Prévisions pour données fonctionnelles

Kenza Graoui

Juin 2019

1 Introduction

Les progrès techniques en matière de stockage et de traitement de données atteignent de nombreux domaines scientifiques au sein desquels les quantités mesurées prennent la forme de courbes aléatoires.
Les méthodes statistiques multivariées usuelles ne sont plus performantes pour exploiter ce type de données.

La statistique pour données fonctionnelles que nous allons étudier s'avère être une approche adéquate. Elle s'articule autour de ces ensembles de données dans lesquelles les observations ne sont plus identifiées par des vecteurs : elles ne traduisent pas les réalisations de variables aléatoires réelles. En effet, ce sont des courbes qui représentent des fonctions : les données produites sont de dimension infinie.

La situation est telle que l'on dispose de données concernant un même phénomène mesuré quantitativement à différents instants t. Les observations aboutissent à des courbes aléatoires dépendant du temps.

Dès lors, le défi que nous souhaitons relever à travers ce travail de recherche est celui de l'estimation d'une fonction à partir d'une variable explicative fonctionnelle.

C'est pourquoi l'on va d'abord présenter le problème de régression auquel on s'intéresse pour ensuite le formaliser à travers un modèle de régression avec variable réponse réelle et variable explicative fonctionnelle.

2 Le problème de régression[1]

2.1 Enoncé du problème

L'approche fonctionnelle que nous allons étudier s'applique dans le cadre expérimental suivant.

[1] La régression désigne une méthode statistique destinée à analyser la relation d'une variable par rapport à une autre

On observe :

$$(X_i, Y_i)_{i=1,...,n} : [0;1]^2 \to \mathbf{R}$$

avec :

$$t \mapsto X_i(t)$$

$$t \mapsto Y_i(t)$$

des fonctions aléatoires.
-X_i : évolution de la température en fonction du temps sur une année dans une ville i.
-Y_i : log des précipitations sur une année dans une ville i

On dispose donc de deux courbes aléatoires:
- la courbe des températures relevées en un point donné, à différents instants.
- la courbe des cumuls mensuels de précipitations en un point donné.

2.2 Enjeu du problème

On suppose :

$$Y_i(t) = \int_{[0,1]} \beta(s,t) X_i(s) ds + \epsilon_i(t)$$

avec β inconnu et ϵ_i bruit

$$\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$$

Le but est de "prédire" la variable aléatoire réelle dite "réponse" notée Y. Pour cela, il faut analyser la relation entre X et Y puisque Y est exprimée en fonction de X. L'enjeu du problème réside donc dans l'estimation de β.

3 Modélisation mathématique

3.1 Le modèle discret

3.1.1 Présentation du modèle

Cette première approche consiste à faire abstraction du caractère fonctionnel des variables X_i. Cela consiste à sélectionner les points de discrétisations informatifs les plus pertinents pour prédire la réponse Y.

On observe $(Y_i(t_k), X_i(t_k))$ variables décorrélées

$$(Y_i, X_i(t_1), ..., X_i(t_k)_{i=1,...,n}$$

On peut vouloir considérer le modèle

$$Y_i = \sum_{k=1}^{K} \phi_k X_i(t_k) + \mu_i$$

avec :
- $(\phi_1, ..., \phi_k) \in \mathbf{R^k} inconnu$
- $\mu_i \sim \mathcal{N}(\mu, \sigma^2)$

3.1.2 Estimation dans le cas du modèle discret

Conformément aux notations de la partie **3.1**, on souhaite estimer les ϕ_k On a :

$$(\widehat{\phi_1}, ..., \widehat{\phi_k}) \in argmin_{(\phi_1,...,\phi_k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \phi_k X_i(t_k$$

La prédiction obtenue est ainsi :

$$\widehat{Y_i} = \sum_{k=1}^{K} \phi_k X_i(t_k)$$

Cette expression résulte du critère des moindres carrés que l'on abordera dans le cas général.

