Estimation type lasso de l'ensemble des variables pertinentes pour des matrices de planification de plein rang

PJC.Tardivel, D.Concordet, C.Canlet, M.Tremblay-Franco, L.Debrauwer et R.Servien

INRA-ENVT, Université de Toulouse, UMR1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France
email : patrick.tardivel@toulouse.inra.fr

Résumé. On considère le modèle linéaire Gaussien

$$Y = X\beta^* + \varepsilon,$$

avec $X \in M_{n,p}(\mathbb{R})$ une matrice de planification de plein rang quelconque (donc $n \geq p$ et X potentiellement non orthogonale). Notre objectif et d'estimer l'ensemble des variables pertinentes (ou *active set*) $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$. Les estimateurs pénalisés de type lasso sont généralement utilisés dans ce but dans le cadre de la grande dimension ($n < p$). En petite dimension ($n \geq p$), nous montrons que ces méthodes fournissent un estimateur $\hat{\mathcal{A}}^{\text{pen}}$ de $\mathcal{A}$ performant. Idéalement nous souhaiterions obtenir $\{\hat{\mathcal{A}}^{\text{pen}} = \mathcal{A}\}$ mais un tel objectif n'est envisageable qu'asymptotiquement. Non-asymptotiquement nous spécifions un paramètre de régularisation à utiliser pour l'estimateur pénalisé et nous donnons un ensemble $\mathcal{E}$ (le plus petit possible) pour que les événements $\{\hat{\mathcal{A}}^{\text{pen}} \subset \mathcal{A}\}$ et $\{\mathcal{A} \backslash \mathcal{E} \subset \hat{\mathcal{A}}^{\text{pen}}\}$ se réalisent avec une probabilité contrôlée. Enfin des arguments théoriques et des simulations montrent que cet estimateur est plus performant que l'estimateur $\hat{\mathcal{A}}^{\text{mle}}$ construit à partir du maximum de vraisemblance. Cette méthodologie sera ensuite appliquée à la détection et à la quantification de métabolites en métabolomique.

Mots-clés. lasso, lasso adaptatif, estimateur du maximum de vraisemblance, estimation des variables pertinentes

Abstract. Let us consider the Gaussian linear model

$$Y = X\beta^* + \varepsilon,$$

with X a full rank design matrix (thus $n \geq p$). Our aim is to estimate the active set $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$.

lasso type estimators are habitually used for this goal in the high-dimensional setting ($n < p$). In the small-dimensional setting ($n \geq p$), we are going to see that these methods provide a performing estimator $\hat{\mathcal{A}}^{\text{pen}}$ of $\mathcal{A}$. Ideally, we wish to obtain $\{\hat{\mathcal{A}}^{\text{pen}} = \mathcal{A}\}$ but this goal is only possible asymptotically. Non-asymptotically, we specify a tuning parameter to use and we give a set $\mathcal{E}$ (as small as possible) such that the events $\{\hat{\mathcal{A}}^{\text{pen}} \subset \mathcal{A}\}$ and $\{\mathcal{A} \backslash \mathcal{E} \subset$

$\hat{\mathcal{A}}^{\text{pen}}\}$ occur with a large probability. Finally, theoretical arguments and simulations experiments illustrate that this estimator is more performing than the estimator $\hat{\mathcal{A}}^{\text{mle}}$ build from the maximum likelihood. An application of this method will be given in metabolomics.

Keywords. lasso, adaptive lasso, maximum likelihood estimator, active set estimation

1 Estimateur type lasso de l'ensemble des variables pertinentes

On considère le modèle linéaire Gaussien

$$Y = X\beta^* + \varepsilon, \tag{1}$$

avec $X \in M_{n,p}(\mathbb{R})$ une matrice de planification de plein rang (i.e $n \geq p$) et ε un vecteur aléatoire gaussien centré de matrice de variance Γ inversible. Notre objectif et d'estimer l'ensemble des variables pertinentes $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$.

L'estimation de cet ensemble pourrait se faire avec des méthodes classiques. Une alternative à ces méthodes est l'utilisation d'estimateurs pénalisés comme le lasso introduit par Tibshirani (1996) ou le lasso adaptatif défini par Zou (2006). Ces estimateurs sont définis par

$$\hat{\beta}^{\text{pen}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \text{pen}(\beta) \right\},$$

avec $\text{pen}(\beta) = \|\beta\|_1$ pour le lasso et $\text{pen}(\beta) = \sum_{\text{i}=1}^{\text{p}} \frac{1}{|\hat{\beta}_{\text{i}}|} |\beta_{\text{i}}|$, où $\hat{\beta}$ est un estimateur de β^* pour le lasso adaptatif.

