Evaluation de l'importance des variables dans la méthode SIR (sliced inverse regression)

Ines Jlassi [1,2] & Jérôme Saracco [3,4,5]

[1] *Laboratoire de Physique-Mathématique, Fonctions Spéciales et Applications (LR11ES35), Université de Sousse, Tunisie*

[2] *Faculté des Sciences de Monastir, Université de Monastir, Tunisie*
jlassi.ines.fsm@gmail.com

[3] *Ecole Nationale Supérieure de Cognitique (ENSC - Bordeaux INP), 109 avenue Roul, 33405 Talence, France*

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

[5] *Inria Bordeaux Sud-Ouest, CQFD team, 200 avenue de la Vieille Tour, 33 405 Talence Cedex, France.*
jerome.saracco@math.u-bordeaux.fr

Résumé. Dans ce travail, nous nous intéressons à un modéle de régression semi-paramétrique entre une variable à expliquer $y \in \Re$ et une covariable multidimensionnelle $x \in \Re^p$. L'approche SIR (sliced inverse regression) permet d'estimer la partie paramétrique de ce modéle et d'obtenir une estimation d'une base de l'espace EDR (effective dimension reduction). Nous proposons dans cette communication une maniére de quantifier l'importance des variables explicatives dans ce modéle (en terme d'impact sur la variable à expliquer y) ne reposant que sur l'estimation de l'espace EDR. Cette approche computationnelle (implémentée en R) permet alors de sélectionner les variables explicatives les plus utiles/importantes du modéle. Nous illustrons le bon comportement numérique de la méthode sur des simulations et sur un jeu de données réelles.

Mots-clés. Réduction de dimension, régression semiparamétrique, régression inverse par tranches (sliced inverse regression), importance des variables, sélection de variables.

Abstract. We are interested in treating the relationship between a dependent variable y and a multivariate covariate $x \in \Re^p$ in a semiparametric regression model. Since the purpose of most social, biological or environmental science research is the explanation, the determination of the importance of the variables is a major concern. It is a way to determine which variables are the most important when predicting y. Sliced inverse regression methods allows to reduce the space of the covariate x by estimating the directions β that form an effective dimension reduction (EDR) space. The aim of this paper is to propose a computational method based on importance variable measure (only relying on the EDR space) in order to select the most useful variables. The numerical behavior

of this new method, implemented in R, is studied on a simulation study. An illustration on a real data is also provided.

Keywords. Dimension reduction, semiparametric regression model, sliced inverse regression, variable importance, variable selection.

1 Introduction

Pour pallier aux problèmes sous-jacents des approches purement paramétriques (bon choix de la famille paramétrique de la fonction de lien) ou nonparamétriques (fléau de la dimension), des modèles semiparamétriques (à indices $x'\beta_k$, avec $k = 1, \ldots, K$ où $1 \leq K < p$) ont été introduits, où les vecteurs β_k sont supposés être linéairement indépendants. Posons $\beta = [\beta_1, \ldots, \beta_K]$. On peut par exemple considérer un modèle de la forme suivante :

$$y = f(\beta' x) + \varepsilon,$$

où f est une fonction inconnue à valeur réelle, la distribution du terme d'erreur aléatoire ε est arbitraire et inconnue, et $\varepsilon \perp x$. Une autre manière d'exprimer le modèle est :

$$y \perp x \mid x'\beta$$

où $v_1 \perp v_2 \mid v_3$ désigne que la variable aléatoire v_1 est indépendante de la variable aléatoire v_2 étant donné toutes les valeurs de la variable aléatoire v_3. Ainsi, la variable x peut être remplacée par une ou plusieurs combinaisons linéaires de ses composantes, $x'\beta$, sans perdre d'information sur la distribution conditionnelle de y sachant x.

Puisque f est inconnue, le paramètre β n'est pas totalement identifiable, par contre le sous-espace engendré par β est identifiable. Ce sous-espace est appelé espace effectif de réduction de dimension (EDR pour Effective Dimension Reduction en anglais) suivant Li [9] dans la présentation originale de la méthode de régression inverse par tranches (SIR pour Sliced inverse regression). L'espace EDR E est donc le sous-espace vectoriel de $\Re^p$, de dimension K, engendré par les β_k : $E = \mathrm{Span}(\beta_1, \ldots, \beta_K)$.

