Sélection de modèle incluant des composantes principales

Alois Kneip, Pascal Sarda

Sélection de modèle incluant des composantes principales

Alois Kneip[1] & Pascal Sarda[2]

[1] *Statistische Abteilung, Department of Economics and Hausdorff, Center for Mathematics, Universität Bonn, Adenauerallee 24-26, 53113 Bonn, Germany*
[2] *Institut de Mathématiques, UMR 5219, Équipe Statistique et Probabilités, 118, Route de Narbonne, 31062 Toulouse Cedex, France*

Résumé. Nous considérons un modèle de régression linéaire de grande dimension et plus précisément le cas d'un modèle factoriel pour lequel le vecteur des variables explicatives se décompose en la somme de deux termes aléatoires décrivant respectivement la variabilité spécifique et commune des prédicteurs. Nous montrons tout d'abord que les procédures de sélection de variables et d'estimation usuelles telles que le lasso ou le sélecteur Dantzig sont performantes dans ce contexte et sous l'hypothèse additionnelle que le vecteur des paramètres est *sparse*. Cette hypothèse peut être cependant restrictive. Nous introduisons ainsi un modèle de régression *augmenté* qui inclut les composantes principales. Nous montrons que ces composantes peuvent être convenablement estimées à partir de l'échantillon et nous nous concentrons ensuite sur les propriétés théoriques du modèle *augmenté*.

Abstract. We consider a high dimensional linear regression model and more precisely the case of a factor model where the vector of explanatory variables can be decomposed as a sum of two random terms representing respectively specific and common variability of the predictors. We show at first that usual prameter estimation and variable selection procedures such as Lasso or Dantzig selector are efficient in this context with the additional assumption that the vector of parameters is sparse. Such an assumption may be however restrictive. We thus introduce an augmented regression model which includes principal components. We show that these components can be accurately estimated from the sample and then we concentrate on the theoretical properties of the augmented model.

Mots clés. Modèle de régression linéaire, grande dimension, sélection de variables, composantes principales, Lasso, sélecteur Dantzig.

1 Introduction

Dans de nombreuses applications le nombre de variables ou de paramètres est très élevé voire plus grand que la taille de l'échantillon. Une large littérature statistique est désormais consacrée à l'étude de problèmes en grande dimension. Un des modèles les plus souvent considérés est le modèle de régression linéaire :

$$Y_i \;=\; \boldsymbol{\beta}^T \mathbf{X}_i \;+\; \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

où $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, sont des couples aléatoires avec $Y_i \in \mathbb{R}$ et $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$. Dans le modèle (1) $\boldsymbol{\beta}$ est un vecteur de paramètres dans $\mathbb{R}^p$ et $(\epsilon_i)_{i=1,\ldots,n}$ sont des v.a.r. i.i.d., indépendantes de $\mathbf{X}_i$, centrées et telles que $Var(\epsilon_i) = \sigma^2$. La dimension p du vecteur des paramètres est ici élevée comparativement à la taille de l'échantillon n.

Le modèle (1) décrit deux situations qui ont donné lieu à deux branches relativement indépendantes de la littérature statistique. La première correspond au cas où $\mathbf{X}_i$ représente un vecteur (de grande dimension) de différents prédicteurs alors qu'une autre situation apparaît lorsque les variables explicatives sont p points de discrétisation d'une même courbe. Dans ce dernier cas le modèle (1) est une version discrète du modèle linéaire fonctionnel. Pour chacune de ces situations des stratégies très différentes ont été adoptées afin d'estimer le vecteur des paramètres $\boldsymbol{\beta}$ et les hypothèses structurelles sous-jacents semblent être incompatibles.

