Imputation par régression dans le modèle linéaire fonctionnel avec valeurs manquantes dans la réponse

Christophe Crambes[1] & Yousri Henchiri[2]

[1] *Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Université de Montpellier, Place Eugène Bataillon, 34000 Montpellier. E-mail: christophe.crambes@univ-montp2.fr*

[2] *Département de Mathématiques, Université du Québec à Montréal, 201 Avenue du Président Kennedy, H2X 3Y7, Montréal, Canada. E-mail: yousri.henchiri@univ-montp2.fr*

Résumé. Nous nous intéressons au modèle linéaire fonctionnel lorsque la variable d'intérêt, réelle, est sujette à des observations manquantes et la variable explicative, fonctionnelle, est complètement observée. Une méthode d'imputation des données manquantes par régression est présentée, en utilisant l'estimation du coefficient fonctionnel du modèle par régression fonctionnelle sur composantes principales. Nous étudions le comportement asymptotique de l'erreur commise lorsque la valeur manquante est replacée par la valeur imputée par régression, dans un cadre de données manquantes "missing at random". Le comportement de la méthode est également étudié en pratique sur des données simulées.

Mots-clés. Modèle linéaire fonctionnel, données manquantes, missing at random, imputation par régression.

Abstract. We are interested in functional linear regression when some observations of the real response are missing, while the functional covariate is completely observed. A regression imputation method of missing data is presented, using functional principal component regression to estimate the functional coefficient of the model. The asymptotic behaviour of the error we commit when the missing data is replaced by the regression imputed value, in a "missing at random" framework. The practical behaviour of the method is also studied on simulated data sets.

Keywords. Functional linear model, missing data, missing at random, regression imputation.

1 Introduction

Le développement de méthodes pour traiter des données fonctionnelles s'est fortement intensifié ces dernières années. Ces données, bien connues maintenant, permettent de considérer des situations pratiques où une ou plusieurs variables peuvent être des fonctions

(du temps par exemple). Lors de la dernière décennie, des ouvrages de référence comme Ramsay et Silverman (2002, 2005), Horváth et Kokoszka (2012), ou encore Ferraty et Vieu (2006) pour les aspects non-paramétriques, ont été développés et donnent une vue d'ensemble sur le sujet.

Un des modèles les plus célèbres dans ce contexte de données fonctionnelles est le modèle linéaire fonctionnel, qui relie une variable d'intérêt $Y \in \mathbb{R}$ à une variable explicative $X \in H$, où H est un espace fonctionnel muni d'un produit scalaire $\langle ., . \rangle$ et sa norme associée $\|.\|$. Habituellement, H est l'espace $L^2([a, b])$ des fonctions de carré intégrable définies sur un intervalle $[a, b]$ et la produit scalaire correspondant est défini par $\langle f, g \rangle = \int_a^b f(t)g(t)\, dt$ pour des fonctions $f, g \in L^2([a, b])$. Sans perte de énéralité, nous considèrerons X centrée. Le modèle linéaire fonctionnel s'écrit alors

$$Y = \langle \theta, X \rangle + \varepsilon, \tag{1}$$

où ε est une variable réelle centrée representant l'erreur du modèle, de variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$, et indépendante de X. Nous considérons un échantillon $(X_i, Y_i)_{i=1,\dots,n}$ indépendant et identiquement distribué de même loi que (X, Y). Cependant, pour diverses raisons, il peut arriver que certaines observations de la variable d'intérêt Y ne soient pas disponibles (défaillance d'un appareil de mesure, ...). De nombreux auteurs se sont penchés sur le problème des données manquantes, et plusieurs ouvrages permettent maintenant d'avoir une vision générale de cette problématique, comme par exemple Little et Rubin (2002) ou Van Buuren (2012). Dans le cadre fonctionnel, le travail le plus proche du notre est celui de Ferraty *et. al.* (2013), bien que se situant dans un cadre non-paramétique.

Nous définissons la variable réelle δ et nous considérons l'échantillon $(\delta_i)_{i=1,\dots,n}$ tel que $\delta_i = 1$ si la valeur Y_i est observée et $\delta_i = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \dots, n$. Nous considérons que les valeurs manquantes sont "missing at random" (MAR), c'est-à-dire, pour $z \in \{0, 1\}$

$$P\left(\delta = z|X, Y\right) = P\left(\delta = z|X\right). \tag{2}$$

Le nombre de données manquantes dans l'échantillon est noté

$$m_n = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i = 0\}}. \tag{3}$$

Dans la suite (section 2), nous présentons l'imputation par régression d'une valeur manquante pour la variable d'intérêt. La valeur imputée a des propriétés de convergence vers la valeur inconnue, que nous donnerons (section 3). Enfin, des simulations (section 4) illustreront le comportement de notre méthode.

