

Habilitation à Diriger des Recherches - Contributions aux modèles fonctionnels de régression

Christophe Crambes

30 juin 2020



Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

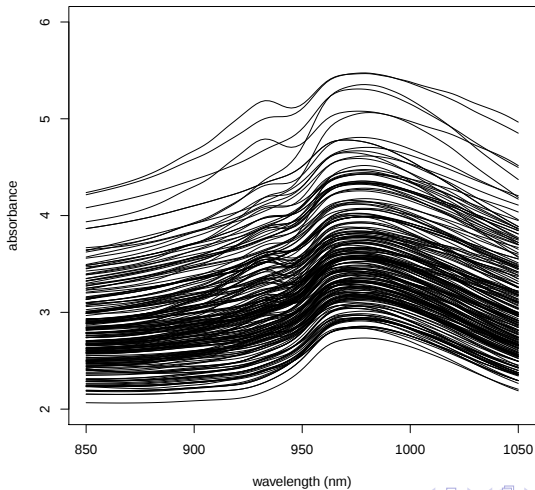
Perspectives

Analyse de données fonctionnelles

- ➡ Donnée fonctionnelle : variable aléatoire à valeurs dans un espace de fonctions (ex : $\mathbb{L}^2([a; b])$)
- ➡ En pratique : observation d'une donnée fonctionnelle en certains points de mesure
- ➡ Point clé : régularité
- ➡ Références : Ramsay et Silverman (2002, 2005), Ferraty et Vieu (2006), Horváth et Kokoszka (2012), Hsing et Eubank (2015), ...

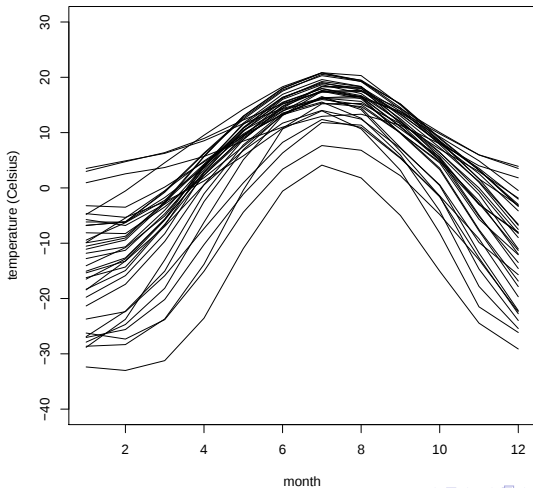
Exemples de données fonctionnelles

Données **spectrométriques**



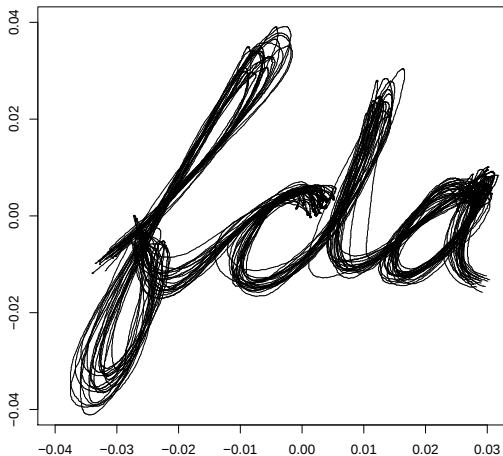
Exemples de données fonctionnelles

Données de températures canadiennes



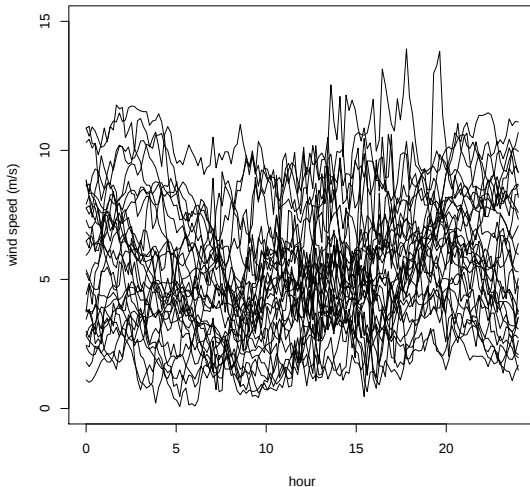
Exemples de données fonctionnelles

Données d'écriture

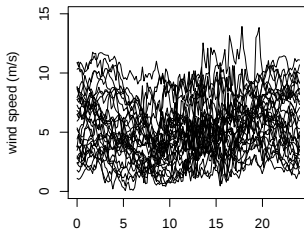
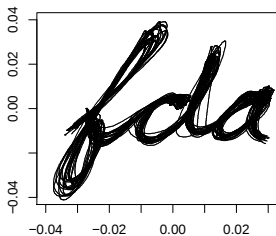
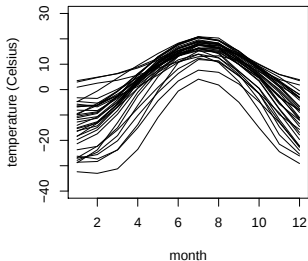
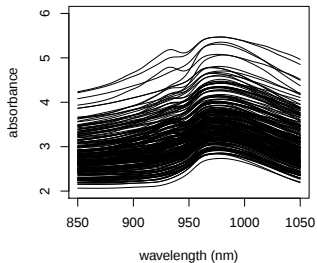


Exemples de données fonctionnelles

Données éoliennes



Exemples de données fonctionnelles



Modèles de régression

➡ **Modèle** général :

