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

Quentin Grimonprez, Alain Celisse, Guillemette Marot

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 aider aussi bien 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

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 comprenant 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 l'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 de groupes donnés 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 sélectionner 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 $\epsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_n)$ où I_n est la matrice identité de taille $n \times n$, $\mathbf{0}_n$ le vecteur nul de longueur n et $\beta \in \mathbb{R}^p$ le vecteur des coefficients.

L'estimateur des moindres carrés $\hat{\beta}^{LS}$ de la régression linéaire est

$$\hat{\beta}^{LS} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 \right\}.$$

Dans le cas $p > n$, la solution des moindres carrés est mal définie (la matrice ${}^tXX$ n'est pas de plein rang et donc non inversible). Les valeurs de β_j^{LS} représentent l'influence des variables. En présence de beaucoup de variables, il n'est pas intéressant d'avoir tous les coefficients de $\hat{\beta}^{LS}$ non nuls pour des soucis d'interprétation notamment.

Le LASSO va résoudre ces problèmes en contraignant 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 $G = |\mathcal{G}|$ 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) (voir [Jain *et al.*, 1999] pour une review). 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.

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.

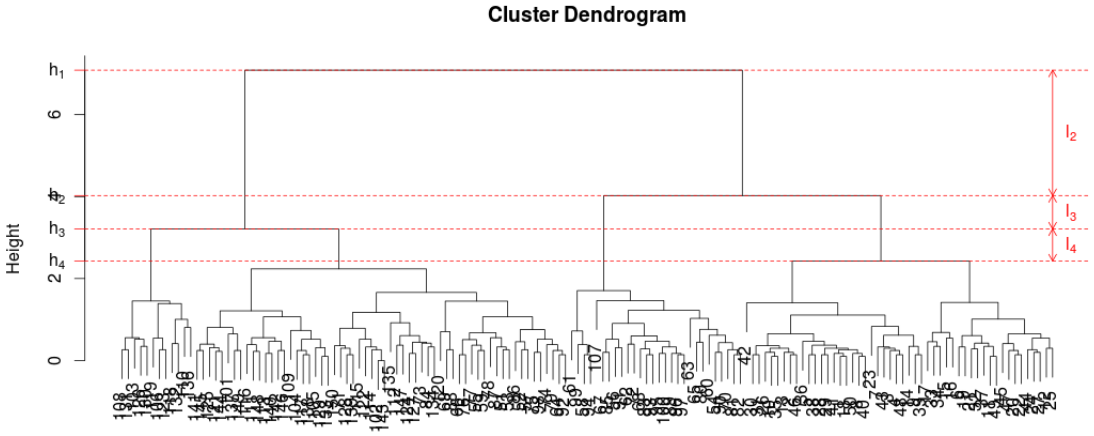

FIGURE 1 – Dendrogramme obtenu après CAH sur les données iris. l_2, l_3, l_4 représentent les longueurs des 3 dernières branches et h_1, h_2, h_3, h_4, les valeurs du critère d'agrégation associées aux 4 derniers niveaux de la CAH.

Premièrement, une CAH est effectuée sur les variables $X^1, \dots, X^p$. À chaque niveau $s = 1, \dots, p$ 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 :

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

où $\mathcal{G}_s$ est l'ensemble des groupes associés au niveau s de la CAH. La matrice $\tilde{X}$ 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 $\frac{1}{l_s}$ 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 Simulations

Chaque individu est généré suivant une loi gaussienne multivariée de moyenne nulle et de matrice de variance Σ_ρ :

$$X_1, \dots, X_n \sim \mathcal{N}(\mathbf{0}_p, \Sigma_\rho) \ .$$

La matrice de variance Σ_ρ définit une structure en un nombre G de groupes de tailles différentes et indépendants entre eux (Fig. 2). Les variables au sein d'un même groupe ont un coefficient de corrélation de ρ entre elles.

