Sélection de groupes de variables corrélées par classification ascendante hiérarchique et group-lasso

Quentin Grimonprez[1] & Alain Celisse[2] & Guillemette Marot[3]

[1] *DGA & Inria Lille-Nord Europe, quentin.grimonprez@inria.fr*
[2] *Inria Lille-Nord Europe & Laboratoire Paul Painlevé, Université Lille 1, alain.celisse@inria.fr*
[3] *Inria Lille-Nord Europe & EA 2694, Université Lille 2, guillemette.marot@inria.fr*

Résumé. Dans un contexte de sélection de variables, utiliser des régressions pénalisées en présence de fortes corrélations peut poser problème. Seul un sous-ensemble des variables corrélées est sélectionné. Agréger préalablement les variables liées entre elles peut aussi bien aider à la sélection qu'à l'interprétation. Cependant, les méthodes de regroupement de variables nécessitent la calibration de paramètres supplémentaires. Nous présenterons une nouvelle méthode combinant classification ascendante hiérarchique et sélection de groupes de variables.

Mots-clés. group-lasso, classification, sélection de variables

Abstract. In a context of variable selection, the use of penalized regressions in presence of high correlations might be problematic. Only a subset of the correlated variables is selected. Firstly aggregating related variables can help both for selection and interpretation. However, clustering methods require calibration of additional parameters. We will introduce a new method combining hierarchical clustering and group selection.

Keywords. group-lasso, hierarchical clustering, variable selection

1 Introduction

Le problème de la sélection de variables est un problème courant notamment en génomique où l'on est, par exemple, intéressé par la sélection de quelques marqueurs d'intérêt (par rapport à une réponse associée) sur un profil ADN comportant plusieurs milliers de variables. Une méthode classiquement utilisée pour répondre à ce problème est le LASSO [Tibshirani, 1994] couplant régression et sélection de variables par l'ajout d'une pénalité l_1 sur les coefficients estimés. Les propriétés du LASSO ont été largement étudiées dans la littérature comme la consistance de la sélection des variables ainsi que ses limitations notamment en présence de variables corrélées [Zhao et Yu, 2006] et la grande dimension. [Wainwright, 2009] a calculé des bornes théoriques sur la taille des données pour assurer la consistance des variables sélectionnées. Différentes versions du LASSO ont été développées

pour dépasser ces limitations, notamment ADAPTIVE LASSO [Zou, 2006], rajoutant des poids pour diminuer l'impact des corrélations, le GROUP-LASSO [Yuan et Lin, 2006] permettant de sélectionner des groupes de variables à partir d'une partition donnée a priori.

Dans la suite, nous proposerons une méthode basée sur le regroupement de variables via la classification ascendante hiérarchique (CAH) et le GROUP-LASSO pour choisir les groupes d'intérêts puis nous la testerons sur données simulées. Regrouper les variables corrélées entre elles et utiliser un GROUP-LASSO peut être un moyen de contourner les limitations du LASSO concernant la corrélation. En effet, en présence de fortes corrélations entre variables, le LASSO va généralement privilégier une variable parmi l'ensemble des variables corrélées entre elles.

2 Méthode

2.1 Group-lasso

Soit $X \in \mathcal{M}_{n,p}(\mathbb{R})$, une matrice contenant en lignes les différents individus, $y \in \mathbb{R}^n$, une réponse associée définie par $y = X\beta^* + \epsilon$ avec $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$ où I_n est la matrice identité de taille n et $\beta^* \in \mathbb{R}^p$ le vecteur des coefficients avec $\|\beta^*\|_0 = k < p$. Le but est ici de retrouver les k variables non nulles.

Le LASSO permet de répondre à cette problématique en résolvant le problème des moindres carrés avec une contrainte sur la norme l_1 des coefficients du vecteur β, ce qui forcera un certain nombre de coefficients de β à être nuls. L'estimateur LASSO est

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

avec $\lambda \geq 0$ le paramètre de régularisation. Plus λ sera grand et plus le problème sera contraint et donc plus de coefficients de $\hat{\beta}_\lambda$ seront nuls.

