Modélisation non paramétrique de la régression pour variables explicatives fonctionnelles avec autocorrélation des erreurs.

Camille Ternynck [1] & Sophie Dabo-Niang [2] & Serge Guillas [3]

[1] *iWater, Masdar Institute, Masdar City, Abu Dhabi, Emirats Arabes Unis;*
cternynck@masdar.ac.ae
[2] *Laboratoire LEM, Université de Lille, Villeneuve d'Ascq, France;*
sophie.dabo@univ-lille3.fr
[3] *Department of Statistical Science, University College London, London, UK;*
s.guillas@ucl.ac.uk

Résumé. Dans cette présentation, nous introduisons une nouvelle approche basée sur l'estimateur à noyau pour estimer le modèle de régression non linéaire en présence de variables réponses réelles et de variables explicatives à valeurs dans un espace fonctionnel. Par ailleurs, le processus résiduel est considéré stationnaire et autocorrélé. La procédure consiste à pré-blanchir la variable dépendante en se basant sur l'autocorrélation estimée. L'idée principale est de transformer le modèle de régression original de sorte que le terme d'erreur du modèle transformé devienne non corrélé. Nous établissons la convergence de l'estimateur de la régression ainsi que sa normalité asymptotique en considérant des variables explicatives α-mélangeantes, le cas le plus général de variables faiblement dépendantes. Bien que, dans la littérature sur les méthodes à noyau, il est généralement préférable d'ignorer entièrement la structure de corrélation, nous montrons ici que la fonction d'autocorrélation du processus des erreurs apporte de l'information utile permettant d'améliorer l'estimation de la fonction de régression. Nous appliquons l'estimateur proposé à des données simulées ainsi qu'à des données de concentration en ozone dans l'air. Lorsque le processus des erreurs présente une forte corrélation, nous constatons que notre procédure permet d'améliorer les résultats obtenus avec l'estimateur classique.

Mots-clés. Régression à noyau, Séries temporelles, Pré-blanchiment, Données fonctionnelles

Abstract. In this talk, a kernel-based procedure of estimation for a nonlinear functional regression is introduced in the case of a functional predictor and a scalar response. More precisely, the explanatory variable takes values in some abstract function space and the residual process is stationary and autocorrelated. The procedure consists in a pre-whitening transformation of the dependent variable based on the estimated autocorrelation. The main idea is to transform the original regression model, so that this transformed regression has a residual term that is uncorrelated. The asymptotic distribution of the proposed estimator is established considering that the explanatory variable is an

α-mixing process, the most general case of weakly dependent variables. Although, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely, it is shown here that the autocorrelation function of the error process has useful information for improving estimators of the regression function. The skills of the methods are illustrated on simulations as well as on application on ozone levels over the US. When the error process is strongly correlated, we note that our procedure allows improving the results obtained with the conventional estimator.

Keywords. Kernel regression, Time series, Pre-whitening, Functional data

1 Introduction

Ce travail concerne l'étude d'un modèle de régression $Y_t = r(X_t) + u_t$, $t = 1, \ldots, T$, en séries temporelles lorsque les variables explicatives X_t sont fonctionnelles, les variables réponses Y_t sont réelles et les termes d'erreurs u_t sont autocorrélés. Plus précisément, les variables explicatives X_t appartiennent à l'espace fonctionnel $(\mathcal{E}, d)$, muni de la semi-métrique d. La particularité de ce travail est de proposer une approche basée sur l'estimateur à noyau qui permet de tenir compte de l'information contenue dans le terme d'erreur. Cette procédure est une généralisation d'un travail existant dans le cadre réel (voir Xiao *et al.* (2003)). En effet, l'estimateur à noyau classique (1), adapté au cadre des données fonctionnelles dans Ferraty et Vieu (2000, 2006), ignore la structure de corrélation, induisant une perte d'information

$$\widehat{r}(x) = \frac{\sum_{t=1}^{T} Y_t K\left(\frac{d(x,X_t)}{h_T}\right)}{\sum_{t=1}^{T} K\left(\frac{d(x,X_t)}{h_T}\right)}, \qquad x \in (\mathcal{E}, d) \tag{1}$$

où K est un noyau et h_T le paramètre de lissage (ou fenêtre) correspondant.

