Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP

Jean Steve Tamo Tchomgui, Julien Jacques, Stéphane Chrétien, Guillaume Fraysse, Vincent Barriac

Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP

Jean Steve Tamo Tchomgui[1] & Julien Jacques[1] & Stéphane Chretien[1] & Guillaume Fraysse[2] & Vincent Barriac[2]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*jean-steve.tamo-tchomgui, julien.jacques, stephane.chretien*}*@univ-lyon2.fr*
[2] *Orange Innovation, France.* {*guillaume.fraysse, vincent.barriac*}*@orange.com*

Résumé: Ce travail propose une étude comparative entre les techniques récentes de l'analyse des données fonctionnelles et celles des signatures pour la résolution d'un problème de prédiction de la Qualité d'Expérience (QoE), une mesure qui reflète la perception de l'utilisateur final d'un service de télécommunications. La QoE ainsi que les facteurs pouvant l'influencer étant mesurés à haute fréquence, le problème que nous considérons trouve donc sa formulation naturelle dans le contexte de la régression linéaire fonctionnelle, où la variable à prédire et les variables explicatives sont toutes fonctionnelles. Notre contribution principale est celle de montrer que l'utilisation des signatures permet de résumer de façon efficace l'information contenue dans les variables explicatives et de produire des prédictions souvent meilleures que celles obtenues par des méthodes classiques.
Mots-clés. Données fonctionnelles, Modèles de régression fonctionnels, Signatures.

Abstract: This work presents a comparative study of the functional data analysis approach and the approach based on the theory of signatures for a prediction problem in a Quality of Experience (QoE) monitoring projet. The problem we consider is a functional linear regression problem where both the dependent variables and the independent variables are functional variables. Our experiments demonstrate the efficiency of the signature approach for prediction in a difficult functional regression problem.
Keywords. Functional data, Functional regression models, Signatures.

Introduction

La plupart des problèmes d'apprentissage rencontrés aujourd'hui n'ont plus de données collectées sous forme classique, c'est-à-dire une sortie $Y \in \mathbb{R}$ décrite par un nombre fini de prédicteurs $X \in \mathbb{R}^p$. On observe plutôt de multiples enregistrements pour chaque prédicteur au cours du temps, autrement dit une fonction du temps de la forme $X : T \longrightarrow \mathbb{R}^p$. L'approche naturelle pour mener une inférence sur ce type de données est d'étendre le modèle linéaire classique à ce cadre plus général : c'est l'Analyse des Données

Fonctionnelles (ADF) et particulièrement la régression linéaire fonctionnelle. Comme le soulignent les auteurs de [1], qui constitue une excellente introduction au domaine, l'analyse fonctionnelle s'applique aux données dont la structure peut être représentée par une ou plusieurs fonctions et repose sur l'hypothèse selon laquelle les données à traiter possèdent une structure sous-jacente plus ou moins apparente dont l'identification et la prise en compte permettent d'étendre efficacement les techniques de l'analyse classique.

En pratique, on n'observe pas directement une fonction mais les valeurs de cette dernière à différents instants. Les données collectées se présentent donc sous forme vectorielle et nécessitent un premier traitement dans l'analyse fonctionnelle, il consiste à reconstruire la nature fonctionnelle des données. Une approche récente a été introduite en analyse fonctionnelle, basée sur la notion de signatures [2, 7] qui sont des séries (infinies) d'intégrales itératives permettant de résumer l'information contenue dans des données fonctionnelles.

Dans cet article, la Section 1 est consacrée à présenter brièvement le cadre d'analyse des données fonctionnelles, la Section 2 celle des signatures. La Section 3 décrit l'inférence menée et présente l'interprétation des résultats obtenus.

