Modèles de régression linéaire pour variables explicatives fonctionnelles

Christophe Crambes

Laboratoire de Statistique et Probabilités

Jeudi 23 Novembre 2006

Plan de l'exposé

Introduction

1-Estimation de quantiles conditionnels

2-Estimation de la moyenne conditionnelle (cas non bruité)

3-Estimation de la moyenne conditionnelle (cas bruité)

4-Application à la prévision de pics de pollution

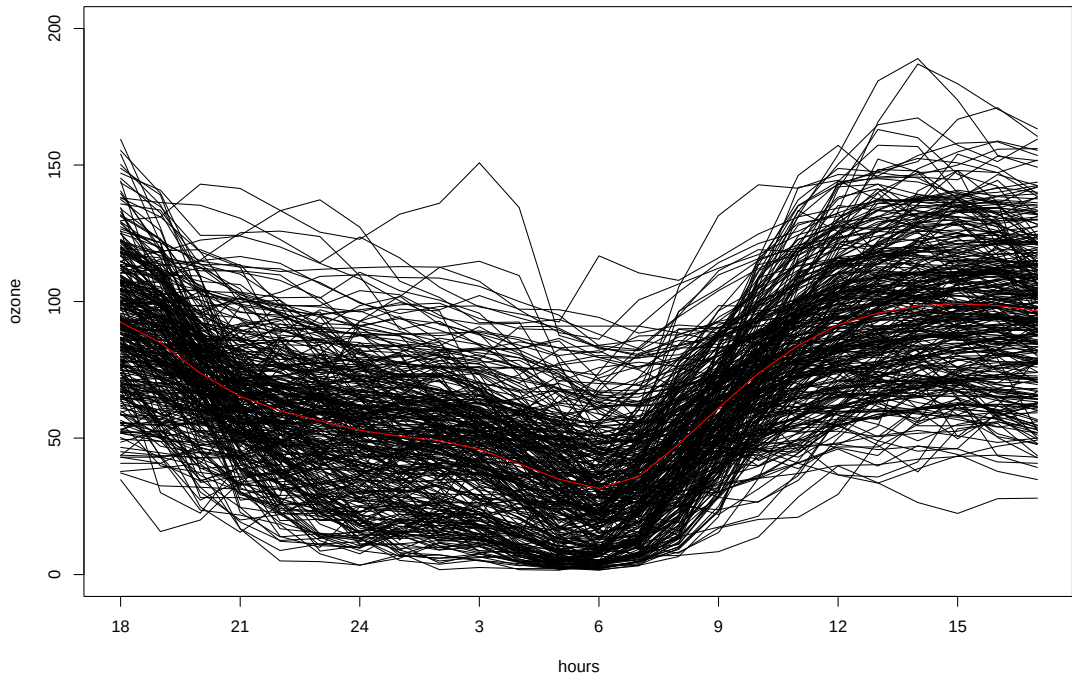
Plan de l'exposé

Introduction sur les données fonctionnelles

- Données fonctionnelles : données assimilables à des courbes ou des surfaces

Introduction sur les données fonctionnelles

- Données fonctionnelles : données assimilables à des courbes ou des surfaces



Introduction sur les données fonctionnelles

- Données fonctionnelles : données assimilables à des courbes ou des surfaces
- Exemples : courbes de croissance, courbes de températures, images satellite, ...

Introduction sur les données fonctionnelles

- Données fonctionnelles : données assimilables à des courbes ou des surfaces
- Exemples : courbes de croissance, courbes de températures, images satellite, ...
- Références : Ramsay et Silverman (2002, 2005)

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i$$

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative
- Y_i variable aléatoire réelle : variable d'intérêt
- ϵ_i variable aléatoire d'erreur

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative
- Y_i variable aléatoire réelle : variable d'intérêt
- ϵ_i variable aléatoire d'erreur
- $\alpha \in L^2([0,1])$ fonction inconnue

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i = \int_0^1 \alpha(t) X_i(t) dt + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative
- Y_i variable aléatoire réelle : variable d'intérêt
- ϵ_i variable aléatoire d'erreur
- $\alpha \in L^2([0,1])$ fonction inconnue

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i = \int_0^1 \alpha(t) X_i(t) dt + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative
- Y_i variable aléatoire réelle : variable d'intérêt
- ϵ_i variable aléatoire d'erreur
- $\alpha \in L^2([0,1])$ fonction inconnue

Références : Ramsay et Dalzell (1991), Cardot, Ferraty et Sarda (1999, 2003)

Modèles considérés

Régression linéaire fonctionnelle : pour $i = 1, \ldots, n$:

$$Y_i = \langle \alpha, X_i \rangle + \epsilon_i = \int_0^1 \alpha(t) X_i(t) dt + \epsilon_i$$

avec :

- X_i variable (aléatoire ou non) à valeurs dans $L^2([0,1])$: variable explicative
- Y_i variable aléatoire réelle : variable d'intérêt
- ϵ_i variable aléatoire d'erreur
- $\alpha \in L^2([0,1])$ fonction inconnue

Références : Ramsay et Dalzell (1991), Cardot, Ferraty et Sarda (1999, 2003)
Extensions : Cuevas, Febrero et Fraiman (2002), Ferraty et Vieu (2006)

Problème lié à la dimension infinie

- Si $\mathbb{E}(\|X\|^2) < +\infty$, on définit l'opérateur de covariance de X :

$$\Gamma_X u = \mathbb{E}\left(\langle X, u\rangle X\right), \qquad u \in L^2([0,1])$$

Problème lié à la dimension infinie

- Si $\mathbb{E}(\|X\|^2) < +\infty$, on définit l'opérateur de covariance de X :

$$\Gamma_X u = \mathbb{E}\left(\langle X, u\rangle X\right), \qquad u \in L^2([0,1])$$

- Les valeurs propres de cet opérateur décroissent rapidement vers 0
- Conséquence : situations liées à des problèmes inverses mal posés

Problème lié à la dimension infinie

- Si $\mathbb{E}(\|X\|^2) < +\infty$, on définit l'opérateur de covariance de X :

$$\Gamma_X u = \mathbb{E}\left(\langle X, u\rangle X\right), \qquad u \in L^2([0,1])$$

- Les valeurs propres de cet opérateur décroissent rapidement vers 0
- Conséquence : situations liées à des problèmes inverses mal posés
- Solution envisagée : régularisation par pénalisation

Plan de l'exposé

Introduction

1-Estimation de quantiles conditionnels

2-Estimation de la moyenne conditionnelle (cas non bruité)

3-Estimation de la moyenne conditionnelle (cas bruité)

4-Application à la prévision de pics de pollution

Plan de l'exposé

Définition des quantiles conditionnels - Modèle

- Données : $(X_i, Y_i)_{i=1,\dots,n}$ couples de variables aléatoires i.i.d. de même loi que (X, Y) avec $Y \in \mathbb{R}$ et $X \in L^2([0,1])$

Définition des quantiles conditionnels - Modèle

- Données : $(X_i, Y_i)_{i=1,\dots,n}$ couples de variables aléatoires i.i.d. de même loi que (X, Y) avec $Y \in \mathbb{R}$ et $X \in L^2([0,1])$
- Soient $\tau \in]0,1[$, $x \in L^2([0,1])$, le quantile conditionnel $\langle \alpha_\tau, x \rangle$ sachant $X = x$ d'ordre τ est défini par :

$$\mathbb{P}\left(Y \leq \langle \alpha_\tau, X \rangle | X = x\right) = \tau$$

Définition des quantiles conditionnels - Modèle

- Données : $(X_i, Y_i)_{i=1,\dots,n}$ couples de variables aléatoires i.i.d. de même loi que (X, Y) avec $Y \in \mathbb{R}$ et $X \in L^2([0,1])$
- Soient $\tau \in]0,1[$, $x \in L^2([0,1])$, le quantile conditionnel $\langle \alpha_\tau, x \rangle$ sachant $X = x$ d'ordre τ est défini par :

$$\mathbb{P}\left(Y \leq \langle \alpha_\tau, X \rangle | X = x\right) = \tau$$

- Propriété :

$$\langle \alpha_\tau, x \rangle = \arg\min_{a \in \mathbb{R}} \mathbb{E}\left(l_\tau(Y - a) | X = x\right)$$

avec :

