Un modèle de mélange pour la réduction de dimension

Laurent Gardes [1] & Jean-Luc Dortet-Bernadet [1]

[1] *Université de Strasbourg & CNRS, IRMA,*
UMR 7501, 7 rue René Descartes,
67 084 Strasbourg Cedex,
gardes@unistra.fr & dortet@unistra.fr

Résumé. La qualité de l'estimation du lien entre une variable Y et une variable explicative X décroit avec la dimension $p \geq 1$ de X. Ce phénomène est bien connu sous le nom de fléau de la dimension. Une solution à ce problème consiste à supposer l'existence d'un sous-espace de dimension inférieure à p, appelé espace DR, tel que la projection de X sur ce sous-espace contient toute l'information disponible sur Y. De nombreux estimateurs de cet espace ont été proposés dans la littérature, le plus connu étant l'estimateur SIR (Sliced Inverse Regression). Cependant, la plupart des méthodes d'estimation existantes sont mal adaptées au cas où la variable réponse est multivariée ou au cas où la fonction lien entre X et Y est symétrique. Nous proposons dans ce travail une nouvelle méthode d'estimation du sous-espace DR qui est efficiente dans ces situations. L'idée de départ est de modéliser la loi jointe du couple (X, Y) par une loi de mélange. La méthode proposée est comparée sur des données simulées aux méthodes d'estimations classiques.

Mots-clés. Réduction de dimension, Estimateurs du maximum de vraisemblance, Lois de mélange, Sliced Inverse regression.

Abstract. The existence of a Dimension Reduction (DR) subspace is a common assumption in regression analysis when dealing with high-dimensional predictors. The estimation of such a DR subspace has received considerable attention in the past few years, the most popular method being undoubtedly the Sliced Inverse Regression. Nevertheless, this method is limited to univariate response variables and is known to fail in presence of regression symmetric relationships. To overcome these limitations, we propose in this paper a new estimation procedure of the DR subspace assuming that the joint distribution of the predictor and the response variables is a finite mixture of distributions. The new method is compared through a simulation study to some classical methods.

Keywords. Dimension reduction, Maximum likelihood estimates, Mixture of distributions, Sliced Inverse Regression.

1 Introduction

Les méthodes de régression étudient la loi conditionnelle d'une variable réponse $Y \in \mathbb{R}^q$ sachant la valeur $X = x$ d'une variable explicative $X \in \mathbb{R}^p$ sur la base de n copies

indépendantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ d'un vecteur aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$. Lorsque la dimension p devient grande les n observations deviennent éparses, rendant l'inférence sur la loi conditionnelle difficile. Une solution possible à ce problème est de supposer l'existence d'un sous-espace de $\mathbb{R}^p$ de dimension $d < p$ contenant toute l'information disponible sur Y. Autrement dit, on suppose qu'il existe une matrice $\Gamma \in \mathbb{R}^{p\times d}$ de rang $d < p$ telle que $X \perp\!\!\!\perp Y | \Gamma^t X$ où la notation $U \perp\!\!\!\perp V | W$ signifie que, conditionnellement à W, U et V sont indépendantes. Notre objectif est donc l'estimation d'un sous-espace vectoriel $\mathcal{S}(\Gamma)$ engendré par les colonnes de Γ appelé espace DR (pour Dimension Reduction).

Une des premières méthodes permettant l'estimation de $\mathcal{S}(\Gamma)$ est la méthode SIR (Sliced Inverse Regression) proposée par Li (1991). Elle est basée sur l'estimation de $\mathrm{Var}(\mathbb{E}(X|Y))$ à l'aide d'un découpage en H tranches disjointes $\{S_h,\ h = 1, \ldots, H\}$ du support de la variable $Y \in \mathbb{R}$. Plus récemment, Cook (2007) a montré que l'estimateur SIR de Γ peut être vu comme un estimateur du maximum de vraisemblance en supposant que pour tout $h = 1, \ldots, H$, la loi conditionnelle X sachant $Y \in S_h$ est Gaussienne avec une espérance dépendant de Γ.

