Estimation conjointe de plusieurs modèles de régression avec des pénalités ℓ_1.

Vivian Viallon [1] & Edouard Ollier [2]

[1] *Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron.* [2] *Université de Lyon, F-69622, Lyon, France.*

Résumé. Nous proposons une nouvelle approche, ainsi que des extensions, reposant sur l'utilisation de pénalisation ℓ_1 avec comme objectif l'estimation conjointe de plusieurs modèles de régression. Ce type de problème survient régulièrement en statistique appliquée, notamment en recherche clinique et en épidémiologie, lorsque les données proviennent de plusieurs strates d'observation. Un des intérêts principaux de notre approche est qu'elle peut être réécrite comme un simple Lasso pondéré sur une transformation des données originales. Son implémentation est de fait directe sous une variété de modèles de régression puisqu'il suffit d'utiliser les packages R disponibles pour l'implémentation du Lasso pondéré. Nous obtenons par ailleurs les propriétés oraculaires asymptotiques pour la version adaptative de notre approche, ainsi que des résultats non-asymptotiques préliminaires. A travers une étude de simulations, nous établissons par ailleurs les bonnes propriétés empiriques de notre approche. Nous l'illustrons enfin sur un jeu de données en épidémiologie du risque d'accident de la route.

Mots-clés. Apprentissage multi-tâche, Lasso, pénalté ℓ_1, analyse stratifiée. ...

Abstract. We propose a new approach, along with refinements, based on ℓ_1 penalties and aimed at jointly estimating several related regression models. Its particularly useful in clinical research and epidemiology when data come from several strata of observations. One of the main interests of our approach is that it can be rewritten as a weighted lasso on a simple transformation of the original data set. In particular, it does not need new dedicated algorithms and is ready to implement under a variety of regression models, *e.g.*, using standard R packages. Moreover, asymptotic oracle properties are derived along with preliminary non-asymptotic results, suggesting good theoretical properties. Our approach is further compared with state-of-the-art competitors under various settings on synthetic data: these empirical results confirm that our approach performs at least similarly to its competitors. As a final illustration, an analysis of road safety data is provided..

Keywords. Multi-task learning, Lasso, ℓ_1-penalization, stratified analysis...

1 Structure du texte long

Estimation conjointe de plusieurs modèles de régression avec des pénalités ℓ_1.

Vivian Viallon [1] & Edouard Ollier [2]

[1] *Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron.*
vivian.viallon@univ-lyon.fr
[2] *Université de Lyon, F-69622, Lyon, France.*

Ce travail se place dans le cadre de l'estimation et de la sélection de variables dans les modèles linéaires généralisés. Les méthodes pénalisées, notamment par la norme ℓ_1, se sont progressivement imposées parmi les méthodes de référence en particulier lorsque le nombre de covariables p excède le nombre d'observations n. La pénalisation ℓ_1 permet d'aboutir à des modèles interprétables (propriété de sélection de variables) et présentant de bonnes propriétés en terme d'estimation et de prédiction (sous certaines hypothèses sur la matrice de design notamment).

Dans de nombreuses applications, les données proviennent de plusieurs groupes ou strates d'une population de référence. Chaque strate peut correspondre à un type ou dosage de traitement, une zone géographique, ou peut être définie en croisant l'âge et le sexe des individus, etc. La relation entre la variable réponse d'intérêt $y \in \mathbb{R}$ et le vecteur de covariables $\mathbf{x} \in \mathbb{R}^p$ est alors à étudier sur chacune des strates. Cette relation est rarement strictement identique sur l'ensemble des strates mais, le plus souvent, une certaine homogénéité est attendue. Ainsi, l'approche naïve qui consiste à construire un modèle de régression sur chaque strate est le plus souvent inadaptée.

Le principe général des approches d'apprentissage multi-tâche est de tirer profit de la similitude attendue des fonctions à estimer, *i.e.*, dans notre cas, de la similitude des vecteurs de paramètres $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$, avec K le nombre de strates. En effet, on peut généralement travailler sous la double hypothèse suivante : (i) chacun des vecteurs $\boldsymbol{\beta}_k^*$ est relativement creux, et (ii) ces vecteurs sont en un certain sens proches $\boldsymbol{\beta}_{k_1}^* \approx \boldsymbol{\beta}_{k_2}^*$. Ainsi, en notant $\mathbf{B}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*) \in \mathbb{R}^{p \times K}$, cette matrice des paramètres est non seulement creuse, mais elle a aussi une certaine structure. Certaines des approches d'apprentissage multi-tâche reposent ainsi sur l'utilisation de pénalité renvoyant des vecteurs à sparsité dite structurée : c'est le cas notamment du group-lasso, etc.