$$l_\tau(u) = |u| + (2\tau - 1)u$$

Définition des quantiles conditionnels - Modèle

- Données : $(X_i, Y_i)_{i=1,\dots,n}$ couples de variables aléatoires i.i.d. de même loi que (X, Y) avec $Y \in \mathbb{R}$ et $X \in L^2([0,1])$
- Soient $\tau \in]0,1[$, $x \in L^2([0,1])$, le quantile conditionnel $\langle \alpha_\tau, x \rangle$ sachant $X = x$ d'ordre τ est défini par :

$$\mathbb{P}\left(Y \leq \langle \alpha_\tau, X \rangle | X = x\right) = \tau$$

- Propriété :

$$\langle \alpha_\tau, x \rangle = \arg\min_{a \in \mathbb{R}} \mathbb{E}\left(l_\tau(Y - a) | X = x\right)$$

avec :

$$l_\tau(u) = |u| + (2\tau - 1)u$$

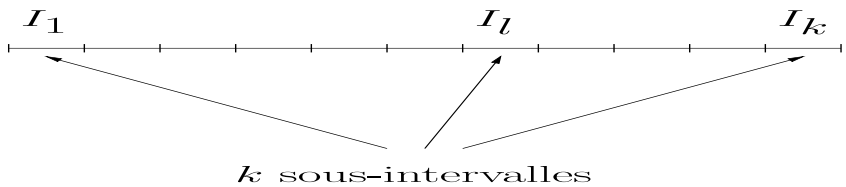
- Référence : Koenker et Bassett (1978)

Estimation

- But : estimation de α_τ par splines de régression

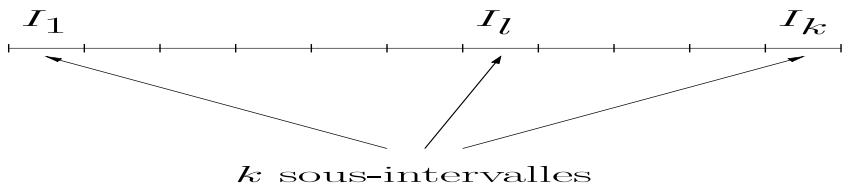
Estimation

- But : estimation de α_τ par splines de régression
- On se donne $k \in \mathbb{N}^\star$, $q \in \mathbb{N}$



Estimation

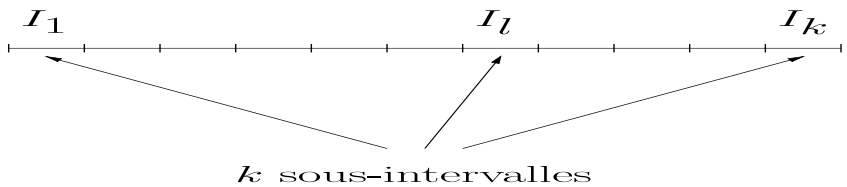
- But : estimation de α_τ par splines de régression
- On se donne $k \in \mathbb{N}^\star$, $q \in \mathbb{N}$



- $\mathbf{B}_{k,q} = (B_1, \ldots, B_{k+q})^T$ base (B-splines)

Estimation

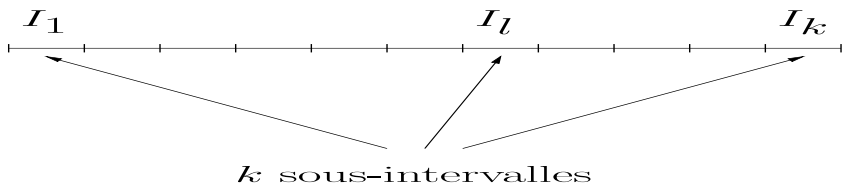
- But : estimation de α_τ par splines de régression
- On se donne $k \in \mathbb{N}^\star$, $q \in \mathbb{N}$



- $\mathbf{B}_{k,q} = (B_1, \ldots, B_{k+q})^T$ base (B-splines)
- $\widehat{\alpha}_\tau = \mathbf{B}_{k,q}^T \widehat{\boldsymbol{\theta}} = \sum_{j=1}^{k+q} \widehat{\theta}_j B_j$

Estimation

- But : estimation de α_τ par splines de régression
- On se donne $k \in \mathbb{N}^\star$, $q \in \mathbb{N}$



- $\mathbf{B}_{k,q} = (B_1, \ldots, B_{k+q})^T$ base (B-splines)
- $\widehat{\alpha}_\tau = \mathbf{B}_{k,q}^T \widehat{\boldsymbol{\theta}} = \sum_{j=1}^{k+q} \widehat{\theta}_j B_j$

Estimation de $\boldsymbol{\theta}$

- $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\tau(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle) + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

Estimation de $\boldsymbol{\theta}$

- $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\tau(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle) + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

Estimation de $\boldsymbol{\theta}$

- $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\tau(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle) + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

Estimation de $\boldsymbol{\theta}$

- $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l_\tau(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle) + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

- Pas de solution explicite (algorithmes de résolution)

Résultat de convergence (1)

Notations :

- $\Gamma_{X,n}$: version empirique de l'opérateur de covariance Γ_X

Résultat de convergence (1)

Notations :

- $\Gamma_{X,n}$: version empirique de l'opérateur de covariance Γ_X
- $\widehat{\mathbf{C}}_\rho = \left(\langle \Gamma_{X,n}(B_j), B_l \rangle + \rho \langle B_j^{(m)}, B_l^{(m)} \rangle \right)_{j,l=1,\ldots,k+q}$

Résultat de convergence (1)

Notations :

- $\Gamma_{X,n}$: version empirique de l'opérateur de covariance Γ_X
- $\widehat{\mathbf{C}}_\rho = \left(\langle \Gamma_{X,n}(B_j), B_l \rangle + \rho \langle B_j^{(m)}, B_l^{(m)} \rangle \right)_{j,l=1,\ldots,k+q}$
- On considère (η_n) telle que

$$\mathbb{P}\left(\lambda_{\min}(\widehat{\mathbf{C}}_{\boldsymbol{\rho}}) > c\eta_n \right) \xrightarrow[n \to +\infty]{} 1$$

Résultat de convergence (2)

Hypothèses :

- $(A.1)\ \|X\|_{L^2} \leq C_0 < +\infty, \quad p.s.$

Résultat de convergence (2)

Hypothèses :

- $(A.1)$ $\|X\|_{L^2} \leq C_0 < +\infty, \quad p.s.$
- $(A.2)$ La fonction α_τ est p' fois dérivable et $\alpha_\tau^{(p')}$ vérifie :

$$\left|\alpha_\tau^{(p')}(t) - \alpha_\tau^{(p')}(s)\right| \leq C_1 |t-s|^\nu, \quad s,t \in [0,1],$$

avec $C_1 > 0$ et $\nu \in [0,1]$. Dans la suite, on pose $p = p' + \nu$ et on suppose que $q \geq p \geq m$.

Résultat de convergence (2)

Hypothèses :

- $(A.1)$ $\|X\|_{L^2} \leq C_0 < +\infty, \quad p.s.$
- $(A.2)$ La fonction α_τ est p' fois dérivable et $\alpha_\tau^{(p')}$ vérifie :

$$\left|\alpha_\tau^{(p')}(t) - \alpha_\tau^{(p')}(s)\right| \leq C_1 |t-s|^\nu, \quad s,t \in [0,1],$$

avec $C_1 > 0$ et $\nu \in [0,1]$. Dans la suite, on pose $p = p' + \nu$ et on suppose que $q \geq p \geq m$.

- $(A.3)$ Les valeurs propres de Γ_X sont strictement positives.

Résultat de convergence (2)

Hypothèses :

- $(A.1)$ $\|X\|_{L^2} \leq C_0 < +\infty, \quad p.s.$
- $(A.2)$ La fonction α_τ est p' fois dérivable et $\alpha_\tau^{(p')}$ vérifie :

$$\left|\alpha_\tau^{(p')}(t) - \alpha_\tau^{(p')}(s)\right| \leq C_1|t-s|^\nu, \quad s,t \in [0,1],$$

avec $C_1 > 0$ et $\nu \in [0,1]$. Dans la suite, on pose $p = p' + \nu$ et on suppose que $q \geq p \geq m$.

