Une nouvelle approche pour les méthodes SIR

Bernard Bercu[1,2], Thi Mong Ngoc Nguyen[1,3], Jérôme Saracco[1,3,4]

[1] *Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Université de Bordeaux, 351 cours de la libération, 33405 Talence Cedex, France.*

[2] *INRIA Bordeaux Sud-Ouest, ALEA team, France.*

[3] *INRIA Bordeaux Sud-Ouest, CQFD team, France.*

[4] *Institut Polytechnique de Bordeaux, 1 avenue du Dr Albert Schweitzer, 33402 Talence.*

e-mail : {Bernard.Bercu,Thi.Mong.Ngoc.Nguyen,Jerome.Saracco}@math.u-bordeaux1.fr

Résumé. Dans cette communication, nous considérons un modèle de régression semi-paramétrique. Nous proposons un estimateur de type SIR (Sliced Inverse Regression) de la partie paramétrique de ce modèle. Cet estimateur est basé sur l'information contenue dans une seule tranche optimale. Nous appelons cette nouvelle méthode "SIRoneslice". Nous donnons aussi une version récursive de cet estimateur. Nous établissons un résultat de convergence presque sûre de cet estimateur. Enfin, nous montrons sur des simulations les bonnes performances numériques de la méthode SIRoneslice récursive ou non, ainsi que l'avantage principal de l'utilisation de la version récursive du point de vue des temps de calcul.
Mots-clés : Estimation récursive, Modèle semi-paramétrique, Sliced Inverse Regression.

Abstract. In this communication, we consider a semiparametric regression model. We propose a SIR estimator of the euclidean parameter of this model. This estimator is based on the use of only one "optimal" slice. We call this new method "SIRoneslice". We give a recursive version of this estimator and we provide the almost sure convergence of our estimator. A simulation study shows good numerical performances of SIRoneslice and recursive SIRoneslice methods and clearly exhibits the main avantage of using recursive version from a computational times point of view.
Key words: Recursive estimation, Semiparametric regression model, Sliced Inverse Regression (SIR).

1 Introduction

Les modèles de régression sont très utiles pour étudier la liaison entre une variable à expliquer y et une variable explicative x. Dans cette communication, nous nous intéressons au modèle de régression semi-paramétrique proposé par Duan et Li (1991) lorsque la variable à expliquer y est à valeurs dans $\mathbb{R}$ et la covariable x appartient à $\mathbb{R}^p$:

$$y = f(\theta' x) + \varepsilon, \tag{1}$$

où le paramètre θ est un vecteur inconnu de $\mathbb{R}^p$, le bruit ε est un terme d'erreur aléatoire indépendant de x (aucune hypothèse n'est faite sur la distribution de ε) et la fonction de lien f est un paramètre fonctionnel à valeurs dans $\mathbb{R}$, inconnu et arbitraire. Notons que dans le cadre de ce modèle, le paramètre θ n'est pas totalement identifiable, seule la direction de θ est identifiable. On parlera alors de direction EDR pour "effective dimension reduction".

Nous rappelons tout d'abord brièvement la méthode SIR qui estime la direction EDR. La méthode SIR est une méthode de régression semi-paramétrique reposant sur un argument géométrique. Elle a été introduite par Li (1991) et Duan et Li (1991). Cette méthode repose sur une propriété de la fonction de régression inverse. Le coût à payer est de rajouter une hypothèse probabiliste sur la distribution de la variable explicative x :

(H) *la variable explicative x possède une distribution de probabilité non dégénérée telle que, pour tout $b \in \mathbb{R}^p$, l'espérance conditionnelle $\mathbb{E}[b'x \mid \theta'x]$ soit linéaire en $\theta'x$.*

Cette hypothèse, encore appelée condition de linéarité (LC), est vérifiée lorsque la variable explicative x suit une distribution elliptique, en particulier lorsque la distribution de x est multinormale. Posons $\mu = \mathbb{E}[x]$ et $\Sigma = \mathbb{V}(x)$. Nous donnons maintenant le théorème de caractérisation de la direction de θ établi par Li (1991).

