Parcimonie par intervalle pour la régression inverse par tranche fonctionnelle

Victor Picheny [1], Rémi Servien [2] & Nathalie Villa-Vialaneix[1]

[1] *INRA, UR 0875 MIAT, 31326 Castanet Tolosan cedex, France*
{victor.picheny,nathalie.villa}@toulouse.inra.fr
[2] *INRA - ENVT, Université de Toulouse, UMR1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France* remi.servien@toulouse.inra.fr

Résumé. Dans cette proposition de communication, nous présentons une approche de sélection de variables par intervalle dans le cadre d'un modèle semi-paramétrique de régression fonctionnelle. L'objectif est de détecter dans un cadre de réduction de dimension, par exemple pour des séries temporelles de grande taille, les intervalles temporels explicatifs pour la variable à régresser. Nous montrons que ce problème revient à résoudre consécutivement deux problèmes de régression pénalisée. Notre approche est illustrée sur un problème jouet.

Mots-clés. régression fonctionnelle, SIR, lasso, régression régularisée

Abstract. In this proposal, a semi-parametric functional model is described which aims at selecting relevant intervals for the prediction in a functional regression framework. For the case of large time series, the purpose is to detect temporal intervals in the predictors for a dimension reduction method which explains a given variable. Our approch is illustrated on a toy example.

Keywords. functional regression, SIR, lasso, ridge regression

1 Introduction

Dans de nombreuses applications, les données, qui se présentent sous la forme de vecteurs de grande dimension, sont en fait des enregistrements en divers points d'évaluation de phénomènes continus. On peut citer, comme exemple de données de ce type, les données météorologiques (courbes de température et précipitation), les séries temporelles financières, les données spectrométriques en chimiométrie ou diverses données issues du séquençage haut débit en biologie. Une introduction à l'*analyse de données fonctionnelle* peut être trouvée dans [Ramsay and Silverman, 1997, Ferraty and Vieu, 2006].

Un problème complexe avec ce type de données est que leur dimension (c'est-à-dire le nombre de points d'échantillonnage, p) est souvent très supérieure au nombre d'observations (c'est-à-dire le nombre de courbes, n) disponibles. Dans cette proposition de communication, nous nous intéressons à un modèle de régression fonctionnelle dans lequel

une variable réelle, Y, doit être prédite à partir d'une variable explicative fonctionnelle, X. Le modèle que nous étudions est un modèle semi-paramétrique qui est une extension de la méthode SIR (Sliced Inverse Regression, [Li, 1991]) au cadre fonctionnel. Le principe de SIR est de trouver un espace de faible dimension pour la projection de X qui explique au mieux Y. SIR nécessite d'inverser la matrice de variance de X, ce qui est impossible en grande dimension ($n < p$) ou dans le cadre fonctionnel, et des adaptations de l'approche initiale, par régularisation ou pénalisation, ont donc été proposées afin de pallier ce problème [Zhong et al., 2005, Li and Yin, 2008, Bernard-Michel et al., 2008, Li and Nachtsheim, 2008, Coudret et al., 2014, Ferré and Yao, 2003].

Ici, nous présentons une approche de sélection de variables pour la SIR qui est adaptée au cadre fonctionnel. En effet, les approches multi-dimensionnelles usuelles de sélection de variables ne sont pas toujours pertinentes dans le cadre fonctionnel : dans la plupart des situations, la variable Y est intrinsèquement dépendante d'un ou plusieurs intervalles (et non pas de points de mesure isolés) bien plus petit que l'ensemble du temps d'enregistrement de la variable X. De plus, des décalages entre courbes font que les parties permettant d'expliquer la variable Y ne peuvent pas être des points de mesure isolés mais des sous-intervalles entiers de l'intervalle de définition des variables fonctionnelles X. Nous proposons ici une approche basée sur une pénalité L_1 qui permet d'identifier de tels intervalles.

