Sélection de modèles pour la classification de données de régression en grande dimension : un résultat théorique.

Emilie Devijver [1]

[1] *Laboratoire de Mathématiques UMR 8628, Université Paris-Sud 11, F-91405 Orsay cedex, emilie.devijver@math.u-psud.fr*

Résumé. Les modèles de mélange en régression sont utilisés pour modéliser la relation qui existe entre la réponse et les prédicteurs, lorsque ces données sont hétérogènes. Avec l'augmentation des données de grande dimension, les modèles doivent aujourd'hui tenir compte des problèmes entrainés. Durant cet exposé, nous proposerons deux procédures de classification non supervisée en grande dimension. Dans chacune, nous construisons une collection de modèles de mélanges en faisant varier la dimension des modèles, pour pallier la grande dimension. Nous estimons les paramètres de chaque modèle par maximum de vraisemblance, sous contrainte de faible rang ou non, puis nous sélectionnons un modèle grâce à l'heuristique de pente introduite par Birgé et Massart. Nous obtenons une inégalité oracle pour chacune de nos procédures, ce qui nous permet de justifier la sélection de modèles par un critère pénalisé.

Mots-clés. Modèle de mélange, régression, grande dimension, sélection de modèles, inégalité oracle, heuristique de pentes.

Mixture models in regression are used to model the relationship between response and regressors, when data are heterogeneous. With the increasing of high-dimensional data, models have to deal with this issue.

During this talk, we will propose two procedures to cluster high-dimensional data. In each of them we will construct a model collection of mixture models, varying dimension. We estimate parameters in each model by Maximum Likelihood Estimator, under rank constraint or not, and we select a model thanks to the slope heuristic, introduced by Birgé and Massart. We get an oracle inequality for each procedure, which ensure utility of penalized criterion to select a model.

Keywords. Model-based clustering, regression, high-dimension, model selection, oracle inequality, slope heuristic.

1 Introduction

Soit $(x_i, y_i)_{1\leq i\leq n} \in \mathbb{R}^p \times \mathbb{R}^q$ un échantillon issu de variables aléatoires notées (X, Y). On veut regrouper les observations pour lesquelles les variables Y, conditionnellement à X, ont un comportement similaire.

Le modèle utilisé est alors un mélange de gaussiennes en régression, développé récemment par Städler et coauteurs par exemple. Pour étudier des données de grande dimension, on considère une collection de modèles, plus ou moins parcimonieux, et avec plus ou moins de classes. Ces procédures sont une généralisation de celle introduite par Maugis et Meynet dans un cadre de régression. De plus, la réestimation par faible rang est une généralisation à des modèles de mélange des techniques étudiées par Giraud et par Bunea et coauteurs.

On développe ici la justification théorique de l'utilisation de l'heuristique des pentes (introduite par Birgé et Massart) pour sélectionner un modèle. On généralise un théorème de Cohen et Le Pennec, décrit, dans le cas d'une famille aléatoire de modèles. On obtient ainsi une inégalité oracle pour deux collections de modèles.

2 Modèles de mélanges Gaussiens en régression

On observe n couples indépendants $(x_i, y_i)_{1\leq i\leq n} \in \mathbb{R}^p \times \mathbb{R}^q$, de densité conditionnelle inconnue s^*. On suppose que ces données proviennent d'un modèle de mélanges de modèles linéaires Gaussien : si le couple de variables aléatoires (X, Y) appartient à la classe k,

$$Y = \beta_k X + \epsilon \tag{1}$$

où $\epsilon \sim \mathcal{N}(0, \Sigma_k)$. On obtient donc que $Y|X = x \sim s_\xi(y|x)dy$, avec

$$s_\xi(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k x)^t \Sigma_k^{-1}(y-\beta_k x)}{2}\right);$$

$$\xi = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \Sigma_1, \ldots, \Sigma_K) \in \Xi = \left(\Pi_K \times (\mathbb{R}^{q\times p})^K \times (\mathbb{S}_q^{++})^K\right);$$

$$\Pi_K = \left\{(\pi_1, \ldots, \pi_K); \pi_k > 0 \text{ pour } k \in \{1, \ldots, K\} \text{ et } \sum_{k=1}^{K} \pi_k = 1\right\};$$

$\mathbb{S}_q^{++}$ est l'ensemble des matrices symétriques définies positives sur $\mathbb{R}^q$.

