Stabilité de la sélection de variables pour la classification de données en grande dimension

Emeline Perthame, Chloé Friguet, David Causeur

Stabilité de la sélection de variables pour la classification de données en grande dimension

Emeline Perthame [1] & Chloé Friguet [2] & David Causeur [1]

[1] *Institut de Recherche Mathématique de Rennes (IRMAR) - CNRS UMR 6625*
Agrocampus Ouest - Applied Mathematics Department,
65 rue de Saint Brieuc 35 042 Rennes Cedex, France
E-mail : perthame@agrocampus-ouest.fr & causeur@agrocampus-ouest.fr
[2] *Laboratoire de Mathématiques de Bretagne-Atlantique (LMBA) - CNRS UMR 6205*
Université de Bretagne-Sud, Campus de Tohannic, 56 000 Vannes, France
E-mail : chloe.friguet@univ-ubs.fr

Résumé. Les données à haut-débit ont motivé le développement de méthodes statistiques pour la sélection de variables. Ces données sont caractérisées par leur grande dimension et par leur hétérogénéité car le signal est souvent observé simultanément à plusieurs facteurs de confusion. Les approches habituelles sont ainsi remises en question car elles peuvent conduire à des décisions erronées. Efron (2007), Leek and Storey (2007, 2008), Friguet et al (2009) montrent l'impact négatif de l'hétérogénéité des données sur le nombre de faux-positifs des tests multiples.

La sélection de variables est une étape importante de la construction d'un modèle de classification en grande dimension car elle réduit la dimension du problème aux variables les plus prédictives. On s'intéresse ici aux performances de classification de la sélection de variables, via la procédure LASSO (Tibshirani (1996)) et à la reproductibilité des ensembles de variables sélectionnés. Des simulations montrent que l'ensemble des variables sélectionnées par le LASSO n'est pas celui des meilleurs prédicteurs théoriques. Aussi, d'intéressantes performances de classification ne sont atteintes que pour un grand nombre de variables sélectionnées.

Notre méthode s'appuie sur la description de la dépendance entre covariables grâce à un petit nombre de variables latentes (Friguet et al. (2009)). La stratégie proposée consiste à appliquer les procédures sur les données conditionnellement à cette structure de dépendance. Cette stratégie permet de stabiliser les variables sélectionnées : d'intéressantes performances de classification sont atteintes pour de plus petits ensembles de variables et les variables les plus prédictives sont détectées.

Mots-clés. Sélection de variables, grande dimension, stabilité

Abstract.

The analysis of high throughput data has renewed the statistical methodology for multiple testing and feature selection in regression issues. Such data are both characterized by their high dimension and their heterogeneity, as the true signal and several confusing

factors are often observed at the same time. In such a framework, the usual statistical approaches are questioned and can lead to misleading decisions. Some papers (Efron 2007, Leek and Storey 2007 and 2008, Friguet et al. 2009) have focused on the negative impact of data heterogeneity on the consistency of the ranking resulting from multiple testing procedures.

This presentation aims at showing that data heterogeneity also affects the stability of supervised classification variable selection which is often used to identify relevant subsets of features. Key characteristics of selection methods are both classification performance and reproducibility of the selected subsets. It is first shown through a simulation study that selected subsets using usual procedures are subject to a high variability. Simulation studies show that most usual methods such as LASSO (Tibshirani, 1996) does not select theoretical best predictors and that interesting performances of classification are performed only when a high number of variables are selected.

As suggested in Friguet et al. (2009), a supervised factor model is proposed to identify a low-dimensional linear kernel which captures data dependence. The deduced strategy is finally shown to improve stability of the usual methods. Indeed, interesting performances of classification are reached for a smaller number of selected variables and predictive variables are more often selected.