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

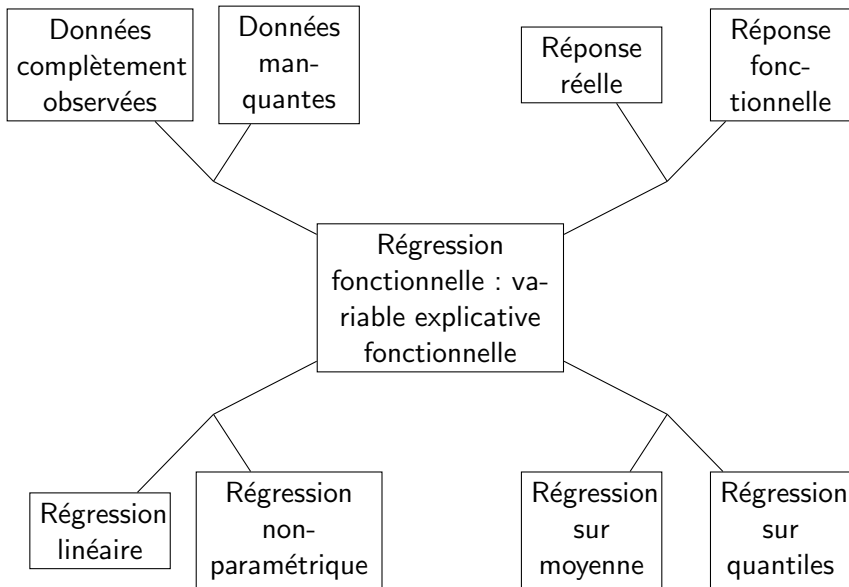
➡ X : variable explicative **fonctionnelle** (à valeurs dans $\mathbb{L}^2([a; b])$)

➡ Y : variable réponse **réelle** ou **fonctionnelle**

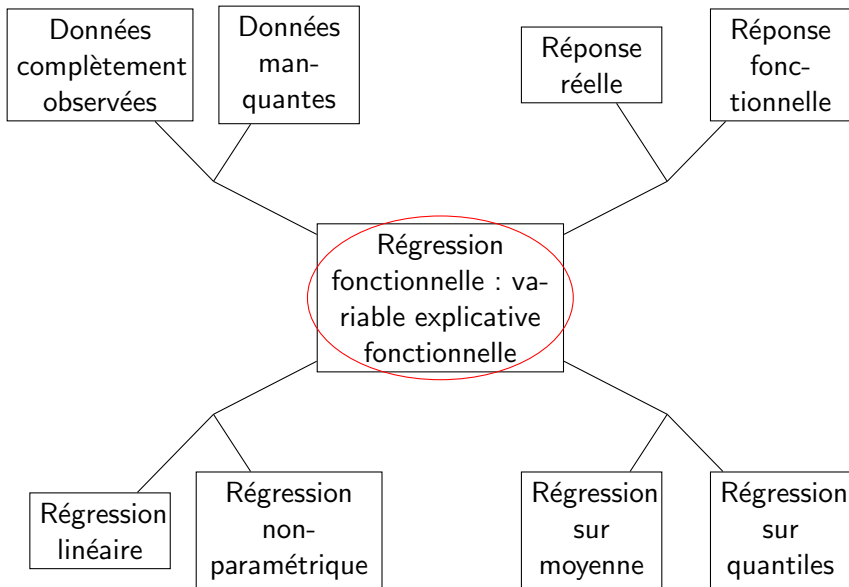
➡ Ψ : opérateur de régression $\mathbb{L}^2([a; b]) \longrightarrow \mathbb{R}$