Dans la littérature statistique, il existe différentes méthodes visant à estimer l'espace EDR. Les approches SIR et SAVE (sliced average variance estimation) sont les plus populaires: voir, par exemple, [1]-[7], [9]-[13].

Dans la suite, nous présentons d'abord l'évaluation de l'importance des variables lorsque la taille de l'échantillon n est supérieur à p. Nous considérons ensuite le cas où $n < p$, couramment appelé "small n, large p" en anglais. Enfin, sur la base de ces mesures d'importance des variables, nous fournissons un moyen de sélectionner les composantes de la covariable x qui ont un effet le plus important sur la variable d'intérêt y.

2 Evaluation de l'importance des variables

Le cas "$n > p$" est d'abord examiné et la mesure d'importance des variables (notée IV ci-après) sera basée sur des estimations de l'espace EDR. Ensuite, nous traitons le cas "$n < p$" pour lequel la mesure IV sera basée sur les estimations des indices $x'\beta$.

Cas de "$n > p$". Définissons la mesure de proximité suivante entre deux sous-espaces de $\Re^p$ de dimension K, $E_1 = \text{Span}(B_1)$ et $E_2 = \text{Span}(B_2)$ avec B_1 et B_2 deux matrices de dimension (p, K) de plein rang colonne :

$$m(E_1, E_2) = 1 - Q_M(E_1, E_2),$$

où $Q_M(E_1, E_2) = \frac{1}{K}\text{Trace}(P_{E_1}P_{E_2})$ avec $P_E = B(B'MB)^{-1}B'M$ (pour $E = \{E_1, E_2\}$ et $B = \{B_1, B_2\}$), avec M désignant une métrique de $\Re^p$. Notons que: $m(E_1, E_2) \in [0, 1]$, $m(E_1, E_2) = 0$ si $E_1 = E_2$, et $m(E_1, E_2) = 1$ si $E_1 \perp_M E_2$. Remarquons également que $m(E, \widehat{E})$ converge vers 0 à une vitesse de convergence $\frac{1}{\sqrt{n}}$ pour toutes les méthodes SIR classiques quand E est le véritable espace EDR et $\widehat{E}$ une estimation de E. Lorsque $K = 1$, cette mesure de proximité peut être vue comme

$$m(E_1, E_2) = 1 - \cos^2(B_1, B_2),$$

où'cos^2" désigne le cosinus carré de l'angle entre les deux vecteurs B_1 et B_2 de $\Re^p$.

Nous allons utiliser cette mesure de proximité pour définir la mesure IV. Considérons un échantillon $\{(x_i, y_i), i = 1, \ldots, n\}$. Les directions EDR estimées $\widehat{B}$ peuvent être obtenues avec l'une des méthodes SIR.

L'idée est de considérer un échantillon perturbé de x_i, avec une permutation aléatoire des valeurs de la $j^{\text{ème}}$ composante x^j de x, notée $\widetilde{x}^j$ dans la suite. L'échantillon perturbé correspondant $\{(y_i, (x_i^1, \ldots, \widetilde{x}_i^j, \ldots, x_i^p)'), i = 1, \ldots, n\}$ est utilisé pour estimer les directions EDR, notée $\widehat{B}^{(j)}$, à l'aide d'une méthode SIR. Si la $j^{\text{ème}}$ composante de x a un effet sur y, cette permutation aléatoire aura un effet sur l'estimation de l'espace EDR et la mesure de proximité $m(\widehat{E}, \widehat{E}^{(j)})$ aura alors une valeur sensiblement différente de zéro, où $\widehat{E} = \text{Span}(\widehat{B})$ et $\widehat{E}^{(j)} = \text{Span}(\widehat{B}^{(j)})$. Par contre, si la $j^{\text{ème}}$ composante de x n'est pas liée à y, cette permutation aléatoire n'affectera pas l'estimation de l'espace EDR et la mesure de proximité $m(\widehat{E}, \widehat{E}^{(j)})$ aura alors une valeur proche de zéro.

Afin d'avoir une idée convenable de l'importance de cette $j^{\text{ème}}$ composante, il est nécessaire de faire un grand nombre de réplications de cette procédure. A partir de ces R réplications, nous pouvons résumer l'importance de cette variable par la moyenne des R valeurs IV, ainsi que par un boxplot de ces valeurs IV.