1 Inférence Fonctionnelle

Nous allons nous extraire dans ce travail du cadre d'analyse classique. Les observations ne seront plus des réalisations indépendantes et identiquement distribuées (i.i.d.) d'un processus aléatoire réel (ou vectoriel plus généralement), mais plutôt d'une variable aléatoire à valeur dans un espace de fonctions : c'est la notion de fonction aléatoire. Faire de l'ADF sur des observations collectées sous forme de séries pour chaque prédicteur revient d'abord à accepter l'hypothèse fondamentale d'existence du processus aléatoire sous-jacent et ensuite considérer chaque série observée comme les valeurs en un nombre fini de pas de temps d'une réalisation de ce dernier. Explicitement, pour une fonction aléatoire Y (également valable pour les variables explicatives X), nous disposons en réalité de n réalisations $\mathrm{Y_i}$ sous la forme :

$$\mathrm{Y_i} = \Big\{(\mathrm{y}_{i,1}, t_{i,1}),\ (\mathrm{y}_{i,2}, t_{i,2}),\ \ldots,\ (\mathrm{y}_{i,m_i}, t_{i,m_i})\Big\},\ \text{pour}\ \begin{array}{c} 1\leq i\leq n \text{ et } 1\leq j\leq m_i \\ t_{i,1}<t_{i,2}<\cdots<t_{i,m_i}\in \mathrm{T} \end{array}$$

Pour i et j fixés, $\mathrm{y}_{i,j}$ est la capture de la courbe $\mathrm{Y_i}$ au point $t_{i,j}$. Et nous devons reconstituer pour chaque i, la trajectoire $\mathrm{Y_i}(.)$ de sorte que :

$$\mathrm{y}_{i,j} = \mathrm{Y}_i(t_{i,j}) + \varepsilon_{i,j} \quad \text{avec } \varepsilon_{i,j} \text{ bruit blanc i.i.d. non observé.}$$

Une méthode courante pour reconstruire $\mathrm{Y_i}$ est de l'exprimer dans une base de fonctions (B-Splines, Fourier, ...).

Dans le cas d'une régression de variable fonctionnelle sur variables fonctionnelles, l'un des modèles proposé par [1] consiste à prendre en compte l'influence de chaque régresseur

fonctionnel sur tout son domaine de définition. Ce modèle appelé *intégral* s'écrit :

$$\mathrm{Y_i}(t_{i,j}) = \beta_0(t_{i,j}) + \sum_{k=1}^{p} \int_{\mathrm{T}} \beta_k(s, t_{i,j})\, \mathrm{X_{k,i}}(s)\, ds. \tag{1}$$

Il est alors d'usage [5] de supposer que les coefficients fonctionnels $\beta_0 : \mathrm{T} \longrightarrow \mathbb{R}$ et $\beta_k : \mathrm{T} \times \mathrm{T} \longrightarrow \mathbb{R}$ s'expriment eux aussi dans une base de fonctions, sous la forme :

$$\beta_0(t_{i,j}) = \sum_{l=1}^{\mathrm{L_0}} \mathrm{b}_{0,l}\, \varphi_{0,l}(t_j) \qquad \text{et} \qquad \beta_k(s, t_{i,j}) = \sum_{l=1}^{\mathrm{L_1^k} \times \mathrm{L_2^k}} \mathrm{b}_{k,l}\, \Phi^k_{1,l}(t_{i,j}) \otimes \Phi^k_{2,l}(t_{i,j})$$

où $(\varphi_{0,l})_l$, $(\Phi^k_{1,l})_l$, et $(\Phi^k_{2,l})_l$ sont des bases de fonctions connues, ou tout du moins choisies par l'utilisateur.
Dans ce cas, l'équation (1) s'écrit en termes de coefficients fonctionnels $\{\mathrm{b}_{kl}\}_{k,l}$ sous forme du modèle de régression multivariée dont le procédé d'estimation est bien connu [5].