A l'instar de la plupart des méthodes d'estimation existantes de $\mathcal{S}(\Gamma)$, la méthode SIR s'adapte difficilement au cas d'une réponse Y multivariée ($q > 1$) principalement parce que le choix des H tranches devient alors problématique. Elle n'est pas adaptée non plus au cas où la fonction lien entre X et Y est symétrique. L'objectif de ce travail est de proposer une nouvelle méthode d'estimation de $\mathcal{S}(\Gamma)$ qui surmonte ces limitations. L'idée principale est de supposer que la distribution du couple (X, Y) est un mélange fini de distributions dépendant de la matrice Γ. Cette modélisation ne nécessite pas de découpage du support de Y, donc s'adapte facilement à la situation où la variable réponse est multivariée. Elle offre de plus une grande flexibilité qui permet d'appréhender une fonction lien symétrique entre X et Y.

Le modèle de mélange que l'on considère est défini dans la paragraphe **??**. Une méthode d'estimation de $\mathcal{S}(\Gamma)$ par maximum de vraisemblance est introduite dans le paragraphe **??** et une comparaison avec d'autres méthodes existantes est proposée dans le paragraphe **??**.

2 Modèle de mélange pour la réduction de dimension

On suppose dans la suite que le vecteur aléatoire $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$ admet une densité $f_{X,Y}(x, y)$ par rapport à la mesure de Lebesgue. Rappelons que notre objectif est l'estimation de l'espace DR $\mathcal{S}(\Gamma)$. Pour ce faire, pour un entier $M \geq d + 1$, on suppose qu'il existe des probabilités $\pi_1, \ldots, \pi_M$ de somme 1 telles que

$$f_{X,Y}(x, y) = \sum_{m=1}^{M} \pi_m f_m(x, y), \tag{1}$$

avec, pour $m = 1, ..., M$, $f_m(x, y) = g(x|\Theta)g_m(\Gamma^t x|\Gamma, \Theta_m)h_m(y|\Gamma, \Theta_m)$ où $g(\cdot|\Theta)$, $g_m(\cdot|\Gamma, \Theta_m)$ et $h_m(\cdot|\Gamma, \Theta_m)$ sont des fonctions connues dépendant de paramètres inconnus.

Théorème 1. *Sous le modèle (??), l'espace $\mathcal{S}(\Gamma)$ engendré par les colonnes de Γ est un espace DR pour la régression de Y sur X.*

Un exemple naturel de densité $f_m(\cdot, \cdot)$ est la densité de la loi Gaussienne : pour $\xi \in \mathbb{R}^p$, $\beta_m \in \mathbb{R}^d$ et une matrice de variance $V \in \mathbb{R}^{p\times p}$, pour $\alpha_m \in \mathbb{R}^q$ et une matrice de variance $W_m \in \mathbb{R}^{q\times q}$, on pose

$$f_m(x, y) = \varphi_p(x|\xi + V\Gamma\beta_m; V)\varphi_q(y|\alpha_m; W_m) \tag{2}$$

où $\varphi_k(\cdot|\mu; \Sigma)$ est la densité d'une loi Gaussienne multivariée de moyenne $\mu \in \mathbb{R}^k$ et de matrice de variance $\Sigma \in \mathbb{R}^{k\times k}$. Dans le paragraphe suivant, nous donnons l'expression de l'estimateur du maximum de vraisemblance de l'espace $\mathcal{S}(\Gamma)$ dans le cas où la densité $f_m(\cdot, \cdot)$ est donnée par (??).

