Modèle de mélanges d'experts pour données fonctionnelles

Jean Steve TAMO TCHOMGUI, Julien Jacques, Stéphane Chrétien, Guillaume Fraysse, Vincent Barriac

Modèle de mélanges d'experts pour données fonctionnelles

Jean Steve Tamo Tchomgui[1,2] & Julien Jacques[1] & Stéphane Chrétien[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: Dans le cadre de la prédiction d'une variable fonctionnelle par une ou plusieurs variables fonctionnelles, nous proposons un modèle de régression de mélange d'experts (ME). Les ME sont utiles à l'analyse des données hétérogènes et les données de nature fonctionnelle auxquelles nous souhaitons appliquer ce modèle sont désormais très présentes dans divers domaines. Notre contribution principale est celle de montrer que ce type de mélange de régression permet de produire des résultats souvent meilleures qu'en utilisant un unique modèle de régression.
Mots-clés. Mélange d'Experts, Modèles de mélange, Données fonctionnelles, Modèles de régression fonctionnels, Algorithme EM.

Abstract: In the present work, we propose a mixture of experts-type (ME) regression model for predicting of a functional variable by regressing on several functional variables. ME's have been found very useful for analysing heterogeneous data and functional data, which are pervasive in various practical applications and often pose subtle modelling and calibration challenges. Our main contribution is to provide novel relevant statistical and numerical approaches to address funtions to function regression problems. Our algorithms are shown to outperform simple regression models throughout various numerical experiments.
Keywords. Mixture of Experts, Mixture models, Functional data, Functional regression models, EM algorithm.

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 $\mathrm{Y} \in \mathbb{R}$ décrite par un nombre fini de prédicteurs $\mathrm{X} = (\mathrm{X}^1, \ldots, \mathrm{X}^p) \in \mathbb{R}^p$. On observe plutôt de multiples enregistrements au cours du temps, à la fois pour la sortie, $Y : \mathrm{T} \to \mathbb{R}$, et les prédicteurs $\mathrm{X}^j : \mathrm{T} \to \mathbb{R}^p$.

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 Ramsay and Silverman (2005), 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. Elle 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 consistant à reconstruire la nature fonctionnelle des données.
En présence de données hétérogènes, imposer une unique structure de régression à l'ensemble du jeu de données devient non pertinent pour avoir des performances correctes. Les modèles de mélanges de régression sont alors indiqués dans une telle situation (DeSarbo and Cron (1988), McLachlan and Peel (2000)). Dans sa formulation originale, dédiée à la modélisation explicative, les covariables sont considérées comme déterministes. Elles ne fournissent pas d'informations sur le groupe auquel un individu est susceptible d'appartenir, c'est-à-dire que la probabilité d'appartenance à un groupe est indépendante de ces covariables. Ne pas savoir à quelle composante du mélange de régression un nouvel individu appartient s'est rapidement avéré problématique dans le contexte de la modélisation prédictive. Pour résoudre ce problème, nous considérons que les proportions du mélange dépendent des covariables. C'est le cadre du modèle de mélange d'experts (Dayton and Macready, 1988).

Dans cet article, la Section 1 est consacrée à présenter brièvement le cadre d'analyse des données fonctionnelles et un modèle linéaire de régression d'une variable fonctionnelle sur plusieurs variables fonctionnelles. La Section 2 introduit le modèle de mélange d'experts proposé et son inférence. La Section 3 présente les résultats obtenus sur un jeu de données simulées.

1 Régression 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. En pratique, les fonctions aléatoires ne sont pas directement observées. Explicitement, pour une réalisation $\mathrm{X}_i^j(t)$ du $j$ème prédicteur, nous disposons en réalité de m_i observations de $\mathrm{X}_i^j(t)$ à des pas de temps $t_{i,1} < \cdots < t_{i,m_i} \in \mathrm{T}$: $\left\{(\mathrm{x}_{i,1}^j, t_{i,1}), \ldots, (\mathrm{x}_{i,m_i}^j, t_{i,m_i})\right\}$, où $\mathrm{x}_{i,\ell}^j$

est l'observation de $\mathrm{X}_i^j(t)$ au temps $t_{i,\ell}$. Nous devons donc reconstituer pour chaque i, la trajectoire $\mathrm{X}_i^j(.)$ en supposant que :

$$\mathrm{x}_{i,\ell}^j = \mathrm{X}_i^j(t_{i,\ell}) + \varepsilon_{i,\ell}^j \quad \text{avec } \varepsilon_{i,\ell}^j \text{ un bruit blanc indépendant et identiquement distribué.}$$