Keywords. Variable selection, high dimension, stability

1 Sélection de variables en grande dimension

Le contexte est celui de la classification d'individus en deux groupes. On note X_j la j-ième variable explicative, $j \in \{1 \dots m\}$, et $X = [X, \dots, X_m]$ la matrice constituée des m variables explicatives. La variable réponse est une variable binaire prenant la valeur 1 (resp. -1) avec une probabilité p_1 (resp. $p_0 = 1 - p_1$). Sachant l'appartenance au groupe 1 (resp. -1), on considère que le vecteur X est distribué selon une loi normale de moyenne μ_1 (resp. μ_0) et de matrice de variance-covariance Σ :

$$X = \mu_i + e, i = 1 \text{ si } Y = 1 \text{ et } -1 \text{ sinon} \tag{1}$$

où e est le terme d'erreur $e \sim \mathcal{N}(0, \Sigma)$. Dans ce contexte, on sait que la régression logistique admet une relation linéaire entre le log-ratio $LR(x)$ des probabilités postérieures de groupe et les covariables :

$$LR(x) = \log \frac{P(Y = +1|X = x)}{P(Y = -1|X = x)} = \beta_0 + x'\beta. \tag{2}$$

Le classifieur de Bayes ($\beta_0^* = \log \frac{p_1}{p_0} - \frac{1}{2}(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0), \beta^* = \Sigma^{-1}(\mu_1 - \mu_0)$) découle de la minimisation de l'erreur de mauvais classement théorique $P(\hat{Y} \neq Y)$. On s'intéresse à cette règle de classification théorique afin d'identifier les variables les plus

prédictives, et d'évaluer les procédures de sélection de variables dans le cadre d'une étude par simulations.
En pratique, l'analyse discriminante de Fisher, par minimisation du critère des moindres carrés et la régression logistique, par maximisation de la vraisemblance du modèle sont deux méthodes bien connues d'estimation du coefficient de pente de cette règle de classification. Cependant, ces deux approches demandent chacune d'inverser la matrice $S = \hat{\Sigma}$, matrice non inversible en grande dimension. La sélection de variables permet dans ce contexte de diminuer la dimension du problème en le réduisant au sous-ensemble des variables les plus prédictives.

Des méthodes parcimonieuses permettent théoriquement d'atteindre de bonnes performances de classification et d'assurer la sélection d'un sous-ensemble de variables inclus dans l'ensemble des prédicteurs théoriques. Or, en grande dimension, l'hétérogénéité des données induit de la dépendance entre les variables. En effet, dans le domaine de l'analyse de données d'expression de gènes ou de génotypage à grande échelle (SNP), il est devenu systématique pour de nombreux auteurs de rechercher des facteurs latents conditionnellement auxquels les variables sont indépendantes (ou moins dépendantes). Celà revient à dire que, non-conditionnellement, la dépendance des variables est structurée par ces facteurs. Ces facteurs latents s'identifient souvent à des effets nuisibles, non-contrôlés dans le plan d'expérience et pour lesquels John Storey et Jeff Leek (2007 et 2008) ont été les premiers à parler de "facteurs d'hétérogénéité ". Ainsi, la présence de cette dépendance affecte les procédures classiques de sélection de modèle. On mettra en évidence les propriétés de stabilité en grande dimension des procédures de sélection de variables sur une méthode usuelle, la sélection LASSO (Tibshirani (1996)).

La Figure 1 montre les valeurs que prend la pente du classifieur de Bayes $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$ dans une configuration de 1000 variables où la moyenne μ_0 est le vecteur nul (groupe $Y = -1$) et la moyenne μ_1 (groupe $Y = 1$) est nulle sauf sur 50 composantes pour lesquelles la même valeur prédictive a été introduite. Ainsi, 50 variables discriminent les deux groupes d'individus et on souhaite détecter le sous-ensemble des variables les plus prédictives. La matrice de covariance est générée selon un modèle en facteurs et représente un cas de dépendance élevée. Sous cette configuration, on peut identifier le couple (resp. triplet) de variables menant à une probabilité de mauvais classement théorique minimale : ces variables sont colorées en bleu (resp. rouge). En simulant 1000 jeux de données selon cette configuration, on s'attend à ce que le LASSO sélectionne régulièrement des sous-ensembles de variables contenant ce meilleur couple (resp. triplet). On verra dans cet exposé que les sous-ensembles sélectionnés ne sont en fait pas stables d'un jeu de données à l'autre.

