Régression additive sous variables imparfaites

Germain Van Bever [1] & Jeong Min Jeon [2]

[1] *Université libre de Bruxelles et Université de Namur, Belgique, germain.van.bever@ulb.be*
[2] *Seoul National University, Corée du Sud, jeongmin.jeon.stat@gmail.com*

Résumé. Dans cet exposé, nous présentons un modèle additif dans lequel la variable réponse prend ses valeurs dans un espace de Hilbert. Les prédicteurs sont multivariés. Toutes les variables peuvent possiblement être imparfaitement mesurées. Le modèle permet de considérer des variables réponses Euclidiennes, fonctionnelles ou même à valeur dans des espaces de densité. L'ajout de variables imparfaites permet de couvrir le cas de variables fonctionnelles à valeurs dans une variété Riemannienne, le cas où seul un échantillon aléatoire d'une densité inconnue est disponible, ou encore le cas où les régresseurs sont les scores obtenus par analyse en composantes principales ou singulières dans un espace de Hilbert. L'estimation des fonctions de régression se fait via la méthode de smooth backfitting (Mammen *et al.*, 1990). Nous étudions le comportement non-asymptotique et asymptotique de ces estimateurs. Plusieurs applications illustrent les méthodes introduites.

Mots-clés. Analyse des données fonctionnelles, Erreurs de mesure, Espaces de Hilbert, Régression additive

Abstract. In this talk, we study an additive model where the response variable is Hilbert-space-valued and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows to cover Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, we can cover functional variables taking values in a Riemannian manifold and the case where only a random sample from a density-valued response is available. Dealing with imperfect predictors allows us to cover various principal component and singular component scores obtained from Hilbert-space-valued variables. For the estimation of the additive model having such variables, we use the smooth backfitting method originated by Mammen et al. (1999). We provide full non-asymptotic and asymptotic properties of our regression estimator and present its wide applications via several simulation studies and real data applications..

Keywords. Additive regression, Functional data analysis, Hilbert valued data, measurement errors

1 Modèle additif hilbertien

Le modèle de régression additive "classique", pour des régresseurs $X_j \in \mathbb{R}$, $j = 1, \ldots, d$, et une variable réponse $Y \in \mathbb{R}$, prend la forme

$$Y = f_0 + \sum_{j=1}^{d} f_j(X_j) + \epsilon,$$

où $f_0 \in \mathbb{R}$ est une constante et les fonctions f_j, $j = 1, \ldots, d$ sont les fonctions-composantes inconnues à estimer. Ce modèle permet une balance entre un modèle paramétrique, peu flexible, et un modèle totalement nonparamétrique du type $Y = f(X_1, \ldots, X_d) + \epsilon$, dans lequel les taux de convergence de l'estimateur de f souffrent du "curse of dimensionality".

Soit $\mathbb{H}$ un espace de Hilbert séparable. Soit également $\oplus$, $\odot$, $\mathbf{0}$, $\langle \cdot, \cdot \rangle$ et $\|\cdot\|$, respectivement, l'addition vectorielle, la multiplication scalaire, le vecteur nul, le produit scalaire et la norme sur $\mathbb{H}$. Notons que $\oplus$, $\odot$, $\mathbf{0}$, $\langle \cdot, \cdot \rangle$ et $\| \cdot \|$ pour $\mathbb{H} = \mathbb{R}^k$ correspondent à $+$, $\times$, $(0, \ldots, 0) \in \mathbb{R}^k$, le produit scalaire et la norme ℓ_2, respectivement. Des exemples d'espaces $\mathbb{H}$ non-euclidiens sont disponibles dans les exemples de la section 3.

Dans ce papier, nous considérons le modèle additif Hilbertien multivarié

$$\mathbf{Y} = \mathbf{f}_0 \oplus \bigoplus_{j=1}^{d} \mathbf{f}_j(\xi_j) \oplus \epsilon, \tag{1}$$

où $\mathbf{Y} \in \mathbb{H}$ est une variable réponse satisfaisant $\mathbb{E}(\|\mathbf{Y}\|^2) < \infty$, $\xi_j = (\xi_{j1}, \ldots, \xi_{jL_j}) \in \mathbb{R}^{L_j}$ avec $L_j \in \mathbb{N}$ sont des prédicteurs multivariés, $\epsilon \in \mathbb{H}$ est un terme d'erreur satisfaisant $\mathbb{E}(\|\epsilon\|^2) < \infty$ et $\mathbb{E}(\epsilon | \xi_1, \cdots, \xi_d) = \mathbf{0}$, $\mathbf{f}_0 \in \mathbb{H}$ est une constante inconnue et $\mathbf{f}_j : \mathbb{R}^{L_j} \to \mathbb{H}$ sont des fonctions-composantes inconnues. Ici, l'espérance conditionnelle $\mathbb{E}(\epsilon | \xi_1, \cdots, \xi_d)$ est définie via une intégrale de Bochner.