2 Description de la méthode Sparse Interval-SIR (SI-SIR)

2.1 Contexte et notations

Dans cette partie, on notera (X, Y) une paire de variables aléatoires telle que X est une variable aléatoire fonctionnelle observées à des points $\tau = \{t_1, \ldots, t_p\}$ supposés donnés et déterministes et Y est une variable aléatoire réelle. n i.i.d. observations de (X, Y), $(x_i, y_i)_{i=1,\ldots,n}$ sont connues sur τ. On note également $\mathbf{x}_i = (x_i(t_j))_{j=1,\ldots,p} \in \mathbb{R}^p$ la i-ème observation, $\mathbf{x}^j = (x_i(t_j))_{i=1,\ldots,n} \in \mathbb{R}^n$ la j-ème variable et x_{ij} l'observation $x_i(t_j)$. Enfin, la matrice $n \times p$, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, est notée $\mathbf{X}$.

Notre objectif est d'estimer un « espace central », $\mathcal{S}_{Y|X}$, qui est le plus petit sous-espace de $\mathbb{R}^p$ tel que la projection de $\mathbf{X}$ sur $\mathcal{S}_{Y|X}$ contient toute l'information sur Y disponible dans X. De manière plus précise, on se place dans le cadre du modèle

$$Y = F(\mathbf{a}_1^T X, \ldots, \mathbf{a}_d^T X, \epsilon),$$

avec $(\mathbf{a}_j)_{j=1,\ldots,d} \in \mathbb{R}^p$, $d < p$, $F : \mathbb{R}^{p+1} \to \mathbb{R}$ est une fonction inconnue et ϵ est un terme d'erreur indépendant de X. On définit alors $\mathcal{S}_{Y|X} = \{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$.

Nous supposons de plus que seuls certains intervalles temporels sont utiles pour la prédiction, ce qui revient à faire l'hypothèse qu'un grand nombre de valeurs consécutives

des vecteurs $\mathbf{a}_j$ sont nulles. De manière plus précise, on considère que l'intervalle de définition de X est découpé en D sous-intervalles, $(\tau_k)_{k=1,\dots,D}$, de telle sorte qu'il existe un nombre restreint d'indices k tels quel $\mathbf{a}_j(t) \neq 0$ si et seulement si $t \in \tau_k$.

[Li, 1991] montre que les $(\mathbf{a}_j)_j$ peuvent être estimés par une décomposition spectrale qui fait intervenir l'espérance conditionnelle $\mathbb{E}(X|Y)$. Cette dernière est estimée en découpant le support de Y en H tranches qui sont des intervalles disjoints et consécutifs, $(\mathcal{S}_h)_{h=1,\dots,H}$. Dans [Chen and Li, 1998], les auteurs proposent différentes reformulations de la SIR, comme des problèmes de régression ou de décompositions spectrales qui peuvent être utilisées comme base pour régulariser ou pénaliser le problème dans le cadre de la grande dimension.

La méthode que nous proposons se déroule en deux étapes : une première est une procédure de pré-estimation régularisée qui est adaptée à la grande dimension. La deuxième est une étape de sélection de variables par intervalles qui est effectuée par introduction de coefficients de rétrécissement (*shrinkage*).

2.2 SIR régularisée

Dans une première étape, nous utilisons l'approche régularisée introduite dans [Bernard-Michel et al., 2008], qui est une correction de la méthode proposée par [Li and Yin, 2008]. Un estimateur de la matrice $A = (\mathbf{a}_1, \dots, \mathbf{a}_d)$ est obtenu par résolution du problème d'estimation régularisé (*ridge*) qui s'exprime comme la minimisation de

$$\mathcal{E}_{r,1}(A, C) = \sum_{h=1}^{H} \hat{p}_h \left\| \left(\overline{X}_h - \overline{X}\right) - \widehat{\Sigma} A C_h \right\|^2_{\widehat{\Sigma}^{-1}} + \mu_2 \sum_{h=1}^{H} \hat{p}_h \|A C_h\|^2, \tag{1}$$