2 Modèle de régression sur les signatures

L'utilisation des signatures dans les modèles d'apprentissage statistique s'est beaucoup développé ces dernières années et semble spécialement approprié aux situations où les données sont définies sur les intervalles de tailles diverses. Les références [2, 3] fournissent une excellente introduction et mettent en exergue le fait que cette approche est très adaptée aux données collectées sous forme de fonctions multidimensionnelles du temps.
Soit $\mathrm{X} : \mathrm{T} = [a, b] \longrightarrow \mathbb{R}^d$ avec $d \geq 2$. On appelle signature de X, la collection (infinie) $\mathrm{S(X)}_{a,b}$ de tenseurs d'ordre croissant définis par les intégrales itérées :

$$\mathrm{S(X)}_{a,b} = \Big\{1\,;\, \underbrace{\mathrm{S(X)}^1_{a,b}\,;\ \ldots\ ;\ \mathrm{S(X)}^d_{a,b}}_{\text{coef. d'ordre 1}}\,;\, \underbrace{\mathrm{S(X)}^{1,1}_{a,b}\,;\ \mathrm{S(X)}^{1,2}_{a,b}\,;\ \ldots\ ;\ \mathrm{S(X)}^{d,d}_{a,b}}_{\text{coef. d'ordre 2}}\,;\ldots\Big\} \tag{2}$$

où les coefficients d'ordre 1 sont définis par :

$$\mathrm{S(X)}^i_{a,b} = \int_{a<s<b} \mathrm{dX}^i(s) = \mathrm{X}^i(b) - \mathrm{X}^i(a), \quad 1 \leq i \leq d,$$

les coefficients d'ordre 2 par :

$$\mathrm{S(X)}^{i,j}_{a,b} = \int_{a<s<b} \mathrm{S(X)}^i_{a,s}\, \mathrm{dX}^j(s) = \int_{a<r<s<b} \mathrm{dX}^i(r)\, \mathrm{dX}^j(s), \quad 1 \leq i, j \leq d,$$

et ainsi de suite, ceux d'ordre k se calculant récursivement par la relation :

$$\mathrm{S(X)}^{i_1,i_2,\ldots,i_k}_{a,b} = \int_{a<s<b} \mathrm{S(X)}^{i_1,i_2,\ldots,i_{k-1}}_{a,s}\, \mathrm{dX}^{i_k}(s), \quad 1 \leq i_1, \ldots i_k \leq d.$$

On définit également la signature tronquée à l'ordre m de X, notée $\mathrm{S}^m(\mathrm{X})_{a,b}$, qui est la série (finie) contenant tous les coefficients de la signature d'ordre inférieur ou égal à m. Cette série est de dimension $s_d(m) = \sum_{l=0}^{m} d^l = \frac{d^{m+1}-1}{d-1}$; cf [3].

Le modèle de régression linéaire sur les signatures définie dans [3] cherche à modéliser une relation linéaire entre une cible $\mathrm{Y} \in \mathbb{R}$ et une variable fonctionnelle multidimensionnelle X. Des conditions garantissent l'existence de $m \in \mathbb{N}$ et $\beta_m^\star \in \mathbb{R}^{s_d(m)}$ telles que :

$$\mathbb{E}\big[\mathrm{Y} \,|\, \mathrm{X}\big] \;\; = \langle\ \beta_m^\star\ ,\ \mathrm{S}^m(\mathrm{X})_{a,b}\ \rangle \quad \text{et} \quad \mathrm{Var}(\mathrm{Y} \,|\, \mathrm{X}) = \sigma^2 \leq \infty \tag{3}$$

Il faut noter une simplification dans notre utilisation des signatures par rapport au modèle fonctionnel (1) où le choix de la base d'expansion est une question naturelle et assez complexe, alors que dans le problème (3) seule se pose la question du choix de l'ordre maximal m considéré. En adoptant les signatures, nous avons donc naturellement beaucoup moins d'hyperparamètres à calibrer que dans le modèle (1).

Un des objectifs de notre travail utilisant les signatures est d'aborder le cas où Y n'est plus scalaire comme dans [3] mais fonctionnel. Nous proposons alors de construire au préalable la forme fonctionnelle de $\mathrm{Y_i}$ pour $1 \leq i \leq n$ dans une base de fonctions connue de la forme $\mathrm{Y_i}(t) = \sum_{k=1}^{\mathrm{K}} \theta_{i,k}\ \phi_k(t)$ [1] de sorte que Y soit complètement identifiable par les coefficients $\big\{\Theta_i = (\theta_{i,1}, \ldots, \theta_{i,\mathrm{K}}),\ 1 \leq i \leq n\big\}$. Nous nous retrouvons ensuite avec un problème de régression multi-output sur les signatures de la forme :

$$\Theta_i = \langle\ \beta_m^\star\ ,\ \mathrm{S}^m(\mathrm{X_i})\ \rangle + \varepsilon_i \quad 1 \leq i \leq n.$$

Les coefficients fonctionnels $(\widehat{\Theta}_i)_i$ estimés pour ce modèle seront utilisés pour reconstruire la fonction cible (output) $\widehat{\mathrm{Y}}_\mathrm{i}$.