L'approche que nous avons développée repose sur la décomposition suivante des vecteurs $\boldsymbol{\beta}_k^*$:

$$\boldsymbol{\beta}_k^* = \overline{\boldsymbol{\beta}}^* + \boldsymbol{\gamma}_k^*, \quad k \in [K],$$

où nous utilisons la notation $[K] = \{1, \ldots, K\}$. Dans cette décomposition, le terme $\overline{\boldsymbol{\beta}}^*$ peut être interprété comme le vecteur des effets communs ou globaux, alors que $\boldsymbol{\gamma}_k^*$ est censée capturer les variations des effets dans la strate k autour de ces effets globaux. Bien sûr, cette décomposition n'est pas unique, mais partant de l'hypothèse selon laquelle les $\boldsymbol{\beta}_k^*$ sont proches les uns des autres, les décompositions minimisant les quantités $\sum_k \|\boldsymbol{\gamma}_k^*\|_q$, pour $q \geq 0$, apparaissent naturelles. Elles correspondent à des choix eux-mêmes naturels pour le vecteur global $\overline{\boldsymbol{\beta}}^* \in \arg\min_{\overline{\boldsymbol{\beta}}} \sum_k \|\boldsymbol{\beta}_k^* - \overline{\boldsymbol{\beta}}\|_q$. En particulier, les choix $q = 0, 1$ et 2

correspondent aux décompositions où pour tout $j \in [p]$, $\overline{\beta}_j^*$ est respectivement un mode, une médiane et la moyenne de $(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$.

Le principe de notre approche est d'obtenir des estimations pour les vecteurs $\widehat{\boldsymbol{\beta}}_k$ de paramètres de chaque strate s'écrivant $\widehat{\boldsymbol{\beta}}_k = \widehat{\overline{\boldsymbol{\beta}}} + \widehat{\boldsymbol{\gamma}}_k$ avec $\widehat{\overline{\boldsymbol{\beta}}}$ et $\widehat{\boldsymbol{\gamma}}_k$ tous deux creux. Cet objectif peut être atteint en ayant recours à des pénalités ℓ_1. Plus précisément, soit n_k le nombre d'observations dans la k-ème strate, et $n = \sum_{k=1}^K n_k$ le nombre total d'observations. Soit $\mathbf{y}^{(k)} \in \mathbb{R}^{n_k}$ et $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times p}$ les observations de la variable d'intérêt et la matrice de design dans la k-ème strate. Pour simplifier les notations, nous nous placerons ici dans le cadre du modèle de régression linéaire, mais notre approche s'étend facilement à l'ensemble des modèles linéaires généralisés. Pour des valeurs appropriées des paramêtres $\lambda_1 \geq 0$ et $\lambda_{2,k} \geq 0$, nous définissons nos estimateurs comme

$$(\widehat{\overline{\boldsymbol{\beta}}}, \widehat{\boldsymbol{\gamma}}_1 \ldots, \widehat{\boldsymbol{\gamma}}_K) \in \underset{\overline{\boldsymbol{\beta}}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K}{\operatorname{argmin}} \left\{ \sum_{k \geq 1} \frac{\|\mathbf{y}^{(k)} - \mathbf{X}^{(k)}(\overline{\boldsymbol{\beta}} + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\overline{\boldsymbol{\beta}}\|_1 + \sum_{k \geq 1} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1 \right\}.$$

Ce problème d'optimisation est bien sûr équivalent à la minimisation du critère suivant

$$\sum_{k \geq 1} \frac{\|\mathbf{y}^{(k)} - \mathbf{X}^{(k)} \boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \{\|\overline{\boldsymbol{\beta}}\|_1 + \sum_{k \geq 1} \frac{\lambda_{2,k}}{\lambda_1} \|\boldsymbol{\beta}_k - \overline{\boldsymbol{\beta}}\|_1\}$$

sur $\boldsymbol{\beta}_k \in \mathbb{R}^p$ pour $k \in [K]$ et $\overline{\boldsymbol{\beta}} \in \mathbb{R}^p$. Il apparaît alors que $\widehat{\overline{\beta}}_j$ est une médiane pondérée et "shrinkée" de $(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$, que l'on notera $\text{WSmedian}_{\mu_{[K]}}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$ avec $\mu_{[K]} = (\mu_1, \ldots, \mu_K)$ et $\mu_k = \lambda_{2,k}/\lambda_1$. Si les μ_k sont égaux et tendent vers 0, alors $\text{WSmedian}_{\mu_{[K]}}$ tend vers la fonction constante $\mathbf{0}_p$: cela correspond à la décomposition $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\gamma}}_k$. D'autre part, si tous les μ_k's sont égaux mais tendent vers l'infini, alors $\text{WSmedian}_{\mu_{[K]}}$ tend vers la fonction "médiane": cela correspond à la décomposition $\widehat{\boldsymbol{\beta}}_k = \widehat{\overline{\boldsymbol{\beta}}} + \widehat{\boldsymbol{\gamma}}_k$, avec $\widehat{\overline{\beta}}_j = \text{mediane}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$. Si les μ_k's sont tous égaux à $1/\tau$, avec $\tau \in \mathbb{N}$, alors $\text{WSmedian}_{\mu_{[K]}}$ est une médiane "shrinkée" au sens où

$$\widehat{\overline{\beta}}_j = \text{WSmedian}_{1/\tau, \ldots, 1/\tau}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}) = \text{median}(\mathbf{0}_\tau^T, \widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}).$$

