Estimation récursive en régression inverse par tranche (sliced inverse regression)

Thi Mong Ngoc Nguyen, Jérôme Saracco

Estimation récursive en régression inverse par tranche (sliced inverse regression)

Thi Mong Ngoc Nguyen[1,2] & Jérôme Saracco[1,2,3]

[1] *Université Bordeaux 1, Institut de Mathématiques de Bordeaux, UMR 5251, 351 cours de la libération, 33405 Talence.*
[2] *Equipe CQFD, INRIA Bordeaux Sud-Ouest, France.*
[3] *Université Montesquieu-Bordeaux 4, GREThA, UMR CNRS 5113, Avenue Léon Duguit,33608 Pessac Cedex,France.*

Résumé. Dans cette communication, nous nous intéressons à la méthode SIR (Sliced Inverse Regression, que l'on peut traduire par régression inverse par tranches) qui permet d'estimer le paramètre θ dans un modèle semi-paramétrique de régression du type $y = f(x'\theta, \varepsilon)$ sans avoir à estimer le paramètre fonctionnel f ni à spécifier la loi de l'erreur ε. Nous proposons un estimateur récursif de la direction de θ dans le cas particulier où l'on considère $H = 2$ tranches. Nous donnons des propriétés asymptotiques de cet estimateur (convergence et normalité asymptotique). Nous illustrons aussi sur des simulations le bon comportement numérique de la méthode proposée.
Mots-clés : estimation récursive, modèle semi-paramétrique, Sliced Inverse Regression.

Abstract. In this communication, we propose a recursive estimation procedure for sliced inverse regression (SIR). When the number H of slices is equal to two, we obtain a recursive estimator of the direction of the parameter θ in the semiparametric regression model $y = f(x'\theta, \varepsilon)$, which does not require the estimation of the link function f. We establish asymptotic properties of our recursive estimator (almost sure convergence and asymptotic normality). A simulation study illustrates the good numerical behavior of the estimator.
Key words: recursive estimation, semiparametric regression model, sliced inverse regression.

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 semi-paramétrique de régression 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 valeur 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 de direction EDR (comme "effective dimension reduction").

Les techniques d'estimation développées jusqu'à présent pour la méthode SIR ne sont pas récursives. D'une manière générale, l'avantage des méthodes récursives est de prendre en compte l'arrivée successive des données et d'affiner ainsi au fil du temps les algorithmes d'estimation mis en œuvre. Un intérêt majeur de ces méthodes est qu'il n'est pas nécessaire de relancer tous les calculs d'estimation des paramètres du modèle à chaque fois que la base des données est complétée par de nouvelles observations. L'idée est ici d'utiliser les estimations calculées sur la base des données initiales et de les remettre à jour en tenant uniquement compte des nouvelles observations arrivant dans la base. Le gain en terme de temps de calcul peut être très intéressant et les applications d'une telle approche sont nombreuses.

2 Estimation récursive de la direction θ

Nous rappelons tout d'abord brièvement la méthode SIR. Ensuite, nous proposons un estimateur récursif θ_N de la direction de θ.

2.1 Méthode SIR

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. L'avantage de cet inversion de rôle est que la dimension du problème a été réduite : nous avons en effet maintenant p problèmes de dimension 1, la régression inverse permettant de régresser chaque coordonnée de x sur y. 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, est vérifiée lorsque la variable explicative x suit une distribution elliptique, en particulier lorsque la distribution de x est multinormale. Posons $\mathbb{E}[x] = \mu$ et $\Sigma = \mathbb{V}(x)$. Nous donnons maintenant le théorème de caractérisation de la direction de θ établi par Li (1991).

Théorème 2.1 *Dans ce 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 (qui correspondra au "tranchage" précisé ultérieurement).*

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 $\tilde{\theta}$ 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 la suite, nous nous focaliserons sur le cas où l'on ne considère que deux tranches s_1 et s_2. La raison essentielle de ce choix est que nous pouvons obtenir facilement la forme analytique pour l'estimateur de la direction de θ. En effet, lorsque $H = 2$, la matrice Γ s'écrit :

$$\Gamma = p_1 z_1 z_1' + p_2 z_2 z_2',$$

où $p_h = \mathbb{P}(y \in s_h)$ et $z_h = m_h - \mu$ avec $m_h = \mathbb{E}[x \mid y \in s_h]$ pour $h = 1, 2$. On peut alors montrer que la valeur propre non nulle λ de $\Sigma^{-1}\Gamma$ et le vecteur propre $\tilde{\theta}$ correspondant s'écrivent sous la forme :

$$\lambda = \frac{p_1}{p_2} z_1' \Sigma^{-1} z_1 \quad \text{et} \quad \tilde{\theta} = \Sigma^{-1}(z_1 - z_2). \tag{2}$$