Nous supposons également que les régresseurs ξ_{jl}, pour $1 \leq j \leq d$ et $1 \leq l \leq L_j$, peuvent provenir de différentes sources. Par exemple, ξ_{jl} peuvent être des prédicteurs scalaires ou des scores obtenus par une analyse en composantes principales ou singulières. Ces prédicteurs scalaires peuvent, parfois, souffrir d'erreurs de mesures (Delaigle, 2008). Les scores principaux sont inobservables en général, la vrai structure de covariance étant inconnue. Pour ces raisons, nous supposons que seule une approximation imparfaite $\tilde{\xi}_{jl} \in \mathbb{R}$ de ξ_{jl} est disponible. Pour les mêmes raisons, nous supposons observer une approximation $\tilde{\mathbf{Y}}$ de $\mathbf{Y}$. Ceci couvre, entre autres, des réponses fonctionnelles reconstruites sur base d'évaluations temporelles, etc.

2 Estimation

L'estimation des fonctions-composantes $\mathbf{f}_j$, $j = 1, \ldots, d$ sur un domaine compact $D_j \subset \mathbb{R}^{L_j}$ se fait via la méthode de smooth backfitting (Mammen, 1999). Soit p la densité de $\xi = (\xi_1, \ldots, \xi_d)$ et soit $p_0^D = \int_D p(\mathbf{x}) d\mathbf{x} > 0$, où $D = \prod_{j=1}^{d} D_j$. Soit $p^D(\mathbf{x}) = p(\mathbf{x})/p_0^D$ pour $\mathbf{x} = (x_1, \ldots, x_d) \in D$. Soit la densité marginale $p_j^D(x_j) = \int_{D_{-j}} p^D(\mathbf{x}) d\mathbf{x}_{-j}$ et $p_{jk}^D(x_j, x_k) =$

$\int_{D_{-jk}} p^D(\mathbf{x}) d\mathbf{x}_{-jk}$, où $D_{-j} = \prod_{m \neq j} D_m$, $D_{-jk} = \prod_{m \neq j,k} D_m$, et $\mathbf{x}_{-j}$ et $\mathbf{x}_{-jk}$ denotent, respectivement, les $(d-1)$- et $(d-2)$-vecteurs obtenus en omettant x_j et (x_j, x_k) in $\mathbf{x}$.

Remarquons que les $\mathbf{f}_j$ ne sont pas identifiables dans le modèle (1) puisque $\bigoplus_{j=0}^d \mathbf{f}_j = \bigoplus_{j=0}^d (\mathbf{f}_j \oplus \mathbf{c}_j)$ pour toutes constantes $\mathbf{c}_j \in \mathbb{H}$ telles que $\bigoplus_{j=0}^d \mathbf{c}_j = \mathbf{0}$. Nous imposons donc que

$$\int_{D_j} \mathbf{f}_j(x_j) \odot p_j^D(x_j) dx_j = \mathbf{0}, \quad 1 \leq j \leq d.$$

Les contraintes ci-dessus déterminent $\mathbf{f}_0$ comme

$$\mathbf{f}_0 = \int_D \mathbb{E}(\mathbf{Y}|\xi = \mathbf{x}) \odot p^D(\mathbf{x}) d\mathbf{x} = (p_0^D)^{-1} \odot \mathbb{E}(\mathbf{Y} \odot \mathbb{I}(\xi \in D)),$$

où $\mathbb{I}(\cdot)$ est la fonction indicatrice. Il est facile de montrer que les fonctions composantes sont alors solutions du système d'équations intégrales

$$\mathbf{f}_j(x_j) = \mathbf{m}_j(x_j) \ominus \mathbf{f}_0 \ominus \bigoplus_{k \neq j} \int_{D_k} \mathbf{f}_k(x_k) \odot \frac{p_{jk}^D(x_j, x_k)}{p_j^D(x_j)} dx_k, \quad 1 \leq j \leq d, \tag{2}$$

où

$$\mathbf{m}_j(x_j) = (p_j^D(x_j))^{-1} \odot \int_{D_{-j}} \mathbb{E}(\mathbf{Y}|\xi = \mathbf{x}) \odot p^D(\mathbf{x}) d\mathbf{x}_{-j}$$

et $\ominus$ est la soustraction vectorielle dans $\mathbb{H}$.