Dans le premier cas les travaux reposent sur l'hypothèse que seul un nombre relativement petit de variables explicatives ont une influence significative sur la réponse Y_i ou qu'en d'autres termes le vecteur des coefficients β_j est *sparse* : $S := \#\{\beta_j | \beta_j \neq 0\} \ll p$. Cette hypothèse s'accompagne d'une condition sur les corrélations entre les différentes variables explicatives qui doivent être "suffisamment" faibles. Les procédures les plus populaires pour identifier et estimer les coefficients non nuls sont le *Lasso* et le *sélecteur Dantzig* : voir par exemple Tibshirani (1996), Bickel et al. (2009) et Candes and Tao (2007). Dans les travaux ayant trait à la statistique fonctionnelle on adopte des hypothèses très différentes. Si on considère le cas le plus simple où $X_{ij} = X_i(t_j)$ pour des fonctions aléatoires $X_i \in L^2([0,1])$ observées en des points équidistants $t_j = \frac{j}{p}$, on a alors $\beta_j := \frac{\beta(t_j)}{p}$, où $\beta(t) \in L^2([0,1])$ et lorsque $p \to \infty$, $\sum_j \beta_j X_{ij} = \sum_j \frac{\beta(t_j)}{p} X_i(t_j) \to \int_0^1 \beta(t) X_i(t) dt$. Par ailleurs, les corrélations entre les variables $X_{ij} = X_i(t_j)$ et $X_{il} = X_i(t_l)$, $j \neq l$, sont très fortes : lorsque $p \to \infty$, $corr(X_i(t_j), X_i(t_{j+m})) \to 1$ pour tout m fixé. Dans ce contexte, aucune variable $X_{ij} = X_i(t_j)$ n'a une influence particulière sur Y_i, et il y a une grand nombre de coefficients β_j qui sont proportionnels à $1/p$. Bien entendu, la réduction de dimension est également présente dans le cadre fonctionnel mais cependant elle s'obtient ici en réécrivant le modèle en termes d'une décompositon des prédicteurs sur une base "sparse", c'est-à-dire sur un petit nombre k de fonctions de base. Il est alors bien connu que la meilleure base possible au sens de l'erreur L^2 est celle fournie par les fonctions propres correspondant aux plus grande valeurs propres de l'opérateur de covariance de X_i. Parmi les nombreuses références sur le modèle linéaire fonctionnel citons Ramsay et Dalzell (1981), Cardot et al. (1999), Cai et Hall (2007), Hall et Horowitz (2007), Cardot et al. (2007) et Crambes et al. (2009).

Le but de notre travail est de montrer qu'une combinaison des idées développées dans les deux approches ci-dessus conduit, pour le modèle "discret" (1), à une procédure d'estimation nouvelle qui peut être utile dans de nombreuses applications. Nous mous plaçons dans un cadre général dans lequel les variables explicatives peuvent provenir ou pas de la discrétisation d'une même courbe. Par ailleurs, nous considérons un *modèle factoriel* de

la forme

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,\; i = 1, \ldots, n, \tag{2}$$

où $\mathbf{W}_i$ et $\mathbf{Z}_i$ sont deux vecteurs aléatoires indépendants de $\mathbb{R}^p$. Nous supposons que W_{ij} et Z_{ij} représentent des parties non négligeables de la variance de X_{ij} : chaque variable X_{ij}, $j = 1, \ldots, p$ possède une variabilité *spécifique* induite par Z_{ij} qui peut expliquer une partie de la réponse Y_i. D'un autre côté le terme W_{ij} représente une variabilité *commune* et les composantes principales, qui quantifient cette variabilité simultanée des régresseurs, peuvent également contribuer aux variations de la réponses. Ces arguments ont motivé l'utilisation d'un modèle de régression "augmenté" qui inclut les composantes principales comme variables explicatives additionnelles.

2 Le cadre de l'étude

Considérons le modèle de régression linéaire (1) avec des variables explicatives $\mathbf{X}_i$ qui peuvent être décomposées selon (2). On suppose de plus que $\mathbb{E}(X_{ij}) = 0$ pour tout $j = 1, \ldots, p$, et que

$$\sup_j \mathbb{E}(X_{ij}^2) \leq D_0 < \infty. \tag{3}$$

La matrice de variances-covariances de $\mathbf{\Sigma}$ de $\mathbf{X}_i$ se décompose sous la forme $\mathbf{\Sigma} \;=\; \mathbf{\Gamma} + \mathbf{\Psi}$, où $\mathbf{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$, alors que $\mathbf{\Psi}$ est une matrice diagonale. On note dans la suite $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T$ la matrice de variances-covariances empirique basée sur l'échantillon $\mathbf{X}_i$, $i = 1, \ldots, n$.

Les variables indépendantes Z_{ij}, $j = 1, \ldots, p$, avec $var(Z_{ij}) = \sigma_j^2$, sont supposées vérifier la condition suivante : il existe deux constantes positives D_1 et D_2 telles que