3 Estimateur du maximum de vraisemblance

On considère à présent un vecteur aléatoire (X, Y) de densité (??) où $f_m(\cdot, \cdot)$ est donnée par (??). Soit $(x, y) := ((x_1, y_1), \ldots, (x_n, y_n))$ les observations de n copies indépendantes de (X, Y). Pour estimer $\mathcal{S}(\Gamma)$ on propose de maximiser en Γ, V, ξ, $\{(\pi_m, \beta_m, \alpha_m, W_m),\ m = 1, \ldots, M\}$ la fonction de vraisemblance

$$\mathcal{L}\left((x, y)\,|\Gamma, V, \xi, (\pi_m, \beta_m, \alpha_m, W_m)_{m=1,...,M}\right) = \prod_{i=1}^{n}\sum_{m=1}^{M} \pi_m\varphi_p(x_i|\xi+V\Gamma\beta_m; V)\varphi_q(y_i|\alpha_m, W_m).$$

Pour ce faire, on utilise l'algorithme EM qui consiste à introduire une variable aléatoire cachée Z prenant ses valeurs dans l'ensemble $\{1, \ldots, M\}$ avec $\mathbb{P}(Z = m) = \pi_m$ de telle sorte que la loi conditionnelle de (X, Y) sachant $Z = m$ admette pour densité $f_m(x, y)$. L'algorithme EM est un algorithme itératif qui à chaque étape calcule l'espérance de la fonction de vraisemblance complète (*i.e.* du vecteur aléatoire (X, Y, Z)) et la maximise. Le résultat suivant donne l'expression de l'estimateur de $\mathcal{S}(\Gamma)$ obtenu après convergence de l'algorithme EM. Tout d'abord, introduisons quelques notations : soient

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t$$

la moyenne et la variance empirique de X, pour $i = 1, \ldots, n$ et $m = 1, \ldots, M$ notons $z_{i,m}$ l'estimateur de $\mathbb{P}(Z = m|(X, Y) = (x_i, y_i))$ obtenu par l'algorithme EM et soit $\hat{C}_n$ la matrice de dimension $p \times p$ définie par

$$\hat{C}_n = \sum_{m=1}^{M} \hat{\pi}_m(\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^t$$

où

$$\hat{\pi}_m = \frac{1}{n}\sum_{i=1}^{n} z_{i,m} \quad \text{et} \quad \bar{x}_m = \frac{1}{n\hat{\pi}_m}\sum_{i=1}^{n} z_{i,m}x_i.$$

Remarquons que $\bar{x}_m$ peut être vu comme un estimateur de $\mathbb{E}(X|Z=m)$ et $\hat{C}_n$ comme un estimateur de la matrice de variance $C := \mathrm{Var}(\mathbb{E}(X|Z))$.

Théorème 2. *Sous le modèle (??), l'estimateur du maximum de vraisemblance de $\mathcal{S}(\Gamma)$ est l'espace engendré par les d vecteurs propres associés aux d plus grandes valeurs propres de la matrice $\hat{\Sigma}_n^{-1}\hat{C}_n$.*

4 Comparaison avec d'autres méthodes de réduction de dimension

Il existe de nombreuses méthodes de réduction de dimension dans la littérature. Nous allons ici comparer notre approche à deux de ces méthodes : la méthode SIR (Li (1991)) qui est la méthode la plus couramment utilisée et la méthode MSIR proposée par Srucca (2011) qui fait intervenir une loi de mélange. Comme dans notre approche, ces deux méthodes proposent d'estimer l'espace DR par une décomposition spectrale.

SIR La méthode SIR estime la matrice Γ par une décomposition spectrale d'un estimateur de $\Sigma^{-1}C^{(SIR)}$ où $C^{(SIR)} = \mathrm{Var}(\mathbb{E}(X|Y))$. L'estimateur SIR de Γ peut être vu comme l'estimateur du maximum de vraisemblance du modèle suivant : pour H tranches disjointes $\{S_1, \ldots, S_H\}$ couvrant le support de Y, on suppose que la loi conditionnelle de X sachant que $Y \in S_h$ est une loi Gaussienne de moyenne $\xi + V\Gamma\beta_h$ et de matrice de variance V. En notant $\hat{\Gamma}^{(SIR)}$ l'estimateur du maximum de vraisemblance de Γ, l'espace DR est estimé par $\mathcal{S}(\hat{\Gamma}^{(SIR)})$ qui est l'espace engendré par les d vecteurs propres associés aux d plus grandes valeurs propres de $\hat{\Sigma}_n^{-1}\hat{C}_n^{(SIR)}$ où $\hat{C}_n^{(SIR)}$ est un estimateur de $C^{(SIR)}$ donné par

$$\hat{C}_n^{(SIR)} = \sum_{h=1}^{H} \frac{n_h}{n}\left(\frac{1}{n_h}\left(\sum_{i:Y_i\in S_h} x_i\right) - \bar{x}\right)\left(\frac{1}{n_h}\left(\sum_{i:Y_i\in S_h} x_i\right) - \bar{x}\right)^t,$$

où n_h est le nombre d'observations dans la tranche S_h, $h = 1, \ldots, H$.

