Power of TLT test build from lasso statistics

Stéphane Mourareau [1], Jean-Marc Azaïs [1] & Yohan de Castro [2]

[1] *Institut de Mathématiques de Toulouse (IMT), Université Paul Sabatier, Toulouse.*
[2] *Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, Orsay.*
stephane.mourareau@math.univ-toulouse.fr
jean-marc.azais@math.univ-toulouse.fr
yohan.decastro@math.u-psud.fr

Résumé. Dans des travaux récents, Taylor, Lockhart et Tibshirani ont proposé une nouvelle statistique de test pour le problème général de détection de signal en utilisant les propriétés de l'algorithme LARS (Least-Angle Regression). Sous l'hypothèse nulle, ils donnent une distribution exacte pour leur statistique de test et ce en dimension quelconque. A notre connaissance, aucun résultat n'a encore été démontré concernant son comportement sous l'alternative. Dans ce papier, nous prouvons que ce test est bien sans biais. De plus, nous comparons son efficacité à celle du test d'adéquation du χ^2 dans de nombreux cas.

Mots-clés. Minimisation l_1, LASSO, test de nullité, puissance.

Abstract. New test statistic for signal detection with exact distribution in finite dimension and using Least-Angle Regression (LARS) have been proposed by Taylor, Lockhart and Tibshirani. To the best of our knowledge, no results have been shown regarding the distribution of their statistic under the alternative. For the first time, this paper investigates the power of this test. In particular, we prove that their test is unbiased. Furthermore, we compare the power of their test to the power of the chi-square test.

Keywords. l_1 minimization, LASSO, global null test, power.

1 Introduction

Dans cet exposé, nous considérons le problème classique de la régression linéaire. On se donne un vecteur de sortie $Y \in \mathbb{R}^n$ et une matrice de prédicteurs (dite matrice de design) $X \in \mathbb{R}^{n \times p}$ telle que

$$Y = X\beta^\star + \varepsilon \quad \text{avec} \quad \varepsilon \sim \mathcal{N}(0, \Sigma)\,,$$

et on cherche à déterminer $\beta^\star \in \mathbb{R}^p$. On définit l'estimateur Lasso par

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

où $\lambda > 0$ est un paramètre d'ajustement qui gouverne le niveau de sparsité de $\hat{\beta}(\lambda)$. Une question importante dans l'estimation Lasso concerne le nombre d'éléments non nuls de $\beta^\star$. Afin de réaliser un test de nullité globale

$$\mathrm{H}_0 : \beta^\star = 0 \quad \text{against} \quad \mathrm{H}_1 : \beta^\star \neq 0\,.$$

Taylor and al. [10] définissent, dans un contexte plus général de problèmes de pénalisation, la statistique de test suivante

$$S := \frac{\Phi(\nu_{i*}^- / R_{i*i*}^{1/2}) - \Phi(\lambda_1 / R_{i*i*}^{1/2})}{\Phi(\nu_{i*}^- / R_{i*i*}^{1/2}) - \Phi(\nu_{i*}^+ / R_{i*i*}^{1/2})}\,, \qquad \text{(TLT statistic)}$$

où Φ désigne la fonction de répartion d'une Gaussienne centrée réduite, λ_1 désigne le premier temps d'entrée d'un paramètre dans $\hat{\beta}$, ν_{i*}^- et ν_{i*}^+ sont des variables définies dans [10]. Dans leur papier, ils démontrent que S suit une loi unforme sur $[0,1]$ sous H_0. De plus, ils définissent la zone de rejet comme étant

$$\mathrm{Reject}_\alpha := \{S \leq \alpha\}\,,$$

pour tout $\alpha \in (0,1)$. En d'autres mots, S est la p-valeur. Malgré tout, ce choix de zone de rejet peut paraitre arbitraire. En effet, de nombreuses transformation de S suivent une loi uniforme. Cependant, l'intuition nous conduit à dire que S prend de petites valeurs sous l'alternative. Dans la suite, nous démontrons que le test TLT est bien sans biais pour cette zone de rejet, nousdonnons une formule pour calculer la puissance du test et nous le comparons à un test d'adéquation simple sur Y.