Une méthode courante pour estimer les fonctions $\mathrm{X}_i^j(.)$ est de supposer qu'elles peuvent se décomposer une base de fonctions (B-Splines, Fourier, ...) :

$$\mathrm{X}_i^j(t) = \sum_{l=1}^{\mathrm{L}_j} x_{i,l}^j \, \mathrm{B}_l^j(t) = \mathrm{B}^j(t)^\top \, x_i^j \quad \text{avec } 1 \leq j \leq p. \tag{1}$$

avec $x_i^j = \{x_{i,l}^j\}_l$ le vecteur de coefficients de base et $\mathrm{B}^j(t) = \left\{\mathrm{B}_l^j(t)\right\}_l$ le vecteur des fonctions de base. Dans le cas d'une régression de variable fonctionnelle sur variables fonctionnelles, l'un des modèles abordés par Ramsay and Silverman (2005) consiste à prendre en compte l'influence de chaque régresseur fonctionnel au même instant que la réponse pour l'expliquer. C'est le modèle concurrentiel donné par :

$$\mathrm{Y_i}(t) = \beta_0(t) \, + \, \sum_{j=1}^{p} \beta_j(t) \, \mathrm{X}_i^j(t) + \varepsilon_i(t). \tag{2}$$

Il est alors d'usage (Ruppert et al., 2003) de supposer que les coefficients fonctionnels $\beta_j : \mathrm{T} \longrightarrow \mathbb{R}$ avec $0 \leq j \leq p$ s'expriment eux aussi dans une base de fonctions, sous la forme :

$$\beta_j(t) \;\; = \;\; \sum_{l=1}^{\mathrm{L_j}} \mathrm{b}_{j,l} \, \mathrm{C}_{j,l}(t) \tag{3}$$

où les $\{\mathrm{C}_{j,l}\}_l$ forment une base de fonctions connues, choisies par l'utilisateur, et les coefficients fonctionnels $\{\mathrm{b}_{j,l}\}_{j,l}$.

2 Modèles de mélange d'experts

Les modèles de mélange de régression constituent le sous-ensemble de la grande catégorie des modèles statistiques connus sous le nom de modèles de mélange fini dont l'objectif principal est de modéliser l'hétérogénéité d'une population par le biais d'un ensemble fini de classes latentes. La classe des modèles de mélange de régressions consiste à supposer qu'il existe K composantes dans la population, chacune d'entre elles suivant une distribution paramétrique. L'appartenance à une composante est indiquée par une variable catégorielle latente Z = 1, 2, ..., K. Il existe à notre connaissance, deux options de modélisation des poids des mélanges. La première est le mélange de régression simple où

les proportions du mélange π_k ne dépendent pas des variables exogènes observées X(.). Cela suppose implicitement que la moyenne de X(.) est identique dans chaque classe latente. La deuxième modélisation, plus générale, est le mélange d'experts qui inclut une partie ou la totalité des prédicteurs dans la prédiction des composantes latentes.

La densité conditionnelle du mélange à K composantes dans ce cas s'écrit :

$$f\Big(\mathrm{Y}_i(t)\,|\,\mathrm{X}_i(t),\ t\in\mathrm{T},\ \Psi\Big) \;=\; \sum_{k=1}^{\mathrm{K}} \pi_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big)\ \underbrace{f_k\big(\mathrm{Y}(t)\,|\,\mathrm{X}(t),\ \theta_k\big)}_{\text{Expert } k} \tag{4}$$

avec

- $\pi_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big)$ le poids de mélange du groupe k, plus communément appelé fonction d'activation de l'expert k ;
- $\Psi=\{\alpha_k,\theta_k\}_{1\leq k\leq \mathrm{K}}$ le vecteur de paramètres à estimer ;
- $f_k\big(\mathrm{Y}(t)\,|\,\mathrm{X}(t),\ \theta_k\big)$ la densité de la distribution conditionnelle $\mathrm{Y}(t)\,|\,\mathrm{X}(t)$.