3.2 Le modèle fonctionnel linéaire

La nature fonctionnelle de la variable X suggère une modélisation particulière. Les données fonctionnelles constituent des échantillons statistiques de très grande dimension. Pour appréhender ce type de données, on considère que les courbes correspondent aux réalisations de processus stochastiques[2] caractérisées par des trajectoires relativement régulières (lisses).

En effet, on définit un échantillon de données fonctionnelles comme étant une famille (X_i) d'observations[3] d'une variable aléatoire X à valeurs dans un espace de dimension infinie.

3.2.1 Analyse fonctionnelle : Espace de Hilbert

Définition du modèle :

Définition 3.1 ***Variable et donnée fonctionnelles*** *[Ferraty et Vieu (2006)].*
Une variable aléatoire est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie.
Une observation d'une variable fonctionnelle est appelée donnée fonctionnelle.

Notre problème consiste en l'étude d'un modèle linéaire doublement fonctionnel. On rappelle que l'observation prend la forme d'un échantillon $(Y_i, X_i)_{i=1,..,n}$ où $\forall i$,X_i et Y_i sont des fonctions sur [0,1] telles que :

$$Y_i(t) = \int_{[0,1]} \beta(s,t) X_i(s) ds + \epsilon(t)$$

[2] Un processus stochastique représente une évolution d'une variable aléatoire
[3] Ces observations ne sont pas nécessairement indépendantes

On suppose que X est à valeurs dans un espace de Hilbert séparable $(\mathbf{H}, <.,.>, ||.||)$, c'est-à-dire :
- un espace vectoriel muni d'un produit scalaire $<.,.>$
- complet pour la norme associée $||.|| = \sqrt{<x,x>}$
- il admet un sous-ensemble dénombrable dense.
Ici $\mathbf{H} = L^2$ (espace des fonctions de carré intégrable) pour I un intervalle de R.

Un résultat théorique fondamental :

Théorème 3.2 *Soit* $(\mathbf{H}, <.,.>)$ *un espace de Hilbert séparable.*
(i) H admet une (des) base(s) hilbertienne(s) dénombrable(s), c'est-à-dire une famille $(\varphi_j)_{j\in\mathbf{N}}$ *d'éléments orthonormés de H telle l'adhérence de l'espace vectoriel engendré par cette famille est égale à l'espace H tout entier. Cela signifie que, tout élément de H se décompose de façon unique sous la forme :*

$$\forall x \in H, x = \sum_{j\in\mathbf{N}} < x, \varphi_j > \varphi_j$$

(ii) (Riesz)Toute forme linéaire sur H, $f : \mathbf{H} \to \mathbf{R}$*, peut se mettre sous la forme* $f =< .,x >$ *pour un élément x de H. On dit que H est isomorphe à son dual* H^**.*

Dès lors, l'approche fonctionnelle du problème de régression consiste à modéliser le lien de dépendance linéaire entre la variable aléatoire réelle Y (réponse) et sa covariable (la variable aléatoire fonctionnelle explicative X).

$$Y =< \beta, X > +\epsilon$$

3.2.2 Probabilités : Processus aléatoires

Les outils propres à la statistique multivariée peuvent être adaptés au cadre fonctionnel[4].

	Cas multivarié	Cas fonctionnel
Variable	$X \in \mathbb{R}^d$ $X = {}^t(X_1, \ldots, X_d)$	$X \in L^2(T)$ $X = \{X(t), t \in T\}$
Espérance	vecteur de moyenne $\mathbb{E}[X] = {}^t(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$	courbe de la moyenne $\mathbb{E}[X] = \{\mathbb{E}[X(t)], t \in T\}$
Covariance	matrice $\Sigma_X = \mathrm{Cov}(X, X)$	fonction de covariance C_X ou opérateur Γ_X $C_X(s,t) = \mathrm{Cov}(X(s), X(t))$