- $(A.3)$ Les valeurs propres de Γ_X sont strictement positives.
- $(A.4)$ Pour $x \in L^2$, la variable aléatoire $\epsilon = Y - \langle \alpha_\tau, X \rangle$ a une densité conditionnelle f_x sachant $X = x$, continue et bornée inférieurement par une constante strictement positive, uniformément par rapport à $x \in L^2$.

Résultat de convergence (3)

Théorème (cf. Cardot, Crambes et Sarda, 2005) :
Sous $(A.1) - (A.4)$, si $k = k_n \sim n^{\beta}$ et $\rho = \rho_n \sim n^{(\gamma-1)/2}$ $(\beta, \gamma \in]0, 1[)$:

Résultat de convergence (3)

Théorème (cf. Cardot, Crambes et Sarda, 2005) :
Sous $(A.1) - (A.4)$, si $k = k_n \sim n^{\beta}$ et $\rho = \rho_n \sim n^{(\gamma-1)/2}$ $(\beta, \gamma \in]0,1[)$:

- $\widehat{\alpha}_\tau$ existe et est unique sur un espace dont la probabilité tend vers 1 lorsque n tend vers $+\infty$.
- On a :

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{1}{n\eta_n} + \frac{\rho_n^2}{k_n\eta_n} + \rho_n k_n^{2(m-p)}\right)$$

avec $\|u\|^2_{\Gamma_X} = \langle \Gamma_X u, u\rangle$

Commentaires et perspectives

- Conséquences : pour $\eta_n \sim \rho_n / k_n$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho_n} + \rho_n + \rho_n k_n^{2(m-p)}\right)$$

Commentaires et perspectives

- Conséquences : pour $\eta_n \sim \rho_n / k_n$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho_n} + \rho_n + \rho_n k_n^{2(m-p)}\right)$$

et pour $k_n \sim n^{1/(4p+1)}$ et $\rho_n \sim n^{-2p/(4p+1)}$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(n^{-2p/(4p+1)}\right)$$

Commentaires et perspectives

- Conséquences : pour $\eta_n \sim \rho_n / k_n$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho_n} + \rho_n + \rho_n k_n^{2(m-p)}\right)$$

et pour $k_n \sim n^{1/(4p+1)}$ et $\rho_n \sim n^{-2p/(4p+1)}$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(n^{-2p/(4p+1)}\right)$$

- Amélioration de la vitesse ?

Commentaires et perspectives

- Conséquences : pour $\eta_n \sim \rho_n / k_n$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho_n} + \rho_n + \rho_n k_n^{2(m-p)}\right)$$

et pour $k_n \sim n^{1/(4p+1)}$ et $\rho_n \sim n^{-2p/(4p+1)}$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(n^{-2p/(4p+1)}\right)$$

- Amélioration de la vitesse ?
- Extensions : cas d'une variable d'intérêt multivariée ou fonctionnelle, cas de X_i dépendants, ...

Commentaires et perspectives

- Conséquences : pour $\eta_n \sim \rho_n / k_n$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho_n} + \rho_n + \rho_n k_n^{2(m-p)}\right)$$

et pour $k_n \sim n^{1/(4p+1)}$ et $\rho_n \sim n^{-2p/(4p+1)}$:

$$\|\widehat{\alpha}_\tau - \alpha_\tau\|^2_{\Gamma_X} = O_{\mathbb{P}}\left(n^{-2p/(4p+1)}\right)$$

- Amélioration de la vitesse ?
- Extensions : cas d'une variable d'intérêt multivariée ou fonctionnelle, cas de X_i dépendants, ...
- Perspectives : autres méthodes d'estimation (Fourier, ondelettes, noyau, ...)

Plan de l'exposé

Introduction

1-Estimation de quantiles conditionnels

2-Estimation de la moyenne conditionnelle (cas non bruité)

3-Estimation de la moyenne conditionnelle (cas bruité)

4-Application à la prévision de pics de pollution

Plan de l'exposé

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i, \qquad i = 1, \ldots, n$$

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i, \qquad i = 1, \ldots, n$$

- $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma_\epsilon^2$, $\mathbb{E}(\epsilon | X) = 0$

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i, \qquad i = 1, \ldots, n$$

- $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma_\epsilon^2$, $\mathbb{E}(\epsilon | X) = 0$
- En pratique, les courbes sont observées en $t_1 < \ldots < t_p$ (équirépartis)

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i, \qquad i = 1, \ldots, n$$

- $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma_\epsilon^2$, $\mathbb{E}(\epsilon | X) = 0$
- En pratique, les courbes sont observées en $t_1 < \ldots < t_p$ (équirépartis)
- But : donner un estimateur de α basé sur les splines de lissage avec les observations $(X_i, Y_i)_{i=1,\ldots,n}$

Fonctions splines naturelles

- $NS^m(t_1, \ldots, t_p)$: espace des splines naturelles de degré $2m-1$ ($m \in \mathbb{N}$) de nœuds en $t_1, \ldots, t_p$
- $(b_1, \ldots, b_p)^T$ base de $NS^m(t_1, \ldots, t_p)$

Fonctions splines naturelles

- $NS^m(t_1, \ldots, t_p)$: espace des splines naturelles de degré $2m-1$ ($m \in \mathbb{N}$) de nœuds en $t_1, \ldots, t_p$
- $(b_1, \ldots, b_p)^T$ base de $NS^m(t_1, \ldots, t_p)$
- $\mathbf{b}(t) = (b_1(t), \ldots, b_p(t))^T$, $\mathbf{B} = (b_i(t_j))_{i,j=1,\ldots,p}$

Fonctions splines naturelles

- $NS^m(t_1, \ldots, t_p)$: espace des splines naturelles de degré $2m-1$ ($m \in \mathbb{N}$) de nœuds en $t_1, \ldots, t_p$
- $(b_1, \ldots, b_p)^T$ base de $NS^m(t_1, \ldots, t_p)$
- $\mathbf{b}(t) = (b_1(t), \ldots, b_p(t))^T$, $\mathbf{B} = (b_i(t_j))_{i,j=1,\ldots,p}$
- Pour tout $\mathbf{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^p$, il existe une unique interpolation spline $s_{\mathbf{w}}$:

$$s_{\mathbf{w}}(t) = \mathbf{b}(t)^T (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{w}$$

Fonctions splines naturelles

- $NS^m(t_1, \ldots, t_p)$: espace des splines naturelles de degré $2m-1$ ($m \in \mathbb{N}$) de nœuds en $t_1, \ldots, t_p$
- $(b_1, \ldots, b_p)^T$ base de $NS^m(t_1, \ldots, t_p)$
- $\mathbf{b}(t) = (b_1(t), \ldots, b_p(t))^T$, $\mathbf{B} = (b_i(t_j))_{i,j=1,\ldots,p}$
- Pour tout $\mathbf{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^p$, il existe une unique interpolation spline $s_{\mathbf{w}}$:

$$s_{\mathbf{w}}(t) = \mathbf{b}(t)^T (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{w}$$

- Référence : Eubank (1988)

Estimation de α (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \frac{1}{p}\mathbf{X}\mathbf{a} \right\|^2 + \rho \int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt \right\}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (X_i(t_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$

Estimation de α (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \frac{1}{p}\mathbf{X}\mathbf{a} \right\|^2 + \rho \int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt \right\}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (X_i(t_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$

Estimation de α (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \frac{1}{p}\mathbf{X}\mathbf{a} \right\|^2 + \rho \int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt \right\}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (X_i(t_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$

- ρ : paramètre de lissage
- $s_{\mathbf{a}}$: interpolation spline de $\mathbf{a} \in \mathbb{R}^p$

Estimation de α (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \frac{1}{p}\mathbf{X}\mathbf{a} \right\|^2 + \rho \int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt \right\}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (X_i(t_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$

- ρ : paramètre de lissage
- $s_{\mathbf{a}}$: interpolation spline de $\mathbf{a} \in \mathbb{R}^p$
- $\displaystyle\int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt = \frac{1}{p}\mathbf{a}^T \mathbf{A}_m^* \mathbf{a}$

Estimation de α (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \frac{1}{p}\mathbf{X}\mathbf{a} \right\|^2 + \rho \int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt \right\}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (X_i(t_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$