Puisque nous nous intéressons aux p composantes de x, cette procédure est naturellement appliquée à $j = 1, \ldots, p$. Pour toutes les réplications $(r = 1, \ldots, R)$ et toutes

les composantes ($j = 1, \ldots, p$), des estimations des directions EDR, notées $\widehat{B}^{(j,r)}$, sont obtenues et les valeurs des mesures de IV correspondantes sont alors calculées :

$$IV^{(j,r)} = m\left(\widehat{E}, \widehat{E}^{(j,r)}\right), \quad \text{avec } \widehat{E}^{(j,r)} = \text{Span}\left(\widehat{B}^{(j,r)}\right).$$

Comme mentionné précédemment, pour résumer l'information sur les valeurs de IV et comparer l'effet de chaque composante de x, la moyenne d'importance des variables peut être calculée pour chaque composante de x de la manière suivante:

$$\overline{IV}^{(j)} = \frac{1}{R}\sum_{r=1}^{R} IV^{(j,r)}, \quad \text{avec } j = 1, \ldots, p.$$

De plus, des boxplots parallèles des valeurs de IV peuvent être tracés pour comparer visuellement l'importance de chaque composante de x.

Cas de "$n < p$". Dans ce cas, il est impossible d'estimer l'espace EDR avec les méthodes SIR usuelles puisque la matrice de variance $\widehat{\Sigma}$ des x_i n'est pas inversible. Il est dépendant possible d'estimer les indices $x'\beta$ avec la méthode SIR-QZ, voir [5]. Ainsi, nous utilisons, pour calculer la mesure IV, des estimations des indices $x_i'\beta$ pour $i = 1, \ldots, n$, au lieu de l'estimation de l'espace EDR comme dans le cas précédent. La mesure IV est donc maintenant basée sur des sous-espaces de $\Re^n$ de dimension K (engendrés par des estimations des indices) plutôt que des sous-espaces de dimension K de $\Re^p$ (engendrés par des estimations de l'espace EDR). Soit X la matrice de taille $n \times p$ contenant les observations x_i. La mesure de proximité des deux indices de dimension K, $X'B_1$ et $X'B_2$ où B_1 et B_2 sont des matrices de taille $p \times K$, est définie comme suit :

$$m(X'B_1, X'B_2) = 1 - Q(X'B_1, X'B_2),$$

où $Q(X'B_1, X'B_2) = \dfrac{1}{K}\text{Trace}(P_1P_2)$ avec $P_l = X'B_l(B_l'XX'B_l)^{-1}B_l'X$ pour $l = 1, 2$. Sans perte de généralité, supposons que la matrice X est centrée (chaque variable est centrée). Lorsque $K = 1$, cette mesure de proximité peut être vue comme suit

$$m(X'B_1, X'B_2) = 1 - \text{cor}^2\left(X'B_1, X'B_2\right),$$

où 'cor" désigne la corrélation linéaire empirique entre les deux indices unidimensionnels.

Comme dans le cas précédent, on utilise cette mesure de proximité pour définir la mesure IV. Considérons un échantillon $\{(x_i, y_i), i = 1, \ldots, n\}$. L'estimation des indices EDR, $X'\widehat{B}$, peut être obtenue par SIR-QZ. Nous examinons ensuite les échantillons perturbés des x_i, avec une permutation aléatoire des valeurs de la $j^{\text{ème}}$ composante x^j de x. Cet échantillon $\{(y_i, (x_i^1, \ldots, \widetilde{x}_i^j, \ldots, x_i^p)'), i = 1, \ldots, n\}$ est utilisé pour calculer les indices EDR estimés, notée $X'\widehat{B}^{(j)}$, avec la méthode SIR-QZ, puis la mesure de proximité $m(X'\widehat{B}, X'\widehat{B}^{(j)})$ est calculée. Plus cette mesure est élevée, plus l'effet de la $j^{\text{ème}}$ composante de x sur l'estimation des indices est important.

Pour avoir une idée raisonnable de l'importance de cette $j^{\text{ème}}$ composante, nous utilisons R réplications ($r = 1, \ldots, R$) de cette procédure. Nous dupliquons ensuite cette procédure à l'ensemble des composantes de la covariable ($j = 1, \ldots, p$). A partir de toutes ces réplications, les valeurs de IV correspondantes sont calculées

$$IV^{(j,r)} = m(X'\widehat{B}, X'\widehat{B}^{(j,r)})$$

pour $j = 1, \ldots, p$ et $r = 1, \ldots, R$. Comme mentionné précédemment, pour résumer de l'information sur les valeurs de IV et comparer l'effet de chaque composante de x, la moyenne de l'importance des variables $\overline{IV}^{(j)}$ peut être calculée pour toutes les composantes de x. De plus, des boxplots parallèles des valeurs de IV peuvent être tracés pour comparer l'importance de chaque composante de x.