Lorsque nous diposons d'un échantillon d'observations $\{(x_i, y_i), i = 1, ..., N\}$ de variables aléatoires indépendantes et identiquement distribuées (x, y) issues du modèle (1), nous pouvons déduire de (2) des estimateurs λ_N et $\tilde{\theta}_N$ de λ et $\tilde{\theta}$:

$$\lambda_N = \frac{p_{1,N}}{p_{2,N}} z_{1,N}' \Sigma_N^{-1} z_{1,N} \quad \text{et} \quad \tilde{\theta}_N = \Sigma_N^{-1}(z_{1,N} - z_{2,N}), \tag{3}$$

où $p_{h,N} = \frac{1}{N}\sum_{i=1}^N \mathbb{I}[y_i \in s_h] = \frac{N_h}{N}$, $z_{h,N} = m_{h,N} - \bar{x}_N$ et $\Sigma_N = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x}_N)(x_i - \bar{x}_N)'$ avec $m_{h,N} = \frac{1}{N_h}\sum_{i \in s_h} x_i$ et $\bar{x}_N = \frac{1}{N}\sum_{i=1}^N x_i$.

2.2 Estimateur récursif de la direction de θ lorsque $H = 2$

Nous avons donné en (3) l'expression analytique des formes non récursives des estimateurs λ_N et $\tilde{\theta}_N$. Nous allons maintenant en déduire leurs formes récursives. Pour la forme récursive des estimateurs, on scinde l'échantillon $\{(x_i, y_i), i = 1, ..., N\}$ en deux parties: le sous-échantillon des $N-1$ premières observations $\{(x_i, y_i), i = 1, ..., N-1\}$ et la nouvelle observation (x_N, y_N). Nous supposons que la nouvelle observation (x_N, y_N) est telle que $y_N \in s_{h^*}$ avec $h^* = 1$ ou 2.

Forme récursive de $\tilde{\theta}_N$. Notons tout d'abord que nous avons les formes récursives suivantes pour les estimateurs $\bar{x}_N$, Σ_N, Σ_N^{-1}, $p_{h,N}$ et $z_{h,N}$. En posant $\Phi_N = x_N - \bar{x}_{N-1}$, $\rho_N = \Phi_N' \Sigma_{N-1}^{-1} \Phi_N$ et $\Phi_{h^*,N} = x_N - m_{h^*,N-1}$, on peut écrire :

$$\bar{x}_N = \bar{x}_{N-1} + \frac{1}{N}\Phi_N, \quad \Sigma_N = \frac{N-1}{N}\Sigma_{N-1} + \frac{N-1}{N^2}\Phi_N \Phi_N',$$

$$\Sigma_N^{-1} = \frac{N}{N-1}\Sigma_{N-1}^{-1} - \frac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi'_N\Sigma_{N-1}^{-1},$$

$$p_{h,N} = \begin{cases} \frac{N-1}{N}p_{h^*,N-1} + \frac{1}{N} & \text{si } h = h^*, \\ \frac{N-1}{N}p_{h,N-1} & \text{sinon.} \end{cases}$$

$$m_{h,N} = \begin{cases} m_{h^*,N-1} + \frac{1}{N_{h^*,N-1}+1}\Phi_{h^*,N} & \text{si } h = h^*, \\ m_{h,N-1} & \text{sinon.} \end{cases}$$

$$z_{h,N} = \begin{cases} z_{h^*,N-1} - \frac{1}{N}\Phi_N + \frac{1}{N_{h^*,N-1}+1}\Phi_{h^*,N} & \text{si } h = h^*, \\ z_{h,N-1} - \frac{1}{N}\Phi_N & \text{sinon.} \end{cases}$$