2 Résultats théoriques

Sans perte de généralité, nous pouvons considérer le cas $\Sigma = I_n$ et $\|X_i\|_2 = 1$ où X_i désigne la i-ème colonne de la matrice de design X. Dans ce cas, le vecteur des corrélations $U = X^\top Y$ satisfait

$$U = (U_1, \ldots, U_p) \sim \begin{cases} \mathcal{N}_p(0, R) & \text{sous l hypothse nulle,} \\ \mathcal{N}_p(\mu^\star, R) & \text{sous l}' \text{ alternative ,} \end{cases}$$

où $R = X^\top X$ et $\mu^\star = R\beta^\star$. Il est bien connu que $\lambda_1 = \|U\|_\infty$. Supposons maintenant que les colonnes de X sont deux à deux non colinéaires. Cela implique, avec probabilité 1, qu'il existe une unique paire $(\hat{\imath}, \hat{\varepsilon})$, $\hat{\imath} \in [1,p] := \{1, \ldots, p\}$, $\hat{\varepsilon} = \pm 1$ telle que $\hat{\varepsilon} U_{\hat{\imath}} = \|U\|_\infty$ et les évènements $\mathcal{E}_{i,\varepsilon} = \{\varepsilon U_i = \|U\|_\infty\}$ sont p.s. disjoints. Notons

$$\lambda_1 = \sum_{i=1}^{p} \sum_{\varepsilon = \pm 1} \varepsilon U_i \mathbb{1}_{\mathcal{E}_{i,\varepsilon}}\,,$$

où $\mathbb{1}$ désigne la fonction indicatrice d'ensemble. En utilisant un argument de regression Gaussienne standard, pour tout $i, j \in [1, p]$

$$U_j = R_{ji} U_i + U_j^i \,,$$

où le reste U_j^i est indépendant de U_i. Définissons, pour tout $1 \leq i \leq p$ et $\varepsilon = \pm 1$,

$$\lambda_2^{i,\varepsilon} = \bigvee_{j \neq i} \Big\{ \frac{\varepsilon U_j^i}{1 - R_{ji}} \vee \frac{-\varepsilon U_j^i}{1 + R_{ji}} \Big\} \,,$$

où $a \vee b = \max(a, b)$. Cette notation est très liée à la précédente car $\mathcal{E}_{i,\varepsilon} = \{\lambda_2^{i,\varepsilon} < \varepsilon U_i\}$. En effet, considérons $j \neq i$, alors

$$\begin{aligned} \{-\varepsilon U_i < U_j < \varepsilon U_i\} &= \{-\varepsilon U_i < \varepsilon U_j < \varepsilon U_i\} \,, \\ &= \{-\varepsilon U_i \,(1 + R_{ji}) < \varepsilon (U_j - R_{ji} U_i) < \varepsilon U_i \,(1 - R_{ji})\} \,, \\ &= \{ \Big\{ \frac{\varepsilon U_j^i}{1 - R_{ji}} \vee \frac{-\varepsilon U_j^i}{1 + R_{ji}} \Big\} < \varepsilon U_i \} \,. \end{aligned}$$

Enfin, on peut en déduire, en utilisant la formulation standard donnée pour λ_2 dans [5], que la variable aléatoire λ_2 s'écrit

$$\lambda_2 = \sum_{i=1}^{p} \sum_{\varepsilon = \pm 1} \lambda_2^{i,\varepsilon} \mathbb{1}_{\mathcal{E}_{i,\varepsilon}} \,.$$

ce qui donne l'écriture de couple

$$(\lambda_1, \lambda_2) = \sum_{i=1}^{p} \sum_{\varepsilon = \pm 1} (\varepsilon U_i, \lambda_2^{i,\varepsilon}) \mathbb{1}_{\{\varepsilon U_i > \lambda_2^{i,\varepsilon}\}} \,. \tag{1}$$ eq:DefJointeKn

Cette formulation est la base du calcul de la puissance. En effet, par argument de régression, en utilisant l'indépendance entre U_j^i et U_i, on peut produifier la densité pour trouver

Proposition 1 *Pour tout $\alpha \in (0, 1)$, on pose*

$$h_\alpha(\ell) = \bar{\Phi}^{-1}\Big(\alpha \bar{\Phi}(\ell)\Big) - \ell, \tag{2}$$ eq:h

où $\bar{\Phi}^{-1}$ désigne la fonction inverse de $\bar{\Phi}$. Alors,

$$\mathbb{P}_{\mu^\star}(S \leq \alpha) = \alpha \, \mathbb{E}_{\mu^\star} \Big\{ \sum_{i=1}^{p} \sum_{\varepsilon = \pm 1} \exp\big[\varepsilon \mu_i^\star h_\alpha(\varepsilon U_i)\big] \mathbb{1}_{\mathcal{C}_{i,\varepsilon}} \Big\} \,. \tag{3}$$ eq:secExpressi

avec $\mathcal{C}_{i,\varepsilon} = \{(u_1, \ldots, u_p) \in \mathbb{R}^p \;;\; \forall j \neq i, \; |u_j| < \varepsilon u_i\} \,.$

De cette proposition, on déduit directement

Corollary 2 *Sous* H_0*, la statistique* S *suit une loi uniforme sur* $[0,1]$.