➡ ε : **bruit** du modèle, indépendant de X

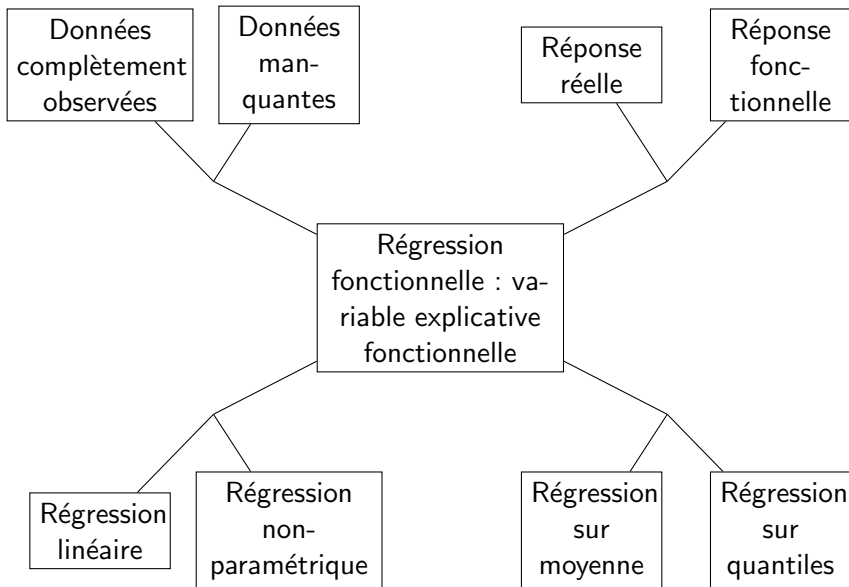
Modèles étudiés



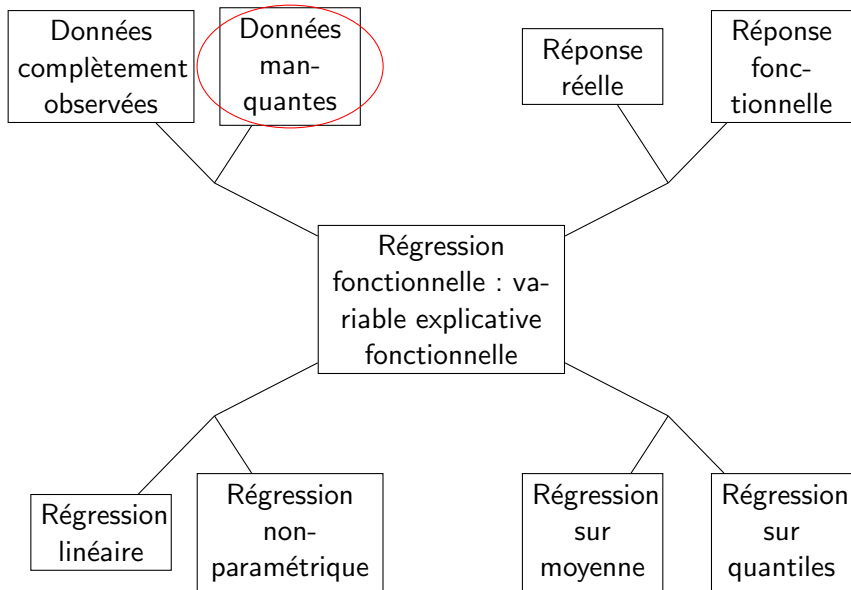
Modèles étudiés



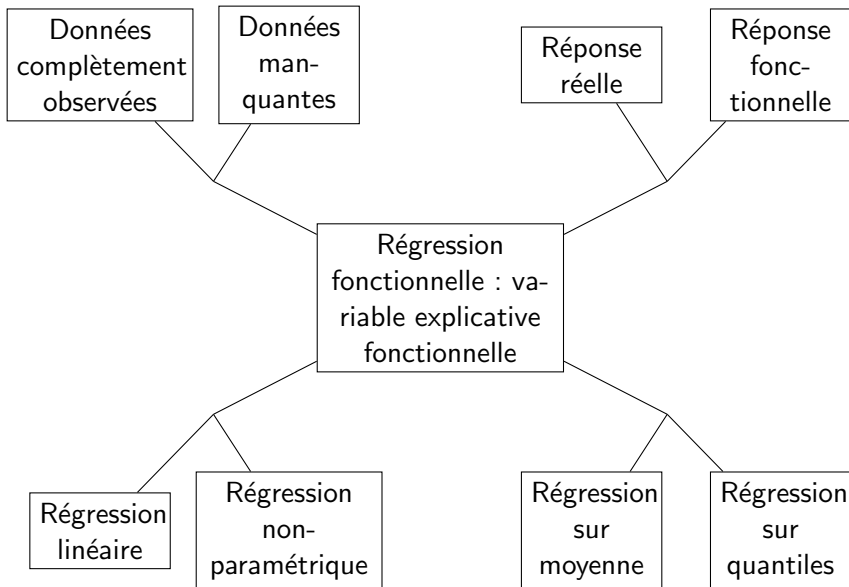
Modèles étudiés



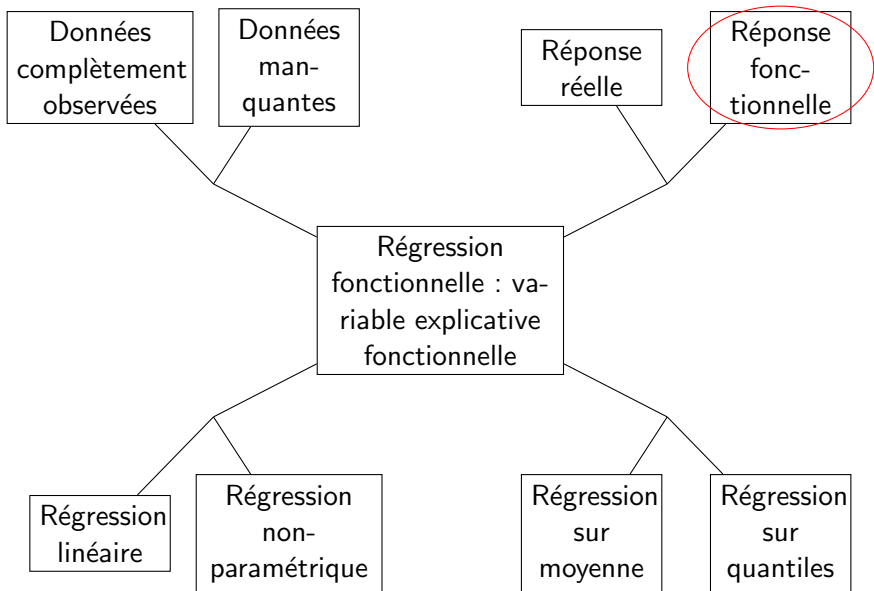
Modèles étudiés



Modèles étudiés



Modèles étudiés



Modèle linéaire complètement fonctionnel

➡ Travail réalisé avec André Mas (2013)

➡ Modèle :

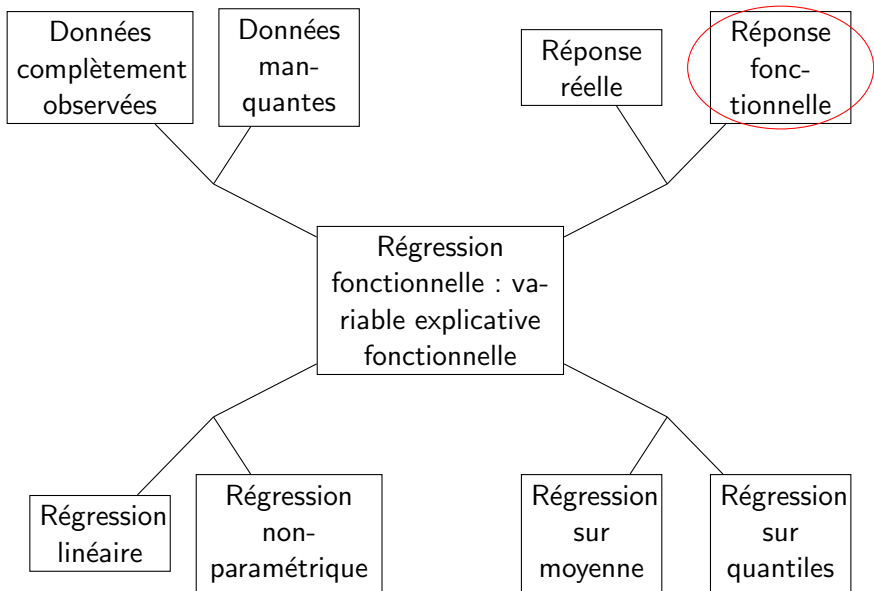
$$Y(t) = \Theta X(t) + \varepsilon(t)$$

➡ Estimateur :