[4] http://gchagny.perso.math.cnrs.fr/CoursFDA_slides_Chap2.pdf

Décomposition de Karhunen-Loeve

Définition 3.3 ***Opérateur de covariance .***
En supposant $\mathbf{E}[||X||^2] < \infty$*, l'opérateur de covariance de* X *est ainsi défini :*

$$\Gamma : f \in \mathbf{H} \mapsto \mathbf{E}[< X, f > X]$$

Définition 3.4 ***Opérateur de covariance .***
Nous noterons $(\varphi_j)_{j\geq 1}$ *les fonctions propres de l'opérateur de covariance. Elles forment une base de Hilbert. Les* $(\lambda_j)_{j\geq 1}$ *sont les valeurs propres associées, rangées dans l'ordre décroissant.*

$$\Gamma : f \in \mathbf{H} \mapsto \mathbf{E}[< X, f > X]$$

Définition 3.5 ***Développement de Karhunen-Loève .***
Le développement de Karhunen-Loève d'une variable X *à valeurs dans un espace de Hilbert telle que* $\mathbf{E}[||X||^2] < \infty$ *est son expression dans la base de fonctions propres de l'opérateur de covariance associé. Avec les notations précédentes, on a :*

$$X = \mathbf{E}[X] + \sum_{j=1}^{\infty} < X_j, \varphi_j > \varphi_j = \mathbf{E}[X] + \sum_{j=1}^{\infty} \sqrt{\lambda_j}\xi_j\varphi_j > \varphi_j$$

où $\xi_j = \frac{<X_j,\varphi_j>}{\sqrt{\lambda_j}}$

C'est une décomposition unique.

[**Preuve**] On va faire la preuve

4 Estimation du paramètre fonctionnel β

La méthode de prédiction de Y est repose sur l'estimateur $\widehat{\beta}$ de β.

4.1 Représentation des fonctions dans une base de Hilbert

4.1.1 Idée

On considère que les variables aléatoires sont à valeurs dans $L^2(\mathbf{T})$ où $\mathbf{T}$ est un intervalle de $\mathbf{R}$. Le but est d'approcher les variables fonctionnelles comme combinaisons linéaires de fonctions d'une base donnée.

Le concept de base hilbertienne est donc l'outil théorique qui va nous permettre d'approcher la variable aléatoire fonctionnelle X (qui appartient à L^2).

Définition 4.1 ***Base Hilbertienne .***
Une famille de fonctions $(\varphi_j)_{j\in\mathbf{N}}$ forme une base hilbertienne si :
- elle est composée de fonctions orthonormales (donc linéairement indépendantes)
- l'espace engendré par cette famille est dense dans L^2. En prenant une combinaison linéaire d'un nombre suffisamment grand de ces fonctions, tout élément de de L^2 peut être approché par une combinaison linéaire de ces fonctions.

Ainsi :

$$X = \sum_{j=1}^{\infty} \theta_j \varphi_j$$

la série étant convergente dans $L^2(\mathbf{T})$.

Pour approcher X par $\widehat{X}$, il faut choisir un certain niveau d, tronquer la série à ce niveau et estimer les coefficients.

$$\widehat{X}(t) = \sum_{j=1}^{d} \widehat{\theta}_j \varphi_j(t), t \in \mathbf{T}$$

Cette approximation $\widehat{X}$ appartient à sous-espace de dimension finie d de $L^2(\mathbf{T})$ qui n'est autre que $Vect(\varphi_1, ..., \varphi_d)$.

On se retrouve donc avec deux paramètres à estimer : les coefficients $\theta = (\theta_d, ..., \theta_d)$ et la dimension de l'espace d'approximation auquel on va se retreindre.