Enfin, en utilisant de manière astucieuse le lemme d'Anderson (voir dessous ou [1]) pour la densité Gaussienne, on peut démontrer le caractère sans biais du test TLT.

Proposition 3 *Anderson's inequality for Gaussian measure*
Soit E *un ensemble convexe de* $\mathbb{R}^p$*, symétrique par rapport à l'origine et* $Z \sim \mathcal{N}(0,\Sigma)$*. Alors, pour tout* $k \in [0,1]$ *et* $\mu \in \mathbb{R}^p$*,*

$$\mathbb{P}(Z + k\mu \in E) \geq \mathbb{P}(Z + y \in E). \tag{4}$$

Anderson

Cette utilisation du lemme d'Anderson est d'ailleurs particulièrement troublante car une preuve assez rapide du caractère sans biais du test en dimension deux découle de son application. Malgré tout, cette preuve s'avère bien plus géométrique que la preuve générale en dimension quelconque.

3 Comparaison numérique avec le χ^2

3.1 A Matlab Toolbox

Afin de calculer numériquement la puissance grâce à la formule (3), nous avons besoin d'outils d'intégration en grande dimension. Dans un premier temps, (3) peut-être vu comme une intégrale Gaussienne en dimension n où $n = \dim(\mathrm{Im}(X))$. En effet,

$$\mathbb{P}_{\mu^\star}(S \leq \alpha) = \alpha \mathbb{E}_{\mu^\star}\left(W(U_1,\ldots,U_n)\right) = \alpha \mathbb{E}_{\epsilon}\left(W(X'X\beta + X'\epsilon)\right) \tag{5}$$

où $\epsilon \sim \mathcal{N}(0,\Sigma)$ et

$$W(U_1,\ldots,U_n) = \sum_{i=1}^{p} \sum_{\varepsilon=\pm 1} \exp(\varepsilon \mu_i h_\alpha(\varepsilon U_i)) \mathbb{1}_{C_{i,\epsilon}}. \tag{6}$$

power:function

Pour calculer numériquement la dernière expression, nous utilisons un algorithme construit par Alan Genz (voir [6] et [2]). Basé sur une réécriture de l'intégrale sur l'hypercube $[0,1]^n$ et sur une méthode d'intégration Monte-Carlo Quasi Monte-Carlo (MCQMC), il s'avère très efficace pour ce type de calculs. On pourra trouver la version légèrement modifiée pour le calcul de la puissance du test TLT sur le site web de Stephane Mourareau.

3.2 χ^2 test versus TLT test

On considère ici un simple test d'adéquation de $H_0 : \beta^\star = 0$ contre $H_1 : \beta^\star \neq 0$. On définit pour cela la statistique $T = \| y \|_2^2$ qui suit une loi $\chi^2(n, \| X\beta \|_2^2)$ où $\chi^2(a, b)$ désigne la loi χ^2 avec a degrés de liberté et un paramètre de décentrement b. Notre but est de comparer ce test à celui proposé par [10] dans différents cas.

Dans les simulations, la matrice de design X est composée de $n \times p$ réalisations indépendantes de loi normale centrée réduites, la moyenne β est tirée selon un loi normale centrée de variance 2 (signal important), variance 1 (signal moyen) ou selon une loi uniforme sur $[0, 1]$ (faible signal). Les résultats sont assez contrastés, le test TLT semble plus efficace dans le cas de grande dimension lorsque un écart significatif existe entre les 2 plus grandes moyennes (voir Figure 2). A l'inverse, quand les β_i sont du même ordre de grandeur, le test du χ^2, basé sur la norme, s'avère bien plus puissant, même dans des cas de très grande sparsité (voir Figure 1).

Pour les simulations, nous avons fait le choix de prendre $\alpha = 0.05$ qui est un choix classique. D'autres simulations pour un niveau $\alpha = 0.01$ ont donné des résultats similaires.