3 Selection des variables basée sur leurs mesures IV

L'objectif est d'identifier les composantes utiles de x, c'est à dire d'être capable de faire la différence entre les variables importantes et celles non pertinentes dans le modèle et donc dans la prédiction de y. La sélection des composantes les plus importantes de x peut être fait par deux méthodes basées sur les mesures IV.

Inspection visuelle. Dans certains cas, l'utilisateur peut facilement visualiser la différence d'importance des variables à partir des boxplots des valeurs de IV. Les boxplots des variables les plus importantes seront clairement distincts (valeurs élevées de l'indice et une forte dispersion de ces valeurs) des autres (valeurs proches de zéro et faible dispersion). En outre, le graphique des moyennes $\overline{IV}^{(j)}$, rangées par ordre décroissant des valeurs, est également utile pour identifier les variables importantes. Toutefois, cette inspection visuelle peut être parfois difficile lorsque l'utilisateur traite un grand ensemble de données (quand la dimension p de x est grande).

La détection des points de changement. L'idée est de trouver la position d'une rupture dans la séquence des moyennes d'importance $\overline{IV}^{(j)}$ classées par ordre décroissant des valeurs. Des nombreux auteurs ont proposé des méthodes de recherche pour détecter les points de rupture. Récemment, Killick et Eckley [8] ont développé un package R `Changepoint` qui permet de détecter l'emplacement des différents points de rupture. Pour la détection d'unique ou multiple points de rupture, l'approche permet d'estimer les points où les propriétés statistiques d'une séquence d'observations changent. Dans ce package, plusieurs méthodes de détection de ruptures en moyenne sont disponibles, avec des méthodes en mettant l'accent sur les détections des ruptures dans la moyenne ou dans la variance, et des méthodes recherchant une rupture à la fois en moyenne et en variance.

Dans les simulations numériques que nous avons faites, le package `Changepoint` est utilisé pour détecter un seul point de rupture en moyenne et en variance des séquences

ordonnées des moyennes $\overline{IV}^{(j)}$. Les variables situées avant ce point de rupture sont alors considérées comme importantes/pertinentes pour le modèle.

4 Simulations et données réelles

Nous illustrerons le bon comportement numérique de l'approche proposée avec des simulations dans les cas où "$n > p$" et "$n << p$". Nous considérerons également un jeu de données réelles afin de montrer le bon fonctionnement pratique de cette méthode.

Bibliographie

[1] Azais, R., Gégout-Petit, A. and Saracco, J. (2012). Optimal quantization applied to Sliced Inverse Regression. *Journal of Statistical Planning and Inference*, 142, 481-492.

[2] Bercu, B., Nguyen, T.M.N. and Saracco, J. (2011). A new approach of recursive and non recursive SIR methods. *Journal of the Korean Statistical Society*, 41, 17-36.

[3] Chen, C.H. and Li, K.C. (1998). Can SIR be as popular as multiple linear regression?. *Statistica Sinica*, 8 (2), 289-316.

[4] Cook, R.D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29, 2109-2121.

[5] Coudret, R., Liquet, B. and Saracco, J. (2014). Comparison of sliced inverse regression approaches for underdetermined cases. *Journal de la SFdS*, 155 (2), 72-96.

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

[6] Gannoun, A. and Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, 13, 297-310.

[7] Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics*, 27 (2), 697-731.

[8] Killick, R., Fearnhead, P. and Eckley, I.A. (2012). Optimal Detection of Changepoints with a Linear Computational Cost. *Journal of the American Statistical Association*, 107 (500), 1590-1598.

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

[10] Li, Y. and Zhu, L. (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics*, 35, 41-69.

[11] Liquet, B. and Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, 37(6),1198-1218.

[12] Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and Methods*, 26, 2141-2717.

[13] Yin, X. and Seymour, L. (2005). Asymptotic distributions for dimension reduction in the SIR-II method. *Statistica Sinica*, 15 (4), 1069-1079.