La log-vraisemblance associée à un échantillon $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{1\leq i\leq n}$ est

$$l(\xi, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y_i-\beta_k x_i)^t \Sigma_k^{-1}(y_i-\beta_k x_i)}{2}\right)\right)$$

et l'estimateur du maximum de vraisemblance est défini par

$$\hat{\xi}_0 := \underset{\xi\in\Xi}{\operatorname{argmin}} \left\{-\frac{1}{n} l(\xi, \mathbf{x}, \mathbf{y})\right\}.$$

On estime dans ces exposé la matrice de covariance par une matrice diagonale : $\Sigma_k = \text{diag}([\Sigma_k]_{1,1}, \ldots, [\Sigma_k]_{q,q})$.

3 Deux procédures pour classifier les données

Cette procédure peut être décomposée en trois étapes principales : on construit une collection de modèles, pour chaque modèle on calcule l'estimateur du maximum de vraisemblance sous contrainte de faible rang (procédure Lasso-Rang) ou non (procédure Lasso-MLE), puis on sélectionne le meilleur parmi ces modèles.

- On construit une collection de modèles notée $\{\mathcal{S}_{(K,J)}\}_{(K,J)\in\mathcal{M}}$ dans laquelle $\mathcal{S}_{(K,J)}$ est défini par l'équation

$$\mathcal{S}_{(K,J)} = \left\{ y \in \mathbb{R}^q | x \in \mathbb{R}^p \mapsto s_\xi^{(K,J)}(y|x) \right\} \tag{2}$$

  où

$$s_\xi^{(K,J)}(y|x) = \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{q/2}\det(\Sigma_k)^{1/2}} \exp\left(-\frac{(y-\beta_k^{[J]}x)^t \Sigma_k^{-1}(y-\beta_k^{[J]}x)}{2} \right),$$

  et

$$\xi = (\pi_1, \ldots, \pi_K, \beta_1^{[J]}, \ldots, \beta_K^{[J]}, \Sigma_1, \ldots, \Sigma_K) \in \Pi_K \times \left(\mathbb{R}^{q\times p}\right)^K \times \left(\mathbb{R}_+^q\right)^K,$$

  où la notation $A^{[J]}$, pour une matrice $A \in \mathcal{M}_{q,p}(\mathbb{R})$, signifie que la matrice a des coefficients nuls sur le complÃ©mentaire de J.

  La collection de modèles est indexée par $\mathcal{M} = \mathcal{K} \times \mathcal{J}$. On note $\mathcal{K} \subset \mathbb{N}^*$ l'ensemble du nombre de composantes possibles, et $\mathcal{J}$ l'ensemble des parties de $\{1, \ldots, q\} \times \{1, \ldots, p\}$. Pour trouver les variables actives, et construire un ensemble $\mathcal{J}^L \subset \mathcal{J}$ de taille raisonnable, on peut par exemple utiliser l'estimateur du Lasso.

- La deuxième étape consiste à approcher l'estimateur du maximum de vraisemblance, restreint aux variables actives,

$$\hat{s}^{(K,J)} = \underset{t\in\mathcal{S}_{(K,J)}}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(t(y_i|x_i)) \right\}$$

  ou sous contrainte de faible rang.