On définit ensuite une réponse $y \in \mathbb{R}^n$ de la forme $y = X\beta^* + \epsilon$, où $\epsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_n)$ et β^* contient K éléments non nuls correspondant à des groupes différents que nous appellerons vraies variables dans

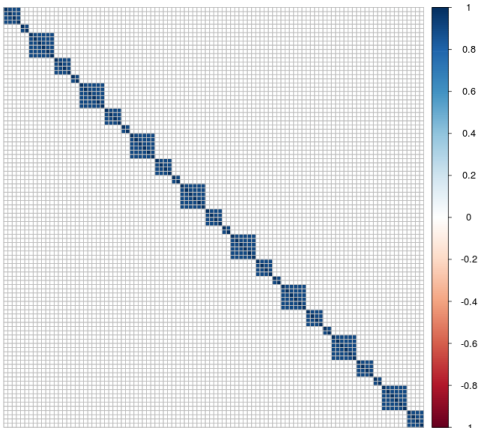

FIGURE 2 – Exemple de matrice de variance Σ_ρ utilisée pour générer les données. La corrélation au sein des groupes est $\rho = 0.9$.

la suite.

La méthode présentée (cf Section **??**, équation 1) est ensuite appliquée sur les données simulées. On compare la capacité de la méthode (CAH + GROUP-LASSO) à retrouver exactement les K vraies variables par rapport au LASSO. Pour ce dernier, les K vraies variables sont exactement retrouvées s'il existe un $\lambda > 0$ tel que seuls les K coefficients de l'estimateur $\hat{\beta}_\lambda$ correspondant aux vraies variables soient non nuls. Pour le GROUP-LASSO, on considérera comme vrai groupe, un groupe contenant une seule des K vraies variables et des variables corrélées à celle-ci. Pour ce faire, des données sont simulées suivant le design présenté ci-dessus avec un nombre total de variables $p = 250$, $K = 5$ vraies variables, $\rho = 0.7$. On fera varier le nombre d'individus n et pour chaque valeur de n, 100 échantillons sont simulés.

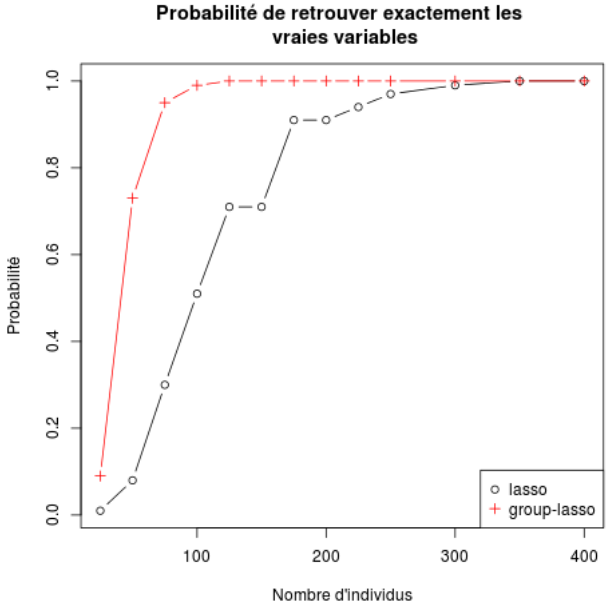

FIGURE 3 – Probabilité de trouver exactement les K vraies variables pour au moins une valeur de $\lambda > 0$ pour le lasso et la méthode proposée (à p fixé). En noir (o), le LASSO, en rouge (+) la méthode proposée (CAH + GROUP-LASSO). Probabilités calculées sur 100 échantillons simulés suivant le design présenté dans la section 3 avec $p = 250$, $\rho = 0.7$, $K = 5$.

Sur la figure 3, on voit clairement que la méthode combinant CAH et GROUP-LASSO permet de retrouver plus souvent les vraies variables que le LASSO classique.

4 Conclusion

Regrouper les variables, par exemple sur la base de corrélation, permet d'améliorer les performances de sélection en retrouvant plus souvent la vraie solution. Cela fait aussi plus sens pour des données où la redondance au sein de groupes de variables est importante. On proposera une méthode permettant de choisir le paramètre λ afin de sélectionner la solution optimale.

Remerciements

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

Références

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

[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.

[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.