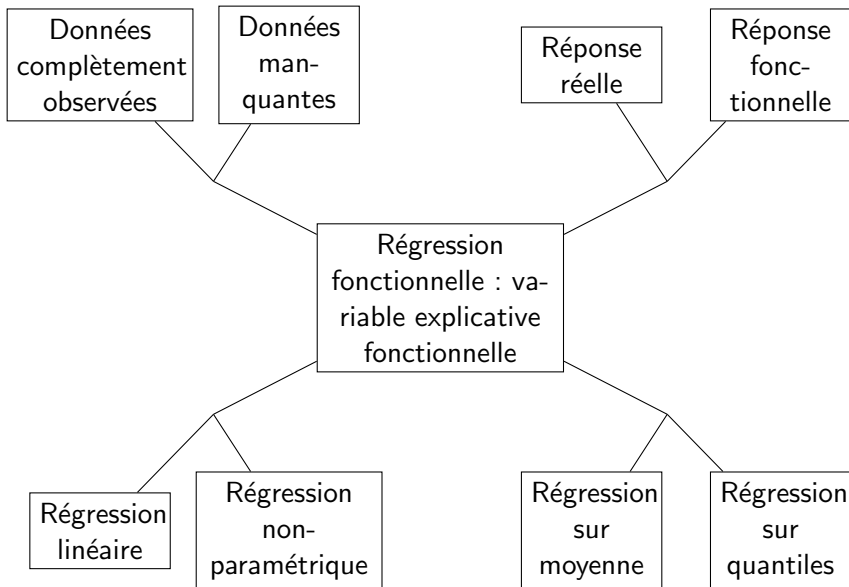
$$\hat{\Theta} = \hat{\Pi}_{k_n} \hat{\Delta}_n \left(\hat{\Pi}_{k_n} \hat{\Gamma}_n \right)^{-1}$$

- $\hat{\Gamma}_n$: opérateur de **covariance** empirique $\hat{\Gamma}_n$
- $\hat{\Pi}_{k_n}$: opérateur de **projection** sur le sous-espace engendré par les k_n premières fonctions propres de $\hat{\Gamma}_n$
- $\hat{\Delta}_n$: opérateur de **covariance croisée** empirique

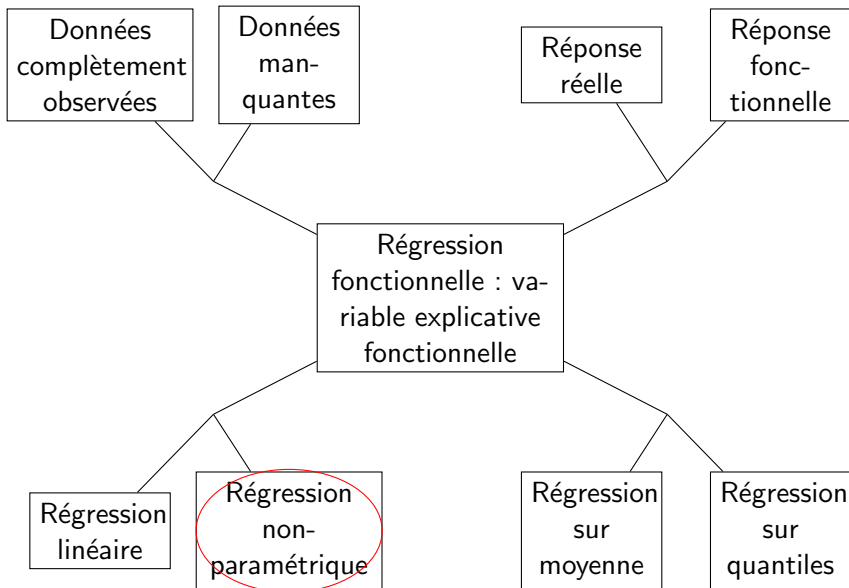
Modèles étudiés



Modèles étudiés



Modèles étudiés



Estimateur non-paramétrique récursif (1)

➡ Travail réalisé avec Aboubacar Amiri et Baba Thiam (2014)

➡ Estimateur :

$$\Psi_n^{[\ell]}(\chi) = \frac{\sum_{i=1}^n \frac{Y_i}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)}{\sum_{i=1}^n \frac{1}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)}$$

- $\chi \in \mathbb{L}^2([a; b])$, $\ell \in [0; 1]$
- $(h_n)_{n \geq 1}$: suite de **fenêtres**
- K : **noyau**
- F : **fonction de répartition** de $\|\chi - X\|$

Estimateur non-paramétrique récursif (2)

➡ Écriture récursive :