Théorème 1.1 *Dans le cadre du modèle (1) et sous l'hypothèse* **(H)***, la courbe de régression inverse centrée, $y \mapsto \mathbb{E}[x \mid \mathbb{T}(y)] - \mu$, appartient au sous-espace linéaire de $\mathbb{R}^p$ engendré par θ, où $\mathbb{T}$ est une transformation monotone de y.*

Une conséquence directe de ce théorème est que la matrice de covariance de cette courbe, $\Gamma = \mathbb{V}(\mathbb{E}[x \mid \mathbb{T}(y)])$, est dégénérée dans toute direction Σ-orthogonale à θ. Ainsi le vecteur propre b associé à la valeur propre non nulle de $\Sigma^{-1}\Gamma$ est colinéaire à θ et est donc une direction EDR.

Afin d'estimer facilement la matrice Γ, Duan et Li (1991) ont proposé un choix particulier pour $\mathbb{T}$: un "tranchage" qui est une discrétisation de y fondée sur un découpage du support de y en H tranches distinctes $s_1, \ldots, s_H$. Dans ce cadre, la matrice Γ s'écrit sous la forme : $\Gamma := \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'$ où $p_h = P(y \in s_h)$ et $m_h = \mathbb{E}[x \mid y \in s_h]$. Ainsi à partir d'un échantillon d'observation (x_i, y_i) et en substituant les moments empiriques aux moments théoriques, nous obtenons l'estimateur de la direction EDR par le vecteur propre associé à la plus grande valeur propre de l'estimateur $\widehat{\Sigma}_n^{-1}\hat{\Gamma}_n$ de $\Sigma^{-1}\Gamma$, où $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)'$ et $\widehat{\Gamma}_n = \sum_{h=1}^{H} \hat{p}_{h,n}(\hat{m}_{h,n} - \bar{x}_n)(\hat{m}_{h,n} - \bar{x}_n)'$, avec $\bar{x}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$, $\hat{p}_{h,n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{I}[y_i \in s_h] = \dfrac{\hat{n}_{h,n}}{n}$ et $\hat{m}_{h,n} = \dfrac{1}{\hat{n}_{h,n}}\displaystyle\sum_{i}^{n} x_i \mathbb{I}[y_i \in s_h]$.

2 Approche SIR basée sur une tranche : SIRoneslice

Dans le cadre du modèle (1) et sous l'hypothèse (H), pour chaque tranche h, nous avons

$$\mathbb{E}[x \mid y \in s_h] = \mu + k_h \Sigma\theta, \quad \text{où} \quad k_h = \frac{\mathbb{E}[(x-\mu)'\theta \mid y \in s_h]}{\theta'\Sigma\theta}.$$

Posons $z_h := \mathbb{E}[x \mid y \in s_h] - \mu = m_h - \mu$ pour $h = 1, \ldots, H$. Nous en déduisons que

$$z_h = k_h \Sigma\theta \quad \text{et} \quad k_h = \frac{z_h'\theta}{\theta'\Sigma\theta}. \tag{2}$$

Ainsi z_h appartient au sous-espace de dimension 1 engendré par $\Sigma\theta$ si $k_h \neq 0$. À partir de (2), pour toute tranche h telle que le scalaire k_h est non nul, nous pouvons estimer la direction EDR d'un échantillon d'observations $\{(x_i, y_i), i = 1, \ldots, n\}$ en utilisant l'estimateur suivant :

$$\hat{b}_{h,n} := \widehat{\Sigma}_n^{-1} \hat{z}_{h,n}, \quad \text{avec} \quad \hat{z}_{h,n} = \hat{m}_{h,n} - \bar{x}_n.$$

Notons que, contrairement à l'approche SIR classique, nous n'utilisons pas l'information des H tranches dans cette approche SIRoneslice. Nous ne nous concentrons que sur une tranche h pour laquelle la valeur de k_h est non nulle. Dans certaines situations, par exemple, lorsque le modèle de régression est partiellement symétrique, l'estimateur $\hat{b}_{h,n}$ obtenu à partir d'une certaine tranche h est plus performant que l'estimateur SIR classique. Définissons maintenant comment trouver une tranche h^o optimale en un certain sens.