- La troisième étape consiste à sélectionner un modèle. On utilise l'heuristique des pentes décrites dans l'article de Birgé et Massart (2007). D'abord, on regroupe les modèles par leur dimension D. La dimension d'un modèle correspond au nombre de paramètres à estimer dans ce modèle. Pour chaque dimension D, notons

$\hat{s}_D$ l'estimateur maximisant la vraisemblance parmi les estimateurs associés aux modèles de dimension D. Alors, la fonction $D/n \mapsto \frac{1}{n}\sum_{i=1}^{n} \log(\hat{s}_D)$ a un comportement linéaire pour les grandes dimensions. On estime la pente, notée $\hat{\kappa}$. Alors, on sélectionne le minimiseur $\hat{D}$ du critère pénalisé $-\frac{1}{n}\sum_{i=1}^{n} \log(\hat{s}_D) + 2\hat{\kappa}D/n$, et l'estimateur sélectionné est $\hat{s}^{(K_{\hat{D}},J_{\hat{D}})}$.

4 Sélection de modèles

Commençons par quelques notations.

On note KL la divergence de Kullback-Leibler, définie par

$$\mathrm{KL}(s,t) = \begin{cases} E_s\left(\log\left(\frac{s}{t}\right)\right) \text{ si } s << t \\ +\infty \text{ sinon.} \end{cases}$$

Comme on travaille en régression, on définit une version tensorisée de la divergence de Kullback-Leibler ;

$$\mathrm{KL}^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(s(.|x_i), t(.|x_i))\right].$$

On a aussi besoin de la divergence de Jensen-Kullback-Leibler, définie par, pour $\rho \in]0,1[$,

$$\mathrm{JKL}_\rho(s,t) = \frac{1}{\rho}\,\mathrm{KL}(s, (1-\rho)s + \rho t);$$

et de sa version tensorisée

$$\mathrm{JKL}_\rho^{\otimes_n}(s,t) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{JKL}_\rho(s(.|x_i), t(.|x_i))\right].$$

Pour obtenir des résultats théoriques, on a besoin de borner nos paramètres. On considère

$$\mathcal{S}^{\mathcal{B}}_{(K,J)} = \left\{ s_\xi^{(K,J)} \in \mathcal{S}_{(K,J)} \,\middle|\, \xi \in \tilde{\Xi}_{(K,J)} \right\} \tag{3}$$

$$\tilde{\Xi}_{(K,J)} = \Pi_K \times ([-A_\beta, A_\beta]^{|J|})^K \times ([a_\Sigma, A_\Sigma]^q)^K \tag{4}$$

On obtient alors une inégalité oracle dans le cadre de la procédure Lasso-MLE.

Théorème 4.1 *Soit $(x_i, y_i)_{1\leq i\leq n}$ les observations, issues d'une densité conditionnelle inconnue s^*. Soit $\mathcal{S}_{(K,J)}$ définie par (2). On considère $\mathcal{J}^L \subset \mathcal{J}$ la sous-collection d'ensembles d'indices construite à travers le chemin de régularisation de l'estimateur du Lasso. Pour $(K, J) \in \mathcal{K} \times \mathcal{J}^L$, notons $\mathcal{S}^{\mathcal{B}}_{(K,J)}$ le modèle défini par (3).*

On considère l'estimateur du maximum de vraisemblance

$$\hat{s}^{(K,J)} = \underset{s_\xi^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(s_\xi^{(K,J)}(y_i|x_i)) \right\}.$$

Notons $D_{(K,J)}$ la dimension du modèle $\mathcal{S}^{\mathcal{B}}_{(K,J)}$, $D_{(K,J)} = K(|J| + q + 1) - 1$. Soit $\bar{s}_\xi^{(K,J)} \in \mathcal{S}^{\mathcal{B}}_{(K,J)}$ telle que

$$\mathrm{KL}^{\otimes_n}(s^*, \bar{s}_\xi^{(K,J)}) \leq \inf_{t \in \mathcal{S}^{\mathcal{B}}_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \frac{\delta_{\mathrm{KL}}}{n}; \tag{5}$$

et soit $\tau > 0$ tel que

$$\bar{s}_\xi^{(K,J)} \geq e^{-\tau} s^*. \tag{6}$$

Soit $\mathrm{pen} : \mathcal{K} \times \mathcal{J} \to \mathbb{R}_+$*, et supposons qu'il existe une constante absolue $\kappa > 0$ telle que, pour tout $(K, J) \in \mathcal{K} \times \mathcal{J}$,*