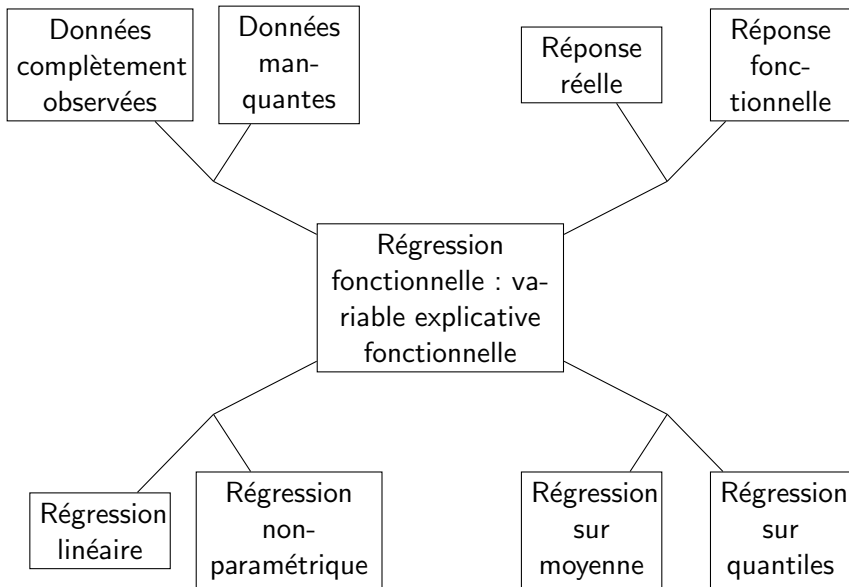
$$\psi_{n+1}^{[\ell]}(x) = \frac{\left[\sum_{i=1}^n F(h_i)^{1-\ell} \right] \varphi_n^{[\ell]}(x) + \left[\sum_{i=1}^{n+1} F(h_i)^{1-\ell} \right] Y_{n+1} K_{n+1}^{[\ell]}(\|x - X_{n+1}\|)}{\left[\sum_{i=1}^n F(h_i)^{1-\ell} \right] f_n^{[\ell]}(x) + \left[\sum_{i=1}^{n+1} F(h_i)^{1-\ell} \right] K_{n+1}^{[\ell]}(\|x - X_{n+1}\|)}$$

$$\varphi_n^{[\ell]}(x) = \frac{\sum_{i=1}^n \frac{Y_i}{F(h_i)^\ell} K\left(\frac{\|x - X_i\|}{h_i}\right)}{\sum_{i=1}^n F(h_i)^{1-\ell}} \quad f_n^{[\ell]}(x) = \frac{\sum_{i=1}^n \frac{1}{F(h_i)^\ell} K\left(\frac{\|x - X_i\|}{h_i}\right)}{\sum_{i=1}^n F(h_i)^{1-\ell}}$$

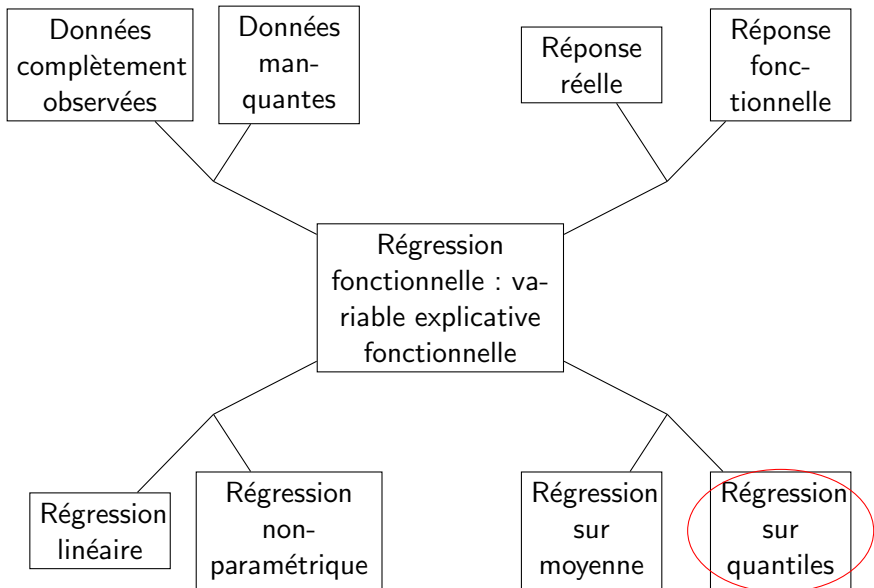
$$K_i^{[\ell]}(\cdot) = \frac{K\left(\frac{\cdot}{h_i}\right)}{F(h_i)^\ell \sum_{j=1}^i F(h_j)^{1-\ell}}$$

➡ Calcul du **biais**, de la **variance**, résultats de **convergence presque sûre** et de **normalité asymptotique**

Modèles étudiés



Modèles étudiés



Estimation de quantiles de régression par méthodes SVM (1)

➡ Travail réalisé avec Yousri Henchiri et Ali Gannoun (2010-2013)

➡ Problème de minimisation

$$\min_{\Psi \in \mathcal{H}_{\mathcal{K}}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} \rho_{\tau}(Y_i - \Psi(X_i)) + \lambda \|\Psi\|_{\mathcal{H}_{\mathcal{K}}}^2 \right\}$$

- $\rho_{\tau}(u) = |u| + (2\tau - 1)u$: fonction de perte
- $\mathcal{K}$: noyau associé à un RKHS $\mathcal{H}_{\mathcal{K}}$

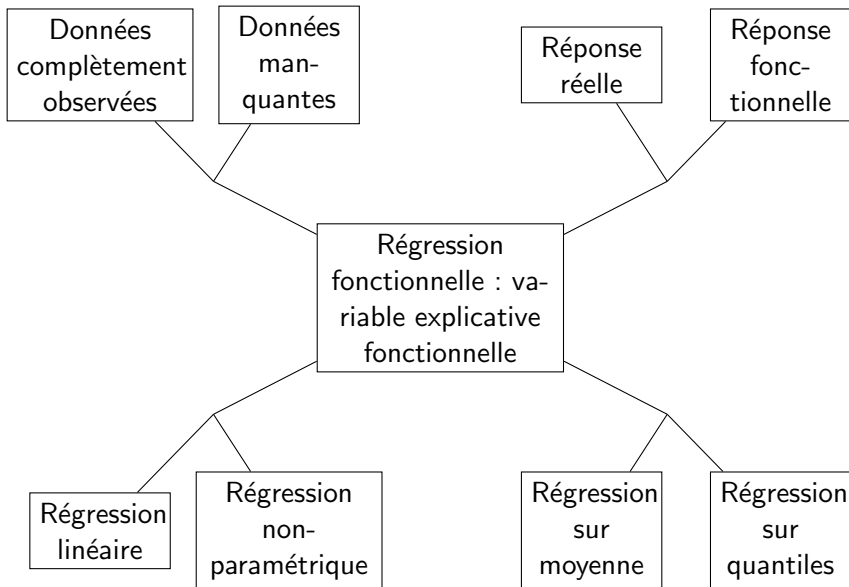
➡ Problème reformulé :

$$\min_{\alpha \in \mathbb{R}^{\ell}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} \rho_{\tau} \left(Y_i - \sum_{j=1}^{\ell} \alpha_j \mathcal{K}^{(d)}(X_i^{(d)}, X_j^{(d)}) \right) + \lambda \alpha' \mathbf{K}^{(d)} \alpha \right\}$$