- ρ : paramètre de lissage
- $s_{\mathbf{a}}$: interpolation spline de $\mathbf{a} \in \mathbb{R}^p$
- $\displaystyle\int_0^1 s_{\mathbf{a}}^{(m)}(t)^2 dt = \frac{1}{p}\mathbf{a}^T \mathbf{A}_m^* \mathbf{a}$
- Solution explicite :

$$\widehat{\alpha}_{FLS,X}^* = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\rho}{p}\mathbf{A}_m^*\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Estimation de α (2)

- $\mathbf{A}_m^*$ a m valeurs propres nulles

Estimation de α (2)

- $\mathbf{A}_m^*$ a m valeurs propres nulles
- E_m : sous-espace propre correspondant
- $\mathbf{P}_m$: matrice de projection sur E_m
- $\mathbf{A}_m = \mathbf{P}_m + \mathbf{A}_m^*$

Estimation de α (2)

- $\mathbf{A}_m^*$ a m valeurs propres nulles
- E_m : sous-espace propre correspondant
- $\mathbf{P}_m$: matrice de projection sur E_m
- $\mathbf{A}_m = \mathbf{P}_m + \mathbf{A}_m^*$
- On considère :

$$\widehat{\alpha}_{FLS,X} = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\rho}{p}\mathbf{A}_m\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Estimation de α (2)

- $\mathbf{A}_m^*$ a m valeurs propres nulles
- E_m : sous-espace propre correspondant
- $\mathbf{P}_m$: matrice de projection sur E_m
- $\mathbf{A}_m = \mathbf{P}_m + \mathbf{A}_m^*$
- On considère :

$$\widehat{\boldsymbol{\alpha}}_{FLS,X} = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\rho}{p}\mathbf{A}_m\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

- Estimation de α :

$$\widehat{\alpha}_{FLS,X} = s_{\widehat{\boldsymbol{\alpha}}_{FLS,X}}$$

Résultat de convergence

Notation :

- $\|\mathbf{u}\|^2_{\Gamma_{X,n,p}} = \frac{1}{p}\mathbf{u}^T \left(\frac{1}{np}\mathbf{X}^T\mathbf{X}\right)\mathbf{u}$

Résultat de convergence

Notation :

- $\|\mathbf{u}\|^2_{\Gamma_{X,n,p}} = \frac{1}{p}\mathbf{u}^T \left(\frac{1}{np}\mathbf{X}^T\mathbf{X}\right)\mathbf{u}$

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$

Résultat de convergence

Notation :

- $\|\mathbf{u}\|^2_{\Gamma_{X,n,p}} = \frac{1}{p}\mathbf{u}^T \left(\frac{1}{np}\mathbf{X}^T\mathbf{X}\right)\mathbf{u}$

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\ldots,n} \sup_{j=1,\ldots,p} |X_i(t_j)| \leq C_1$$

Résultat de convergence

Notation :

- $\|\mathbf{u}\|^2_{\Gamma_{X,n,p}} = \frac{1}{p}\mathbf{u}^T \left(\frac{1}{np}\mathbf{X}^T\mathbf{X}\right)\mathbf{u}$

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\ldots,n} \sup_{j=1,\ldots,p} |X_i(t_j)| \leq C_1$$

Théorème (cf. Cardot, Crambes, Kneip et Sarda, 2006) :
Sous $(H.1) - (H.2)$, pour p assez grand :

$$\|\widehat{\boldsymbol{\alpha}}_{FLS,X} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(\frac{1}{n\rho} + \rho\right)$$

- Conséquence : pour $\rho \sim n^{-1/2}$

$$\|\widehat{\boldsymbol{\alpha}}_{FLS,X} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

Commentaires et perspectives

- Conséquence : pour $\rho \sim n^{-1/2}$

$$\|\widehat{\boldsymbol{\alpha}}_{FLS,X} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

- Amélioration de la vitesse
- Résultats de vitesse sur $\widehat{\alpha}_{FLS,X}$ et prise en compte de l'approximation du produit scalaire dans le modèle ($\langle \alpha, X_i \rangle = \frac{1}{p}\sum_{j=1}^{p} \alpha(t_j)X_i(t_j) + d_i$)

} (travaux en cours avec A. Kneip et P. Sarda)

Plan de l'exposé

Introduction

1-Estimation de quantiles conditionnels

2-Estimation de la moyenne conditionnelle (cas non bruité)

3-Estimation de la moyenne conditionnelle (cas bruité)

4-Application à la prévision de pics de pollution

Plan de l'exposé

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i$$

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i$$

$$W_i(t_j) = X_i(t_j) + \delta_{ij}$$

- $(\delta_{ij})_{i=1,\dots,n, j=1,\dots,p}$ est une suite de variables aléatoires i.i.d., $\mathbb{E}(\delta_{ij}) = 0$, $\mathbb{E}(\delta_{ij}^2) = \sigma_\delta^2$

Présentation du modèle

- Modèle :

$$Y_i = \langle X_i, \alpha \rangle + \epsilon_i$$

$$W_i(t_j) = X_i(t_j) + \delta_{ij}$$

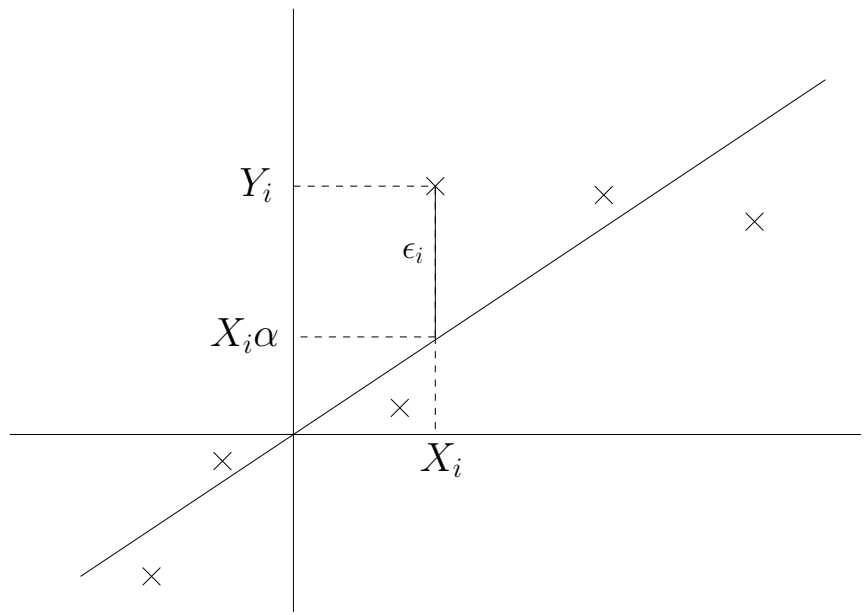
- $(\delta_{ij})_{i=1,\dots,n, j=1,\dots,p}$ est une suite de variables aléatoires i.i.d., $\mathbb{E}(\delta_{ij}) = 0$, $\mathbb{E}(\delta_{ij}^2) = \sigma_\delta^2$
- But : donner un estimateur de α basé sur les splines de lissage avec les observations $(W_i, Y_i)_{i=1,\dots,n}$

Cas multivarié (1) : $\mathbf{X}_i \in \mathbb{R}^p$

- problème de minimisation (cas non bruité) :

$$\min_{\mathbf{a}\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T \mathbf{a} \right)^2 \right\}$$

MC Ord

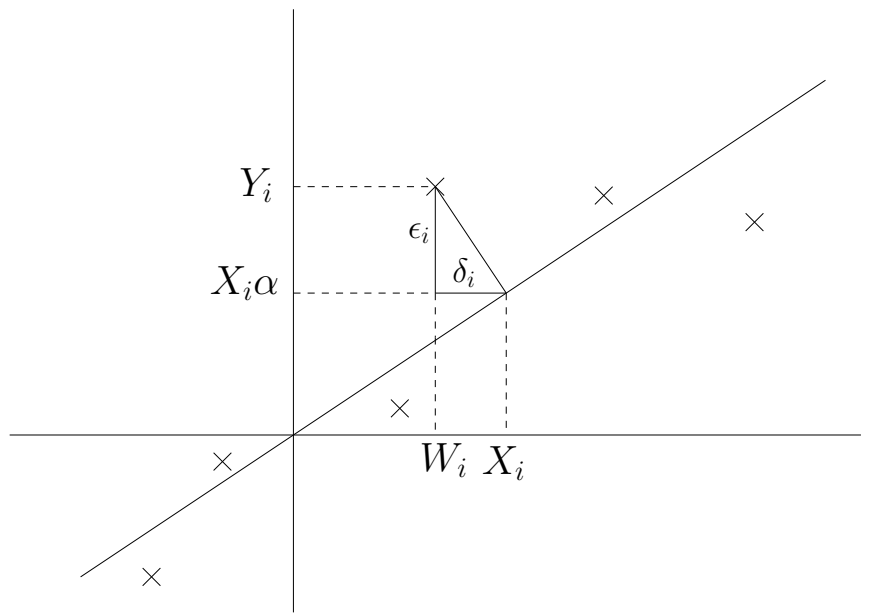


Cas multivarié (1) : $\mathbf{X}_i \in \mathbb{R}^p$

- problème de minimisation (cas bruité) :

$$\min_{\mathbf{a}\in\mathbb{R}^p, \mathbf{X}_i\in\mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\left(Y_i - \mathbf{X}_i^T \mathbf{a} \right)^2 + \|\mathbf{W}_i - \mathbf{X}_i\|^2 \right] \right\}$$