Dans certains cas, la sélection simultanée d'un ensemble de variables fait plus sens que la sélection de variables une à une. Par exemple, en génomique, on peut imaginer regrouper les marqueurs correspondant à une même voie métabolique. Une des variantes du LASSO, le GROUP-LASSO, permet à l'aide d'une partition donnée a priori de sélectionner un certain nombre de groupes en fonction du paramètre de régularisation. On définit $\mathcal{G}$ un ensemble de groupes formant une partition en $K = |\mathcal{G}|$ groupes des p variables et $\beta_g \in \mathbb{R}^{|g|}$ avec $g \in \mathcal{G}$, le vecteur β restreint aux variables du groupe g. L'estimateur du GROUP-LASSO est :

$$\hat{\beta}_\lambda^{\mathcal{G}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \sqrt{|g|} \|\beta_g\|_2 \right\}.$$

Le poids $\sqrt{|g|}$ permet de pénaliser plus fortement les groupes de grande taille qui auront tendance à être favorisés pour la sélection par rapport à des groupes de petite taille.

2.2 Regroupement de variables

Différentes méthodes de regroupement de variables existent, les plus connues sont les $k-$means et la classification ascendante hiérarchique (CAH) [Jain *et al.*, 1999]. La CAH a l'avantage d'être un algorithme stable (indépendant d'une initialisation) et permet une meilleure interprétation grâce à la structure hiérarchique fournie.

Soient $X^1, \ldots, X^p$, p variables et d une mesure de dissimilarité (une distance sans la propriété de l'inégalité triangulaire), généralement la distance euclidienne. L'algorithme part de p classes (une pour chaque variable) et à chaque étape les 2 classes les plus proches (selon un critère d'agrégation basé sur la dissimilarité d) vont être réunies, et ce jusqu'à obtenir une seule classe contenant l'ensemble des variables. La hiérarchie formée peut être représentée dans un dendrogramme (Fig. 1), où les hauteur des branches représentent les valeurs du critère auxquelles les groupes se rejoignent. De grandes longueurs de branches indiquent le regroupement de classes peu ressemblantes.

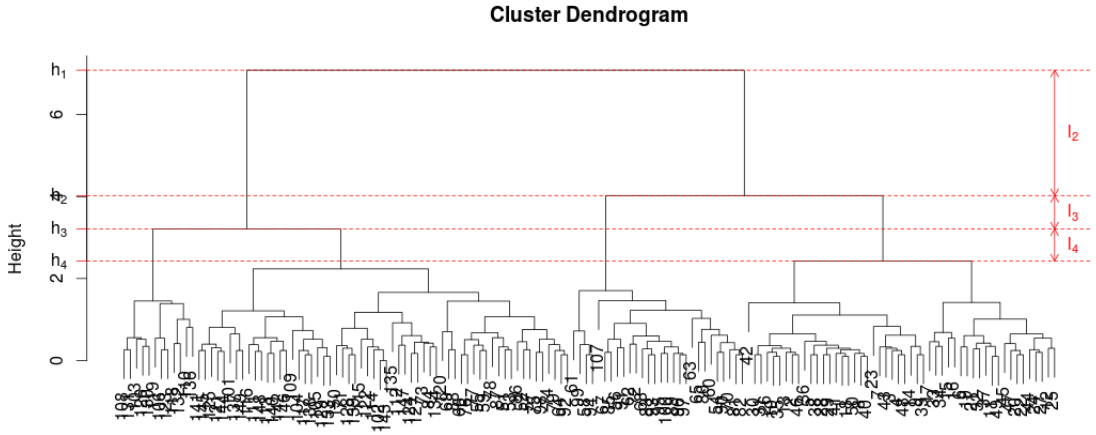

Figure 1: Dendrogramme obtenu après CAH sur les données iris. l_2, l_3, l_4 représentent les longueurs des 3 dernières branches et h_1, h_2, h_3, h_4, les valeurs du critère d'agrégation associées aux 4 derniers niveaux de la CAH.