En grande dimension (i.e $n < p$), ces estimateurs sont notamment utilisés par Meinshausen et Yu (2009) et Wasserman et Roeder (2009) pour estimer l'ensemble $\mathcal{A}$. Cependant, deux contraintes rendent cette estimation délicate. Pour le lasso, Meinshausen et Bühlmann (2006), Zou (2006) et Zhao et Yu (2006) montrent que la condition d'irreprésentabilité sur la matrice de planification X est nécessaire et suffisante pour avoir un estimateur convergent de $\mathcal{A}$. Géométriquement, cette condition implique qu'une colonne quelconque X_i de la matrice X avec $i \notin \mathcal{A}$ est quasiment orthogonale à l'espace vectoriel engendré par la famille $\{X_j\}_{j \in \mathcal{A}}$. Le lasso adaptatif nécessite quant à lui un estimateur $\hat{\beta}$ convergent pour le paramètre β^*. En petite dimension ces deux contraintes ne sont pas restrictives. En effet, il est possible d'appliquer une transformation linéaire U à chacun des membres du modèle (1) de telle sorte que la condition d'irreprésentabilité soit vérifiée pour la matrice UX. Par ailleurs, l'estimateur des moindres carrés ordinaire est un estimateur convergent. Ainsi, en petite dimension, le lasso et le lasso adaptatif fournissent un estimateur $\hat{\mathcal{A}}^{\text{pen}}$ convergeant vers $\mathcal{A}$. Cependant, les résultats asymptotiques sont difficile à utiliser, puisqu'ils préconisent l'usage d'un paramètre de régularisation λ_n

pour lequel on connaît uniquement l'ordre de grandeur (par exemple $\lambda_n/\sqrt{n} \to \infty$ et $\lambda_n/n \to 0$). L'obtention de résultats non asymptotiques pour l'estimation de l'ensemble des variables pertinentes via un estimateur de type lasso demeure à notre connaissance un sujet inexploré. Dans cette communication, nous proposons une telle étude. Les résultats obtenus seront appliqués à des données de métabolomique.

2 Estimation non asymptotique

Nous souhaitons tout d'abord garantir que l'événement $\{\hat{\mathcal{A}}^{\text{pen}} \subset \mathcal{A}\}$ se réalise. Pour cela nous démontrons que le choix du paramètre de régularisation λ peut être fait pour que cet événement ait une probabilité contrôlée. Néanmoins, lorsque le paramètre β_i^* est trop petit l'élément i est difficilement détectable (i.e la probabilité de l'événement $\{i \in \hat{\mathcal{A}}^{\text{pen}}\}$ est faible). Pour pallier cette difficulté nous donnons un ensemble $\mathcal{E} = \{i \in \mathcal{A} \mid \beta_i^* \leq c_i\}$ tel que $\mathcal{A} \setminus \mathcal{E}$ soit au moins détecté par l'estimateur $\hat{\mathcal{A}}^{\text{pen}}$ (i.e la probabilité de l'événement $\{\mathcal{A} \setminus \mathcal{E} \subset \hat{\mathcal{A}}\}$ soit grande).

Lorsque la matrice de planification a des colonnes orthogonales (i.e $X^T X$ est diagonale), l'estimateur lasso à une expression explicite données notamment par Bühlmann et Van de Geer (2011). Ainsi, il est possible de choisir le paramètre de régularisation λ et des seuils de détection $(c_i)_{1 \leq i \leq p}$ permettant de contrôler respectivement la probabilité des événements $\{\hat{\mathcal{A}}^{\text{pen}} \subset \mathcal{A}\}$ et $\{\mathcal{A} \setminus \mathcal{E} \subset \hat{\mathcal{A}}^{\text{pen}}\}$. Lorsque X est une matrice de plein rang quelconque, il est possible de se ramener au cas précédent en appliquant à chacun des membres du modèle (1) une transformation linéaire U qui orthogonalise la matrice de planification X (i.e $(UX)^T UX$ diagonale). Chaque transformation linéaire U fournit des seuils $(c_i(U))_{1 \leq i \leq p}$. Logiquement, plus ces seuils sont bas, plus le cardinal de l'ensemble $\mathcal{E}$ associé à ces seuils est petit. Ainsi, il y a un enjeu à déterminer une transformation linéaire U^* pour laquelle les seuils $(c_i(U^*))_{1 \leq i \leq p}$ sont les plus petits possible (pour une norme donnée). Nous montrons que ces seuils sont liés à la variance de l'estimateur des moindres carrés ordinaire $\hat{\beta}^{\text{mco}}(U)$ du modèle

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon}, \text{ avec } \tilde{Y} = UY, \tilde{X} = UX \text{ et } \tilde{\varepsilon} = U\varepsilon.$$

Plus précisément, la transformation linéaire U^* doit être cherchée dans l'ensemble des transformations linéaires U_δ qui orthogonalisent X et pour lesquelles l'estimateur $\hat{\beta}^{\text{mco}}(U_\delta)$ est efficace.