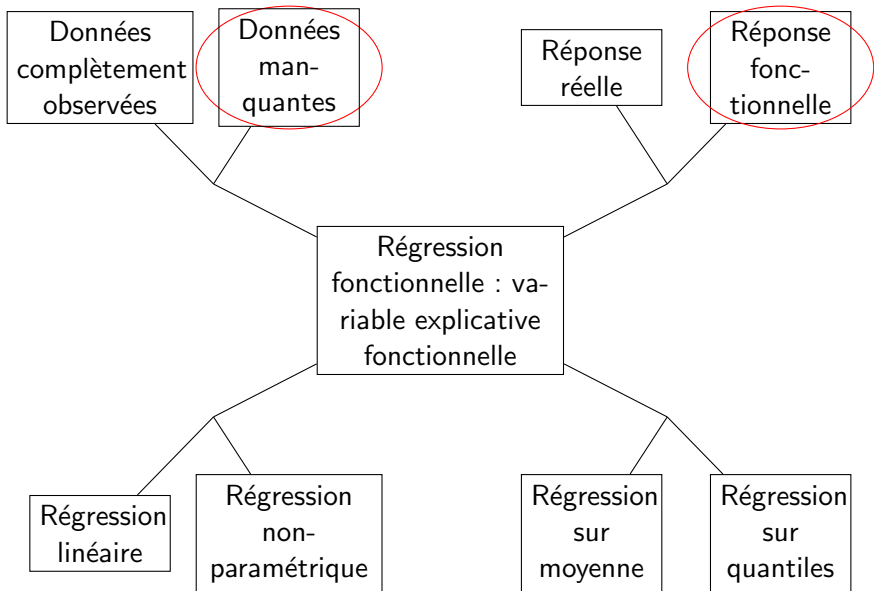
Estimation de quantiles de régression par méthodes SVM (2)

- ➡ Résolution du problème de minimisation par **moindres carrés itérés**
- ➡ Extension au cas de plusieurs variables explicatives avec un **modèle additif**
- ➡ Résultats de **convergence en probabilité** avec une **vitesse de convergence**

Modèles étudiés



Modèles étudiés



Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Modèle linéaire fonctionnel

➡ Travail réalisé avec Yousri Henchiri (2019)

➡ Modèle :

$$Y = \langle \theta, X \rangle + \varepsilon$$

➡ X : variable explicative fonctionnelle (à valeurs dans $\mathbb{L}^2([a; b])$)

➡ Y : variable réponse réelle

➡ θ : fonction de régression appartenant à $\mathbb{L}^2([a; b])$

➡ ε : bruit du modèle, indépendant de X

Données manquantes sur la réponse

- ➡ Indicateur de données observées :

$$\delta_i = \begin{cases} 0 & \text{si } Y_i \text{ est manquant} \\ 1 & \text{si } Y_i \text{ est observé} \end{cases}$$

- ➡ Cadre Missing At Random (MAR) :

$$\mathbb{P}(\delta = 1 \mid X, Y) = \mathbb{P}(\delta = 1 \mid X)$$

- ➡ Nombre de données manquantes dans l'échantillon :

$$m_n = \sum_{i=1}^n \mathbb{1}_{\{\delta_i=0\}}$$

- ➡ Objectif : reconstruction de l'échantillon, estimation de θ et prévision d'une nouvelle valeur de la réponse

Imputation des données manquantes par régression

➡ **Diagonalisation** de l'opérateur de covariance empirique :

valeurs propres : $\hat{\lambda}_1, \dots, \hat{\lambda}_{k_n}$

fonctions propres : $\hat{v}_1, \dots, \hat{v}_{k_n}$

➡ **Imputation** :

$$Y_{\ell, imp} = \frac{1}{n - m_n} \sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i, \hat{v}_j \rangle \langle X_{\ell}, \hat{v}_j \rangle \delta_i Y_i}{\hat{\lambda}_j}$$

Estimation et prévision

➡ Réponse complétée :

$$Y_i^* = Y_i \delta_i + Y_{i,imp}(1 - \delta_i)$$

➡ Estimation de θ :

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i, \hat{v}_j \rangle Y_i^*}{\hat{\lambda}_j} \hat{v}_j$$

➡ Prévision pour une nouvelle entrée X_{new} :

$$\hat{Y}_{new} = \langle X_{new}, \tilde{\theta} \rangle = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_n} \frac{\langle X_i, \hat{v}_j \rangle \langle X_{new}, \hat{v}_j \rangle Y_i^*}{\hat{\lambda}_j}$$

Hypothèses

➡ (A.1) Il existe une fonction convexe λ telle que $\lambda(j) = \lambda_j$ pour tout $j \geq 1$ qui interpole de façon continue les valeurs propres de l'opérateur de covariance

➡ (A.2) Il existe une constante $C > 0$ telle que :

$$\mathbb{E} \left(\|X\|^4 \right) \leq C$$

➡ (A.3) On a :

$$\lim_{n \rightarrow +\infty} \lambda_{k_n} k_n = 0$$

Consistance de l'imputation

Sous les hypothèses (A.1)-(A.3), on a :

$$\mathbb{E} (Y_{\ell, imp} - \langle \theta, X_{\ell} \rangle)^2 = \sum_{j=k_n+1}^{+\infty} \left(\Theta \Gamma^{1/2} v_j \right)^2 + \frac{\sigma_{\varepsilon}^2 k_n}{n - m_n} + o \left(\frac{k_n}{n - m_n} \right)$$