Notre objectif est de montrer que l'information contenue dans le terme d'erreur u_t permet d'améliorer l'estimation de la fonction de régression. L'idée principale est de transformer le modèle de régression original de sorte que le terme d'erreur du modèle transformé devienne non corrélé.

2 Procédure d'estimation

Dans la suite, nous supposons que le processus u_t admet une représentation autorégressive d'ordre 1, $u_t = \epsilon_t - a_1\epsilon_{t-1}$ où ϵ_t est un processus i.i.d. de moyenne nulle mais la méthode peut être généralisée à des ordres supérieurs. Le modèle transformé s'écrit

$$\underline{Y}_t = r(X_t) + \epsilon_t$$

où $\underline{Y}_t = Y_t - a_1 \left(Y_{t-1} - r(X_{t-1})\right)$ est la série filtrée. Nous proposons d'estimer la fonction de régression $r(\cdot)$ par

$$\overline{r}(x) = \frac{\sum_{t=1}^{T} \underline{Y}_t K_0 \left(\frac{d(x,X_t)}{h_0}\right)}{\sum_{s=1}^{T} K_0 \left(\frac{d(x,X_s)}{h_0}\right)}, \qquad x \in (\mathcal{E}, d)$$

où K_0 est un noyau et h_0 le paramètre de lissage (ou fenêtre) correspondant.

Cependant, en pratique, $\underline{Y}_t$ est inconnu ce qui ne permet pas de calculer $\overline{r}(x)$. Pour contourner cette difficulté, nous proposons d'approcher $\underline{Y}_t$ par $\widehat{\underline{Y}}_t$ basé sur l'estimation de a_1. La procédure d'approximation est la suivante:

1. Obtenir un estimateur consistant de r par la régression de Y_t sur X_t, noté $\widehat{r}$, et calculer les résidus estimés $\widehat{u}_t = Y_t - \widehat{r}(X_t)$
2. Estimer le coefficient a_1 de l'autorégression de $\widehat{u}_t$: $\widehat{u}_t = \widehat{a}_1 \widehat{u}_{t-1} + \eta$ avec η un bruit i.i.d.
3. Approcher $\underline{Y}_t$, $t = 2, \ldots, T$, c'est à dire $\widehat{\underline{Y}}_t = Y_t - \widehat{a}_1(Y_{t-1} - \widehat{r}(X_{t-1}))$.

Ainsi, l'estimateur de la fonction de régression r que nous proposons est défini par

$$\widetilde{r}(x) = \frac{\sum_{t=2}^{T} \widehat{\underline{Y}}_t K_1 \left(\frac{d(x,X_t)}{h_1}\right)}{\sum_{s=2}^{T} K_1 \left(\frac{d(x,X_s)}{h_1}\right)}, \qquad x \in (\mathcal{E}, d)$$

où K_1 est un noyau et h_1 est le paramètre de lissage (ou fenêtre) correspondant.
Les propriétés asymptotiques des deux estimateurs précédents sont étudiées dans la section suivante.

3 Hypothèses et résultats de convergence

Tout d'abord, nous énonçons les hypothèses permettant d'obtenir les résultats de convergence des estimateurs $\overline{r}(x)$ et $\widetilde{r}(x)$.

H1 *(Conditions de régularité)*

1. r est une fonction de Lipschitz bornée: $|r(u) - r(v)| \leq c_3 d(u,v)^\beta$ pour tout $u, v \in (\mathcal{C}, d)$ et $\beta > 0$.
2. $g_2(u) = Var[\underline{Y}_t | X_t = u]$, $u \in (\mathcal{E}, d)$, est indépendant de t et continu dans un certain voisinage de x.
   Supposons que $\mathbb{E}|\underline{Y}_t|^\nu < \infty$ et $\mathbb{E}[|\epsilon_t|^\nu] < \infty$ pour un certain $\nu > 2$ et
   $$g_\nu(u) = \mathbb{E}[|\underline{Y}_t - r(x)|^\nu | X_t = u]$$
   est continu dans un certain voisinage de x.

3. Pour $t \neq s$ et $u, v \in (\mathcal{E}, d)$,

$$g(u, v; x) = \mathbb{E}[(\underline{Y}_t - r(x))(\underline{Y}_s - r(x))|X_t = u, X_s = v],$$

ne dépend pas de t, s et est continu dans un certain voisinage de (x, x).