2.3 Méthode proposée

La méthode proposée combine la CAH avec le GROUP-LASSO. Toutes les partitions des variables obtenues aux différents niveaux de la CAH sont fournies au GROUP-LASSO.

Premièrement, une CAH est effectuée sur les variables $X^1, \ldots, X^p$. À chaque niveau $s = 1, \ldots, p$ de la CAH, on obtient une partition en s groupes des p variables, on parlera de partition du niveau s de la CAH. Pour chaque niveau s, on associe la longueur de branche l_s qui correspond à la différence des hauteurs des niveaux $s-1$ et s, $l_s = h_{s-1} - h_s$ (cf Fig. 1).

Ensuite, on définit l'estimateur suivant :

$$\hat{\beta}_\lambda^G = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - \tilde{X}\beta\|_2^2 + \lambda \sum_{s=2}^{p-1} \frac{1}{\sqrt{l_s}} \sum_{g \in \mathcal{G}_s} \sqrt{|g|}\|\beta_g\|_2 \right\}, \quad (1)$$

avec $G = \cup_s \mathcal{G}_s$ où $\mathcal{G}_s$ est l'ensemble des groupes associés au niveau s de la CAH. $\tilde{X}$ est une matrice où les colonnes de X ont été dupliquées autant de fois que les variables apparaissent dans les différents niveaux de la CAH obtenue. Utiliser $\frac{1}{\sqrt{l_s}}$ comme poids pour les différents niveaux s de la CAH va favoriser les niveaux ayant de grandes longueurs de branches. En effet, une grande longueur de branche indique le regroupement de 2 classes peu ressemblantes au niveau suivant.

L'objectif est que le critère choisisse la meilleure partition et les meilleurs groupes de cette partition pour un λ approprié.

3 Post-traitement

Comme démontré dans [Wainwright, 2009], le Lasso peut retrouver exactement les k vraies variables non nulles de β^* sous certaines conditions (sur les corrélations, taille des données, ...). Certaines de ces conditions ne sont pas vérifiables en pratique et il est alors courant d'appliquer une méthode de post-traitement sur les solutions du Lasso pour en améliorer la qualité [Meinshausen et Yu, 2009] [Wasserman et Roeder, 2009].

Pour le Group-Lasso, un résultat de consistance des groupes sélectionnés est démontré dans [Bach, 2008]. De la même manière que pour le Lasso, l'utilisation d'une méthode de post-traitement est envisageable pour améliorer la qualité de la sélection.

Dans le cadre du Lasso, une possibilité est de réestimer les coefficients des variables sélectionnées par les moindres carrés afin d'éliminer le biais introduit (ce qui est possible car le nombre de variables sélectionnées par le Lasso est inférieur au nombre d'individus n) puis d'effectuer un test de nullité des coefficients (test de Student par exemple). Pour le Group-Lasso, cette procédure ne peut être appliquée telle quelle car le nombre de coefficients non nuls peut être plus grand que n. Mais le nombre de groupes sélectionnés est quant à lui inférieur à n [Liu et Zhang, 2009]. L'idée est donc de représenter chaque groupe par une unique variable puis d'appliquer un test. On choisit ici de représenter chaque groupe par la première composante de l'ACP sur ses variables, car celle-ci conservera au plus la variance du groupe.

Notons $\hat{S}_\lambda^G = \{g \in G \mid \|(\hat{\beta}_\lambda^G)_g\|_2 \neq \mathbf{0}\}$, l'ensemble des groupes sélectionnés par l'estimateur (1). La procédure est :

1. Représenter chaque groupe $g_i \in \hat{S}_\lambda$, $i = 1, \ldots, k_\lambda$ par $\dot{X}^i$ la première composante principale de l'ACP de $\tilde{X}^{g_i}$;
2. Calculer $\tilde{\beta}$ l'estimateur des moindres carrés de y sachant $\dot{X} = [\dot{X}^1, \ldots, \dot{X}^{k_\lambda}]$;

3. Tester la significativité des coefficients estimés par un test de nullité avec correction de Benjamini-Hochberg [Benjamini et Hochberg, 1995] pour les tests multiples afin de contrôler le False Discovery Rate (FDR) au niveau α.

