Compromis entre risque prédictif et taux de fausses découvertes pour la régression linéaire Gaussienne en grande dimension.

Perrine Lacroix [1,2]

[1] *Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, Rue Michel Magat Bâtiment 307, 91405 Orsay, France.*
[2] *Université Paris-Saclay, CNRS, INRAE, Institut des Sciences des Plantes de Paris-Saclay (IPS2), Rue Noetzlin Bâtiment 630, 91190 Gif-sur-Yvette, France.*
perrine.lacroix@universite-paris-saclay.fr

Résumé. Dans un contexte de grande dimension, une approche classique pour estimer le paramètre inconnu en régression linéaire gaussienne est de minimiser les moindres carrés pénalisés. Pour obtenir une inégalité oracle sur le risque prédictif, la théorie développée par Birgé et Massart (2001) fournit des formes de pénalité connue à constante multiplicative près. Cette constante est actuellement fixée à 2 via des considérations d'optimalité asymptotique sur le risque. Dans cet exposé, je définirai la notion de variables actives et inactives et j'expliquerai que contrôler la prédiction n'est pas suffisant pour limiter la sélection de variables inactives. Pour pallier à ce problème, notre idée a été de rajouter une contrainte supplémentaire qui est le contrôle du taux de fausses découvertes (false discovery rate (FDR)) sur la procédure de sélection de modèle. Pour cela, notre approche consiste à modifier la constante multiplicative et étudier l'impact de cette variation sur le FDR et le risque prédictif, ceci d'un point de vue théorique (sous un modèle très simplifié) et expérimental.
Mots-clés. Grande dimension, sélection de modèles, prédiction, taux de faux positifs.

1 Cadre

1.1 Le modèle Statistique

Nous considérons le modèle de régression linéaire Gaussien suivant :

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

La variable réponse Y, constituée de n observations, est régressée par p variables, chacune de taille n, notées $X_1, \cdots, X_p$ et résumée dans la matrice de design fixe X de taille $n \times p$. Le bruit ε suit une distribution gaussienne centrée et de variance σ^2. Le paramètre inconnu β^* est de taille p. Si $\beta^*_j \neq 0$, alors la variable associée X_j est impliquée dans la

relation linéaire. Elle est alors appelée variable active. En opposition, les coefficients nuls de β^* correspondent aux variables non impliquées, appelées variables inactives. L'objectif est d'identifier l'ensemble des variables actives, ceci à travers l'estimation de β^*. Nous nous intéressons au cadre de grande dimension, où la dimension p est grande mais pas excessivement pour éviter le cadre de la ultra haute dimension (voir Verzelen (2012)). De plus, p peut être du même ordre de grandeur que n, voire le dépasser. Dans ce contexte de grande dimension, les outils statistiques classiques basés sur une estimation empirique ne parviennent pas à retrouver les variables actives. Des hypothèses de régularité doivent être ajoutées au modèle. Ici, nous choisissons la parcimonie, c'est-à-dire que nous supposons que seules quelques variables sont actives par rapport à p.

1.2 L'application biologique

Ces travaux ont émergé d'une problématique biologique. Dans un organisme, des interactions entre certains gènes sont nécessaires pour démarrer la transcription d'un gène cible. Les gènes candidats pour former ce groupe d'acteurs du processus sont les facteurs de transcription (FT). Déterminer ces interactions pour tous les gènes de l'organisme permettrait de réguler leur expression, par exemple pour le rendre plus résistant à certains stress. A cause du coût expérimental et du nombre de gènes, tester une à une les interactions entre un FT et un gène cible est impossible à la paillasse. Nous utilisons les statistiques pour proposer des interactions candidates, ceci en résolvant des régressions linéaires Gaussiennes. Ainsi, Y modélise le vecteur d'expression d'un gène cible et X la matrice d'expression des p FT. Les coefficients non nuls de $\hat{\beta}$ donneront la liste des FT candidats du gène cible. Nous travaillons sur *Arabidopsis thaliana*, plante modèle chez les biologistes et pour laquelle nous disposons d'un jeu de données, déjà pré-traité, de 19844 gènes dont $p = 1935$ FT et de $n = 2215$ données d'expression indépendantes pour chaque gène.

2 Idées clés

2.1 Les travaux antérieurs

Actuellement, la plupart des procédures pour la régression linéaire Gaussienne en grande dimension peuvent être classées en deux groupes : celles qui contrôlent le risque prédictif (RP) et celles qui contrôlent le taux de fausses découvertes (FDR). Le RP est la perte L_2 moyenne de la différence entre Y et $X\hat{\beta}$, pour $\hat{\beta}$ un estimateur de β^*. Le FDR est la proportion moyenne de variables inactives parmi les variables sélectionnées. Ces deux groupes donnent des ensembles de variables différents. Pour un contrôle du RP, les variables sélectionnées sont prédictives. Cependant, de nombreuses variables prédictives sont disponibles pour construire un modèle de régression. Par conséquent, l'ensemble des variables prédictives peut contenir des variables inactives. Au contraire, pour un contrôle

du FDR, l'accent est seulement mis sur les variables sélectionnées qui sont garanties être actives. Par conséquent, des variables actives peuvent ne pas être sélectionnées. Idéalement, nous voulons sélectionner toutes et uniquement les variables actives. Dans cette idée, nous proposons d'étudier la méthode de sélection de modèle, connue pour un contrôle optimal du RP, dans l'objectif d'obtenir une information supplémentaire pertinente sur le FDR. L'idée étant de pouvoir contrôler les deux métriques simultanément.