$(A.1)$ $$0 \;<\; D_1 \;<\; \sigma_j^2 \;<\; D_2.$$

Nous supposons par ailleurs l'hypothèse suivante

$(A.2)$ Il existe $C_0 < \infty$ tel que les événements

$$\sup_{1 \leq j,l \leq p} |\frac{1}{n}\sum_{i=1}^n W_{ij} W_{il} - cov(W_{ij}, W_{il})| \leq C_0 \sqrt{\frac{\log p}{n}}, \tag{4}$$

$$\sup_{1 \leq j,l \leq p} |\frac{1}{n}\sum_{i=1}^n Z_{ij} Z_{il} - cov(Z_{ij}, Z_{il})| \leq C_0 \sqrt{\frac{\log p}{n}}, \tag{5}$$

$$\sup_{1 \leq j,l \leq p} |\frac{1}{n}\sum_{i=1}^n Z_{ij} W_{il}| \leq C_0 \sqrt{\frac{\log p}{n}}, \tag{6}$$

$$\sup_{1 \leq j,l \leq p} |\frac{1}{n}\sum_{i=1}^n X_{ij} X_{il} - cov(X_{ij}, X_{il})| \leq C_0 \sqrt{\frac{\log p}{n}}, \tag{7}$$

sont réalisés simulatanément avec la probabilité $A(n,p) > 0$, où $A(n,p) \to 1$ as $n, p \to \infty$, $\frac{\log p}{n} \to 0$.

On montre que si les composantes $\mathbf{W_i}$ et $\mathbf{Z}_i$ de $\mathbf{X}_i$ vérifient $\mathbf{W}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$ et $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, alors $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma} + \mathbf{\Psi})$ et l'hypothèse (A.2) est alors satisfaite.

Concernant les variables W_{ij}, nous supposerons par ailleurs qu'un petit nombre de vecteurs propres de $\mathbf{\Gamma}$ suffit à bien approximer $\mathbf{W}_i$ (voir conditon (A.3) ci-dessous).

Envisageons pour le moment le cas d'un vecteur de paramètres $\boldsymbol{\beta}$ sparse :

$$\sharp\{\beta_j | \beta_j \neq 0\} \leq S \text{ for some } S \leq \tfrac{p}{2}.$$

Les procédures les plus populaires pour identifier et estimer les coefficients non nuls β_j sont Lasso et le sélecteur Dantzig. Dans un article récent Bickel et al. (2009) analysent ces méthodes. Ils donnent des conditions, *restricted eigenvalue assumptions*, portant sur les corrélations entre les variables X_{ij} et X_{il}, $j \neq l$, sous lesquelles ils obtiennent entre autres des bornes pour l'erreur L^q, $1 \leq q \leq 2$. On montre qu'une version de ces conditions, $RE(S, S, c_0)$, $c_0 = 1, 3$, est vérifiée par les variables explicatives ayant la structure (2) et lorsque (A.1) et (A.2) sont vérifiées. Notons que les variables X_{ij} sont préalablement normalisées de telle sorte que les éléments diagonaux de la matrice de variances-covariances empririque soient égaux à 1.

Nous avons remarqué plus haut que les variables W_{ij} peuvent également avoir une influence spécifique sur la réponse Y_i au travers d'un vecteur de paramètres non *sparse*. Dans ce cadre, nous pouvons intégrer les composantes principales aux variables explicatives. Nous présentons dans la section suivante le modèle augmenté résultant.

3 Le modèle augmenté

Nous notons dans la suite $\lambda_1 \geq \lambda_2 \geq \ldots$ les valeurs propres de $\frac{1}{p}\mathbf{\Gamma}$, $\mu_1 \geq \mu_2 \geq \ldots$, les valeurs propres de $\frac{\mathbf{1}}{\boldsymbol{p}}\mathbf{\Sigma}$ et $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq$ les valeurs propres de la matrice de variances-covariances $\frac{1}{p}\widehat{\mathbf{\Sigma}}$, alors que $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots$ et $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ sont des vecteurs propres orthonormés correspondant.

Le modèle incluant les composantes principales s'écrit

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r \xi_{ir} + \boldsymbol{\beta}^{*T}\mathbf{X}_i \;+\; \epsilon_i, \quad i = 1, \ldots, n, \tag{8}$$

où $\xi_{ir} = \boldsymbol{\delta}_r^T \mathbf{X}_i / \sqrt{p\mu_r}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T \in \mathbb{R}^k$ et $\boldsymbol{\beta}^* \in \mathbb{R}^p$ sont des vecteurs de paramètres. Nous supposons en outre que le vecteur $\boldsymbol{\beta}^*$ est sparse.