$$\mathrm{pen}(K, J) \geq \kappa \frac{D_{(K,J)}}{n} \left[B^2(A_\beta, A_\Sigma, a_\Sigma, q) - \log\left(\frac{D_{(K,J)}}{n} B^2(A_\beta, A_\Sigma, a_\Sigma, q) \wedge 1 \right) + (1 \vee \tau) \log\left(\frac{4epq}{(D_{(K,J)} - q^2) \wedge pq} \right) \right] ;$$

où les constantes $A_\beta, A_\Sigma, a_\Sigma$ sont définies dans (4), et $B(A_\beta, A_\Sigma, a_\Sigma, q)$ est une constante explicite dépendant seulement de $A_\beta, A_\Sigma, a_\Sigma, q$. Si on sélectionne le modèle indicé par $(\hat{K}, \hat{J})$, où

$$(\hat{K}, \hat{J}) = \underset{(K,J) \in \mathcal{K} \times \mathcal{J}^L}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log\left(\hat{s}^{(K,J)}(y_i|x_i) \right) + \mathrm{pen}(K, J) \right\};$$

alors l'estimateur $\hat{s}^{(\hat{K},\hat{J})}$ vérifie, pour tout $\rho \in]0, 1[$,

$$E\left[\mathrm{JKL}_\rho^{\otimes_n}(s^*, \hat{s}^{(\hat{K},\hat{J})}) \right] \leq C E\left(\inf_{(K,J) \in \mathcal{K} \times \mathcal{J}^L} \left(\inf_{t \in S_{(K,J)}} \mathrm{KL}^{\otimes_n}(s^*, t) + \mathrm{pen}(K, J) \right) + \frac{\Sigma^2}{n} \right);$$

pour une constante C absolue.

Ce théorème propose une pénalité minimale pour laquelle le modèle minimisant la log-vraisemblance pénalisée est aussi bon que l'oracle. Comme la collection de modèles considère des mélanges de Gaussiennes en régression, si on considère un nombre de classes suffisant, l'oracle et donc le modèle sélectionné approchent bien s^*. On obtient le même genre de résultat pour la procédure Lasso-Rang.

Pour obtenir ce résultat, on utilise des inégalités de concentration sur la somme des variables aléatoires et leur moyenne, et on a besoin de borner les fonctions, c'est pourquoi on utilise la divergence de Jensen-Kullback-Leibler pour $\rho \in]0,1[$ plutôt que la divergence de Kullback-Leibler.

Pour prouver une telle inégalité, sur une collection aléatoire, on plonge la collection aléatoire dans la collection complète, en contrôlant les termes sur la sous-collection aléatoire plutôt que sur la collection complète. Ceci nécessite l'utilisation de l'inégalité de Bernstein, et donc l'existence de $\bar{s}_\xi^{(K,J)}$ et de $\tau > 0$ tels qu'on ait (5) et (6).

Bibliographie

[1] Birgé, Massart (2007), Minimal penalties for Gaussian model selection, *Probability Theory Related Fields.*

[2] Bunea, She, Wegkamp (2012), Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, *Ann. Statist. 40, no. 5, 2359–2388. doi:10.1214/12-AOS1039.*

[3] Cohen, Le Pennec (2011), Conditional Density Estimation by Penalized Likelihood Model Selection and Applications, *Rapp. tech. INRIA.*

[4] Giraud (2011), Low rank multivariate regression, *Electron. J. Statist. 5 , 775–799. doi:10.1214/11-EJS625.*

[5] Meynet, Maugis (2012), A sparse variable selection procedure in model-based clustering, *Rapport de recherche INRIA.*

[6] Städler, Bühlman, Van de Geer (2000), ℓ_1-penalization for mixture regression models, *Test*, 19(2):209-256.