2.2 La sélection de modèles

Les méthodes de sélection de modèle sont basées sur trois étapes. La première consiste à générer une collection $\mathcal{M}$ contenant quelques sous-ensembles de variables de tailles variées. Cette étape évite l'exploration de tous les sous-ensembles de variables possible parmi les p variables et dont le nombre grandit exponentiellement avec p. Les variables finales sélectionnées seront celles d'un des sous-ensembles de la collection. La seconde étape consiste simplement en la minimisation des moindres carrés à l'intérieur de chaque sous-ensemble afin d'obtenir une collection d'estimateurs $\hat{\beta}_m := \underset{\beta \in m}{\arg\min}\{||Y - X\beta||_2^2\}$. Enfin, la dernière étape consiste à sélectionner $\hat{\beta}_{\hat{m}}$ par la résolution de l'équation :

$$\hat{m} = \underset{m \in \mathcal{M}}{\arg\min}\Big\{||Y - X\hat{\beta}_m||_2^2 + \text{pen}(D_m)\Big\}, \tag{2}$$

où D_m désigne la dimension du modèle m et pen est une fonction de pénalité dépendant de D_m. Birgé et Massart (2007) proposent des formes de pénalité en fonction de la richesse de la collection de telle sorte qu'un contrôle optimal non-asymptotique sur le RP est obtenu. Ceci donne plusieurs fonctions de pénalité telles que $\text{pen}(D_m) = 2\sigma^2 D_m$ pour une collection fixe de modèles emboîtés (voir deLeeuw (1992)) ou la pénalité LinSelect (voir Giraud, Huet et Verzelen (2012)) et les pénalités dépendantes aux données (voir Baudry, Maugis et Michel (2012)) pour une collection de modèles aléatoire.

2.3 L'idée clé

Simplifions dans un premier temps le problème en supposant que la variance est connue et fixée à 1 et que la collection de modèle est donnée par :

$$\mathcal{M} = \Big\{m_0 = \emptyset, m_1 = \text{Vect}(X_1), m_2 = \text{Vect}(X_1, X_2), \cdots, m_q = \text{Vect}(X_1, X_2, \cdots, X_q)\Big\}. \tag{3}$$

où $q = \min(n, p)$. Cette collection fixe de modèles emboîtés peut être utilisée en pratique si un ordre d'importance au sens de la régression (1) sur les X_j est connu. Supposons enfin que $m^* = \text{Supp}(\beta^*)$ appartient à $\mathcal{M}$. Puisque la taille de la collection (3) est linéaire en p, un choix approprié de pénalité pour obtenir une qualité prédictive optimale est $\text{pen}(D_m) = KD_m$ avec $K > 1$ (d'après Birgé et Massart (2007)). Le choix du paramètre

libre K a été longuement discuté et $K = 2$ est utilisée en pratique car cette constante apparaît lors de considérations asymptotiques. Notons $\hat{m}(K)$ le modèle sélectionné dans (2) avec $\text{pen}(D_m) = KD_m$. En nous basant sur des simulations, nous observons que le RP associé à $\hat{m}(2)$ est petit, cependant ce n'est pas le cas pour le FDR. Ceci suggère que certaines variables sélectionnées sont inactives. Rendre la procédure plus conservative est une idée pour éviter leur sélection et une solution est d'augmenter la pénalisation. Dans $\text{pen}(D_m) = KD_m$, le seul paramètre libre est K que nous proposons d'augmenter. Nous avons en effet observer que certaines valeurs de $K > 2$ permettent de conserver les performances prédictives obtenues avec $K = 2$ tout en baissant drastiquement la valeur du FDR. Ainsi, un choix de K peut être guidé par l'étude simultanée des deux fonctions : $\Big[K \longrightarrow \text{RP}\big(\hat{m}(K)\big)\Big]$ et $\Big[K \longrightarrow \text{FDR}\big(\hat{m}(K)\big)\Big]$.

3 Compromis entre RP et FDR

3.1 Un contrôle théorique du FDR

Bien que la procédure de sélection de modèle a été construite pour un contrôle du RP, nous proposons une borne inférieure et une borne supérieure non-asymptotiques de la fonction $\big[K \longrightarrow \text{FDR}(\hat{m}(K))\big]$. Ces bornes sont satisfaisantes car elles ne demandent plus d'estimation, contrairement au FDR.