dans lequel $\hat{p}_h = \frac{n_h}{n}$, où n_h est le nombre d'observations dans $\mathcal{S}_h$, la tranche numéro h, $\overline{X}_h$ est la moyenne des observations $\mathbf{x}_i$ dans la tranche $\mathcal{S}_h$, $\overline{X}$ est la moyenne empirique des observations $\mathbf{x}_i$, $\widehat{\Sigma}$ est la matrice de variance empirique des $\mathbf{x}_i$, $C = (C_1, \dots, C_H)$ et les C_h sont des vecteurs de dimension d. [Bernard-Michel et al., 2008] montrent que la solution, en A, de l'équation (1) est obtenue par décomposition spectrale de la matrice $\left(\widehat{\Sigma} + \mu_2 \mathbb{I}_p\right)^{-1} \widehat{\Gamma}$, avec $\mathbb{I}_p$ la matrice identité de taille p et $\widehat{\Gamma}$ l'estimateur SIR de la variance de $\mathbb{E}(X|Y)$, $\widehat{\Gamma} = \sum_{h=1}^{H} \hat{p}_h \left(\overline{X}_h - \overline{X}\right)\left(\overline{X}_h - \overline{X}\right)^T$.

2.3 SI-SIR

Dans une seconde étape, de manière similaire à [Li and Nachtsheim, 2008, Li and Yin, 2008], nous utilisons l'estimation *ridge* obtenue à l'étape précédente pour proposer un espace central parcimonieux par intervalle.

Ainsi, si $\hat{A}$ est l'estimateur obtenu par optimisation de l'équation (1), on peut définir

des estimations de la projection de $(\widehat{\mathbb{E}}(X|Y=y_i))_{i=1,\dots,n}$ dans l'espace central par :

$$\mathcal{P}_{\hat{A}}(\widehat{\mathbb{E}}(X|Y=y_i)) = (\overline{X}_h - \overline{X})^T \hat{A} \qquad \text{avec } h \text{ tel que } y_i \in \mathcal{S}_h,$$

où $\widehat{\mathbb{E}}(X|Y=y_i) = \overline{X}_h$ pour h tel que $y_i \in \mathcal{S}_h$. Dans la suite, on notera $\mathbf{P}_i = (\mathcal{P}_i^1, \dots, \mathcal{P}_i^d)^T \in \mathbb{R}^d$ cette quantité. On notera également $\mathbf{P}^j$ (pour $j = 1, \dots, d$) les observations des j-èmes coefficients pour toutes les projections : $\mathbf{P}^j = (\mathcal{P}_1^j, \dots, \mathcal{P}_n^j)^T \in \mathbb{R}^n$.

Dans l'esprit de [Li and Nachtsheim, 2008], nous proposons une estimation basée sur une reformulation en problème de régression linéaire multiple de la SIR qui est donnée par le fait que les vecteurs $\mathbf{a}_j$ peuvent aussi être vus comme minimisant

$$\mathcal{E}(\mathbf{a}_j) = \sum_{i=1}^{n} \left[\mathcal{P}_{\mathbf{a}_j}(X|y_i) - (\mathbf{a}_j)^T \mathbf{x}_i\right],$$

où $\mathcal{P}_{\mathbf{a}_j}(X|y_i)$ est la projection de $\mathbb{E}(X|Y=y_i)$ sur $\mathbf{a}_j$. Une estimation parcimonieuse des $\mathbf{a}_j$ peut être obtenue en résolvant d problèmes *lasso* indépendants ($\min_{\mathbf{a}_j} \mathcal{E}(\mathbf{a}_j) + \mu_1 \|\mathbf{a}\|_{L_1}$, pour $j = 1, \dots, d$).

Cependant, cette approche ne permet pas d'obtenir une parcimonie identique pour toutes les dimensions de l'espace central estimé, ni de gérer une parcimonie « par intervalle ». Nous lui préférons donc une idée proche de celle présentée dans [Li and Yin, 2008] qui introduit la contrainte de parcimonie via des coefficients de rétrécissement.

En s'appuyant sur les D intervalles $(\tau_k)_{k=1,\dots,D}$ qui partitionnent l'intervalle de définition de X, on introduit $\boldsymbol{\alpha} \in \mathbb{R}^D$. On cherche alors à résoudre :

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^D} \sum_{j=1}^{d} \|\mathbf{P}^j - (\mathbf{X}\Delta(\hat{\mathbf{a}}_j))\,\boldsymbol{\alpha}\|^2 + \mu_1 \|\boldsymbol{\alpha}\|_{L_1},$$

avec $\Delta(\hat{\mathbf{a}}_j)$ la matrice $(p \times D)$ telle que $\Delta_{kl}(\hat{\mathbf{a}}_j) = \hat{a}_{jl}$ si $t_l \in \tau_k$ et 0 sinon. Or, ce problème peut s'écrire sous la forme pénalisée de type *lasso* suivante