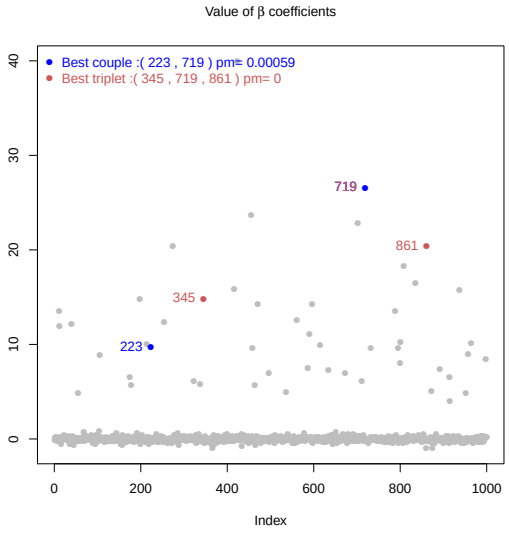

Figure 1: Exemple de valeurs du coefficient de pente du classifieur de Bayes et identification du meilleur couple (en bleu) et triplet (en rouge) théoriques. Leur probabilité de mauvais classement théorique est notée pm

2 Stabilisation de la sélection de variables en grande dimension

2.1 Modélisation de la structure de dépendance

En pratique, le vrai signal et des facteurs non observables sont parfois enregistrés simultanément, introduisant de la dépendance entre les covariables. Afin de capturer cette hétérogénéité, on considère un modèle en facteurs pour les covariables. Ce modèle permet de modéliser la structure de dépendance des covariables en décomposant le modèle (1) en un effet fixe, des effets non-observables et un bruit blanc :

$$X = \mu_i + ZB' + \varepsilon. \tag{3}$$

Les variables latentes $Z = [Z_1, \ldots, Z_q]$ capturent la dépendance dans un espace de dimension $q \ll m$. Les composantes ε_j du terme d'erreur ε suivent une loi normale centrée, de variance ψ_j et sont indépendantes (Leek and Storey (2008)).

Le modèle (3) est équivalent au modèle (1), mais pour lequel la matrice Σ admet une décomposition de la forme $\Sigma = BB' + \Psi$, où $\Psi = diag(\psi_1, \ldots, \psi_m)$ est une matrice diagonale représentant la variance spécifique et BB' représente la variance commune aux données. Friguet et al (2009) propose un algorithme EM pour estimer les paramètres Ψ, B et extraire les facteurs dans un contexte de grande dimension.

2.2 Méthode proposée

La méthode proposée consiste à retrancher l'effet aléatoire aux données et à considérer les données ajustées:

$$X_{FA} = X - ZB'. \tag{4}$$

Par le même raisonnement que celui menant au classique classifieur de Bayes, on définit une règle de classification en travaillant conditionnellement aux facteurs Z. On montre (Friguet et al, (2013)) sous cette hypothèse d'indépendance conditionnelle aux facteurs latents, que ce nouveau classifieur a de meilleures propriétés que le classifieur de Bayes (non conditionnel), en terme de probabilité de mauvais classement et de stabilité en sélection de variables. L'intérêt de l'approche conditionnelle est donc d'obtenir, sous une hypothèse supplémentaire sur la structure de Σ, un classifieur théoriquement plus performant que la règle de classification de Bayes usuelle.