Ce modèle peut être considéré comme un sous-modèle du modèle de classe latente proposé par Dayton and Macready (1988), appelé modèle de classe latente à variables concomitantes. Il existe plusieurs choix pour modéliser la fonction d'activation, mais la version la plus courante et originale proposée par l'auteur est obtenue en reliant les covariables par un modèle logistique multinomial :

$$\begin{aligned}\pi_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big) &= \mathbb{P}\Big(\mathrm{Z}=k\,|\,\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\Big)\\ &= \frac{\exp\Big(h_k\big(\mathrm{X_i}(t)\ t\in\mathrm{T},\ \alpha_k\big)\Big)}{1+\sum_{k'=1}^{\mathrm{K}-1}\exp\Big(h_{k'}\big(\mathrm{X_i}(t)\ t\in\mathrm{T},\ \alpha_{k'}\big)\Big)},\end{aligned} \tag{5}$$

où la fonction h_k est défini par :

$$\begin{aligned}h_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big) &= \int_{\mathrm{T}}\Big(\alpha_{k,0}(t)+\sum_{l=1}^{p}\alpha_{k,l}(t)\,\mathrm{X}_i^l(t)\Big)\,dt\\ &= \tilde{\alpha}_{k,0}+\int_{\mathrm{T}}\sum_{j=1}^{p}\alpha_{k,j}(t)\,\mathrm{X}_i^j(t)\,dt=\tilde{\alpha}_{k,0}+\int_{\mathrm{T}}\alpha_k^{\top}(t)\,\mathrm{X}_i(t)\,dt\end{aligned}$$

$$\text{avec}\qquad \alpha_k(t)=\begin{pmatrix}\alpha_{k,1}(t)\\ \alpha_{k,2}(t)\\ \vdots\\ \alpha_{k,p}(t)\end{pmatrix}\quad\text{et}\quad \mathrm{X}_i(t)=\begin{pmatrix}\mathrm{X}_i^1(t)\\ \mathrm{X}_i^2(t)\\ \vdots\\ \mathrm{X}_i^p(t)\end{pmatrix}.$$

En exprimant les paramètres fonctionnels $\alpha_k(t)$ dans une base de fonctions de la forme $\alpha_{k,j}(t) = \sum_{l=1}^{\mathrm{L}_\alpha} a_{k,l}^j \phi_l^j(t) = \phi^j(t)^\top a_k^j$, la fonction h_k s'écrit sous forme matricielle et par extension $\pi_k(.)$ également.
Une fois la fonction d'activation modélisée, il reste à modéliser les densités conditionnelles des K experts. Elle découle du modèle de régression (2) et s'écrit pour chaque i :

$$\mathrm{Y}_i(t) = \beta_{k,0}(t) + \sum_{j=1}^{p} \beta_{k,j}(t)\mathrm{X}_i^j(t) + \varepsilon_i(t) \quad \text{si } \mathrm{Z}_i = k \tag{6}$$

Les $\beta_{k,j}(t)$ avec $1 \leq k \leq \mathrm{K}$ et $1 \leq j \leq p$ sont les paramètres, et sont supposés être décomposables dans une base de fonctions choisie a priori.
En combinant les expressions obtenus pour la fonction d'activation et pour les densités conditionnelles des K experts, on obtient la densité conditionnelle du mélange :

$$f\Big(\mathrm{Y}_i(t) \,|\, \mathrm{X}_i(t),\ t \in \mathrm{T},\ \Psi\Big) = \sum_{k=1}^{\mathrm{K}} \pi_k\big(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_k\big)\, \Phi\Big(\mathrm{Y}_i(t)\,;\, \beta_k(t)\mathrm{X}_i(t), \sigma_k^2\Big), \tag{7}$$

où $\Phi(y; \mu, \sigma^2)$ est la densité de la loi normale de moyenne μ, de variance σ^2 en y. Les paramètres sont estimés par maximum de vraisemblance via l'algorithme EM (Dempster et al., 1977). Le choix du nombre de classes K est obtenu par le critère BIC.

