Données fonctionnelles avec erreur hétéroscédastique

Steven Golovkine[1] & Nicolas Klutchnikoff[2] & Valentin Patilea[3]

[1] *Renault, steven.s.golovkine@renault.com*
[2] *Université Rennes 2, nicolas.klutchnikoff@univ-rennes2.fr*
[3] *ENSAI, valentin.patilea@ensai.fr*

Résumé. Avec les récentes avancées technologiques, de plus en plus d'objets sont équipés de capteurs leur permettant, par exemple, de connaître la position d'autres objets dans son environnement. Ces capteurs fournissent un grand nombre de signaux pouvant être modélisés comme des données fonctionnelles multivariées entachées d'un bruit. Dans ce travail, nous supposons que ces données sont enregistrées avec un bruit hétéroscédastique d'échelle inconnue. Nous nous intéressons donc à l'estimation adaptatif du signal.

Mots-clés. Données fonctionnelles, Hétéroscédasticité, Estimation non-paramétrique

Abstract. With recent technological advances, more and more objects are equipped with sensors that allow them, for example, to know the position of other objects in their environment. These sensors provide a large amount of data that can be modelled as multivariate functional data. We assume that these data are recorded with a heteroscedastic noise of unknown scale. In this work, we are therefore interested in estimating the signal of interest.

Keywords. Functional data, Heteroscedasticity, Non-parametric estimation

Les capteurs sont de plus en plus présents dans notre vie quotidienne. Ceux-ci fournissent un grand nombre de données pouvant être modélisées comme données fonctionnelles multivariées. Comme ces capteurs ne sont pas parfait, il est raisonable de supposer que les données enregistrées le soient avec un bruit. Ce bruit n'étant pas forcément homoscédastique. Par exemple, lorsque le capteur reçoit la donnée de position d'un objet dans son alentour, il est naturel que l'erreur du capteur soit dépendant de la distance au capteur. En effet, plus l'objet à détecter sera loin du capteur, plus l'approximation du capteur sera grande. Cependant, la précision des capteurs étant de plus en plus élevée, il est important d'envisager que la variance du bruit puisse être très faible.

Nous nous intéressons donc à l'étude de réalisations indépendantes d'un processus stochastique

$$\mathbf{Z} = \left(Z^{(p)}(t) : t \in \mathcal{T}, 1 \leq p \leq P\right) \in \mathbb{R}^P,$$

defini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et indexé par $t \in \mathcal{T}$, un intervalle compact sur la droite des réels. Le processus $\mathbf{Z}$ est P-dimensionnel avec P un entier positif

donné. Nous supposons le modèle suivant pour des réalisations indépendantes $\mathbf{Z}_i = \left(Z_i^{(p)}(\cdot) : 1 \leq p \leq P\right)$, $1 \leq i \leq n$ de $\mathbf{Z}$:

$$Z_i^{(p)}(t) = \eta^{(p)}(t) + v_i^{(p)}(t) + \epsilon_i^{(p)}(t), \quad t \in \mathcal{T}. \tag{1}$$

Les trajectoires P-dimensionnelles, $\left(v_i^{(1)}(\cdot), \ldots, v_i^{(P)}(\cdot)\right)$ et $\left(\epsilon_i^{(1)}(\cdot), \ldots, \epsilon_i^{(P)}(\cdot)\right)$, sont des réalisations indépendantes non-observées des processus stochastiques multivariés de moyenne nulle $\mathbf{v} = \left(v_i^{(p)}(\cdot) : 1 \leq p \leq P\right)$ et $\mathbf{e} = \left(\epsilon_i^{(p)}(\cdot) : 1 \leq p \leq P\right)$. Nous supposons que ces processus sont indépendants. Les processus $\eta^{(p)}, 1 \leq p \leq P$, sont des processus déterministes.

Définissons

$$\gamma_{\mathbf{v}}(s,t) = C_{p,p'}(s,t) = \mathbb{E}(v^{(p)}(s)v^{(p')}(t)), \quad s,t \in \mathcal{T}, 1 \leq p, p' \leq P,$$

la structure de covariance du processus multivarié $\mathbf{v}$. Ainsi, les fonctions sous-jacentes $f_i^{(p)}(t) = \mathbb{E}\left(Z_i^{(p)}(t) \mid v_i^{(p)}(t)\right) = \eta^{(p)}(t) + v_i^{(p)}(t)$ sont des réalisations indépendantes du processus stochastique sous-jacent $f^{(p)}(t) = \eta^{(p)}(t) + v^{(p)}(t)$ pour tout p. De plus, nous faisons l'hypothèse que les P composantes de $\mathbf{e}$ sont des processus deux à deux indépendants, avec un opérateur de covariance commun

$$\gamma_{\mathbf{e}} = \sigma_\epsilon^2\left(f_i^{(p)}(t)\right)\mathbf{1}_{\{s=t\}}, \quad s,t, \in \mathcal{T},$$

où $\sigma_\epsilon^2(\cdot)$ est une fonction inconnue qui peut tendre vers 0 avec la taille de l'échantillon n.