MC Orth



Cas multivarié (2)

- Écriture explicite de $\widehat{\boldsymbol{\alpha}}_{TLS}$:

$$\widehat{\boldsymbol{\alpha}}_{TLS} = \left(\mathbf{W}^T\mathbf{W} - \sigma_k^2\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

Cas multivarié (2)

- Écriture explicite de $\widehat{\boldsymbol{\alpha}}_{TLS}$:

$$\widehat{\boldsymbol{\alpha}}_{TLS} = \left(\mathbf{W}^T\mathbf{W} - \sigma_k^2 \mathbf{I}_p\right)^{-1} \mathbf{W}^T\mathbf{Y}$$

- $-\sigma_k^2 \mathbf{I}_p$: correction

Cas multivarié (2)

- Écriture explicite de $\widehat{\boldsymbol{\alpha}}_{TLS}$:

$$\widehat{\boldsymbol{\alpha}}_{TLS} = \left(\mathbf{W}^T\mathbf{W} - \sigma_k^2\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

- $-\sigma_k^2\mathbf{I}_p$: correction
- σ_k^2 est la plus petite valeur propre non nulle de la matrice $(\mathbf{W},\mathbf{Y})^T(\mathbf{W},\mathbf{Y})$

Cas multivarié (2)

- Écriture explicite de $\widehat{\boldsymbol{\alpha}}_{TLS}$:

$$\widehat{\boldsymbol{\alpha}}_{TLS} = \left(\mathbf{W}^T\mathbf{W} - \sigma_k^2\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

- $-\sigma_k^2\mathbf{I}_p$: correction
- σ_k^2 est la plus petite valeur propre non nulle de la matrice $(\mathbf{W},\mathbf{Y})^T(\mathbf{W},\mathbf{Y})$
- Référence : Van Huffel et Vandewalle (1991)

Extension au cas fonctionnel (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i - \frac{1}{p}\mathbf{X}_i^T\mathbf{a}\right)^2 + \frac{1}{p}\left\|\mathbf{X}_i - \mathbf{W}_i\right\|^2\right] + \frac{\rho}{p}\mathbf{a}^T\mathbf{A}_m\mathbf{a}\right\}$$

où $\mathbf{X}_i = (X_i(t_1),\ldots,X_i(t_p))^T$, $\mathbf{W}_i = (W_i(t_1),\ldots,W_i(t_p))^T$

Extension au cas fonctionnel (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left[\left(Y_i - \frac{1}{p}\mathbf{X}_i^T\mathbf{a} \right)^2 + \frac{1}{p}\left\|\mathbf{X}_i - \mathbf{W}_i\right\|^2 \right] + \frac{\rho}{p}\mathbf{a}^T\mathbf{A}_m\mathbf{a} \right\}$$

où $\mathbf{X}_i = (X_i(t_1),\ldots,X_i(t_p))^T$, $\mathbf{W}_i = (W_i(t_1),\ldots,W_i(t_p))^T$

Extension au cas fonctionnel (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i - \frac{1}{p}\mathbf{X}_i^T\mathbf{a}\right)^2 + \frac{1}{p}\left\|\mathbf{X}_i - \mathbf{W}_i\right\|^2\right] + \frac{\rho}{p}\mathbf{a}^T\mathbf{A}_m\mathbf{a}\right\}$$

où $\mathbf{X}_i = (X_i(t_1),\ldots,X_i(t_p))^T$, $\mathbf{W}_i = (W_i(t_1),\ldots,W_i(t_p))^T$

Extension au cas fonctionnel (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i - \frac{1}{p}\mathbf{X}_i^T\mathbf{a}\right)^2 + \frac{1}{p}\left\|\mathbf{X}_i - \mathbf{W}_i\right\|^2\right] + \frac{\rho}{p}\mathbf{a}^T\mathbf{A}_m\mathbf{a} \right\}$$

où $\mathbf{X}_i = (X_i(t_1),\ldots,X_i(t_p))^T$, $\mathbf{W}_i = (W_i(t_1),\ldots,W_i(t_p))^T$

Extension au cas fonctionnel (1)

- Problème de minimisation :

$$\min_{\mathbf{a}\in\mathbb{R}^p,\mathbf{X}_i\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i-\frac{1}{p}\mathbf{X}_i^T\mathbf{a}\right)^2+\frac{1}{p}\left\|\mathbf{X}_i-\mathbf{W}_i\right\|^2\right]+\frac{\rho}{p}\mathbf{a}^T\mathbf{A}_m\mathbf{a}\right\}$$

où $\mathbf{X}_i=(X_i(t_1),\ldots,X_i(t_p))^T$, $\mathbf{W}_i=(W_i(t_1),\ldots,W_i(t_p))^T$

- Solution :

$$\widehat{\boldsymbol{\alpha}}^*_{FTLS}=\frac{1}{np}\left(\frac{1}{np^2}\mathbf{W}^T\mathbf{W}+\frac{\rho}{p}\mathbf{A}_m-\sigma_k^2\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

où σ_k^2 est la plus petite valeur propre de la matrice
$\frac{1}{n}\left(\frac{\mathbf{W}}{p},\mathbf{Y}\right)^T\left(\frac{\mathbf{W}}{p},\mathbf{Y}\right)+\frac{\rho}{p}\begin{pmatrix}\mathbf{A}_m & \mathbf{0}\\ \mathbf{0} & 0\end{pmatrix}$

Extension au cas fonctionnel (2)

- Problème (numérique) : instabilité de σ_k^2

Extension au cas fonctionnel (2)

- Problème (numérique) : instabilité de σ_k^2
- Proposition : on a

$$\frac{1}{np^2}\mathbf{W}^T\mathbf{W} = \frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\sigma_\delta^2}{p^2}\mathbf{I}_p + \mathbf{R}$$

avec $\|\mathbf{R}\| = O_P\left(\frac{1}{n^{1/2}p}\right)$

Extension au cas fonctionnel (2)

- Problème (numérique) : instabilité de σ_k^2
- Proposition : on a

$$\frac{1}{np^2}\mathbf{W}^T\mathbf{W} = \frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\sigma_\delta^2}{p^2}\mathbf{I}_p + \mathbf{R}$$

avec $\|\mathbf{R}\| = O_P\left(\frac{1}{n^{1/2}p}\right)$

- Estimateur :

$$\widehat{\alpha}_{FTLS} = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{W}^T\mathbf{W} + \frac{\rho}{p}\mathbf{A}_m - \frac{{\sigma_\delta}^2}{p^2}\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

Extension au cas fonctionnel (2)

- Problème (numérique) : instabilité de σ_k^2
- Proposition : on a

$$\frac{1}{np^2}\mathbf{W}^T\mathbf{W} = \frac{1}{np^2}\mathbf{X}^T\mathbf{X} + \frac{\sigma_\delta^2}{p^2}\mathbf{I}_p + \mathbf{R}$$

avec $\|\mathbf{R}\| = O_P\left(\frac{1}{n^{1/2}p}\right)$

- Estimateur :

$$\widehat{\alpha}_{FTLS} = \frac{1}{np}\left(\frac{1}{np^2}\mathbf{W}^T\mathbf{W} + \frac{\rho}{p}\mathbf{A}_m - \frac{\widehat{\sigma}_\delta{}^2}{p^2}\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}$$