3 Application sur données simulées

3.1 Simulation des données

Nous simulons 2 jeux de données de tailles $n_1 = 500$ et $n_2 = 1000$ observations. Chaque jeu de données est constitué de 5 prédicteurs fonctionnels générés sur une grille de $m = 50$ points d'observations $(t_\ell)_\ell$ sur l'intervalle $[0, 2\pi]$ selon la relation

$$\mathrm{X}_i^l(t_\ell) = \xi_{i,1}^l + \Big(\log(10 + t_\ell)\Big)^{\xi_{i,2}^l} + \xi_{i,3}^l \sin\Big(\frac{2\pi t_\ell}{\xi_{i,4}^l}\Big)$$

où $\xi_{i,r}^l$ sont des constantes tirées selon la loi $\mathcal{U}\big([\text{-}1, 1]\big)$ ($1 \leq r \leq 4,\ 1 \leq l \leq 5$ et $1 \leq i \leq n_1$ ou n_2). Nous simulons par la suite les paramètres fonctionnels $\big(\beta_j(t)\big)_{0 \leq j \leq 5}$ données par :

$$\begin{cases} \beta_{0,k}(t_\ell) & = \Big(\log(10 + t_\ell)\Big)^{\rho_{0,k}} \\ \beta_{j,k}(t_\ell) & = \rho_1^{j,k} \sin\Big(\frac{2\pi t_\ell}{\rho_2^{j,k}}\Big) \end{cases} \quad \text{où} \quad \begin{cases} \rho_{0,k},\ \rho_1^{j,k},\ \rho_2^{j,k} \text{ des constantes.} \\ \text{avec } 1 \leq j \leq 5 \text{ et } 1 \leq k \leq \mathrm{K}. \end{cases}$$

Le nombre de classes K est fixé ici à 4 et est équilibré sur le nombre d'observation total. Une fois les prédicteurs et paramètres simulés, on construit la variable réponse selon le modèle (6) avec $\varepsilon_i(t_\ell) \sim \mathcal{N}(0,1)$ et nous répétons ces expériences $N = 50$ fois.

3.2 Critères d'évaluation

Nous allons utiliser trois critères pour évaluer la qualité d'estimation de notre modèle : le premier est le ratio du nombre de modèles ayant obtenu le bon nombre de classes $\mathrm{K} = 4$ sur l'ensemble de nos simulations ; Ensuite, pour les modèles ayant le bon nombre de classes, nous évaluerons la qualité d'estimation des paramètres à travers la racine de l'erreur quadratique moyenne (REQM) donnée par :

$$\mathrm{REQM}\Big(\beta_{j,k}(.)\Big) \quad = \quad \left[\frac{1}{m}\sum_{\ell=1}^{m}\Big(\beta_j(t_\ell) - \widehat{\beta}_j(t_\ell)\Big)^2\right]^{1/2}. \tag{8}$$

Enfin, nous évaluerons la qualité de prédiction des modèles sur un échantillon test de taille $n_{\text{test}} = 2000$ qui n'aura servi à entraîner aucun des modèles. Le critère utilisé sera l'Erreur Quadratique Relative Moyenne (EQRM) donné par :

$$\mathrm{EQRM}_i \quad = \quad \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n}\Big(\mathrm{Y}_i(t_j) - \widehat{\mathrm{Y}}_i(t_j)\Big)^2}{\sum_{i=1}^{n}\mathrm{Y}_i(t_j)^2}\right). \tag{9}$$

3.3 Résultats

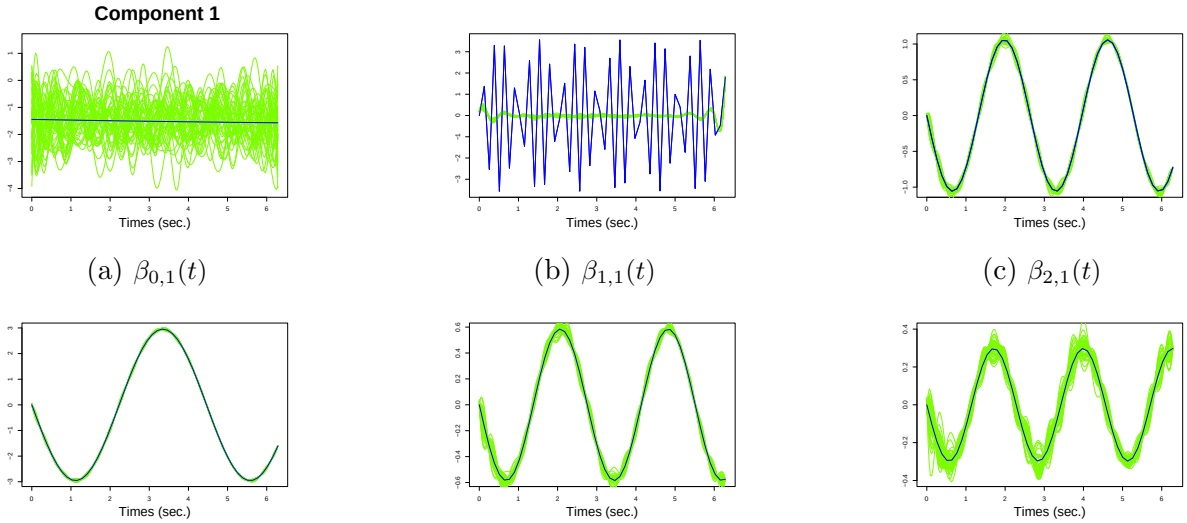
L'efficacité du critère BIC pour le choix du nombre de classes K est évalué sur chaque échantillon en testant différentes valeurs de $\mathrm{K} \in \{2,\ldots,6\}$. La Table 1 donne les fréquences de choix de chacun des nombres de clusters testés en fonction des tailles d'échantillon.