Dans la suite, nous supposons que θ est tel que $|| \theta ||_\Sigma = 1$. Ainsi, nous avons $k_h = z_h'\theta$. Ce terme dépend du paramètre inconnu θ. Toutefois, en utilisant $|| \theta ||_\Sigma = 1$, nous déduisons de (2) que

$$(k_h)^2 = || z_h ||^2_{\Sigma^{-1}} .$$

À partir de ce résultat, nous pouvons proposer une tranche optimale pour l'approche SIRoneslice définie par :

$$h^o = \arg\max_h \ (k_h)^2.$$

La version sur population de l'estimateur Σ-normalisé correspondant est la suivante :

$$\theta_{h^o} := \frac{\Sigma^{-1} z_{h^o}}{|| z_{h^o} ||_{\Sigma^{-1}}}.$$

Nous obtenons ensuite la version sur échantillon de cet estimateur :

$$\hat{\theta}_{\hat{h}_n^o,n} := \frac{\widehat{\Sigma}_n^{-1} \hat{z}_{\hat{h}_n^o,n}}{|| \hat{z}_{\hat{h}_n^o,n} ||_{\widehat{\Sigma}_n^{-1}}}, \quad \text{où} \quad \hat{h}_n^o = \arg\max_h \ (\hat{k}_{h,n})^2 \quad \text{avec} \quad (\hat{k}_{h,n})^2 = || \hat{z}_{h,n} ||^2_{\widehat{\Sigma}_n^{-1}} .$$

Par la Loi des Grands Nombres, nous pouvons démontrer aisément que $\hat{\theta}_{\hat{h}_n^o,n}$ converge presque sûrement vers $\pm\theta$. Nous pouvons aussi considérer un estimateur alternatif de la direction θ (qui n'est pas $\widehat{\Sigma}_n$-normalisé) :

$$\hat{b}_{\hat{h}_n^o,n} := \widehat{\Sigma}_n^{-1}\hat{z}_{\hat{h}_n^o,n},$$

Dans le théorème 4.1, nous démontrons que cet estimateur converge presque sûrement vers $b_{h^o} = \Sigma^{-1}z_{h^o}$ qui est colinéaire à θ et nous précisons la vitesse de convergence.

3 Version récursive de SIRoneslice

Nous disposons d'un échantillon d'observations $\{(x_i, y_i), i = 1, \ldots, n\}$ de variables aléatoires indépendantes et identiquement distribuées issues du modèle (1). Pour obtenir la forme récursive de l'estimateur, on scinde cet échantillon en deux parties : le sous-échantillon $\{(x_i, y_i), i = 1, \ldots, n-1\}$ et la "nouvelle" observation (x_n, y_n). Nous pouvons alors donner une version récursive de l'estimateur $\hat{b}_{\hat{h}_n^o,n}$:

$$\begin{aligned}\hat{b}_{\hat{h}_n^o,n} &= \tfrac{n}{n-1}\widehat{\Sigma}_{n-1}^{-1}\hat{z}_{\hat{h}_n^o,n-1} - \tfrac{1}{n-1}\widehat{\Sigma}_{n-1}^{-1}\Phi_n - \tfrac{1}{(n-1)(n+\rho_n)}\widehat{\Sigma}_{n-1}^{-1}\Phi_n\Phi_n'\widehat{\Sigma}_{n-1}^{-1}(n\hat{z}_{\hat{h}_n^o,n-1} - \Phi_n)\\ &+ \tfrac{1}{n-1}\Phi_{\hat{h}_n^o,n}'\left(\widehat{\Sigma}_{n-1}^{-1} - \tfrac{1}{(n+\rho_n)}\widehat{\Sigma}_{n-1}^{-1}\Phi_n\Phi_n'\widehat{\Sigma}_{n-1}^{-1}\right)\mathbb{I}_{[h^*=\hat{h}_n^o]},\end{aligned}$$

où $\Phi_n = x_n - \bar{x}_{n-1}$, $\Phi_{h,n} = x_n - \widehat{m}_{h,n-1}$, $\rho_n = \Phi_n'\widehat{\Sigma}_{n-1}^{-1}\Phi_n$ et où h^* désigne la tranche contenant la nouvelle observation (x_n, y_n).