MSIR Pour la méthode MSIR de Scrucca (2011) la loi conditionnelle de X sachant que $Y \in S_h$ correspond à un mélange de J_h lois Gaussiennes

$$f_h(x) = \sum_{j=1}^{J_h} q_{h,j}\varphi_p(x|\mu_{h,j};\Sigma_{h,j})f_Y(y),$$

pour $q_{h,j},\ j = 1, \ldots, J_h$, des probabilités dont la somme vaut 1 et pour des vecteurs $\mu_{h,j} \in \mathbb{R}^p$ et des matrices de variance $\Sigma_{h,j} \in \mathbb{R}^{p \times p}$, $j = 1, \ldots, J_h$. L'espace DR est estimé par l'espace engendré par les d vecteurs propres associés aux d plus grandes valeurs propres de $\hat{\Sigma}_n^{-1} \hat{C}_n^{(MSIR)}$ avec

$$\hat{C}_n^{(MSIR)} = \sum_{h=1}^{H} \sum_{j=1}^{J_h} \frac{n_h}{n} \hat{q}_{h,j} \left(\hat{\mu}_{h,j} - \hat{\mu}\right) \left(\hat{\mu}_{h,j} - \hat{\mu}\right)^t \text{ pour } \hat{\mu} = \sum_{h=1}^{H} \sum_{j=1}^{J_h} \frac{n_h}{n} \hat{q}_{h,j} \hat{\mu}_{h,j},$$

et où, pour $h = 1, \ldots, H$ et $j = 1, \ldots, J_h$, $\hat{q}_{h,j}$ et $\hat{\mu}_{h,j}$ sont les estimateurs du maximum de vraisemblance du modèle. Cette matrice $\hat{C}_n^{(MSIR)}$ peut être vue comme un estimateur de $C^{(MSIR)} := \mathrm{Var}(\mathbb{E}(X|Y, Z^*))$ où Z^* est la variable cachée qui donne la composante du mélange à l'intérieur de chaque tranche.

La Figure **??** illustre le fonctionnement de ces méthodes sur un exemple simple où l'espace DR est de dimension 1 et où la variable réponse est réelle et simulée selon le modèle $Y = (\Gamma^t X)^2 + \varepsilon$ pour un bruit gaussien ε. La méthode SIR approxime $\mathbb{E}(X|Y)$ en ajustant à l'intérieur de chaque tranche S_h une unique loi gaussienne sur les observations de X. Cette méthode a donc des difficultés pour estimer l'espace DR dans le cas où, comme ici, la liaison entre $\Gamma^t X$ et Y est symétrique puisque $\mathbb{E}(X|Y)$ est constante. L'approche MSIR permet d'utiliser à l'intérieur de chaque tranche un mélange de lois gaussiennes. Cette méthode est donc plus flexible que la méthode SIR classique et permet de traiter les cas des liaisons symétriques. Cependant, à l'instar de SIR, elle est toujours basée sur un découpage en tranches du support de la variable Y ce qui rend la généralisation au cas où la variable réponse est multivariée difficile. Notre méthode n'utilise pas de tranches, donc s'adapte aisément à ce cas multivarié, et, par l'utilisation d'un mélange de lois, est plus flexible que SIR.

Bibliographie

[1] Cook, R.D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, **22(1)**, 1–26.

[2] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.

[3] Scrucca, L. (2011). Model-based SIR for dimension reduction. *Computational Statistics and Data Analysis*, **55**, 3010–3026.