Un dernier exemple intéressant est lorsque pour $\ell \in [K]$ $\mu_\ell \to \infty$ et les autres μ_k sont finis, pour tout $k \neq \ell$. Dans ce cas, $\text{WSmedian}_{\mu_{[K]}}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}) \to \widehat{\beta}_{\ell,j}$, ce qui correspond à la décomposition $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\beta}}_\ell + \widehat{\boldsymbol{\gamma}}_k$, avec $\widehat{\boldsymbol{\gamma}}_\ell = \mathbf{0}_p$. Cela correspond à la stratégie classique en pratique où l'on considère la strate ℓ comme strate de référence. En résumé, chaque choix particulier des ratios $\lambda_{2,k}/\lambda_1$ "identifie" un vecteur global $\widehat{\overline{\boldsymbol{\beta}}}$ avec $\widehat{\overline{\beta}}_j$ défini comme $\text{WSmedian}_{\mu_{[K]}}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$, et notre approche encourage les solutions $(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ avec un vecteur global $\widehat{\overline{\boldsymbol{\beta}}}$ et des vecteurs de différences $\widehat{\boldsymbol{\beta}}_k - \widehat{\overline{\boldsymbol{\beta}}}$ typiquement creux.

Une propriété très intéressante de notre approche est qu'elle peut être implémentée facilement à l'aide de packages disponibles pour le Lasso pondéré, tel que le package glmnet de R par exemple. En effet, introduisons $\mathscr{Y} = ({\mathbf{y}^{(1)}}^T, \ldots, {\mathbf{y}^{(k)}}^T)^T \in \mathbb{R}^n$ et les quantités

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \ldots & \mathbf{X}^{(K)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta} = \begin{pmatrix} \overline{\boldsymbol{\beta}} \\ \boldsymbol{\gamma}^{(1)} \\ \vdots \\ \boldsymbol{\gamma}^{(K)} \end{pmatrix},$$

qui sont des éléments de $\mathbb{R}^{n\times(K+1)p}$ et $\mathbb{R}^{(K+1)p}$ respectivement. Introduisons de plus le vecteur de poids $\boldsymbol{\omega} = (\mathbf{1}_p^T, (\lambda_{2,1}/\lambda_1)\mathbf{1}_p^T, \ldots, (\lambda_{2,K}/\lambda_1)\mathbf{1}_p^T)^T \in \mathbb{R}^{(K+1)p}$. Alors, en notant $\|\boldsymbol{\theta}\|_{1,\boldsymbol{\omega}} = \sum_{j=1}^{(K+1)p} \omega_j|\theta_j|$, le critère à minimiser dans notre approche se réécrit simplement

$$\frac{\|\mathscr{Y} - \boldsymbol{\mathcal{X}}\boldsymbol{\theta}\|_2^2}{2n} + \lambda_1\|\boldsymbol{\theta}\|_{1,\boldsymbol{\omega}} \tag{1}$$

qui correspond bien à un Lasso pondéré.

Nous établissons le lien entre notre approche et diverses autres méthodes proposées dans la littérature. En particulier, une version adaptive de notre approche peut être vue comme une amélioration de la stratégie qui consiste à sélectionner une strate de référence et inclure des termes d'interaction entre les covariables et les indicatrices d'appartenance aux autres strates. Notre approche présente deux atouts principaux : (i) la strate de référence est choisie automatiquement à partir d'estimateurs initiaux et (ii) la strate de référence peut être différente d'une covariables à l'autre. Ce dernier point est particulièrement intéressant pour les propriétes d'estimation et de prédiction puisqu'il assure que notre approche construit des modèles typiquement moins complexes (en terme de nombre de paramètres non nuls à estimer). Un lien est également établi avec les stratégies reposant sur le Generalized Fused Lasso, ainsi que les Dirty Models de Jalali et al. (2013).

Concernant les propriétés statistiques de nos estimateurs, nous établissons les propriétés oraculaires asymptotiques de la version adaptative de notre approche et certains résultats non-asymptotiques préliminaires. En particulier, reprenant le cadre considéré par Jalali et al. (2013) nous montrons que dans le cas $K = 2$, et sous des designs gaussiens, la "sample complexity" (*i.e.* le nombre d'observations suffisant pour assurer les propriétés de sélection de modèles avec grande probabilité) est plus faible pour notre approche que pour leur Dirty Models, sous l'hypothèse supplémentaire $\beta^*_{1,j}\beta_{2,j} \geq 0$ pour tout $j \in [p]$.

Nos résultats de simulation confirment les bonnes propriétés de notre approche.

Bibliographie

[1] A. Jalali, P. Ravikumar, and S. Sanghavi. (2013), A dirty model for multiple sparse regression, *IEEE Transactions on Information Theory*, , 59, 7947–7968.