3 Application: la prédiction de la QoE

Nous nous tournons à présent vers l'implémentation proprement dite des modèles de régression présentés ci-dessus. Les données d'application sont issues du simulateur de réseau OMNeT[8] sur des conversations téléphoniques d'environ 50 secondes pour 800 clients mobiles à l'intérieur d'une cellule 4G. Différentes conditions de simulation ont été réalisées pour un total d'environ 10 000 clients. Tout au long de chaque conversation, dix métriques réseaux ont été collectées à une fréquence inférieure à la seconde et une note MOS (Mean Opinion Score) réelle comprise entre 1 et 5 (normalisé dans $[0, 1]$) traduisant cette QoE est produite après chaque prise de parole de l'auditeur de la conversation à l'aide du modèle de la norme UIT-T G.107.2 [9]. La fréquence d'enregistrement étant propre à la métrique et au client, nous avons pour chaque client $\mathrm{x_i} = (\mathrm{x_{i,1}}, \ldots, \mathrm{x_{i,10}})$ avec $\mathrm{x_{i,j}} \in \mathbb{R}^{l_{i,j}}$ avec $l_{i,j}$ le nombre d'enregistrements du client i sur la métrique j.

Dans le cas du modèle (1) nous avons opté pour la représentation en base de splines de nos différentes variables fonctionnelles. Le nombre K des fonctions de base obtenu par validation croisée leave-one-out est égal à 7. Une fois les fonctions représentées, nous lançons aisément la fonction `pffr` (Penalized Flexible Functional Regression) du package R `refund` [4].
Dans le cas du modèle (3) de régression sur les signatures, c'est le choix de l'ordre m qui contrôle le nombre de coefficients et la faisabilité du modèle. Son estimation telle que présentée dans [3] ne peut se faire rigoureusement ici du fait que le jeu de données à notre disposition ne soit pas exactement sous la même forme ie régulièrement échantillonnée sur les variables pour un même individu. La réflexion en cours sur ce problème, les valeurs de la signatures que nous utiliserons ici seront dépourvues d'indices croisés. Nous ne travaillerons donc qu'avec des termes $\mathrm{S(X)}_{a,b}^{i_1,i_2,\ldots,i_k}$ tels que $i_1 = i_2 = \cdots = i_k \in \{1 \ldots d\}$.
Nous avons effectué une validation croisée pour obtenir $m = 5$.
Pour comparer les performances de ces deux modèles, nous calculons l'erreur quadratique moyenne de prédiction (RMSPE) sur différents échantillons test grâce à la validation croisée K-fold ($K = 10$). Nous obtenons une erreur moyenne de 0.20 pour le modèle fonctionnel traditionnel et de 0.11 pour le modèle utilisant les signatures.
Une observation de quelques trajectoires prédites représentée dans la Figure 1 ci-dessous confirme le RMSPE observé qui montre les meilleurs performances du modèle sur les signatures. On note principalement sa plus grande flexibilité dans l'ajustement des grosses variations de MOS.

