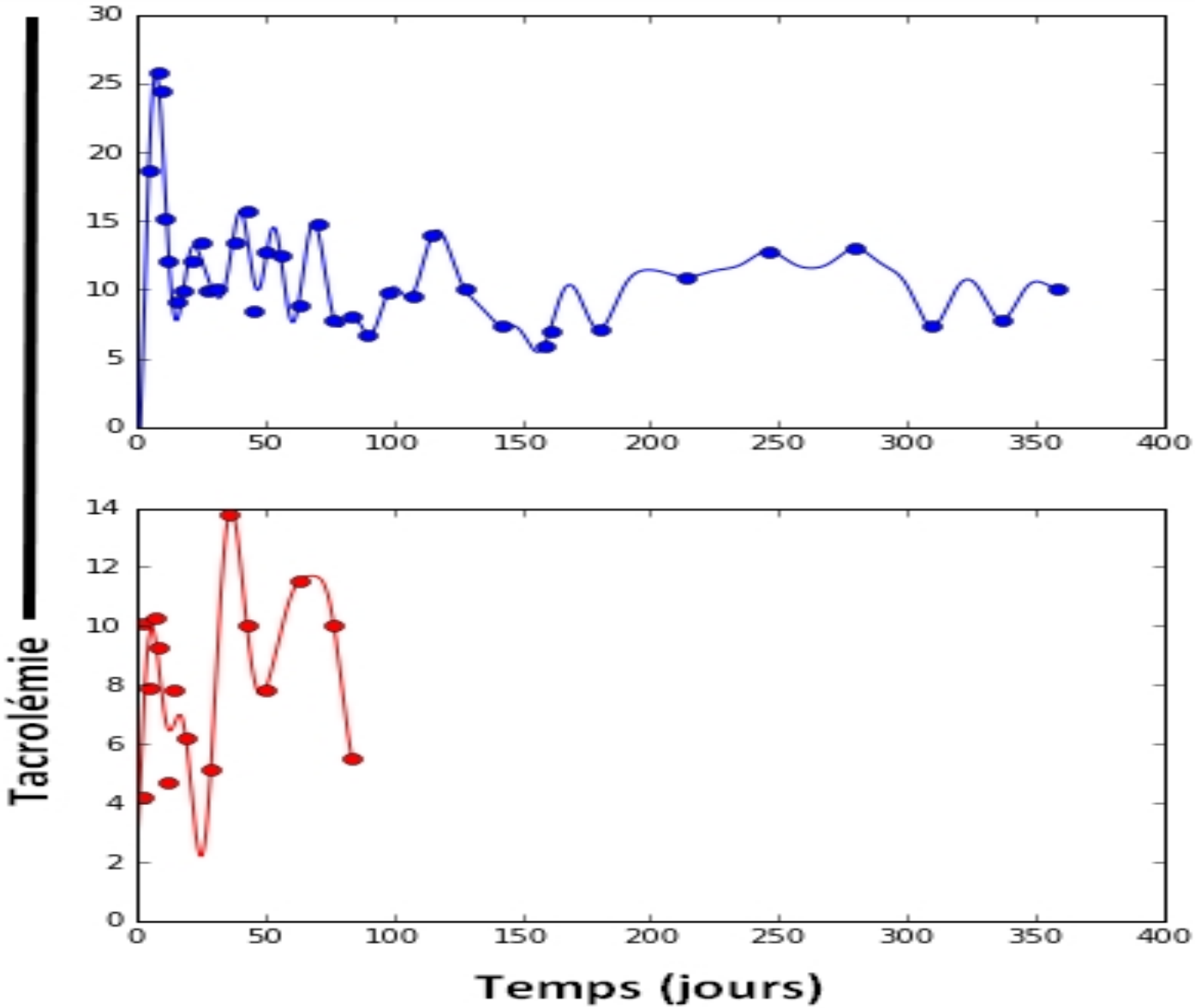
à θ, γ fixés, $\forall t^* \in \mathbb{R}$ la loi de $X(t*)$ conditionnellement à X est normale.

$$\mathbb{E}(X(t*)|X) = K^{*T}(K + \hat{\gamma}^2 I_m)^{-1} y$$

$$\mathbb{V}(X(t*)|X) = k(t^*, t^*) - K^{*T}(K + \hat{\gamma}^2 I_m)^{-1} K^*$$

Où $K* = (k(t_1, t*), ..., k(t_m, t*))^T$.
De plus, l'espérance de la densité prédictive, $s \to \mathbb{E}(X(s)|X)$ est un estimateur consistant en m de la trajectoire de F.