Vitesses particulières

➡ On note $\varphi(j) = \frac{(\Theta \Gamma^{1/2} v_j)^2}{\|\Theta \Gamma^{1/2}\|_\infty^2}$

➡ Vitesses pour $\mathbb{E} (Y_{\ell, \text{imp}} - \langle \theta, X_\ell \rangle)^2$ suivant les données manquantes :

	$\varphi = \varphi_{\text{pol}} = C_\alpha j^{-(2+\alpha)}$
$m_n := a_n n = o(n)$	$\underset{n \rightarrow +\infty}{\sim} K_\alpha n^{-(1+\alpha)/(2+\alpha)}$
$m_n = \rho n$	$\underset{n \rightarrow +\infty}{\sim} K_\alpha (1 - \rho)^{1/(2+\alpha)} n^{-(1+\alpha)/(2+\alpha)}$
$u_n := n - m_n = o(n)$	$\underset{n \rightarrow +\infty}{\sim} K_\alpha u_n^{-(1+\alpha)/(2+\alpha)}$

Résultat sur la prévision

Sous les hypothèses (A.1)-(A.3), si de plus $m_n = o(n)$ et $m_n^2 k_n = O(n)$, alors

$$\mathbb{E}\left(Y_{new} - \langle \theta, X_{new} \rangle\right)^2 = \sum_{j=k_n+1}^{+\infty} \left(\Theta \Gamma^{1/2} v_j\right)^2 + O\left(\frac{k_n}{n}\right)$$

Perspectives (1)

- ➡ Travail actuellement en collaboration avec Chayma Daayeb et Ali Gannoun
- ➡ Données manquantes sur la variable explicative : reconstruction de courbes (Kneip et Liebl, 2020)
- ➡ Opérateur linéaire de reconstruction :

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s)$$

- ➡ Estimation de L à l'aide des parties **observées** des courbes

Perspectives (2)

- ➡ Résultats obtenus : analogues à ceux présentés ici en tenant en compte de la reconstruction des données manquantes sur les courbes
- ➡ Autres perspectives :
 - Amélioration de l'imputation par **imputation multiple**
 - Généralisation à une variable d'intérêt **fonctionnelle**

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Modèle fonctionnel de convolution

- ➡ Travail réalisé avec Tito Manrique et Nadine Hilgert (2013-2016)

- ➡ Modèle :

$$Y(t) = \int_0^t \theta(s)X(t-s)ds + \varepsilon(t)$$

- ➡ X : variable explicative fonctionnelle

- ➡ Y : variable réponse fonctionnelle

- ➡ θ : fonction de régression

- ➡ ε : bruit du modèle, indépendant de X

Transformation de Fourier

➡ Transformée de Fourier :

$$\mathcal{F}(f) : \xi \longmapsto \mathcal{F}(f)(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-i\xi x} dx$$

➡ Transformée de Fourier inverse :

$$\mathcal{F}^{-1}(g) : x \longmapsto \mathcal{F}^{-1}(g)(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} g(\xi) e^{ix\xi} d\xi$$

Modèle fonctionnel concurrent

➡ **Modèle :**

$$\mathcal{Y}(\xi) = \beta(\xi)\mathcal{X}(\xi) + \mathcal{E}(\xi)$$

➡ $\mathcal{X} = \mathcal{F}(X)$: variable explicative

➡ $\mathcal{Y} = \mathcal{F}(Y)$: variable réponse

➡ $\beta = \mathcal{F}(\theta)$: fonction de régression

➡ $\mathcal{E} = \mathcal{F}(\varepsilon)$: bruit du modèle, indépendant de $\mathcal{X}$

Estimation

➡ Estimation de β :

$$\hat{\beta}_n := \frac{\frac{1}{n} \sum_{i=1}^n \mathcal{Y}_i \mathcal{X}_i^*}{\frac{1}{n} \sum_{i=1}^n |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}$$

➡ Estimation de θ :

$$\hat{\theta}_n := \mathcal{F}^{-1}(\hat{\beta}_n) = \mathcal{F}^{-1} \left(\frac{\frac{1}{n} \sum_{i=1}^n \mathcal{Y}_i \mathcal{X}_i^*}{\frac{1}{n} \sum_{i=1}^n |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}} \right)$$

Hypothèses

- ➡ (A.1) $\overline{\text{supp}(|\mathcal{F}(\theta)|)} \subseteq \overline{\text{supp}(\mathbb{E}(|\mathcal{F}(X)|))}$
- ➡ (A.2) $\mathbb{E}\left(|\mathcal{X}|^2\right) < \infty$
- ➡ (A.3) $\left\| \frac{|\mathcal{F}(\theta)|}{\mathbb{E}(|\mathcal{F}(X)|^2)} \mathbf{1}_{\overline{\text{supp}(\mathcal{F}(\theta))} \setminus \partial(\text{supp}(\mathbb{E}(|\mathcal{F}(X)|)))} \right\| < +\infty$
- ➡ (A.4) Il existe des nombres réels positifs $\alpha > 0$, $M_0, M_1, M_2 > 0$ tels que pour tout $p \in C_{\theta, \partial X}$, il existe un voisinage ouvert $J_p \subset \text{supp}(|\mathcal{F}(\theta)|)$ pour lequel
 - Pour tout $t \in J_p$, $\mathbb{E}(|\mathcal{F}(X)|^2(t)) \geq |t - p|^\alpha$ et

$$\left\| \frac{1}{\mathbb{E}(|\mathcal{F}(X)|^2)} \right\|_{L^2(J_p \setminus \{p\})} \leq M_0$$