4.1.2 Base de Fourier

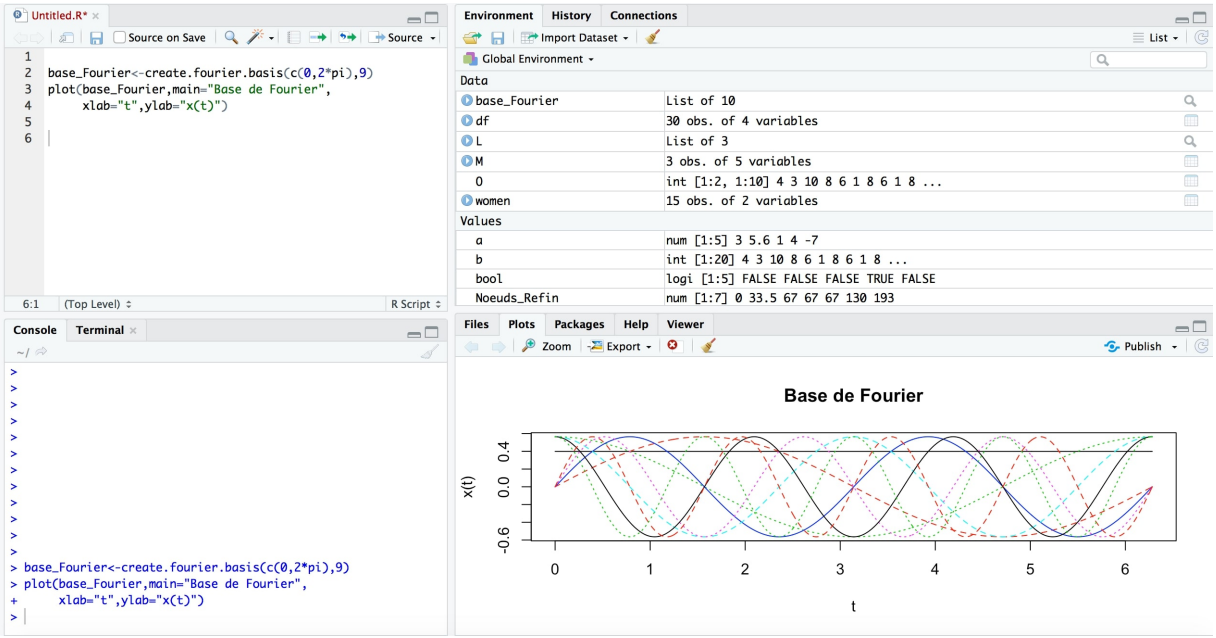
La base de Fourier, ou base trigonométrique, est une base hilbertienne encline à approcher des courbes aléatoires dont le comportement est périodique, ce qui est le cas pour les phénomènes d'évolution de température que nous avons choisis d'étudier.

On a par exemple : $\forall t$,

$$\psi_0(t) = 1, \psi_{2k-1}(t) = \sin(k\omega t), \psi_{2k}(t) = \cos(k\omega t), k \in \mathbf{N} \setminus \{0\}$$

Ainsi, dans l'approche fonctionnelle, on va reconstruire la fonction en la projetant sur une base de Fourier.

L'implémentation sur R d'une base de Fourier se fait à l'aide de l'outil : create.fourier.basis.



4.2 Le lissage des données par moindres carrés

Une courbe aléatoire n'est jamais observée de façon globale : les instruments de mesure ont une vitesse d'enregistrement limitée (bien qu'ils soient de plus en plus performants) et il n'est pas possible de stocker un nombre infini (non dénombrable) de valeurs. En réalité, les processus aléatoires sont insaisissables. De prime abord, une donnée fonctionnelle se présente sous forme vectorielle : elle est constituée d'un certain nombre de valeurs discrètes qui ont été mesurées sur une grille suffisamment fine, et enregistrées.

On distingue deux cas. Lorsque les observations sont obtenues sans erreur, l'expression des Y_i en fonction des X_i résume l'observation de la courbe. Pour trouver β, on fait de l'interpolation : on reconstruit la courbe à partir des données d'un nombre fini de points.

Le second cas est plus fréquent et c'est celui qui vas nous intéresser : les observations sont bruitées.

Le bruit ϵ_i englobe les erreurs de mesure : c'est un une perturbation qui contribue au caractère brut des données. Dès lors, il faut procéder au lissage des données, pour ôter l'erreur de mesure, c'est-à-dire prendre en compte le bruit qui se superpose au signal, le filtrer.

4.2.1 Principe des moindres carrés

:

Le principe des moindres carrés consiste à réduire la dimension de l'espace fonctionnel (ici on appelle m la dimension finie du sous-espace considéré) et à estimer le paramètre d'intérêt β. Remarque : On reprend les notations de **4.1**.