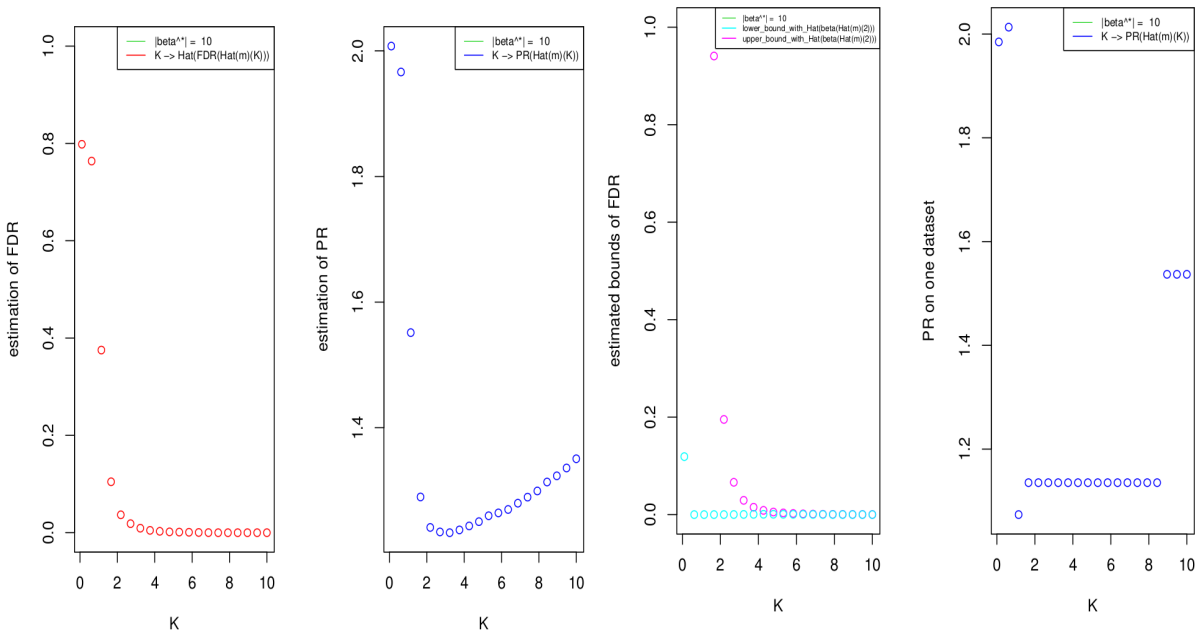
3.2 Le choix de K

Les bornes obtenues précédemment dépendent du paramètre β^*, inconnu en pratique. Nous proposons de le remplacer par l'un de ses estimateurs. Ainsi, à partir d'un jeu de données, nous proposons d'étudier les bornes supérieures et inférieures évaluées en un estimateur bien choisi de β^* simultanément avec l'estimation de la fonction $\big[K \longrightarrow \text{PR}(\hat{m}(K))\big]$, obtenue en séparant classiquement le jeu de données en deux : le jeu d'entraînement sur lequel $\hat{m}(K)$ est calculé pour chaque K et un jeu test sur lequel la métrique est évaluée.

Nous donnons un exemple à la figure (1). A partir des courbes de droite, c'est-à-dire celles disponibles en pratique, nous suggérons de choisir K le plus petit possible de telle sorte que le RP estimé soit aussi petit que celui associé à $K = 2$ et que la borne supérieure estimée du FDR soit non nulle et assez petite (par exemple < 0.05, raisonnable pour un contrôle du FDR). Ce choix de K correspond à une petite valeur des RP et FDR empiriques, fonctions que l'on peut considérer comme la vérité.

Bibliographie

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle, *Breakthroughs in statistics*, pp. 599–609.



Estimation empirique

Évaluation à partir d'un jeu de données

FIGURE 1 – A gauche : Courbes des estimations empiriques $\Big[K \longrightarrow \mathrm{FDR}\big(\hat{m}(K)\big)\Big]$ et $\Big[K \longrightarrow \mathrm{PR}\big(\hat{m}(K)\big)\Big]$. A droite : courbes des bornes estimées du FDR et de la courbe estimée du PR

Birgé, L. et Massart, P. (2001). Gaussian model selection, *Journal of the European Mathematical Society*, 3, pp. 203–268.

Birgé, L. et Massart, P. (2007). Minimal penalties for Gaussian model selection, *Probability theory and related fields*, 138, pp. 33–73.

Verzelen, N. (2012). Minimax risks for sparse regressions : Ultra-high dimensional phenomenons, *Electronic Journal of Statistics*, 6, pp. 38–90.

Giraud, C., Huet, S. et Verzelen, N. (2012). High-dimensional regression with unknown variance, *Statistical Science*, 27, pp. 500–518.

Baudry, J.P., Maugis, C. et Michel, B. (2012). Slope heuristics : overview and implementation, *Statistics and Computing*, 22, pp. 445–470.