2 Imputation par régression

À partir de cette section, nous allons considérer le point de vue opératoriel du modèle 1, de la façon suivante

$$Y = \Theta X + \varepsilon, \tag{4}$$

où $\Theta : H \longrightarrow \mathbb{R}$ est un opérateur linéaire continu défini par $\Theta u = \langle \theta, u \rangle$ pour toute fonction $u \in H$. Nous reprenons l'estimateur de Θ défini par Cardot *et. al.* (1999), que nous rappelons ici. Considérons l'opérateur de covariance de X défini sous la condition $\mathbb{E}\left(\|X\|^2\right) < +\infty$ (que nous supposons satisfaite dans la suite) par

$$\Gamma u = \mathbb{E}\left(\langle X, u \rangle X\right),$$

pour tout $u \in H$ et sa version empirique

$$\widehat{\Gamma}_n u = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, u \rangle X_i.$$

Nous notons $(\lambda_j)_{j\geq 1}$ (resp. $\left(\widehat{\lambda}_j\right)_{j\geq 1}$) la suite des valeurs propres de Γ (resp. $\widehat{\Gamma}_n$) et $(v_j)_{j\geq 1}$ (resp. $(\widehat{v}_j)_{j\geq 1}$) la suite des fonctions propres de Γ (resp. $\widehat{\Gamma}_n$). Considérons enfin l'opérateur $\widehat{\Delta}_n$ de covariance croisée défini par $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^{n} \langle X_i, u \rangle Y_i$ pour tout $u \in H$. Alors, l'estimateur de Θ est défini par

$$\widehat{\Theta} = \widehat{\Pi}_{k_n} \widehat{\Delta}_n \left(\widehat{\Pi}_{k_n} \widehat{\Gamma}_n\right)^{-1}, \tag{5}$$

où $(k_n)_{n\geq 1}$ est une suite de nombres entiers et $\widehat{\Pi}_{k_n}$ est l'operateur de projection sur le sous-espace engendré par $(\widehat{v}_1, \ldots, \widehat{v}_{k_n})$.

Nous pouvons à présent définir l'imputation par régression pour une donnée manquante, en suivant par exemple Little et Rubin (2002). Soit ℓ un nombre entier compris entre 1 et n, nous supposons que la valeur Y_ℓ est manquante. La valeur imputée pour Y_ℓ est définie par

$$Y_{\ell,imp} = \widehat{\Pi}_{k_n,obs} \widehat{\Delta}_{n,obs} \left(\widehat{\Pi}_{k_n,obs} \widehat{\Gamma}_{n,obs}\right)^{-1} X_\ell, \tag{6}$$

où $\widehat{\Gamma}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i X_i$, $\widehat{\Delta}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i Y_i$ et $\widehat{\Pi}_{k_n,obs}$ est l'opérateur de projection sur le sous-espace $\mathrm{span}(\widehat{v}_{1,obs}, \ldots, \widehat{v}_{k_n,obs})$ où $\widehat{v}_{1,obs}, \ldots, \widehat{v}_{k_n,obs}$ sont les k_n premières fonctions propres de l'opérateur de covariance $\widehat{\Gamma}_{n,obs}$.

3 Résultats de convergence

Afin d'établir les résultats de convergence pour la valeur imputée de Y_ℓ, nous considérons les hypothèses suivantes.

$(A.1)$ soit λ la fonction définie par $\lambda(j) = \lambda_j$ pour tout $j \geq 1$, qui interpole continûment les valeurs λ_j entre j et $j+1$. Nous supposons que

$$\lambda \text{ est convexe.}$$

$(A.2)$ Nous supposons que

$$\theta \in L^1([0,1]) = \left\{ f : [0,1] \longrightarrow \mathbb{R} \ / \int_0^1 |f(t)| \, \mathrm{d}t < +\infty \right\}.$$

$(A.3)$ Il existe une constante positive C telle que

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

Théorème 3.1 *Sous les hypothèses* $(A.1)$-$(A.3)$*, si de plus* $\lambda_{k_n} k_n$ *tend vers zéro lorsque* n *tend vers l'infini, nous avons*

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

Afin de préciser la vitesse de convergence de la valeur imputée $Y_{\ell,imp}$ vers la valeur $\langle \theta, X_\ell \rangle$, nous notons de plus, pour toute fonction $\varphi : \mathbb{R}_+^\star \longrightarrow \mathbb{R}_+^\star$ et pour tout nombre réel positif L

$$\mathcal{C}(\varphi, L) = \left\{ T : H \longrightarrow \mathbb{R} \ / \ \forall j \geq 1, Tv_j \leq L\sqrt{\varphi(j)} \right\}.$$