La première étape d'estimation du paramètres $(\alpha_1, \ldots, \alpha_k, \beta_1^*, \ldots, \beta_p^*)$ consiste à projeter le modèle à l'aide de la matrice de la projection sur l'espace engendré par les vecteurs

propres correspondant aux k plus grandes valeurs propres de $\frac{1}{p}\widehat{\mathbf{\Sigma}}$

$$\widehat{\mathbf{P}}_k \;=\; \mathbf{I}_p - \sum_{r=1}^{k} \widehat{\boldsymbol{\psi}}_r \widehat{\boldsymbol{\psi}}_r^T.$$

Le modèle (8) s'écrit alors pour $i = 1, \ldots, n$

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r^* \widehat{\xi}_{ir} + \sum_{j=1}^{p} \beta_j^{**} \frac{(\widehat{\mathbf{P}}_k \mathbf{X}_i)_j}{\left(\frac{1}{n}\sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2\right)^{1/2}} \;+\; \epsilon_i^* + \epsilon_i, \tag{9}$$

où $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T \mathbf{X}_i / \sqrt{p\widehat{\lambda}_r}$, $\alpha_r^* = \alpha_r + \sqrt{p\widehat{\lambda}_r}\sum_{j=1}^{p} \widehat{\psi}_{rj}\beta_j^*$, $\beta_j^{**} = \beta_j^* \left(\frac{1}{n}\sum_{i=1}^{n} (\widehat{\mathbf{P}}_k \mathbf{X}_i)_j^2\right)^{1/2}$ and $\epsilon_i^* = \sum_{r=1}^{k} \alpha_r (\xi_{ir} - \widehat{\xi}_{ir})$.

Nous montrons tout d'abord dans la proposition ci-dessous que les valeurs et vecteurs propres théoriques sont bien approximés par leurs versions empiriques. Nous faisons les hypothèses additionnelles suivantes

(A.3) Il existe $1 \geq v(k) \geq 3C_0(\log p/n)^{1/2}$ tel que les valeurs propres de $\frac{1}{p}\mathbf{\Gamma}$ sont telles que

$$\min_{j,l \leq k, j \neq l} |\lambda_j - \lambda_l| \geq v(k), \quad \min_{j \leq k} \lambda_j \geq v(k).$$

Enfin nous supposons que n et p sont suffisamment grands pour que l'hypothèses suivante soit vérifiée

(A.4) $C_0(\log p/n)^{1/2} \geq \frac{D_0}{pv(k)}$.

Proposition 1 *Sous les hypothèses (A.2)-(A.4) et sous les événemnts (4) - (7) on a pour tout $r \leq k$ et tout $j = 1, \ldots, p$*

$$|\lambda_r - \widehat{\lambda}_r| \leq \frac{D_2}{p} + C_0(\log p/n)^{1/2}, \; |\mu_r - \widehat{\lambda}_r| \leq C_0(\log p/n)^{1/2} \tag{10}$$

$$\|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 5\frac{\frac{D_2}{p} + C_0(\log p/n)^{1/2}}{v(k)}, \; \|\boldsymbol{\delta}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 3\frac{C_0(\log p/n)^{1/2}}{v(k)} \tag{11}$$

$$\psi_{rj}^2 \leq \frac{D_0 - D_1}{p\lambda_r} \leq \frac{D_0 - D_1}{pv(k)}, \tag{12}$$

$$\widehat{\psi}_{rj}^2 \leq \frac{D_0 + C_0(\log p/n)^{1/2}}{p\widehat{\lambda}_r} \leq 3\frac{D_0 + C_0(\log p/n)^{1/2}}{pv(k)}. \tag{13}$$

Nous sommes maintenant en mesure d'estimer les paramètres $\alpha*_r$ et β_j^{**} à l'aide du Lasso ou encore du sélecteur Dantzig. On en déduit ensuite des estimateurs de α_r et β_j^*. La condition $RE(k+S, k+S, c_0)$, $c_0 = 1, 3$, est satisfaite sous les hypothèses (A.2)-(A.4) ci-dessus. On en déduit alors, en utilisant les résultats de Bickel et al. (2009) des bornes pour la convergence L^q des estimateurs pour $1 \leq q \leq 2$.

Bibliographie

[1] Bickel, P.J., Ritov, Y. and Tsybakov, A. (2009) Simulataneous analysis of Lasso and Dantzig selector, *Ann. Statist.*, **37**, 1705-1732.
[2] Cai, T. and Hall, P. (2007). Prediction in functional linear regression, *Ann. Statisti.*, **34**, 2159-2179.
[3] Candès, E. and Tao, T. (2007). The Dantzig selector : statistical estimation when p is much larger than n, *Ann. Statist.*, **35**, 2013–2351, MR2382644.
[4] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model, *Statist. Prob. Letters*, **45**, 11-22.
[5] Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models, *Prob. Theory Related Fields*, **138**, 325-361.
[6] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing spline estimators for functional linear regression, *Ann. Statist.*, **37**, 35-72.
[7] Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression, *Ann. Statist.*, **35**, 70-91.
[8] Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis (with discussion), *J. Roy. Statist. Soc. Ser B*, **53**, 539-572.
[9] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.