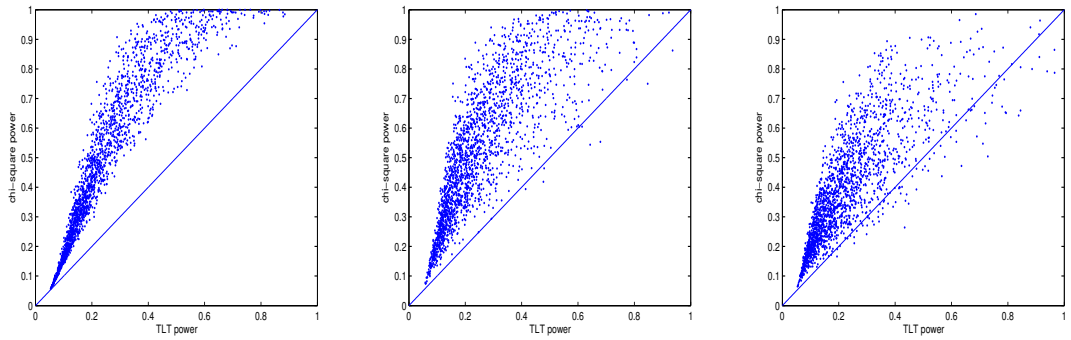

Figure 1: De gauche à droite, 2.000 simulations présentant la puissance du test TLT contre celle du test du χ^2 dans plusieurs cas sparses $(s, n, p) = (5, 10, 50)$, $(10, 50, 100)$ and $(10, 100, 200)$. Dans ce cas, le signal est considéré comme grand (voir détail au dessus). Le test du χ^2 s'avère plus puissant dans 91, 98 et 99 % des cas.

`sparse`

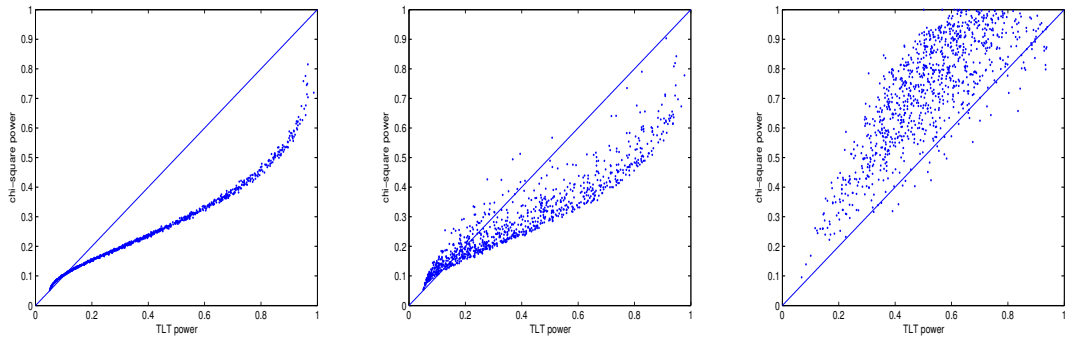

Figure 2: Sur la première figure (gauche), $(s, n, p) = (1, 100, 400)$ et la moyenne est issue d'une $\mathcal{N}(\sqrt{2\log(p)}, 1)$. Dans le second cas (centre), une moyenne est tirée selon une $\mathcal{N}(\sqrt{2\log(p)}, 1)$ et les autres selon une $\mathcal{N}(0, 1)$. Dans le dernier cas, $(s, n, p) = (3, 100, 400)$ et toutes les moyennes sont issues d'une $\mathcal{N}(\sqrt{2\log(p)}, 1)$. Quand une moyenne domine, le test TLT semble plus efficace. Cependant, quand l'écart entre les 2 plus grandes moyennes diminue, le test du χ^2 semble plus robuste.

`asymp`

Bibliographie

[1] Anderson, T.W. (1955), *The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities*, Proceedings of the American Mathematical Society, Vol 6, 170-176.

[2] Azaïs, J.M. et Genz, A. (2013), *Computation of the Distribution of the Maximum of Stationary Gaussian Processes*, Methodology and Computing in Applied Probability, Vol 15, Issue 4, 969-985.

[3] Azaïs, J.M. et Wschebor, M. (2009), *Level sets and extrema of random processes and fields*, John Wiley and sons, Washington DC.

[4] Buhlmann, P., Meier, L. et Van de Geer, S. (2014), *Discussion: "A significance test for the Lasso"*, Annals of Statistics, Vol 42, no. 2, 469-477.

[5] Efron, B., Hastie, T. et Johnstone, I. et Tibshirani, R. (2004), *Least angle regression*, Annals of Statistics, Vol 32, no. 2, 407-499.

[6] Genz, A. (1992), *Numerical computation of Multivariate Normal Probabilities*, J. Comp. Graph Stat, Vol 1, 141-149.

[7] Lee, J.D., Sun, D.L., Sun, Y. et Taylor, J.E. (2015), *Exact post-selection inference with application to the Lasso.*, arXiv:1311.6238v5.

[8] Lockhart, R., Taylor, J., Tibshirani, R.J. et Tibshirani, R. (2014), *A significance test for the Lasso*, Annals of Statistics, Vol 42, no. 2, 413-468.

[9] Loftus, J. et Taylor, J. (2014), *A significance test for forward stepwise model selection*, arXiv:1405.3920v1.

[10] Loftus, J., Taylor, J. et Tibshirani, R. (2014), *Tests in adaptive regression via the Kac-Rice formula*, arXiv:1308.3020v3.