Nous pouvons alors en déduire la forme récursive de l'estimateur $\tilde{\theta}_N$ de $\tilde{\theta}$:

$$\begin{aligned} \tilde{\theta}_N &= \tfrac{N}{N-1}\tilde{\theta}_{N-1} - \tfrac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi'_N\tilde{\theta}_{N-1} \\ &- \tfrac{(-1)^{h^*}N}{(N_{h^*,N-1}+1)(N-1)}\left(\Sigma_{N-1}^{-1} - \tfrac{1}{N+\rho_N}\Sigma_{N-1}^{-1}\Phi_N\Phi'_N\Sigma_{N-1}^{-1}\right)\Phi_{h^*,N}. \end{aligned} \tag{4}$$

Forme récursive de λ_N. Il est aussi possible d'obtenir la forme récursive de l'estimateur λ_N de λ, mais son expression ne peut pas se simplifier énormément :

$$\lambda_N = \frac{p_{1,N-1} + \frac{1}{N-1}\mathbb{I}[h^*=1]}{p_{2,N-1} + \frac{1}{N-1}\mathbb{I}[h^*=2]}v'_N\left(\frac{N}{N-1}\Sigma_{N-1}^{-1} - \frac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi'_N\Sigma_{N-1}^{-1}\right)v_N,$$

avec $v_N = z_{1,N-1} - \frac{1}{N}\Phi_N + \frac{1}{N_{1,N-1}+1}\Phi_{1,N}\mathbb{I}[h^*=1]$. Il est possible d'écrire cet estimateur sous la forme $\lambda_N = \frac{N}{N-1}\lambda_{N-1} + F(x_N, y_N, N, \dots)$. Nous ne donnons pas volontairement l'expression de $F(.)$ qui, malgré quelques simplifications, reste relativement lourde à écrire.

2.3 Résultats asymptotiques pour $\tilde{\theta}_N$

Avant de donner des résultats asymptotiques pour $\tilde{\theta}_N$, nous supposons les deux hypothèses suivantes :

(A1) Les observations $(x_i, y_i), i = 1, ..., N$, sont échantillonnées de manière indépendante à partir du modèle (1).

(A2) Le support de y est partionné en deux tranches fixes s_1 et s_2 telles que $p_h \neq 0$.

Nous avons les résultats de convergence suivants.

Théorème 2.2 *Sous les hypothèses* (A_1) *et* (A_2)*, nous avons, pour* $N \to \infty$ *:*

- $|| \tilde{\theta}_N - \tilde{\theta} || = \mathcal{O}\left(\sqrt{\frac{\log(\log N)}{N}}\right)$ *p.s.*
- $\sqrt{N}(\tilde{\theta}_N - \tilde{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{-1}\Delta_3\Sigma^{-1})$ *où l'expression de* Δ_3 *est donnée dans Nguyen et Saracco (2010).*

3 Quelques résultats de simulation

Nous avons étudié, sur des simulations, le comportement numérique de l'estimateur récursif $\tilde{\theta}_N$ que nous avons proposé. En particulier, nous nous focaliserons sur la convergence de $\tilde{\theta}_N$ vers la vraie direction de θ du modèle. Théoriquement nous avons montré que $\tilde{\theta}_N \xrightarrow{p.s.} \tilde{\theta}$ où $\tilde{\theta}$ est colinéaire à θ. Ainsi la qualité de l'estimation sera mesurée par le cosinus carré de l'angle entre $\tilde{\theta}_N$ et θ défini par $\cos^2(\tilde{\theta}_N, \theta) = \frac{(<\tilde{\theta}_N, \theta>)^2}{||\tilde{\theta}_N|| \times ||\theta||}$. Plus $\cos^2(\tilde{\theta}_N, \theta)$ est proche de 1, meilleure est la qualité de l'estimation.
Dans les simulations présentées ici, nous avons considéré le modèle de régression suivant :

$$y = (\theta' x) \exp(-\theta' x/2) + (\theta' x)\varepsilon, \tag{5}$$

où x suit la loi multinormale $\mathcal{N}_p(0, I_p)$, $\theta = (1, -1, 0, \ldots, 0)' \in \mathbb{R}^p$ et ε suit la loi normale $\mathcal{N}(0, 1)$. Nous présentons à la Figure 1 un nuage de points $\{(x_i'\theta, y_i), i = 1, \ldots, N\}$ simulés