3 Comparaison théorique et simulations

Un estimateur de l'ensemble $\mathcal{A}$ peut également être obtenu grâce à l'estimateur du maximum de vraisemblance $\hat{\beta}^{\text{mle}}$. Cet estimateur permet de tester les hypothèses $\mathcal{H}_i : \beta_i^* = 0$ et $\hat{\mathcal{A}}^{\text{mle}}$ est l'ensemble des hypothèses rejetées. Une procédure de Bonferonni permet de

contrôler la probabilité de l'événement $\{\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}\}$. Nous montrons que dans le cas particulier où les composantes de l'estimateur $\hat{\beta}^{\text{mle}}$ sont i.i.d, les deux estimateurs $\hat{\mathcal{A}}^{\text{lasso}}$ et $\hat{\mathcal{A}}^{\text{mle}}$ ont même loi. Par contre, lorsque les composantes $\hat{\beta}^{\text{mle}}$ sont corrélées, l'estimateur $\hat{\mathcal{A}}^{\text{lasso}}$ est plus performant notamment lorsque le nombre p de composantes est grand et que les corrélations sont fortes.

4 Application en métabolomique

La métabolomique est une science qui s'intéresse à la caractérisation et la quantification de métabolites, ces petites molécules que l'on retrouve dans les cellules, les tissus, les fluides biologiques et les organismes. La technique la plus utilisée pour obtenir cette caractérisation est la résonance magnétique nucléaire des protons (RMN). Afin d'identifier ces métabolites, les experts utilisent une bibliothèque personnelle qui contient les spectres des métabolites purs et comparent ces spectres à la main à celui du mélange biologique à analyser. Plus précisément, lorsqu'un expert veut savoir si un métabolite particulier est présent dans un mélange, il vérifie si tous les pics du spectre de ce métabolite se retrouvent dans le spectre du mélange. Cette méthode dépend donc grandement des connaissances de l'expert, notamment du nombre de spectres de métabolites qu'il connaît. Cette identification peut également être rendue délicate par le chevauchement de certains des pics des métabolites présents dans le mélange. Étudions un exemple d'identification de métabolite dans un spectre avec la Figure 1.

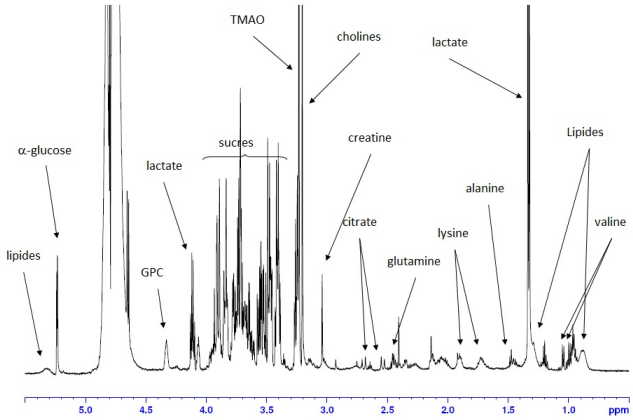

Figure 1: On remarque par exemple que certains pics associés aux lipides se confondent avec ceux associés à la valine et au lactate. Ceci rend la détection de ce métabolite difficile.

Certaines méthodes automatiques d'identification et de quantification des métabolites ont été proposées récemment mais elles restent perfectibles. En effet, MetaboHunter

(Tulpan, 2011) est très rapide mais ne permet pas gérer le chevauchement des pics et donc la compétition entre métabolites. De plus, elle ne fournit qu'un score de présence d'un métabolite lié au nombre de ses pics qui ont été identifiés dans le mélange. D'un autre côté le fort coût computationnel de BATMAN (Hao, 2012) ne permet pas de rechercher des dizaines de métabolites. De plus ces méthodes n'ont pas de performance statistique mesurée en terme de détection de métabolites. Par conséquent, il n'existe pas encore une méthode de référence dans ce domaine.

A partir d'une bibliothèque de métabolites de référence, nous modélisons ce problème de détection et de quantification de métabolites dans un mélange complexe comme un problème de détection de variables pertinentes (les métabolites) en petite dimension. En utilisant notre méthode sur des mélanges complexes, nous démontrons qu'elle permet de détecter et de quantifier les métabolites dans un temps raisonnable.

Remerciements

Ce projet bénéficie du soutien financier du Ministère de l'Écologie, du Développement Durable et de l'Énergie dans le cadre du programme national de recherche Risk'OGM ainsi que de l'IDEX Toulouse "Transversalité 2014".

Bibliographie

[1] Bühlmann, P. et Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
[2] Hao, J., Astle, W., De Iorio, M. and Ebbels, T. (2012). BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15), 2088-2090.
[3] Meinshausen, N. et Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436-1462.
[4] Meinshausen, N. et Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 246-270.
[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
[6] Tulpan, D., Léger, S., Belliveau, L., Culf, A. and Čuperlović-Culf, M. (2011). MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1),400.
[7] Wasserman, L. et Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A), 2178.
[8] Zhao, P. et Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7, 2541-2563.

[9] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.