Théorème 3.2 *Soit* $L = \left\|\Theta\Gamma^{1/2}\right\|_\infty$ *et* φ *la fonction définie par* $\varphi(j) = \frac{\left(\Theta\Gamma^{1/2}v_j\right)^2}{L^2}$ *pour tout* $j \geq 1$*, qui interpole continûment les valeurs* $\varphi(j)$ *entre* j *et* $j+1$*. Sous les hypothèses* $(A.1)$-$(A.3)$*, l'opérateur* $\Theta\Gamma^{1/2}$ *appartient à* $\mathcal{C}(\varphi, L)$ *et*

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\sigma_\varepsilon^2 \frac{k_n^\star}{n - m_n},$$

où $k_n^\star$ *est solution de l'équation en* x

$$\int_x^{+\infty} \varphi(s) \, \mathrm{d}s = \frac{\sigma_\varepsilon^2}{L^2(n-m_n)} x.$$

4 Simulations

Dans cette partie, nous présentons un exemple de simulation qui a été effectué pour étudier le comportement de la méthode d'imputation en pratique. Le modèle considéré est

$$Y = \int_0^1 \sin(4\pi t) X(t)\ \mathrm{d}t + \varepsilon,$$

où X est un mouvement Brownien standard sur l'intervalle $[0,1]$ et $\varepsilon \sim \mathcal{N}(0, 0.2)$. Les courbes sont discrétisées sur une grille de 100 points équidistants sur l'intervalle $[0,1]$. Nous nous plaçons dans le cadre MAR: les observations de la variable Y sont manquantes lorsque la courbe X correspondante se situe en dehors d'une certaine bande de confiance autour des courbes $X_1, \ldots, X_n$. Pour la construction de l'estimateur, le nombre de composantes principales de l'opérateur de covariance de X a été choisi par validation croisée généralisée. Trois critères ont été utilisés pour évaluer l'erreur commise en remplaçant une valeur manquante par sa valeur imputée par notre méthode (sur $S = 500$ simulations)

$$\overline{MSE} = \frac{1}{\boldsymbol{S}} \frac{1}{m_n} \sum_{s=1}^{\boldsymbol{S}} \sum_{\ell=1}^{m_n} (Y_\ell^{(s)} - Y_{\ell,imp}^{(s)})^2,$$

$$\overline{MAE} = \frac{1}{\boldsymbol{S}} \frac{1}{m_n} \sum_{s=1}^{\boldsymbol{S}} \sum_{\ell=1}^{m_n} \left| Y_\ell^{(s)} - Y_{\ell,imp}^{(s)} \right|,$$

$$\overline{CR3} = \frac{1}{\boldsymbol{S}} \sum_{s=1}^{\boldsymbol{S}} \sum_{\ell=1}^{m_n} (Y_\ell^{(s)} - Y_{\ell,imp}^{(s)})^2 / \sum_{\ell=1}^{m_n} \left(\varepsilon_\ell^{(s)} \right)^2.$$

Les valeurs de ces critères sont présentées dans la table 1 pour différentes valeurs de n et différents pourcentages de données manquantes.

On constate que l'erreur diminue lorsque la taille de l'échantillon augmente. De même, l'erreur augmente lorsque le pourcentage de données manquantes augmente.

Bibliographie

[1] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice*. Springer, New-York.

[2] Ferraty, F., Sued, M. and Vieu, P. (2013). Mean estimation with data missing at random for functional covariables. *Statistics*, 47, 688–706.

[3] Horvath, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.

[4] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2^{nd} Ed.)*. Wiley, Chichester.

échantillon	n = 100			
données manquantes	22%	35%	45%	61%
Critère 1: $[\overline{MSE} \times 10^2]$	4.9072	5.0387	5.2793	5.8767
	(2.3709)	(2.0245)	(2.0432)	(2.3758)
Critère 2: $[\overline{MAE} \times 10^2]$	17.5065	17.7815	18.1559	19.1239
	(1.9866)	(1.6035)	(1.5037)	(1.4432)
Critère 3: $[\overline{CR3}]$	1.1637	1.2070	1.2685	1.4253
	(0.3588)	(0.2868)	(0.2948)	(0.4246)
échantillon	n = 500			
données manquantes	13%	28%	39%	60%
Critère 1: $[\overline{MSE} \times 10^2]$	4.1620	4.1676	4.2068	4.3643
	(0.9666)	(0.7975)	(0.7583)	(0.7338)
Critère 2: $[\overline{MAE} \times 10^2]$	16.2022	16.2329	16.319	16.6411
	(1.9866)	(1.6035)	(1.5037)	(1.4432)
Critère 3: $[\overline{CR3}]$	1.0329	1.0346	1.0442	1.0843
	(0.1739)	(0.1148)	(0.0952)	(0.0873)

Table 1: Valeurs moyennes (et écart-types) des trois critères pour différentes tailles d'échantillon et différents pourcentages de données manquantes.

[5] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis*. Springer, New York.
[6] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis (2^{nd} Ed.)*. Springer, New York.
[7] Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC Press, Boca Raton.