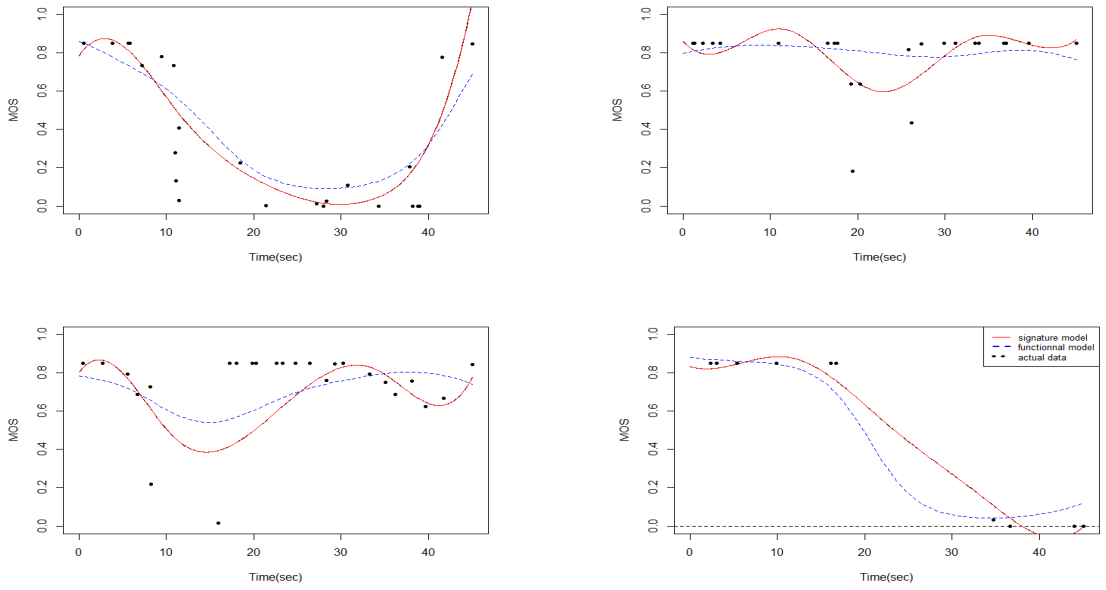

Figure 1: Prédiction des trajectoires de 4 clients. Observations (points noirs), prédiction avec le modèle fonctionnel (pointillés bleu), avec le modèle sur signatures (rouge)

Conclusion

L'objet du présent travail était l'étude comparée d'un modèle fonctionnel traditionnel et d'un modèle basé sur les signatures pour la prédiction de la qualité d'un conversation téléphonique. Nous avons principalement constaté que les signatures ont le bénéfice de nous affranchir de l'étape de la représentation des fonctions de l'ADF alors que cette étape reste incontournable dans l'approche fonctionnelle classique. Les résultats de prédiction nous conduisent à penser que les signatures sont une très bonne piste à poursuivre pour les problèmes de prédiction fonctionnelle avec covariables fonctionnelles. Les signatures présentent également l'avantage de permettre l'analyse de données irrégulièrement échantillonnées au sein d'un même individu.
L'un des défis à affronter dans la suite de notre travail est de tirer profit de la structure des données à analyser pour comprendre comment passer à l'échelle lorsque nous auront à traiter de nombreuses cohortes conjointement.

References

[1] J. O. Ramsay and B. W. Silverman. Functional Data Analysis 2nd ed., 2005. ser. Springer Series in Statistics. New York: Springer-Verlag.

[2] I. Chevyrev and A. Kormilitzin. A primer on the signature method in machine learning, 2016. http ://arxiv.org/abs/1603.03788 arXiv :1603.03788.

[3] A. Fermanian. Learning time-dependent data with the signature transform, PhD Thesis, 2021, Sorbonne Université, Paris.

[4] A. E. Ivanescu, A.-M. Staicu, F. Scheipl, and S. Greven. Penalized function-on-function regression, 2015 Comput Stat 30:539–568

[5] D. Ruppert, M. P. Wand and R. J. Carrol. Semiparametric Regression, 2003. Cambridge University Press, Cambridge, UK.

[6] S. N. Wood. Generalized additive models: An introduction with R, 2006. Chapman & Hall/CRC, New York

[7] K.-T. Chen. Integration of paths - a faithful representation of paths by non-commutative formal power series, 1958. Transactions of the American Mathematical Society 89 395–407.

[8] A. Varga and OpenSim Ltd. OMNeT++ Simulation Manual version 6.5.1, 2016

[9] Union Internationale des Télécommunications, Recommandation G.107.2 "Fullband E-model", 06/2019