Espérance prédictive, illustration



Modèle proposé

Y label de régression, Z données scalaires, F processus gaussien :

$$Y = Z^T\alpha + \frac{1}{T}\int_0^T F(s)\beta(s,T)ds + \epsilon$$

Propriété

$(\frac{1}{T}\int_0^T F(s)\beta(s,T)ds, X(t_1), ..., X(t_m))^T$ est un vecteur gaussien d'espérance nulle et de matrice de variance-covariance :

$$\begin{pmatrix} \frac{1}{T^2}\int_{[0,T]^2} k(s,t)\beta(s,T)\beta(t,T)dsdt & \frac{1}{T}\int_{[0,T]} K_*^T(s)\beta(s,T)ds \\ \frac{1}{T}\int_{[0,T]} \beta(s,T)K_*(s)ds & K_t + \gamma^2 I_m \end{pmatrix}$$

Méthode des moindres carrés

On évalue une approche plus simple, basée sur le critère des moindres carrés.

$$\hat{Y}_i = \mathbb{E}(Y_i|X^i) = Z_i\alpha + \frac{1}{T_i}\int_0^{T_i} \mathbb{E}(F_i(s)|X^i, \theta)\beta(s, T_i)ds$$

Estimation basée sur la minimisation des écarts quadratiques de prédiction :

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Régularisation

Problème : Sur-ajustement de la fonction paramètre en grande dimension, une pénalisation de la courbure du paramètre permet de le garder sous contrôle :

$$
\begin{aligned}
Pen(\beta) &:= J_{22}(\beta) \\
&= \int_0^{Tmax} \int_0^t \sum_{\nu_1+\nu_2=2} \frac{2}{\nu_1!\,\nu_1!} \left(\frac{\partial^2 \beta(s,t)}{\partial s^{\nu_1} \partial t^{\nu_2}} \right)^2 ds\, dt \\
&= b^T \left(\int\int \frac{\partial^2 \phi}{\partial s^2} \frac{\partial^2 \phi^T}{\partial s^2} + 2 \int\int \frac{\partial^2 \phi}{\partial s \partial t} \frac{\partial^2 \phi^T}{\partial s \partial t} + \int\int \frac{\partial^2 \phi}{\partial t^2} \frac{\partial^2 \phi^T}{\partial t^2} \right) b \\
&= b^T I_p b
\end{aligned}
$$

Estimateur

Pour $\lambda > 0$, critère pénalisé :

$$||Y - \hat{Y}||_2^2 + \lambda b^T I_p b$$

- Fortement convexe sous conditions simples, estimateur unique et calculable.

Estimateur :

$$\hat{\mathcal{P}}^{mcp} = \underset{\alpha \in \mathbb{R}^p, b \in \mathbb{R}^{K_\beta}}{\text{argmin}} ||Y - \hat{Y}||_2^2 + \lambda b^T I_p b = (M^T M + \lambda H)^{-1} M^T Y$$

Où $M := Z \oplus L$, $H := 0_{pp} \oplus I_p$.

Procédé de simulation de données

Deux paramètres, la densité d'échantillonnage $n \in \mathbb{N}*$, et la taille de l'échantillon $N \in \mathbb{N}*$.
$\forall i \in [|1, N|]$,

- On tire $T_i \sim \beta(5, 5)$.
- Si $nT_i - 2 > 0$, on tire $m \sim \mathcal{P}(nT_i - 2)$, sinon on tire $m \sim P(0.1)$. Puis, on prend $m_i = m + 2$.
- On tire indépendamment $t_1^i, ..., t_{m_i}^i \sim U(0, T_i)$.

Puis, $X_j^i = u^i + \sum_{k=0}^{10} v_{1k}^i \, sin\left(2\pi k \, t_j^i\right) + v_{2k}^i \, cos\left(2\pi k \, t_j^i\right)$.