4 Résultat asymptotique

Afin d'établir la convergence presque sûre de $\hat{b}_{\hat{h}_n^o,n}$, nous supposons les hypothèses :

(A_1) Les observations $(x_i, y_i), i = 1, ..., n$, sont indépendantes et identiquement distribuées.

(A_2) Le support de y est partionné en H tranches fixes s_h telles que $p_h \neq 0$.

(A_3) $\exists! h_o$ tel que $k_{h^o} > k_h$.

Théorème 4.1 *Sous les hypothèses (H), (A_1), (A_2) et (A_3), nous avons pour n assez grand*

$$|| \hat{b}_{\hat{h}_n^o,n} - b_{h^o} ||= \mathcal{O}\left(\sqrt{\frac{\log(\log n)}{n}}\right) \qquad p.s.$$

La démontration de ce théorème est donnée dans Bercu et al. (2011).

5 Quelques résultats de simulation

Dans Bercu et al. (2011), nous étudions, sur des simulations, le comportement numérique de l'estimateur de la direction EDR avec la méthode SIRoneslice. Dans les simulations présentées ici, nous considérons le modèle de régression suivant :

$$y = (x'\theta)^2 \exp(x'\theta/A) + \varepsilon \tag{3}$$

où x suit la loi multinormale $\mathcal{N}_p(0, \Sigma)$, $\theta = (1, -1, 2, -2, 0, \ldots, 0)'$ et ε suit la loi normale $\mathcal{N}(0, 1.5)$. Le paramètre A a une influence sur la dépendance entre l'indice $\theta'x$ et y.

Tout d'abord, nous comparons les temps de calcul des méthodes SIR et SIRoneslice (récursive ou non). Nous observons dans la Table 1 que les versions récursives sont les plus rapides.

		$p = 5$	$p = 10$	$p = 15$	$p = 20$
$H = 5$	SIR	4.888 (0.053)	5.824 (0.152)	7.365 (0.060)	9.063 (0.014)
	recursive SIR	1.421 (0.013)	1.554 (0.017)	1.750 (0.018)	2.077 (0.020)
	SIRoneslice	3.534 (0.010)	4.686 (0.007)	5.976 (0.010)	7.318 (0.009)
	recursive SIRoneslice	0.481 (0.001)	0.496 (0.005)	0.504 (0.002)	0.540 (0.005)
$H = 20$	SIR	9.046(0.014)	12.251 (0.028)	15.646 (0.032)	19.231 (0.061)
	recursive SIR	1.922 (0.014)	2.082 (0.021)	2.325 (0.021)	2.728 (0.011)
	SIRoneslice	9.628 (0.013)	12.660 (0.042)	15.760 (0.013)	19.006 (0.083)
	recursive SIRoneslice	1.394 (0.005)	1.432 (0.006)	1.459 (0.009)	1.544 (0.007)

Table 1: Temps de calculs en secondes des estimateurs $\hat{b}_n$ de la direction θ (pour n allant de $N_0 = 30$ à 900) par les méthodes SIR, SIR récursive, SIRoneslice et SIRoneslice récursive avec différentes valeurs de H : moyennes et écart-type entre parenthèses calculés sur $\mathcal{B} = 100$ échantillons du modèle (3) avec $A = 2.5$ et différentes valeurs de p

Rappelons ici que les méthodes SIR et SIR récursive (resp. SIRoneslice et SIRoneslice récursive) appliquées sur les mêmes données fournissent les mêmes estimations, seule la manière de calculer diffère. Pour cette raison, nous comparons uniquement à la Figure 1 la qualité des estimateurs de la direction EDR obtenus avec SIR et SIRoneslice. On observe que la méthode SIRoneslice semble être meilleure que SIR lorsque le modèle présente une dépendance symétrique modérée ($A = 2.5$) ou forte ($A = 5$).