Nous construisons alors des solutions approchées des équations intégrales dans (2) sur base de données $(\tilde{\xi_{i1}}, \ldots, \tilde{\xi}_{id}, \tilde{\mathbf{Y}}_i)$, $i = 1, \ldots, n$. Les propriétés non-asymptotiques et asymptotiques des estimateurs obtenus sont alors étudiées.

3 Deux exemples

Afin d'illustrer la méthodologie ci-dessous, nous présentons deux exemples. Dans le premier, la variable réponse $\mathbf{Y} \in \mathcal{S}_1^3 = \{(p_1, p_2, p_3) | p_1 + p_2 + p_3 = 1\}$, le simplexe unité de $\mathbb{R}^3$. Dans le second, $\mathbf{Y} \in L^2(S_1^2)$, l'ensemble des fonctions de carrés intégrables sur la sphère unité S_1^2 de $\mathbb{R}^3$.

3.1 Données compositionnelles

Il est maintenant accepté que les caractéristiques démographiques d'une population et les courants politiques sous-jacents sont des facteurs importants permettant, en partie, de déterminer les résultats d'une élection. Nous illustrons ceci lors des élections présidentielles américaines de 2020 et montrons comment la proportion de personnes possédant un bachelier (ξ_1), les revenus par individus (ξ_2) et l'âge médian (ξ_3) affectent la composition des résultats électoraux dans chaque état. Nous mesurons leur impact sur le vecteur compositionnel $\mathbf{Y} = (Y_1, Y_2, Y_3) \in \mathcal{S}_1^3$, où les Y_j mesurent, respectivement, les proportions de votes pour les démocrates, républicains et autres partis.