Nombre de classes		3	**4**	5
% de modèles	$n_{\text{train}} = n_1$	6%	**86%**	8%
	$n_{\text{train}} = n_2$	0%	**92%**	8%

Table 1: Résultats de nos $N = 50$ simulations en termes de nombre de classes.

La Table 1 nous montre que lorsque la taille de l'échantillon d'apprentissage augmente, le bon nombre de classe est plus souvent obtenu. Le BIC est un critère pertinent pour le choix du nombre de classes.

La Figure 1, quant à elle, montre pour les modèles ayant retenu le bon nombre de classes (K = 4) les paramètres estimés versus les vrais paramètres pour les modèles lorsque $n_{\text{train}} = n_1$.

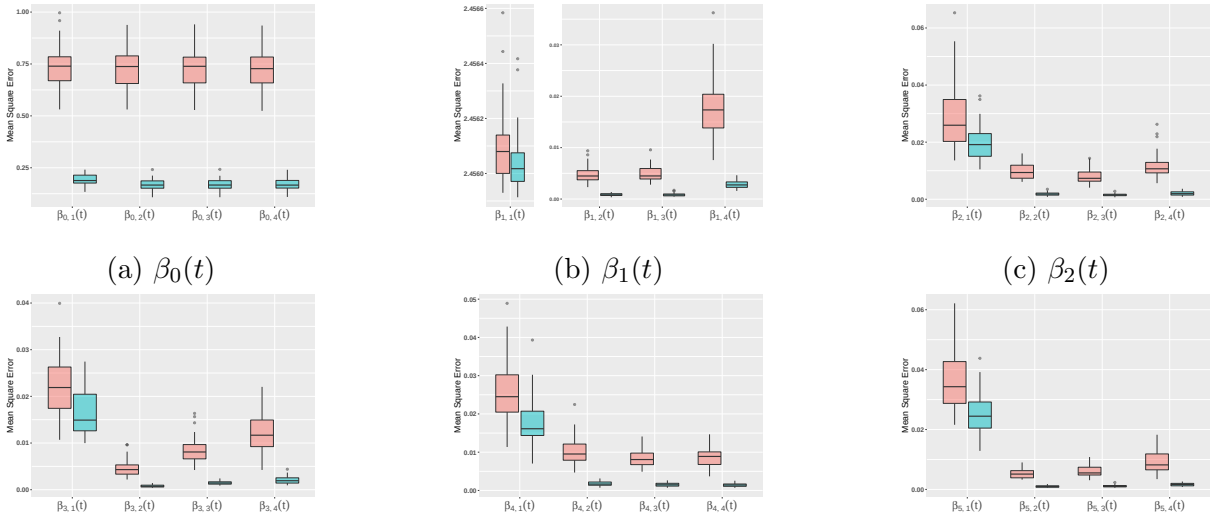


(a) $\beta_{0,1}(t)$ (b) $\beta_{1,1}(t)$ (c) $\beta_{2,1}(t)$

(d) $\beta_{3,1}(t)$ (e) $\beta_{4,1}(t)$ (f) $\beta_{5,1}(t)$

Figure 1: Paramètres estimés (vert) et vrais (bleu) dans nos simulations pour $n_{train} = n_1$ pour la classe 1.