Dans la Figure 2, nous illustrons le comportement de SIRoneslice récursif en fonction de N. Nous voyons clairement que plus le nombre N d'observations est important, plus la qualité d'estimation est bonne. Les résultats obtenus montrent que l'estimateur SIRoneslice récursif se comporte particulièrement bien pour des tailles d'échantillon raisonnables ($N \geq 300$).

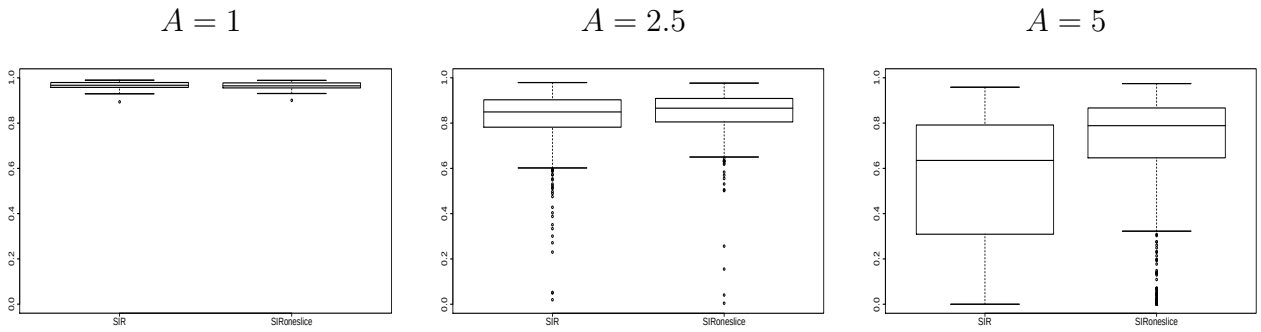

Figure 1: Boxplots des $\cos^2(\hat{b}_n, \theta)$ obtenus avec SIR et SIRoneslice et calculés sur 500 échantillons issus du modèle (3) avec $n = 300$, $p = 10$ et différentes valeurs de A

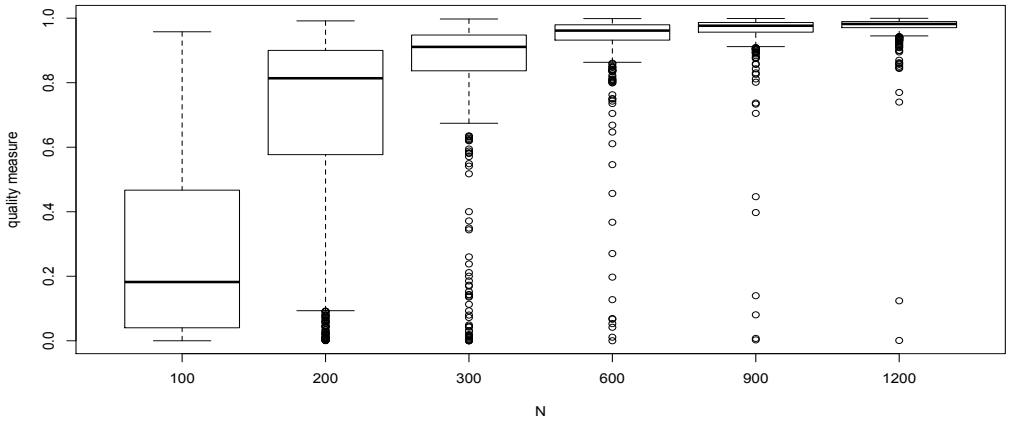

Figure 2: Boxplots des $\cos^2(\hat{b}_n, \theta)$ obtenus avec SIRoneslice récursif et calculés sur 500 échantillons du modèle (3) avec $p = 10$ et $A = 2.5$, pour différentes valeurs de N

Bibliographie

[1] Duan, N. and Li, K. C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, **19**, 505-530.

[2] Li, K. C. (1991). Sliced inverse regression for dimension reduction, with dicussion. *Journal of the American Statistical Association*, **86**, 316-342.

[3] Bercu, B., Nguyen, T.M.N. et Saracco, J. (2011). A new approach on recursive and non recursive SIR. *Soumis pour publication.*