Un problème survient lors de l'application de cette procédure : les éléments de $\hat{S}^G_\lambda$ peuvent être d'intersection deux à deux non nulle car notre estimateur (1) a la possibilité de sélectionner des groupes issus de différentes partitions. Par exemple, on peut avoir $\hat{S}^G_\lambda = \{g_1, g_2\}$ avec $g_1 = \{X^1, X^2, X^3\}$ et $g_2 = \{X^1, X^2\}$, g_1 et g_2 étant issus de 2 partitions associées à des niveaux différents, $\mathcal{G}_{s_1}$ et $\mathcal{G}_{s_2}$. On souhaite que les groupes obtenus après la phase de test soient 2 à 2 disjoints par souci d'interprétation (dans l'exemple précédent, on peut se demander si la variable 3 joue un rôle important ou non). Pour cela, nous appliquons une stratégie de test hiérarchique [Meinshausen, 2008]. Ce test permet de sélectionner, au sein d'une hiérarchie, les plus petits groupes ayant un effet sur la réponse tout en contrôlant le Family-Wise Error Rate (FWER). Ce test hiérarchique avec contrôle du FWER sera utilisé à l'étape 3 de l'étape de post-traitement.

4 Conclusion

Nous proposons une méthode regroupant CAH et GROUP-LASSO afin de sélectionner des groupes de variables corrélées. La méthode de post-traitement permet d'obtenir des groupes de variables corrélées expliquant ua mieux la réponse associée. La méthode de post-traitement permet d'obtenir un contrôle du nombre de faux positifs pour le sgroupes sélectionnés. L'estimateur proposée ainsi que la méthode de post-traitement seront éprouvées sur données simulées.

Remerciements

Merci à la Direction Générale de l'Armement et à Inria pour le financement direct de ce travail.

References

[Bach, 2008] Bach, F. (2008) Consistency of the Group Lasso and Multiple Kernel Learning, *J. Mach. Learn. Res.*, 9, 1179-1225.

[Benjamini et Hochberg, 1995] Benjamini, Y. and Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society, Series B (Methodological)*, 57, 289-300.

[Jain *et al.*, 1999] Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) Data Clustering: A Review.

[Liu et Zhang, 2009] Jiu, H. and Zhang, J. (2009) Estimation Consistency of the Group Lasso and its Applications.

[Meinshausen, 2008] Meinshausen, N. (2008) Hierarchical testing of variable importance, *Biometrika*, 95(2), 265-278.

[Meinshausen et Yu, 2009] Meinshausen, N. and Yu, N. (2009) Lasso-type recovery of sparse representations for high-dimensional data, *Annals of Statistics*, 246-270.

[Tibshirani, 1994] Tibshirani, R. (1994) Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[Wainwright, 2009] Wainwright, Martin J. (2009) Sharp Thresholds for High-dimensional and Noisy Sparsity Recovery Using L1-constrained Quadratic Programming (Lasso), *JIEEE Trans. Inf. Theor.*, 55, 2183-2202.

[Wasserman et Roeder, 2009] Wasserman, L. and Roeder, K. (2009) High-dimensional variable selection, *The Annals of Statistics*, 37, 2178-2201.

[Yuan et Lin, 2006] Yuan, M. and Lin, Y. (2009) Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, 68, 49-67.

[Zhao et Yu, 2006] Zhao, P. et Yu, B. (2006) On model selection consistency of lasso, *J. Mach. Learn. Res.*, 7, 2541-2563.

[Zou, 2006] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, 101, 476, 1418-1429.