On note globalement une assez bonne estimation des paramètres lorsque la forme est assez régulière (cf. Figures 2c, 2d, 2e et 2f). Lorsque la forme du paramètre est très simple (2a) ou très complexe (2b), on a du mal à avoir une bonne correspondance. Ceci est dû au choix du nombre de fonctions de base L_j que l'on décide d'allouer a priori à chaque paramètre dans l'Equation (3). Ne connaissant pas en application réelle la forme du paramètre, ce choix peut s'avérer être très déterminant puisqu'il conditionne également la complexité du problème à résoudre. Des techniques de régularisation des paramètres (Leurgans et al. (1993), James et al. (2009)) existent et permettent plus ou moins de régler ce problème.



(a) $\beta_0(t)$ (b) $\beta_1(t)$ (c) $\beta_2(t)$

(d) $\beta_3(t)$ (e) $\beta_4(t)$ (f) $\beta_5(t)$

Figure 2: Boxplot de REQM pour tous les paramètres estimés (dans les 4 classes) sur les deux scénarios de simulation considérés : $n_{\text{train}} = n_1$ (rouge) et $n_{\text{train}} = n_2$ (bleu).

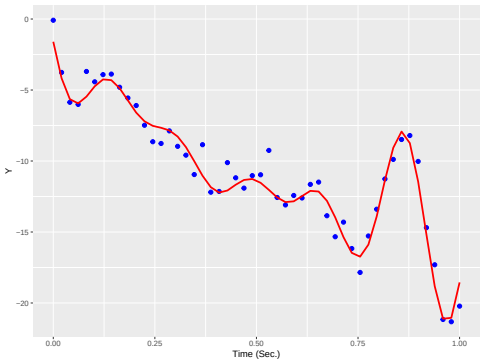
La Figure 2 quant à elle montre les boxplots de REQM obtenus sur les modèles pour les deux tailles d'échantillon d'apprentissage considérés. Les valeurs obtenues sont assez faibles ce qui est plutôt correct et le constat assez clair qui en ressort est que lorsque n_{train} augmente, la REQM baisse.

Scénarios	EQRM
$n_{\text{train}} = n_1$	0.0231 (0.0065)
$n_{\text{train}} = n_2$	0.0221 (0.0040)

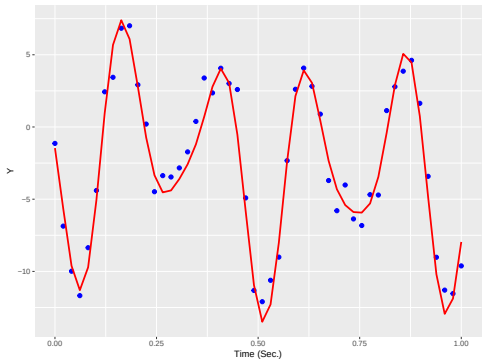
Table 2: EQRM moyen (et ecart-type) obtenu sur l'échantillon test.

Le dernier critère à regarder pour juger de la bonne estimation du modèle est la qualité de prédiction. Le Tableau 2 montre l'erreur relative de prédiction obtenue sur l'échantillon test pour nos différentes simulations lorsque $n_{\text{train}} = n_1$ et lorsque $n_{\text{train}} = n_2$. Le même constat que sur l'estimation des paramètres est observé. La prédiction s'améliore lorsque le nombre de données en entraînement augmente. Ce qui montre la convergence de notre estimation.

La Figure 3 enfin montre pour 2 individus choisis aléatoirement la prédiction fonctionnelle $\hat{Y}_i(t)$ obtenue. Cette dernière s'ajuste bien aux variations non régulières des données observées.



(a) prédiction 1



(b) prédiction 2

Figure 3: Prédiction vs vrais valeurs de la réponse fonctionnelle pour 2 individus choisis aléatoirement.

Conclusion

Dans ce travail préliminaire, nous avons présenté le modèle de régression de mélange d'experts pour la prédiction d'une variable fonctionnelle à partir de plusieurs prédicteurs fonctionnels. Nous avons montré au moyen d'une étude par simulations que les estimations obtenues étaient convergentes. La suite de ce travail consistera

- à considérer un modèle plus général pour modéliser les experts à l'instar du modèle intégral par exemple,
- à améliorer l'estimation des paramètres en introduisant une pénalisation adaptative qui prenne en compte la régularité intrinsèque des fonctions que nous décomposons dans les bases de Splines, et enfin
- à comparer aux méthodes existantes sur des jeux de données réelles pour évaluer la pertinence de notre algorithme, en particulier en contexte industriel.

Bibliographie

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, 55(3):725 – 740.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.