En pratique, les données ne sont pas enregistrées de façon continue. Notons $\mathcal{G}_i = \{t_{i,j} : 1 \leq j \leq m_i\}$ l'ensemble des points d'échantillonage de $\mathbf{Z}_i$. Ainsi, la discrétisation de l'équation (1) s'écrit :

$$Z_{i,j}^{(p)} = \eta^{(p)}(t_{i,j}) + v_i^{(p)}(t_{i,j}) + \epsilon_{i,j}^{(p)}, \quad 1 \leq j \leq m_i, 1 \leq i \leq n,$$

avec $Z_{i,j}^{(p)} = Z_i^{(p)}(t_{i,j})$ et $\epsilon_{i,j}^{(p)} = \epsilon_i^{(p)}(t_{i,j})$. L'analyse de ces données aura pour but d'estimer les fonctions moyennes $\eta^{(p)}, 1 \leq p \leq P$ et d'identifier les structures de covariance $\gamma_{\mathbf{v}}(s,t)$ et $\gamma_{\mathbf{e}}(s,t)$. Dans certaines applications, les valeurs de m_i sont beaucoup plus petite que n ce qui amenerait à une estimation des structures de covariance $\gamma_{\mathbf{v}}$ et $\gamma_{\mathbf{e}}$ de très faible précision. Néanmoins, si σ_ϵ^2 est petit, nous allons améliorer l'estimation des signaux $f_i^{(p)}(\cdot)$.

Notre cadre est une extension de celui considéré par Zhang et Chen (2007) qui ont étudié un modèle correspondant au notre avec $P = 1$ et une fonction particulière pour σ_ϵ^2. Zhang et Chen considèrent le paradigme « *smoothing first, then estimation* ». Leur idée est donc de débruiter chaque courbe prise séparement, et ensuite d'estimer les différentes fonctions $\eta(t), \gamma_{\mathbf{v}}(s,t)$ et $\sigma_\epsilon^2(t)$ et d'en étudier leur comportement asymptotique. Supposons que $P = 1$ et considérons le modèle de régression non-paramétrique suivant :

$$Z_{i,j} = f_i(t) + \epsilon_{i,j}, \quad 1 \leq j \leq m_i, 1 \leq i \leq n.$$

Zhang et Chen se proposent d'estimer chaque $f_i(\cdot)$ par $\widehat{f}_i(\cdot)$ obtenue par polynômes locaux. Ensuite, l'estimation de $\eta(t)$ et de $\gamma_{\mathbf{v}}(s,t)$ se fait en utilisant les estimateurs classiques :

$$\widehat{\eta}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_i(t) \quad \text{et} \quad \widehat{\gamma}_{\mathbf{v}}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n} \left(\widehat{f}_i(s) - \widehat{\eta}(s)\right)\left(\widehat{f}_i(t) - \widehat{\eta}(t)\right).$$

La vitesse de convergence de $\widehat{f}_i(\cdot)$ dépend de l'ordre de grandeur des m_i pour lesquels il est supposé que m_i soit supérieur à Cn^{δ}, pour tout $1 \leq i \leq n$. Ici, C et δ sont des constantes positives. Zhang et Chen montrent que l'estimation des fonctions $\eta(t)$ et $\gamma_{\mathbf{v}}(s,t)$ peut se faire à une vitesse paramétrique si δ est plus grand que 1. Lorsque les m_i sont plus petits que n, c'est-à-dire lorsque δ est plus petit que 1, la vitesse de $\widehat{f}_i(\cdot)$ se détériore et $\eta(t)$ et $\gamma_{\mathbf{v}}(s,t)$ ne peuvent plus s'estimer à une vitesse paramétrique.

Dans notre cadre étendu et non-asymptotique, nous souhaitons laisser la fonction $\sigma_\epsilon^2(\cdot)$ dépendre du signal $f_i^{(p)}$ mais également de la taille de l'échantillon n. Notamment, si la norme uniforme de la fonction $\sigma_\epsilon^2(\cdot)$ tend vers 0 avec la taille de l'échantillon des signaux, la vitesse de convergence des $\widehat{f}_i^{(p)}$ pourra s'améliorer et ainsi, nous pouvons encore obtenir des vitesses paramétriques pour estimer $\eta^{(p)}(t)$ et $\gamma_{\mathbf{v}}(s,t)$. Pour ce faire, dans un premier temps, nous avons besoin d'une estimation de la norme uniforme de la fonction $\sigma_\epsilon^2(\cdot)$. Ensuite, nous procéderons à un lissage par polynômes locaux du signal bruité $Z_i^{(p)}(t)$ avec un choix adaptatif de la fenêtre en suivant l'approche de Goldenshluger et Lepski (2011).

Zhang et Chen (2007) ont proposé un estimateur de $\sigma_\epsilon^2(t)$ sous la forme :

$$\widehat{\sigma}_\epsilon^2(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\left(Z_{i,j} - \widehat{f}_i(t_{ij})\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)},$$

où $K_h(\cdot) = K(\cdot/h)$ avec K un noyau et h un paramètre de lissage. Le choix du paramètre de lissage se fait par validation croisée. Afin d'éviter une procédure trop complexe, nous proposons une approche alternative pour estimer la vitesse de la norme uniforme de $\sigma_\epsilon^2(\cdot)$. Plus précisément, nous proposons d'estimer cette norme uniforme par :

$$\widehat{\tau}_\epsilon^2 = \max_{k=1,\ldots,|\mathcal{T}|} \frac{1}{2n}\sum_{i=1}^{n} \left(Z_i(t_{i,k+1}) - Z_i(t_{i,k})\right)^2.$$

Nous étudions la concentration de cet estimateur et en déduisons des conditions plus faible sur les valeurs de δ sous lesquelles les fonctions $\eta^{(p)}(t)$ et $\gamma_{\mathbf{v}}(s,t)$ peuvent encore s'estimer à une vitesse paramétrique. En particulier, nos conditions permettent que les m_i soient plus petits que n, c'est-à-dire que δ soit plus petit que 1.

Bibliographie

Zhang, J.-T. et Chen, J. (2007), *Statistical Inferences For Functional Data.*, The Annals of Statistics, Vol. 35, No. 3, 1052-1079.

Goldenshluger, A. and Lepski, O. (2011), *Bandwidth selection in kernel density estimation : Oracle inequalities and adaptive minimax optimality.*, The Annals of Statistics, Vol. 39, No. 3, 1608-1632