Règles de sélection de variables pour accélerer la localisation de sources en MEG et EEG sous contrainte de parcimonie

Olivier FERCOQ[1], Alexandre GRAMFORT[1,2] Joseph SALMON[1]

[1]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France

[2]NeuroSpin, CEA Saclay, Bat. 145, Gif-sur-Yvette Cedex, France

olivier.fercoq@telecom-paristech.fr,
alexandre.gramfort@telecom-paristech.fr,
joseph.salmon@telecom-paristech.fr

Résumé – La localisation de sources par électroencéphalographie (EEG) et magnétoencéphalographie (MEG) permet d'identifier les zones d'activités cérébrales. Le problème est une régression en grande dimension qui peut être résolue avec des a priori de parcimonie de type Lasso. A l'aide de certificats d'optimalité vérifiés par les solutions du Lasso il est possible d'écarter au cours de l'optimisation certaines des sources non pertinentes. Ce faisant on peut accélérer drastiquement les algorithmes. Nous proposons de nouvelles règles de pré-sélection qui reposent sur le saut de dualité. Elles s'appuient sur la création de régions dites de sécurité, dont le diamètre tend vers zéro, sous l'hypothèse que l'on dispose d'un algorithme convergent. Cette propriété permet à la fois de dépister plus de sources non pertinentes, et de considérer de plus grandes plages pour le paramètre de régularisation. Nous démontrons la pertinence de notre approche avec une méthode de descente par coordonnées particulièrement adaptée au problème. Des gains de temps de calcul importants sont ainsi obtenus sur données réelles MEG et EEG.

Abstract – Source localization with electroencephalography (EEG) and magnetoencephalography (MEG) allows to identify active brain regions. The problem to solve is a high dimensional regression problem that can be solved using sparsity promoting regularization such as Lasso. Using optimality certificates verified by Lasso solutions it is possible during the optimization to discard certain sources. By doing so, solvers can be significantly sped up. Here we propose new selection rules which rely on duality gap control. We create so called safe regions, whose diameter tends to zero for any converging algorithm. This property allows to discard more sources and offers benefits for small regularization parameters. We demonstrate the relevance of our approach using a coordinate descent algorithm particularly adapted to the problem. Significant computing time reductions are obtained with respect to previous safe rules when tested on actual MEG and EEG data.

1 Introduction

Depuis le milieu des années 90, les statistiques en grande dimension ont attiré beaucoup d'attention, en particulier dans le contexte de la régression linéaire lorsque le nombre de régresseurs est supérieur au nombre d'observations : ce cadre est maintenant connu sous le nom de "$p > n$". Dans ce cas, le problème de moindre carrés avec une régularisation ℓ_1 (connu sous le nom de Lasso en statistiques [1] ou "Basis Pursuit" en traitement du signal [2]), a été particulièrement étudié. L'estimateur de Lasso bénéficie de garanties théoriques [3] et d'un intérêt pratique : il fournit des solutions dites parcimonieuses où seul un faible nombre de régresseurs ont un coefficient non nul et il bénéficie d'algorithmes d'optimisation convexe rapides. Ceci a fait du Lasso un outil classique en analyse de données avec des applications en apprentissage de dictionnaire [4], en biostatistiques [5], sans oublier la neuroimagerie [6].

Dans le cadre du problème de localisation de sources par MEG et EEG, on cherche à identifier des régions actives du cerveau à partir de mesures non-invasives du champ électromagnétique induit par l'activité neuronale. La physique du problème donnée par les équations de Maxwell garantit un problème linéaire et lors du traitement d'une tâche cognitive il est naturel de faire l'hypothèse que seules quelques régions du cerveau sont actives. Cette dernière observation rend naturel les régularisations induisant de la parcimonie. Si l'on considère des données MEG et EEG à un instant donné, par exemple 100 ms après la présentation d'un stimuli, une donnée est un vecteur de n observations, qui correspondent aux mesures des n capteurs, et chaque régresseur ou variable dans le problème inverse correspond à une source candidate dans le cerveau. On se ramène donc à un problème de type Lasso. La littérature contient un certain nombre de contributions utilisant des régularisations de type ℓ_1 pour la MEG et l'EEG [6–8]. Toutefois, bien que de nombreux algorithmes existent, rendre l'utilisation de ces algorithmes plus rapides reste un défi, notamment lors d'une utilisation dans une analyse de données interactive.