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^D} \|\mathbf{P} - \left(\mathbf{X}\Delta(\hat{A})\right)\boldsymbol{\alpha}\|^2 + \mu_1 \|\boldsymbol{\alpha}\|_{L_1}$$

avec $\mathbf{P} = \begin{pmatrix} \mathbf{P}^1 \\ \vdots \\ \mathbf{P}^d \end{pmatrix}$, un vecteur de taille dp et $\Delta(\hat{A}) = \begin{pmatrix} \Delta(\hat{\mathbf{a}}_1) \\ \vdots \\ \Delta(\hat{\mathbf{a}}_p) \end{pmatrix}$, une matrice de dimension $(dp) \times D$.

On pose enfin $\tilde{\mathbf{a}}_j = \Lambda \hat{\mathbf{a}}_j$, où $\Lambda = \text{Diag}\left(\alpha_1 \mathbb{I}_{|\tau_1|}, \dots, \alpha_D \mathbb{I}_{|\tau_D|}\right) \in \mathcal{M}_{p \times p}$. Une fois les vecteurs $(\tilde{\mathbf{a}}_j)_{j=1,\dots,d}$ obtenus, une orthonormalisation de Hilbert-Schmidt est appliquée pour les rendre $\widehat{\Sigma}$-orthonormaux.

3 Illustration

Le modèle utilisé pour générer les données est

$$Y = \log|\langle X, \mathbf{a}_1\rangle| + \varepsilon$$

où $X \sim \mathcal{GP}(m(.), c(.,.))$ est un processus gaussien de moyenne $m(t) = -5+4t-4t^2$ et de covariance $c(t,t') = \frac{1}{10}(1+15|t-t'|+\exp(-15|t-t'|))$ (covariance de Matérn de paramètre $\nu = 3/2$) et où $\mathbf{a}_1(t) = \sin\left(\frac{3\pi t}{2}\right)\mathbb{I}_{[0.1,0.2]}$ (voir Figure 1, a-c). La taille de l'échantillon est $n = 100$ et les fonctions sont observées en $p = 300$ points répartis sur une grille uniforme dans $[0,1]$.