Résultat de convergence

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\ldots,n} \sup_{j=1,\ldots,p} |X_i(t_j)| \leq C_1$$

Résultat de convergence

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\dots,n} \sup_{j=1,\dots,p} |X_i(t_j)| \leq C_1$$

- $(H.3)$ il existe $C_2 > 0$ telle que $\mathbb{E}(\delta_{ij}^4) \leq C_2$

Résultat de convergence

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\dots,n} \sup_{j=1,\dots,p} |X_i(t_j)| \leq C_1$$

- $(H.3)$ il existe $C_2 > 0$ telle que $\mathbb{E}(\delta_{ij}^4) \leq C_2$
- $(H.4)$ Y_i et δ_{ij} sont indépendants

Résultat de convergence

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\dots,n} \sup_{j=1,\dots,p} |X_i(t_j)| \leq C_1$$

- $(H.3)$ il existe $C_2 > 0$ telle que $\mathbb{E}(\delta_{ij}^4) \leq C_2$
- $(H.4)$ Y_i et δ_{ij} sont indépendants
- $(H.5)$ il existe $C_3 > 0$ telle que, pour n et p assez grands, on a $p^{1/2} \left\| \frac{1}{np^2} \mathbf{X}^T \mathbf{X} \boldsymbol{\alpha} \right\| \geq D_6$

Résultat de convergence

Hypothèses :

- $(H.1)$ α est m fois dérivable et $\alpha^{(m)} \in L^2([0,1])$
- $(H.2)$ il existe $C_1 > 0$ telle que (en probabilité) :

$$\sup_{i=1,\dots,n} \sup_{j=1,\dots,p} |X_i(t_j)| \leq C_1$$

- $(H.3)$ il existe $C_2 > 0$ telle que $\mathbb{E}(\delta_{ij}^4) \leq C_2$
- $(H.4)$ Y_i et δ_{ij} sont indépendants
- $(H.5)$ il existe $C_3 > 0$ telle que, pour n et p assez grands, on a $p^{1/2} \left\| \frac{1}{np^2} \mathbf{X}^T \mathbf{X} \boldsymbol{\alpha} \right\| \geq D_6$

Théorème (cf. Cardot, Crambes, Kneip et Sarda, 2006) :
Sous $(H.1) - (H.5)$:

$$\|\widehat{\alpha}_{FTLS} - \widehat{\alpha}_{FLS,X}\|_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(\frac{1}{n^{1/2} p^{1/2} \rho^{1/2}} + \frac{1}{n^{1/2}} \right)$$

- Conséquence : pour p assez grand :

$$\|\widehat{\boldsymbol{\alpha}}_{FTLS} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

Commentaires et perspectives

- Conséquence : pour p assez grand :

$$\|\widehat{\boldsymbol{\alpha}}_{FTLS} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

- Travail réalisé aussi pour l'estimateur par splines de régression
- Simulations : résultats assez satisfaisants

Commentaires et perspectives

- Conséquence : pour p assez grand :

$$\|\widehat{\boldsymbol{\alpha}}_{FTLS} - \boldsymbol{\alpha}\|^2_{\Gamma_{X,n,p}} = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

- Travail réalisé aussi pour l'estimateur par splines de régression
- Simulations : résultats assez satisfaisants
- Autre piste : lisser les courbes X_i (lissage à noyau) pour supprimer le bruit (travail en cours)

Plan de l'exposé

Introduction

1-Estimation de quantiles conditionnels

2-Estimation de la moyenne conditionnelle (cas non bruité)

3-Estimation de la moyenne conditionnelle (cas bruité)

4-Application à la prévision de pics de pollution

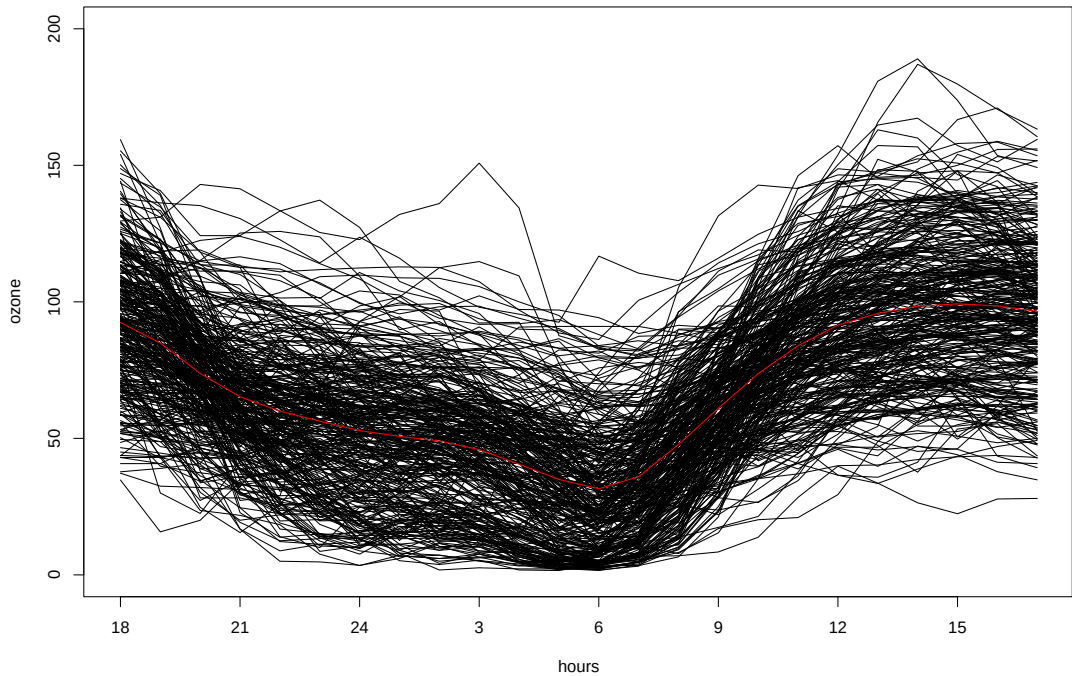
Plan de l'exposé

Présentation des données (1)

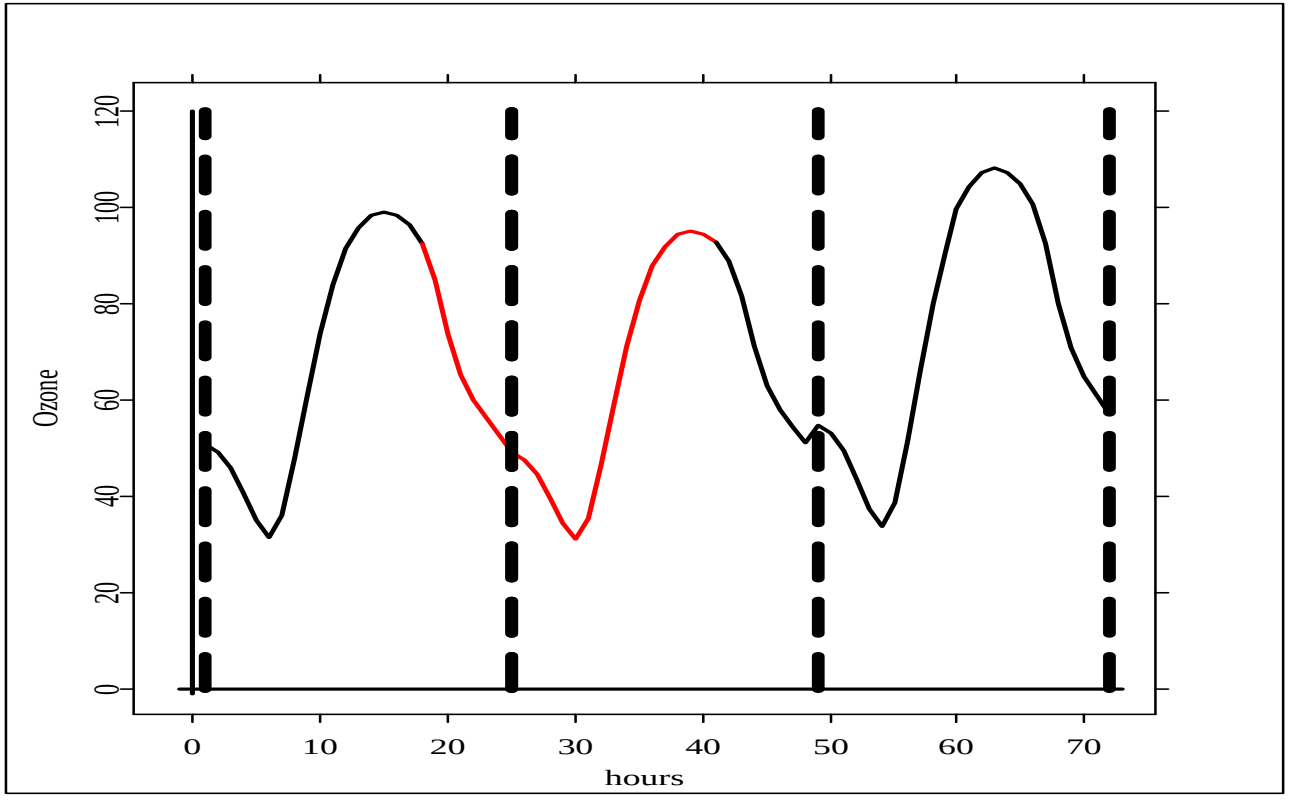
- 9 variables : $NO, N2, O3, DV, VV, \ldots$ (mesures horaires)
- 6 stations de mesure
- 4 années : 1997 – 2000 (15 Mai - 15 Septembre)