H2 *(Conditions sur les noyaux)*
Les noyaux K_i ($i = 0$ ou 1) sont des noyaux bornés, symmétriques, non négatifs avec support compact $[0, 1]$ satisfaisant

1. $\int K_i(u)du = 1$ et $c_1\mathbf{1}_{[0,1]} < K_i < c_2\mathbf{1}_{[0,1]}$, c_1 et c_2 sont deux constantes finies.
2. Pour $j = 1, 2$, nous avons $I_j(h_i) \to C_j$ lorsque $h_i \to 0$, pour une certaine constante positive C_j, avec

$$I_j(h_i) = \frac{1}{\phi(h_i)/h_i} \int_0^1 K_i^j(u)\phi'(h_i v)dv$$

où $\phi(\cdot)$ est définie ci-après.

Soit $\mathcal{B}(x, h)$ une boule de rayon h centrée en $x \in (\mathcal{E}, d)$ et

$$\begin{aligned} F_x(h) &= \mathbb{P}[X_t \in \mathcal{B}(x, h)] \\ F_{x,x}^{s,t}(h) &= \mathbb{P}[(X_t, X_s) \in \mathcal{B}(x, h) \times \mathcal{B}(x, h)] \\ F_{x,y}^{s,t}(h) &= \mathbb{P}[(X_t, X_s) \in \mathcal{B}(x, h) \times \mathcal{B}(y, h)] \end{aligned}$$

où $F_x(h)$ correspond à la notion de *probabilités de petites boules*. Soient f_k, $k = 1$, 2 et 3, des fonctions finies non-négatives (uniformément bornées).

H3 *(Distributions)*

1. $F_x(h) = \phi(h)f_1(x)$ lorsque $h \to 0$, où $\phi(0) = 0$, $\phi(h)$ est absolument continue au voisinage de l'origine, $\sup_t f_1(X_t) < \infty$.
2. $\sup_{t \neq s} F_{x,x}^{s,t}(h) \leq \psi_1(h)f_2(x)$ lorsque $h \to 0$, où $\psi_1(h) \to 0$ lorsque $h \to 0$ et $\sup_t f_2(X_t) < \infty$. On suppose que $\psi_1(h)/\phi^2(h)$ est borné et que $\exists \zeta_1 \in (0, 1)$, $0 < F_{x,x}(h) = O(\phi(h)^{1+\zeta_1})$.
3. $\sup_{t \neq s} F_{x,y}^{s,t}(h) \leq \psi_2(h)f_3(x, y)$ lorsque $h \to 0$, où $\psi_2(h) \to 0$ lorsque $h \to 0$ et $\sup_{t,s} f_3(X_t, X_s) < \infty$. On suppose que $\psi_2(h)/\phi^2(h)$ est borné.

H4 *(Mélange)*

$$\sum_{l=1}^{\infty} l^{\delta}[\alpha(l)]^{1-2/\nu} < \infty$$

pour un certain $\nu > 2$ et $\delta > 1 - 2/\nu$. (ν est l'ordre du moment dans **H1**(2))

H5: On pose $h_i \to 0$, $h_0/h_1 \to 0$ et $\frac{\log T}{T^{1/2}\phi(h_0)} \to 0$ lorsque $T \to \infty$. Soit $\{v_T\}$ une séquence positive d'entiers satisfaisant $v_T \to \infty$ telle que $v_T = o((T\phi(h_0))^{1/2})$ et $(T/\phi(h_0))^{1/2}\alpha(v_T) \to 0$, $Th_0^{2\beta} \to 0$ lorsque $T \to \infty$.

Sous certaines conditions, nous montrons la convergence en moyenne d'ordre q de $\overline{r}(x)$ dont les vitesses de convergence sont données dans le théorème suivant.

Théorème 1 *Sous **H1**(1), **H2**(1), **H3**(1) et **H4** (où dans **H4**, $\delta > \max\{p/2 - 1, 1 - 2/\nu\}$), Y_t est bornée, $\overline{r}(x)$ converge en moyenne d'ordre p ($p > 1$) vers $r(x)$ et*

$$\|\overline{r}(x) - r(x)\|_p = O(h_0^{\beta}) + O\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/2}\right).$$

Dans le Théorème 2, nous obtenons la normalité asymptotique de $\overline{r}(x)$.