- $\sum_{p \in C_{\theta, \partial X}} \|\theta\|_{C_0(J_p)}^2 < M_1$
- $\frac{|\mathcal{F}(\theta)|}{\mathbb{E}(|\mathcal{F}(X)|^2)} \mathbf{1}_{\text{supp}(|\mathcal{F}(\theta)|) \setminus J} < M_2$, où $J := \bigcup_{p \in C_{\theta, \partial X}} J_p$

Résultat de convergence

Sous les hypothèses (A.1)-(A.4), si de plus $\lambda_n := n^{1-\frac{1}{4\alpha+2}}$, on a

$$\left\| \hat{\theta}_n - \theta \right\|_{L^2} = O_P(n^{-\gamma})$$

où $\gamma := \min \left[\frac{1}{2(2\alpha+1)}, \frac{1}{2} - \frac{1}{2(2\alpha+1)} \right]$ et $n^{-\gamma} = \max \left[\frac{\lambda_n}{n}, \frac{\sqrt{n}}{\lambda_n} \right]$

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Plan de l'exposé

Introduction

Imputation par régression dans le modèle linéaire fonctionnel

Modèle fonctionnel de convolution

Perspectives

Modèles à variables latentes

		Variables latentes	
		Qualitatives	Quantitatives
Variables observées	Qualitatives	Modèles de classes latentes	Modèles de traits latents
	Quantitatives	Modèles de profils latents	Modèles d'équations structurelles

Problème : définition de la notion de densité pour des données fonctionnelles

Pseudo-densité

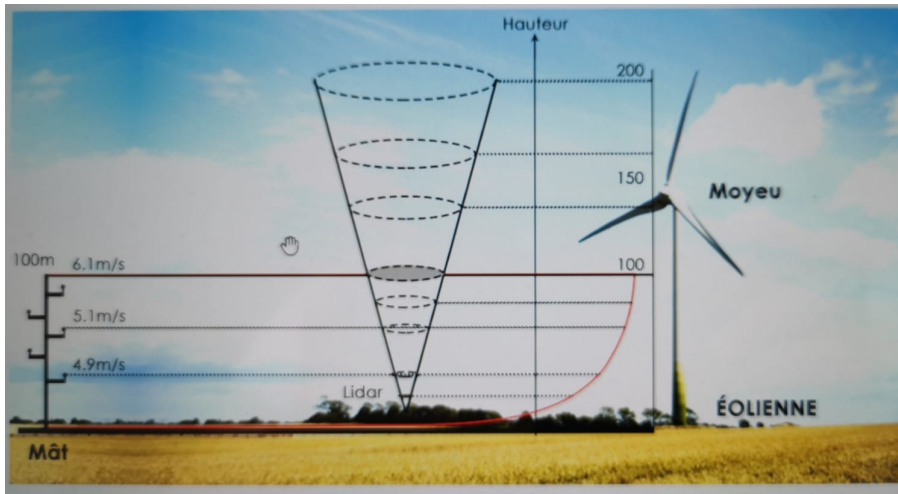
- ➡ Scores de la décomposition K-L de X : $(\xi_j)_{j \geq 1}$
- ➡ Densités des scores : $(f_j)_{j \geq 1}$
- ➡ Probabilités de petites boules : $p(x|h) = \mathbb{P}(\|X - x\| \leq h)$
- ➡ Résultat de Delaigle et Hall (2010) :

$$\log p(x|h) = C_1(k, \lambda) + \sum_{j=1}^k \log f_j(x_j) + o(k)$$

- ➡ Pseudo-log-densité :

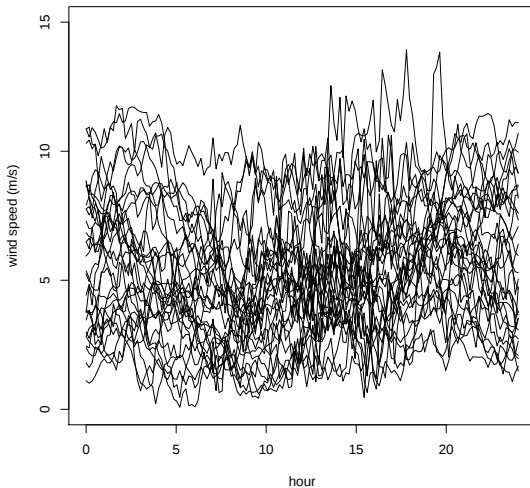
$$\ell(x|k) = \frac{1}{k} \sum_{j=1}^k \log f_j(x_j)$$

Données éoliennes (1)



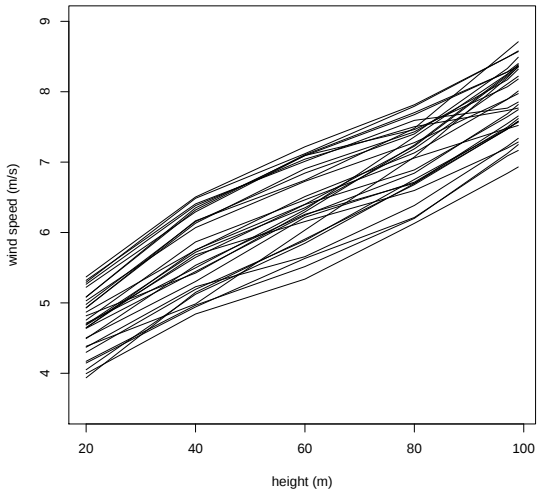
Données éoliennes (2)

Vitesse du vent fonction du **temps**



Données éoliennes (2)

Vitesse du vent fonction de la hauteur



Données éoliennes (3)

➡ Stage de M2 de Nizar Soilihi (2019) :

- Modèle : $VV^{150} = \int_{20}^{100} \theta(h) VV(h) dh + \varepsilon$
- Estimateur : type **ridge** avec pas de mesure **non constants**

➡ Perspectives :

- Autres **modèles** : modèle non linéaire, modèle complètement fonctionnel, ...
- Variables **additionnelles** : température, pression, ...
- Données **manquantes**
- Données **circulaires**