Présentation des données (1)

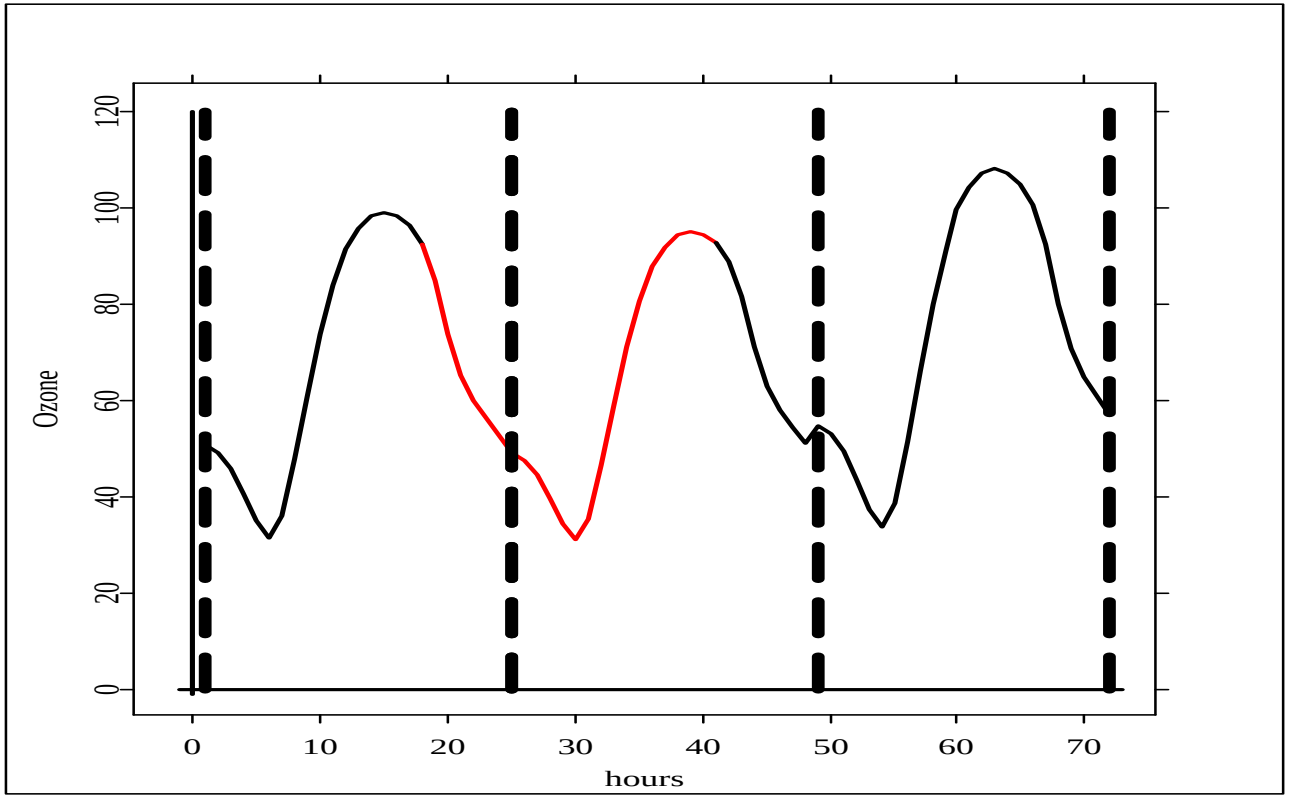
- 9 variables : $NO, N2, O3, DV, VV, \ldots$ (mesures horaires)
- 6 stations de mesure
- 4 années : 1997 – 2000 (15 Mai - 15 Septembre)



Présentation des données (2)



Présentation des données (2)



But : prévision du maximum de $O3$ (variable d'intérêt) connaissant une ou plusieurs variables la veille

Prévision par la moyenne conditionnelle (1)

- On a utilisé les splines de régression

Prévision par la moyenne conditionnelle (1)

- On a utilisé les splines de régression
- On construit $\widehat{\alpha} = \mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}}$ avec $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle)^2 + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

Prévision par la moyenne conditionnelle (1)

- On a utilisé les splines de régression
- On construit $\widehat{\alpha} = \mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}}$ avec $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation :

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \mathbf{B}_{k,q}^T\boldsymbol{\theta}, X_i\rangle)^2 + \rho \left\| (\mathbf{B}_{k,q}^T\boldsymbol{\theta})^{(m)} \right\|_{L^2}^2 \right\}$$

Référence : Cardot, Ferraty et Sarda (2003)

Prévision par la moyenne conditionnelle (2)

- Solution explicite : $\widehat{\theta} = \frac{1}{n}(\frac{1}{n}\mathbf{D}^T\mathbf{D} + \rho\mathbf{G})^{-1}\mathbf{D}^T\mathbf{Y}$

$$\mathbf{D} = (\langle B_j, X_i \rangle_{i=1,\ldots,n, j=1,\ldots,k+q})$$

$$\mathbf{G} = \left(\langle B_j^{(m)}, B_l^{(m)} \rangle_{j,l=1,\ldots,k+q}\right)$$

Prévision par la moyenne conditionnelle (2)

- Solution explicite : $\widehat{\theta} = \frac{1}{n}(\frac{1}{n}\mathbf{D}^T\mathbf{D} + \rho\mathbf{G})^{-1}\mathbf{D}^T\mathbf{Y}$

$$\mathbf{D} = (\langle B_j, X_i \rangle_{i=1,\ldots,n, j=1,\ldots,k+q})$$

$$\mathbf{G} = \left(\langle B_j^{(m)}, B_l^{(m)} \rangle_{j,l=1,\ldots,k+q}\right)$$

- Choix des paramètres : $k = 8,\ q = 3,\ m = 2$

Prévision par la moyenne conditionnelle (2)

- Solution explicite : $\widehat{\theta} = \frac{1}{n}(\frac{1}{n}\mathbf{D}^T\mathbf{D} + \rho\mathbf{G})^{-1}\mathbf{D}^T\mathbf{Y}$

$$\mathbf{D} = (\langle B_j, X_i \rangle_{i=1,\ldots,n, j=1,\ldots,k+q})$$

$$\mathbf{G} = \left(\langle B_j^{(m)}, B_l^{(m)} \rangle_{j,l=1,\ldots,k+q}\right)$$

- Choix des paramètres : $k = 8$, $q = 3$, $m = 2$
- Choix de ρ : validation croisée généralisée (cf. Wahba, 1990)

Prévision par la moyenne conditionnelle (2)

- Solution explicite : $\widehat{\theta} = \frac{1}{n}(\frac{1}{n}\mathbf{D}^T\mathbf{D} + \rho\mathbf{G})^{-1}\mathbf{D}^T\mathbf{Y}$

$$\mathbf{D} = (\langle B_j, X_i \rangle_{i=1,\ldots,n, j=1,\ldots,k+q})$$

$$\mathbf{G} = \left(\langle B_j^{(m)}, B_l^{(m)} \rangle_{j,l=1,\ldots,k+q}\right)$$