3 Étude par simulations

Les résultats de la méthode proposée sont illustrés par une étude sur simulations dans laquelle six scénarios de dépendance sont considérés. Les covariables sont distribuées selon une loi $\mathcal{N}(\mu_i, \Sigma)$ où μ_1 et μ_0 sont construits selon l'exemple de la Figure 1. A la manière de l'étude par simulations menée par Meinhausen and Bülhmann (2010), la matrice Σ diffère selon le scénario de dépendance envisagé: matrice identité, matrice composée de 5 blocs indépendants simulée selon la méthode proposée par Langfelder and Horvath (2008), matrice de Toeplitz type AR(1) ou matrice générée selon un modèle en facteurs représentant des situations générales de faible, moyenne et forte dépendance. Les résultats obtenus sur ces simulations montrent que, pour l'ensemble de ces structures, l'ajustement sur les facteurs stabilise le sous-ensemble de variables identifiées par la procédure LASSO (voir Figure 2) et améliore les performances moyennes de classification des modèles sélectionnés.

4 Conclusion

La sélection de variables est une étape importante lors de la construction d'un modèle de classification en grande dimension car elle permet de réduire la dimension du problème à l'ensemble des variables les plus prédictives. Des méthodes comme la procédure LASSO supposent une faible structure de corrélation mais les données de grande dimension vérifient rarement ce postulat. Nous proposons un cadre général de la prise en compte de la dépendance, basé sur la modélisation de la structure de corrélation par un modèle en facteurs. La stratégie de sélection de variables consiste ensuite à appliquer les procédures usuelles sur les données conditionnellement à cette structure de dépendance. Cette stratégie permet de stabiliser l'ensemble des variables sélectionnées. En effet, d'intéressantes

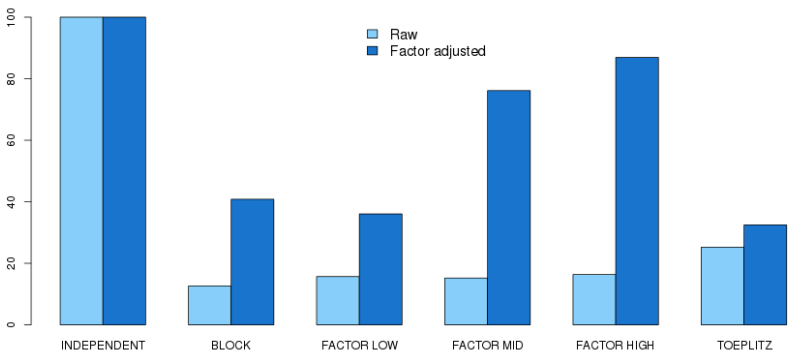

Figure 2: Proportions de sélection des meilleurs prédicteurs théoriques pour le LASSO appliqué aux données brutes (Raw) et aux données facteur-ajustées (Factor adjusted)

performances de classification sont atteintes pour de plus petits ensembles de variables sélectionnées et les variables les plus prédictives sont plus souvent détectées. La stratégie proposée n'est pas une nouvelle méthode de sélection de variables. En ajustant les données sur les facteurs, on propose de replacer les données dans un contexte favorable au LASSO.

Bibliographie

[1] Efron B. (2007), Correlation and large-scale simultaneous testing, *Journal of the American Statistical Association*, 102:93-103.
[2] Friguet C., Kloareg M. and Causeur D. (2009), A factor model approach to multiple testing under dependence, *Journal of the American Statistical Association*, 104:488:1406-1415.
[3] Friguet C., Perthame E. and Causeur D. (2013), Stability of variable selection for high-dimensional data, *Journal de la Société Française de Statistique*, en révision.
[4] Langfelder P. and Horvath S. (2008), WGCNA : an R package for weighted correlation network analysis, *BMC Bioinformatics*, 9:559.
[5] Leek J.T. and Storey J. (2007), Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genetics*, 3(9):e161.
[6] Leek J.T. and Storey J. (2008), A general framework for multiple testing dependence, *Proceedings of the National Academy of Sciences*, 105:18718-18723.
[7] Meinshausen N. and Bülhmann P. (2010), Stability selection, *Journal of the Royal Statistical Society, serie B*, 72.
[8] Tibshirani R. (1996), Regression shrinkage and selection via lasso, Journal of the Royal Statistical Society, serie B, 58:267-288.