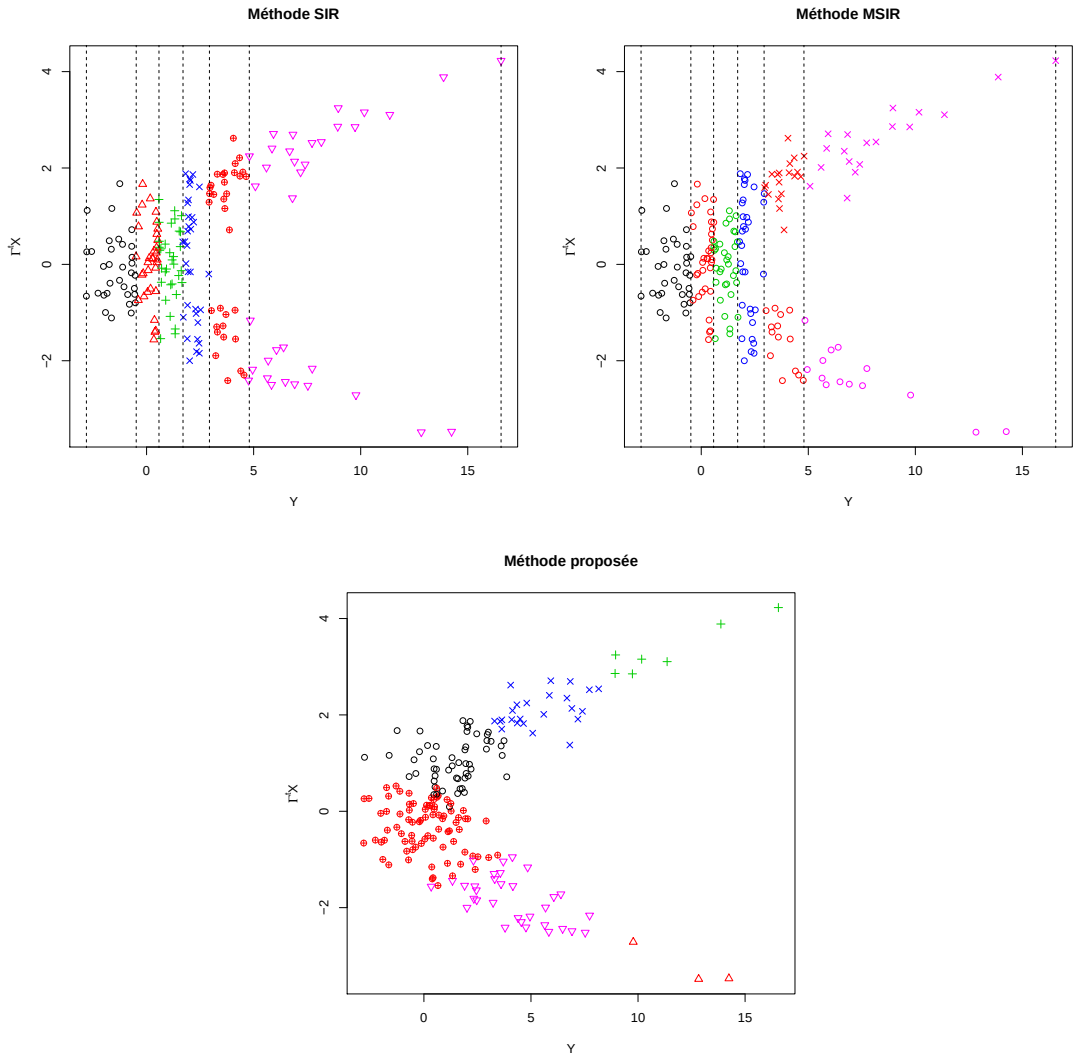

Figure 1: Représentation des observations de $(Y, \Gamma^t X)$. La méthode SIR utilise des tranches du support de Y (délimitées par des tirets) ici construites à partir des quantiles empiriques des observations de Y. La méthode MSIR permet d'utiliser à l'intérieur de chaque tranche un mélange de lois (les symboles différents correspondent aux classifications données par ces mélanges). La méthode proposée ne s'appuie pas sur des tranches et utilise un mélange de loi pour (X, Y).