- Choix des paramètres : $k = 8$, $q = 3$, $m = 2$
- Choix de ρ : validation croisée généralisée (cf. Wahba, 1990)
- Modèle avec plusieurs variables explicatives : algorithme backfitting

Prévision par les quantiles conditionnels

- On a utilisé les splines de régression

Prévision par les quantiles conditionnels

- On a utilisé les splines de régression
- Pas de solution explicite : algorithme de moindres carrés itérés pondérés

Prévision par les quantiles conditionnels

- On a utilisé les splines de régression
- Pas de solution explicite : algorithme de moindres carrés itérés pondérés
- Choix des paramètres : comme pour l'estimation par la moyenne conditionnelle

Prévision par les quantiles conditionnels

- On a utilisé les splines de régression
- Pas de solution explicite : algorithme de moindres carrés itérés pondérés
- Choix des paramètres : comme pour l'estimation par la moyenne conditionnelle
- Modèle avec plusieurs variables explicatives : algorithme backfitting

Qualité des modèles

- Échantillon d'apprentissage : $(X_{a_i}, Y_{a_i})_{i=1,\ldots,n_a}$ $(n_a = 332)$
- Échantillon de test : $(X_{t_i}, Y_{t_i})_{i=1,\ldots,n_t}$ $(n_t = 142)$

Qualité des modèles

- Échantillon d'apprentissage : $(X_{a_i}, Y_{a_i})_{i=1,\dots,n_a}$ $(n_a = 332)$
- Échantillon de test : $(X_{t_i}, Y_{t_i})_{i=1,\dots,n_t}$ $(n_t = 142)$
- Prévision par la moyenne conditionnelle :

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y}_{t_i})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_a)^2}$$

Qualité des modèles

- Échantillon d'apprentissage : $(X_{a_i}, Y_{a_i})_{i=1,\dots,n_a}$ $(n_a = 332)$
- Échantillon de test : $(X_{t_i}, Y_{t_i})_{i=1,\dots,n_t}$ $(n_t = 142)$
- Prévision par la moyenne conditionnelle :

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y}_{t_i})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_a)^2}$$

- Prévision par la médiane conditionnelle :

$$C_2 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t} l_{0.5}(Y_{t_i} - \widehat{Y}_{t_i})}{\frac{1}{n_t}\sum_{i=1}^{n_t} l_{0.5}(Y_{t_i} - q_{0.5}(Y_a))}$$

Résultats (1)

Modèles	Variables	C_1	C_2
1 variable	N2	0.814	0.906
	O3	0.414	0.656
	VV	0.802	0.902
2 variables	O3, NO	0.413	0.643
	O3, N2	0.413	0.637
	O3, VV	0.414	0.635
3 variables	O3, NO, N2	0.412	0.644
	O3, N2, DV	0.409	0.645
	O3, N2, VV	0.410	0.642
4 variables	O3, NO, N2, VV	0.400	0.634
5 variables	O3, NO, N2, DV, VV	0.401	0.639

Résultats (1)

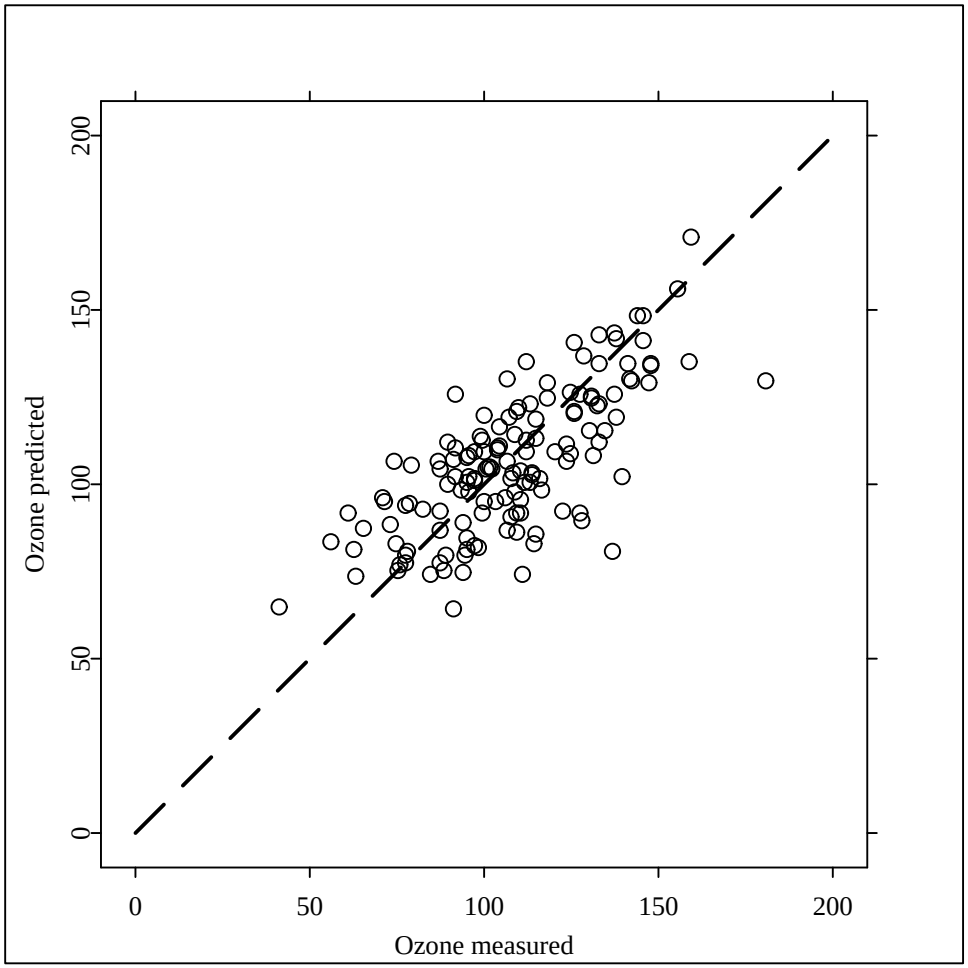
Modèles	Variables	C_1	C_2
1 variable	N2	0.814	0.906
	O3	**0.414**	**0.656**
	VV	0.802	0.902
2 variables	O3, NO	0.413	0.643
	O3, N2	0.413	0.637
	O3, VV	0.414	0.635
3 variables	O3, NO, N2	0.412	0.644
	O3, N2, DV	0.409	0.645
	O3, N2, VV	0.410	0.642
4 variables	O3, NO, N2, VV	0.400	0.634
5 variables	O3, NO, N2, DV, VV	0.401	0.639

Résultats (1)

Modèles	Variables	C_1	C_2
1 variable	N2	0.814	0.906
	O3	0.414	0.656
	VV	0.802	0.902
2 variables	O3, NO	0.413	0.643
	O3, N2	0.413	0.637
	O3, VV	0.414	0.635
3 variables	O3, NO, N2	0.412	0.644
	O3, N2, DV	0.409	0.645
	O3, N2, VV	0.410	0.642
4 variables	**O3, NO, N2, VV**	**0.400**	**0.634**
5 variables	O3, NO, N2, DV, VV	0.401	0.639

Résultats (2)

Maximum de *O3* prédit (médiane conditionnelle) en fonction du maximum de *O3* mesuré (variables explicatives : O3, NO, N2, VV)



Commentaires et perspectives

- Référence : Cardot, Crambes et Sarda (2006)
- Résultats satisfaisants (à comparer avec d'autres méthodes)

Commentaires et perspectives

- Référence : Cardot, Crambes et Sarda (2006)
- Résultats satisfaisants (à comparer avec d'autres méthodes)

- Utilisation des splines de lissage
- Prise en compte du bruit dans les variables explicatives

} (travaux en cours avec A. Kneip et P. Sarda)

Bibliographie

❶ Cardot, H., Crambes, C., Kneip, A. and Sarda, P. (2006). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis*, special issue on functional data analysis, to appear.

❷ Cardot, H., Crambes, C. and Sarda, P. (2005). Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics*, **17**, 841-856.

❸ Cardot, H., Crambes, C. and Sarda, P. (2006). Conditional quantiles with functional covariates : an application to ozone pollution forecasting. In *Applied Biostatistics : Case Studies and Interdisciplinary Methods*, Xplore e-book, to appear.