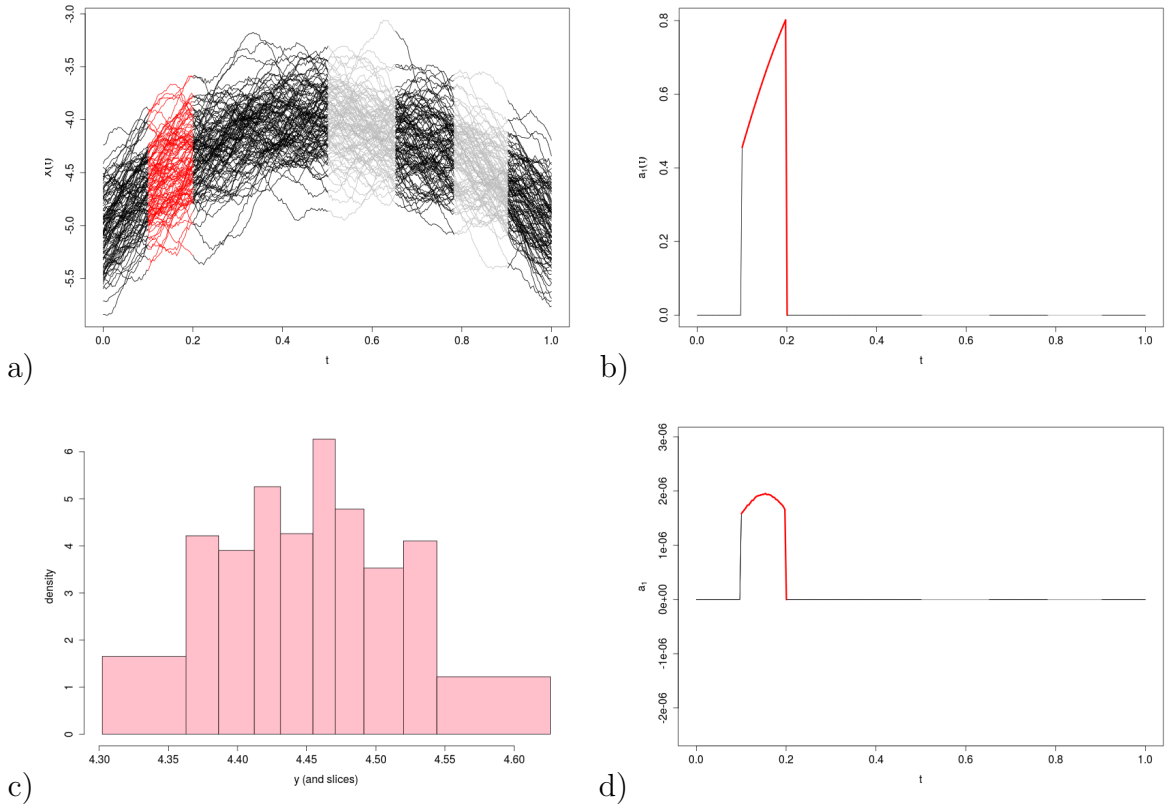

FIGURE 1 – a) $n = 100$ observations de X. Les intervalles $(\tau_k)_{k=1,\dots,D}$ sont mis en valeur par l'alternance des couleurs (noir, rouge, gris) avec $D = 7$. Le seul intervalle actif pour la prédiction de Y est l'intervalle rouge ; b) $\mathbf{a}_1$; c) Distribution de Y et tranches pour l'estimation de l'espérance conditionnelle $\mathbb{E}(X|Y)$; d) $\tilde{a}_1$. L'intervalle cible réel, utile pour la prédiction, est en rouge.

SI-SIR est mis en œuvre sur ces données avec $H = 10$ et $p = 1$ en utilisant le package R **glmnet**. L'intervalle de définition $[0,1]$ est découpé en 7 intervalles $[0, 0.1]$, $[0.1, 0.2]$,

$[0.2, 0.5]$, $[0.5, 0.65]$, $[0.65, 0.78]$, $[0.78, 0.9]$ et $[0.9, 1]$, parmi lesquels se trouve le seul intervalle actif $[0.1, 0.2]$. Enfin, les paramètres μ_1 et μ_2 sont sélectionnés par validation croisée. Les résultats obtenus pour l'estimation du coefficient $\tilde{a}_i$ sont donnés dans la figure 1, d.

L'intervalle cible, utile pour la prédiction est bien détecté par la méthode et les intervalles inactifs sont également correctement détectés comme tels. Les valeurs de $\tilde{a}_1$ ont été Σ-normés donc ne sont pas directement comparables aux valeurs de $\mathbf{a}_1$ mais la forme de la fonction de projection est raisonnable. Les perspectives de ce travail sont, actuellement, la mise au point d'une méthode itérative permettant d'identifier les intervalles pertinents sans *a priori* sur leur nombre ou leur forme ainsi que l'utilisation de l'estimation de la projection de X sur l'espace central dans un objectif de prédiction.

Références

[Bernard-Michel et al., 2008] Bernard-Michel, C., Gardes, L., and Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64(3) :982–986.

[Chen and Li, 1998] Chen, C. and Li, K. (1998). Can SIR be as popular as multiple linear regression ? *Statistica Sinica*, 8 :289–316.

[Coudret et al., 2014] Coudret, R., Liquet, B., and Saracco, J. (2014). Comparison of sliced inverse regression aproaches for undetermined cases. *Journal de la Société Française de Statistique*, 155(2) :72–96.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *NonParametric Functional Data Analysis*. Springer.

[Ferré and Yao, 2003] Ferré, L. and Yao, A. (2003). Functional sliced inverse regression analysis. *Statistics*, 37(6) :475–488.

[Li, 1991] Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414) :316–342.

[Li and Nachtsheim, 2008] Li, L. and Nachtsheim, C. (2008). Sparse sliced inverse regression. *Technometrics*, 48(4) :503–510.

[Li and Yin, 2008] Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64 :124–131.

[Ramsay and Silverman, 1997] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer Verlag, New York.

[Zhong et al., 2005] Zhong, W., Zeng, P., Ma, P., Liu, J., and Zhu, Y. (2005). RSIR : regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21 :4169–4175.