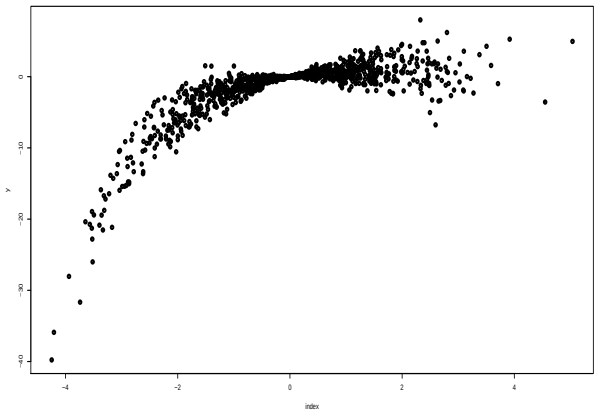

Figure 1: Nuage des points $\{(x_i'\theta, y_i), i = 1, \ldots, 1000\}$ simulés à partir de (5) pour $p = 20$

à partir du modèle (5) pour $p = 20$, $N = 1000$. Ce graphique permet de visualiser le niveau de bruit dans le modèle.

Nous avons simulé, à partir de ce modèle, $\mathcal{B} = 500$ échantillons de taille 1000 avec successivement $p = 10$ et $p = 20$. Pour chaque échantillon simulé, nous avons calculé pour $N = N_0$ jusqu'à $N = 1000$, l'estimateur récursif $\tilde{\theta}_N$ ainsi que la qualité correspondante, à savoir $\cos^2(\tilde{\theta}_N, \theta)$, la valeur N_0 étant égale à $p + 1$.

A la Figure 2, nous présentons les boxplots des $\cos^2(\tilde{\theta}_N, \theta)$ en fonction de certaines valeurs de N pour la dimension $p = 20$. Nous voyons clairement que plus le nombre N d'observations est important, plus la qualité d'estimation est bonne. Dès que la taille de l'échantillon dépasse 200, la médiane des $\cos^2(\tilde{\theta}_N, \theta)$ est supérieure à 0.8.

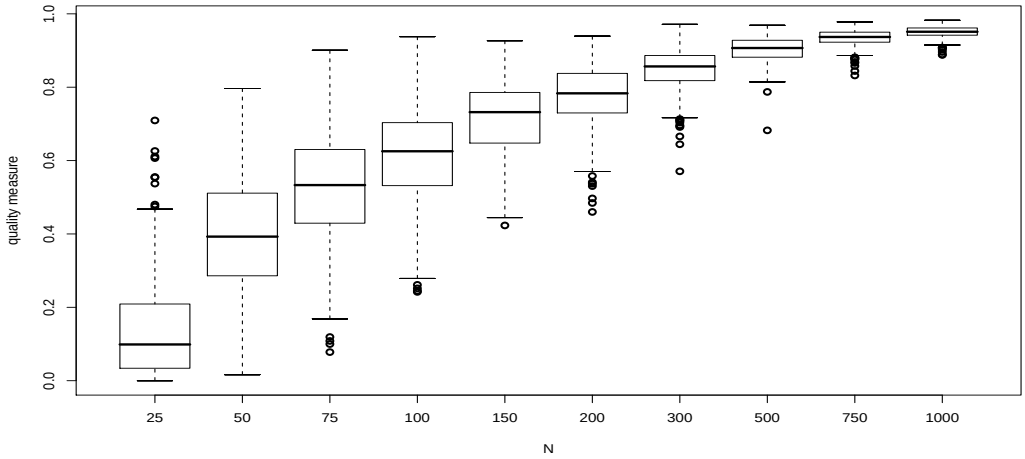

Figure 2: Boxplots des $\cos^2(\tilde{\theta}_N, \theta)$ en fonction de N pour $p = 20$

Une étude de simulation plus complète est disponible dans Nguyen et Saracco (2010). Dans cet article, on considère différents modèles et différentes lois pour x (en particulier avec une matrice Σ non diagonale) et pour ε. Les résultats obtenus montrent que l'estimateur proposé se comporte bien pour des tailles d'échantillon raisonnables ($N \geq 200$) et même lorsque la dimension p de x est importante ($p \leq 50$). La normalité de l'estimateur a aussi été illustrée dans ce papier. Enfin, notons que pour calculer les estimations $\tilde{\theta}_{N_0}, \tilde{\theta}_{N_0+1}, \ldots, \tilde{\theta}_N$, l'approche SIR récursive proposée est 10 à 20 fois plus rapide (selon les valeurs de p) que la méthode de SIR usuelle.

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] Nguyen, T.M.N. et Saracco, J. (2010). Estimation récursive en régression inverse par tranche (sliced inverse regression). *sousmis pour publication.*