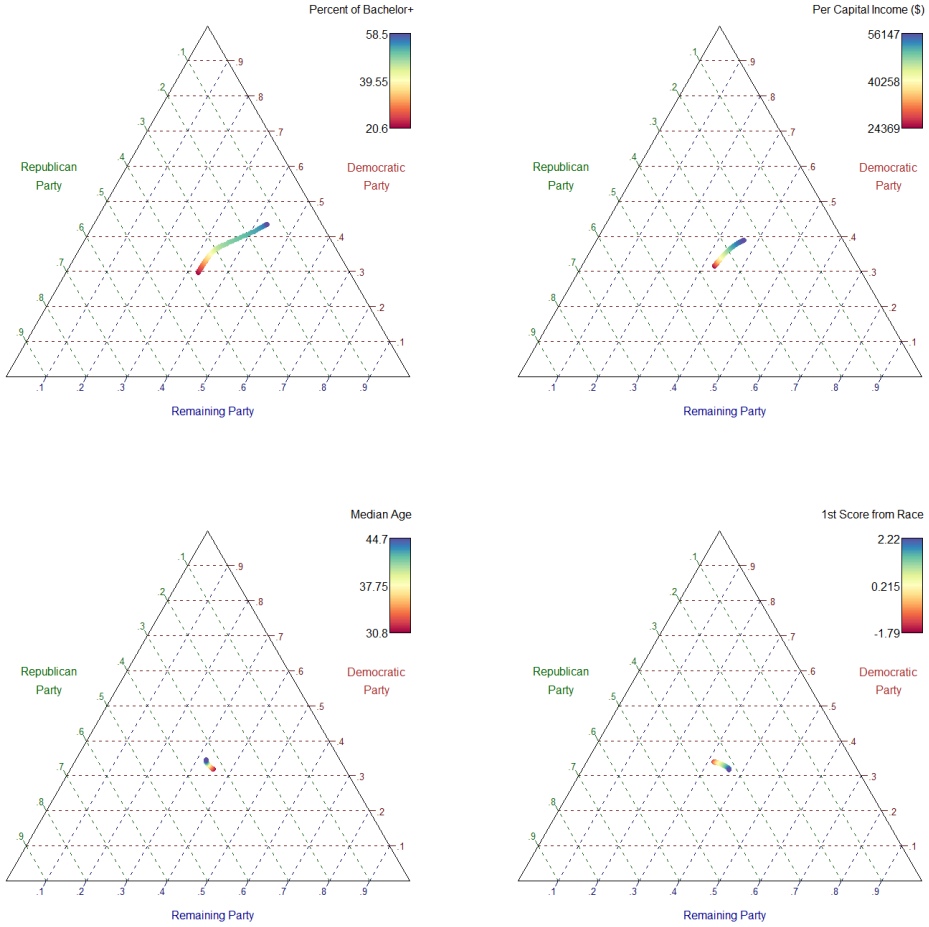

Figure 1: Estimation des fonctions-composantes $\widehat{\mathbf{f}}_j$ pour différentes mesures démographiques par état et leur influence sur les proportions de vote par parti. Pour chaque point dans ces plots ternaires, les proportions pour le parti démocratique, républicain ou autres s'obtiennent en suivant respectivement les lignes rouges, vertes ou bleues.

L'estimation d'un modèle additif pour ce type de données permet d'illustrer l'influence sur les votes de chacun des facteurs ξ_j, comme illustré sur la figure ci-dessous.

3.2 Données fonctionnelles sur la sphère

Les typhons dévastent chaque année de nombreux pays. Pouvoir prévoir leur itinéraire dès les premières heures de leur existence est crucial. L'administration coréenne météorologique (voir https://data.kma.go.kr/data/typhoonData/typInfoTYList.do?pgmNo=689) maintient une base de données contenant de nombreuses informations sur les typhons apparus en Asie du Sud-Est depuis 2001. Sur base de la pression atmosphérique centrale (ξ_1), la vitesse de l'air centrale maximale (ξ_2), la vitesse de déplacement (ξ_3) et les positions initiales (ξ_4), moving direction au temps T_0, nous estimons, à l'aide d'un modèle additif, le trajet de divers

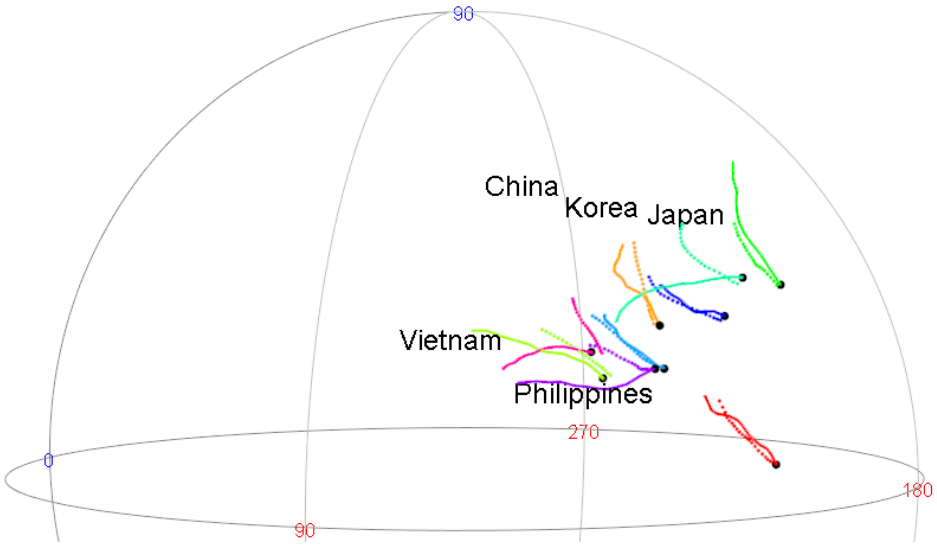

Figure 2: Vraies trajectoires (lignes pleines) et prédites (pointillées) des typhons apparus en 2022.

typhons de l'année 2022 (sur base des $n = 265$ typhons observés de 2001 à 2021).

Bibliographie

Delaigle, A. (2008). An alternative view of the deconvolution problem, *Statistica Sinica*, 18, pp. 1025-1045.

Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics*, 27, pp. 1443-1490.