Soit $(\psi_j)_{j\geq 1}$ la base de Fourier de $L^2([0;1])$. On a $Vect(\psi_1, ..., \psi_j)$ dense dans $L^2([0;1])$. Comme $\beta \in L^2([0;1]), \exists b_1, ..., b_n$ tel que:

$$\sum_{j=1}^{m} b_j \psi_j \rightarrow^{L^2}_{m\rightarrow\infty} \beta$$

$$\beta_m = \sum_{j=1}^{m} < \beta, \psi_j > \psi_j$$

correspond à la projection orthogonale de β sur $S_m = Vect(\psi_1, ..., \psi_m)$. On peut montrer que $||\beta - \beta_m|| \leq ||\beta - f|| \forall f \in L^2$.

Le critère des moindres carrés correspond à la résolution d'un problème d'optimisation. On estime $(b_1, ..., b_m)$ en minimisant

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{m} (b_j X_i j)^2)$$

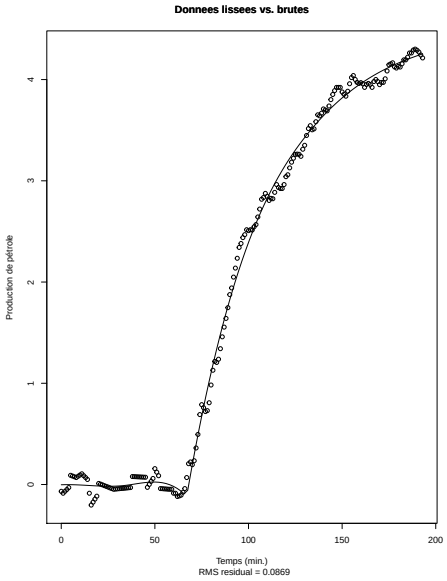
et on trouve finalement :

$$\widehat{\beta} = ({}^t XX)^{-1\,t} XY$$

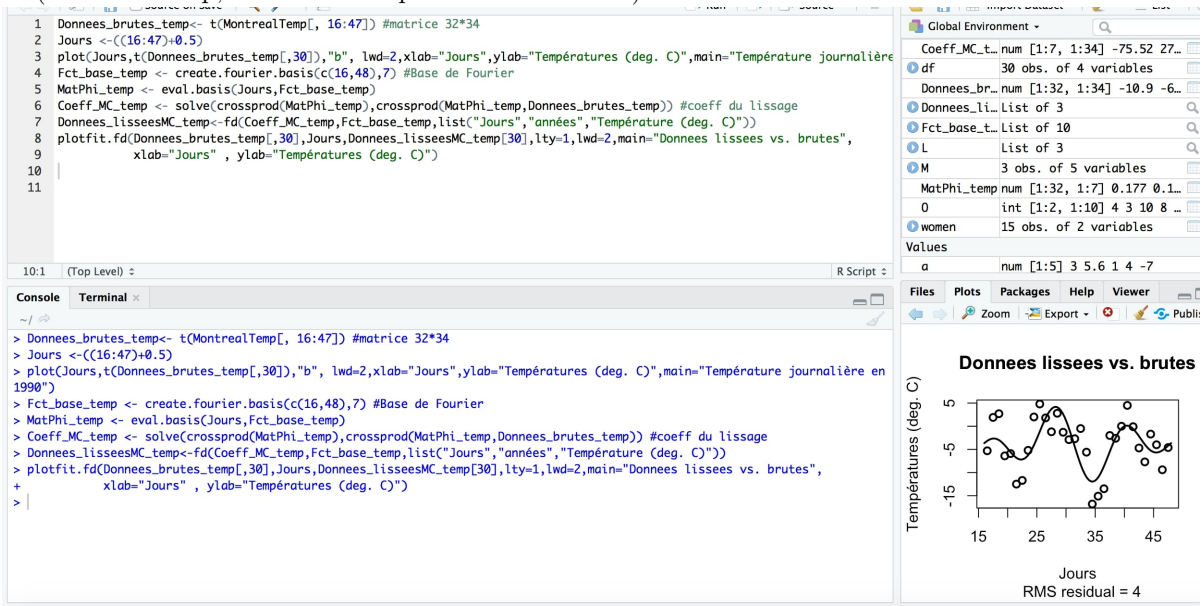
4.2.2 Le lissage des données : Simulation numérique