Théorème 2 *(Masry 2005): Sous **H1**, **H2**, **H3**(1), **H3**(2), **H4** et **H5**,*

$$(T\phi(h_0))^{1/2}\left[\overline{r}(x) - r(x) - B_T(x)\right] \xrightarrow{L} \mathcal{N}(0, \sigma^2(x))$$

avec $\sigma^2(x) = \dfrac{C_2 g_2(x)}{C_1^2 f_1(x)}$, $x \in (\mathcal{E}, d)$ *quel que soit* $f_1(x) > 0$ *et* $B_T(x) = \mathbb{E}[\overline{r}(x)] - r(x)$.

Nous montrons également la normalité asymptotique de $\widetilde{r}(x)$ dont les résultats sont exposés dans le Théorème 3.

Théorème 3 *Sous **H1**- **H5**,*

$$(T\phi(h_1))^{1/2}[\widetilde{r}(x) - r(x) - B_T(x)] \xrightarrow{L} \mathcal{N}(0, \sigma^2(x))$$

avec $\sigma^2(x) = \dfrac{C_2 g_2(x)}{C_1^2 f_1(x)}$, $x \in (\mathcal{E}, d)$ *quel que soit* $f_1(x) > 0$ *et* $B_T(x) = \mathbb{E}[\widetilde{r}(x)] - r(x)$.

Nous avons énoncé les résultats de convergence de l'estimateur proposé. Dans la suite, nous allons étudier son comportement sur des données simulées et réelles.

4 Applications

Nous avons d'abord considéré l'application de notre approche sur des données simulées. Les observations fonctionnelles X_t (avec $t = 1, \ldots, T$) sont définies par $X_t(w) = 1 + 10e_{0,t} + 3e_{1,t}w^2 + 4e_{2,t}(1-w)^3$, $w \in [0,1]$ où $e_{0,t}$, $e_{1,t}$ et $e_{2,t}$ sont i.i.d. de loi $\mathcal{N}(0,1)$. Nous posons $r(x) = \sqrt{|0.5\int_0^1 x\,dx|}$. Le processus des erreurs u_t est un processus $AR(1)$, tel que $u_t = \epsilon_t + \rho\epsilon_{t-1}$. Différentes valeurs du paramètre ρ sont considérées. Pour chaque cas étudié, le

nombre de réplications est de 200. La sélection des paramètres de lissage (fenêtres) est faite par validation croisée. Nous reportons l'efficacité relative, calculée comme le ratio du carré des erreurs de l'estimateur proposé $\widetilde{r}(x)$ et de celui de l'estimateur conventionnel $\widehat{r}(x)$. Nous constatons que l'amélioration obtenue avec l'estimateur proposé est importante, jusqu'à 25% en présence de niveaux élevés d'autocorrélation. Les améliorations moyennes de notre estimateur sont toujours positives sauf quand $\rho = 0.1$, c'est à dire pour un faible niveau d'autocorrélation.

Dans un second temps, nous avons adapté l'estimateur de la régression à la prédiction, adaptation que nous avons appliquée à des données réelles. Notre objectif est de prédire la concentration en ozone à une certaine date non observée à partir du passé. Nous disposons de la concentration horaire en ozone du 2 Juin au 31 Août 2005 (soit 91 jours) pour différentes stations. Nous nous intéressons aux prédictions horaires en ozone sur une journée. Pour illustrer notre objectif, nous allons prédire les concentrations en ozone du 31 Août à partir des observations faites les 90 jours précédents. Nous montrons, sur plusieurs stations, que notre méthode permet d'améliorer les prédictions obtenues avec l'estimateur à noyau classique de la régression.

5 Conclusion

Nous montrons que la procédure présentée permet d'améliorer l'estimateur à noyau classique de la fonction de régression, en présence de variables fonctionnelles et d'erreurs autocorrélées. Dans un travail en cours, cette procédure fait l'objet d'une extension plus générale. En effet, nous l'étendons au cadre des modèles à indice fonctionnel.

Bibliographie

[1] Ferraty, F. et Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Compte Rendus de l'Académie des Sciences, Series I, Mathematics*, 330:403–406.

[2] Ferraty, F. et Vieu, P. (2006). Nonparametric Functional Data Analysis. *Springer Series in Statistics*.

[3] Masry E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115:155–177.

[4] Xiao, Z., Linton, O. B., Carroll, R. J., et Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association*, 98(464):980–992.