Performance de l'estimation fonctionnelle

$$\text{ISE} \quad : \quad \int_0^{T_m} \int_0^t (\hat{\beta}(s,t) - \beta(s,t))^2 dsdt$$

$$\text{Approximation} \quad : \quad \frac{T_m}{J^2} \sum_{i=1}^{J} \sum_{j=1}^{i} (\hat{\beta}(\frac{jT_m}{J}, \frac{iT_m}{J}) - \beta(\frac{jT_m}{J}, \frac{iT_m}{J}))^2$$

$$\text{MISE} \quad : \quad \frac{1}{n_{rep}} \sum_{i=1}^{n_{rep}} ISE_i$$

Intérêt de la pénalisation : étude en simulation

	Inference	MISE	std.	rMSE
N=100 et	MC	4500	3600	3.3
n=10	MC pénalisé	8.4	0.53	0.85
N=200 et	MC	150.	80	1.6
n=10	MC pénalisé	8.3	0.38	0.78

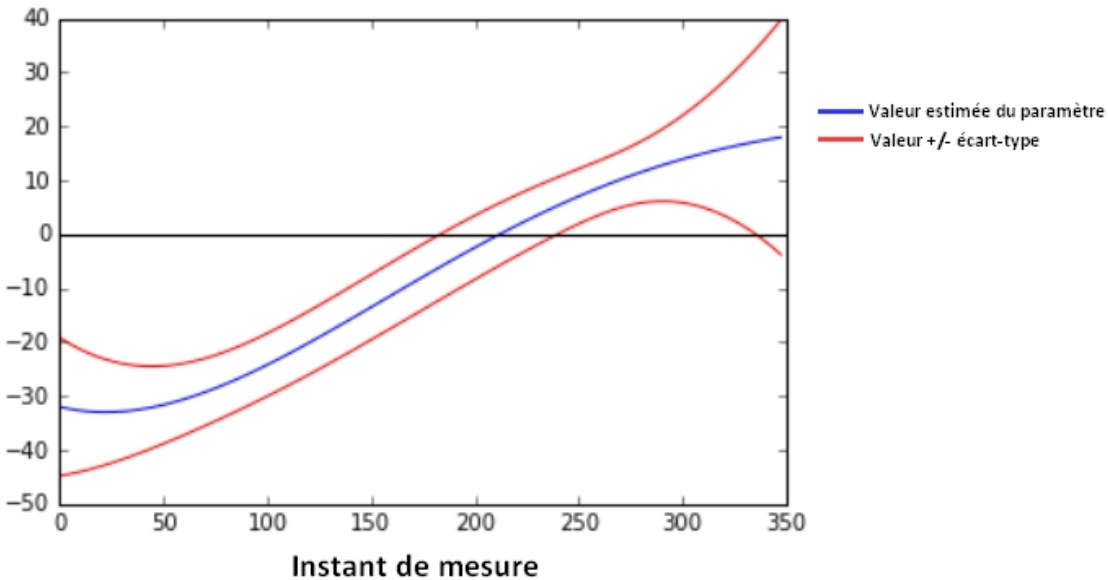
Application aux données ADEQUATE

Données issues d'une étude clinique sur une cohorte de 217 patients ayant subi une opération de transplantation de rein. Modèles évalués :

- Modèle linéaire avec 5 variables explicatives : Âge, Sexe, moyenne de la tacrolémie avant 150 jours, entre 150 et 250, après 250.
- Méthode proposée avec 2 variables scalaires (Âge, Sexe) et une variable fonctionnelle à durée variable (tacrolémie).

Résultats

Modèle	rMSE-ajustement	rMSE-validation
Nul	88.8	89.6
Linéaire	84.1	86.4
Fonctionnel	64.4	66.5



Restriction du paramètre estimé à une durée de 350 jours

Limites et pistes...

- L'optimisation de la méthode du maximum de vraisemblance est à améliorer.
- La base fonctionnelle choisie n'est pas très adaptée.
- Applicable à de nombreuses situations, extension simple à plusieurs variables fonctionnelles, à des labels censurés.

Remerciements

à la SFds et au groupe B&S pour avoir sélectionnée ma présentation,
Aux investigateurs de l'essai clinique ADEQUATE,
à mes encadrants de stage : Agathe Guilloux, Anne-Sophie Jannot, et Simon Bussy.

Bibliographie

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Ramsay, J. O. (2006). Functional data analysis. John Wiley et Sons, Inc..

Gellar, J.E., Colantuoni, E., Needham, D.M., Crainiceanu, C.M. (2014). Variable-domain functional regression for modeling ICU data. Journal of the American Statistical Association, 109(508), 1425-1439.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., Reich, D. (2012). Penalized functional regression. Journal of Computational and Graphical Statistics.

Pimentel, M., Clifton, D., Clifton, L., Tarassenko, L. (2013). Modelling patient time-series data from electronic health records using Gaussian processes. In Advances in Neural Information Processing Systems : Workshop on Machine Learning for Clinical Data Analysis (pp. 1-4).

Gatault, P. et al. (2016). Reduction of extended-release tacrolimus dose in low immunological risk kidney transplant recipients increases risk of rejection and appearance of DSA - a randomized study.