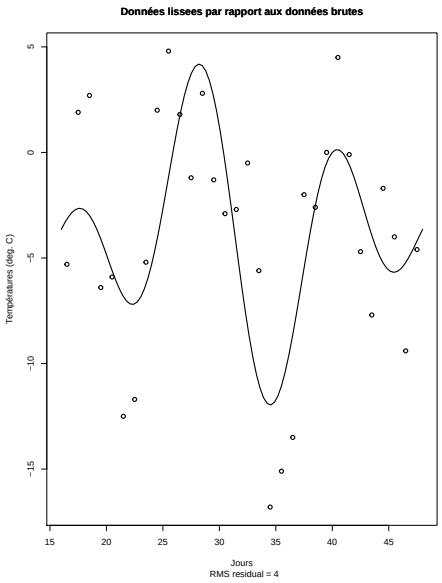
Voici un exemple de lissage de données (base "refinery" issue du package R "fda") utilisant une base de fonctions de Fourier (même si en réalité, une autre base aurait été plus appropriée, dans ce cas - la base "splines"-).

```
install.packages("fda")
library(fda)

Noeuds_Refin<-c(seq(0,67,length.out=3),67,seq(67,193,length.out=3)) #7 noeuds
Fct_base_Refin <- create.fourier.basis(c(0,193),norder=4,breaks=Noeuds_Refin) #9 fct base
plot(Fct_base_Refin,xlab="Temps (min.)")
MatPhi_Refin <-predict(Fct_base_Refin,refinery$Time) #matrice 194 lignes 9 colonnes
# ou M<-eval.basis(refinery$Time,Fct_base_Refin)
Coeff_MC_Refin<- solve(crossprod(MatPhi_Refin),crossprod(MatPhi_Refin,refinery$Tray47))
Donnees_lisseesMC_Refin <-fd(Coeff_MC_Refin,Fct_base_Refin)
plotfit.fd(refinery$Tray47,refinery$Time,Donnees_lisseesMC_Refin,lty=1,lwd=1,main="Données lissées vs. brutes",col=1,
           xlab="Temps (min.)",ylab="Production de pétrole")

```



Voici un exemple représentatif de la situation que nous avons choisie d'étudier : les données liées à la température à partir de la base de données disponible sur R (MontrealTemp, relevés de température à Montréal).





5 Conclusion

Les données fonctionnelles se distinguent par leur complexité structurelle : leur champ d'étude requiert une modélisation particulière qui repose l'analyse fonctionnelle dans les Espaces de Hilberts.
A cela s'ajoutent les extensions fonctionnelles de la statistique multivariée qui permettent d'obtenir des décompositions utiles pour l'estimation du paramètre d'intérêt. Dès lors, l'on procède par lissage de données : le critère des moindres carrés nous permet ainsi d'obtenir un estimateur consistant et de résoudre le problème.

6 Sitographie

- http://gchagny.perso.math.cnrs.fr/CoursFDA_slides_Chap2.pdf.
- http://maths.cnam.fr/IMG/pdf/STA201_AnalyseFonctionelle_Preda_cle0acdf9.pdf.
- http://gchagny.perso.math.cnrs.fr/CoursFDA.pdf
-http://www.modulad.fr/archives/numero-43/VIEU/2-Vieu.pdf
-https://www.ceremade.dauphine.fr/~roche/These.pdf
-https://link-springer-com-s.proxy.bu.dauphine.fr/content/pdf/10.1007%2Fb98888.pdf
-https://lipn.univ-paris13.fr/A3/AAFD10/slides/AAFD10_Saporta.pdf