Même si le Lasso n'est pas à ce jour toujours la méthode la plus efficace en grande dimension, notamment face à des approches non-convexes tels que SCAD [9] ou MCP [10] ou encore [11], la résolution de problème convexes de type Lasso reste souvent une étape cruciale de ces algorithmes.

Parmi les solutions algorithmiques qui existent pour résoudre le Lasso on trouve les méthodes d'homotopie [12] ou LARS [13] qui fournissent les solutions pour tout le chemin de régularisation, *i.e.,* pour toutes les valeurs du paramètre de régularisation λ. Plus récemment, en particulier quand $p > n$, les approches de type descente par coordonnées [11, 14] ont montré d'excellentes performances sur des problèmes en très grande dimension.

Dans la continuité du travail de [15], des approches de sélections de variables ont été proposées afin d'exploiter la connaissance du fait qu'une solution du Lasso se doit d'être parcimonieuse. L'idée est de fournir des règles qui permettent d'exclure de l'optimisation des variables, c'est-à-dire des sources, tout en garantissant que la solution sera optimale. De telles techniques sont connues sous de nom de *safe rules* (*cf.* [16] pour une présentation des approches actuelles). En mettant à zéro ces coefficients, l'algorithme peut se focaliser sur les sources qui sont les plus probables, réduisant ainsi les temps de calcul. Il existe d'autres approches ne pouvant garantir l'exclusion définitive des sources. C'est par exemple le cas de la strategie connue sous le nom de *strong rules* [17].

Les *safe rules* originales fonctionnent de la façon suivante : pour une valeur fixée du paramètre de regularisation λ, et avant même de lancer un algorithme de minimisation, on teste si une variable peut être ignorée ou non. De telles règles sont dites statiques. Le test est effectué en construisant une région dite de sûreté, *i.e.,* une région contenant une solution duale optimale du problème de Lasso.

Notre objectif est d'améliorer le critère de sélection de variables, et donc de sources, en raffinant au cours des itérations de l'algorithme la région de sûreté. En suivant les travaux de [18, 19], nous appelons de telles règles de sélection des méthodes dynamiques. L'ambition est d'accélérer la convergence des algorithmes lorsque des solutions pour des petites valeurs de λ sont demandées ou si plusieurs solutions sur le chemin de régularisation sont souhaitées.

En nous basant sur des arguments théoriques d'optimisation convexe, nous proposons d'utiliser le saut de dualité du problème de Lasso pour construire un nouvelle règle de sélection dynamique. Nous appellerons cette méthode *GAP SAFE rule*.

2 Le Lasso : modèle et notations

Notre vecteur d'observations est $y \in \mathbb{R}^n$ et la matrice des variables explicatives $X = [x_1, \cdots, x_p] \in \mathbb{R}^{n \times p}$ a p variables, une par colonne. Notre but est d'approcher y par une combinaison linéaire d'un petit nombre de variables x_j. Ainsi, on cherche à exprimer y comme $X\beta$, où $\beta \in \mathbb{R}^p$ a peu d'éléments non nuls. Dans le cadre de la MEG et l'EEG, X est la matrice de problème direct obtenue par résolution numérique des équations de Maxwell et β contient l'amplitude des sources.

Pour estimer β, nous considérons le problème du Lasso. Il dépend d'un paramètre $\lambda > 0$ à calibrer qui contrôle le compromis entre l'attache aux données et la parcimonie de la solution. Un estimateur Lasso $\hat{\beta}^{(\lambda)}$ est une des solutions du problème d'optimisation primal

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2} \|X\beta - y\|_2^2 + \lambda \|\beta\|_1}_{=P_\lambda(\beta)} . \tag{1}$$

Notons $\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \left|x_j^\top \theta\right| \leqslant 1, \forall j \in [p]\right\}$ l'ensemble admissible dual. Une formulation duale du Lasso s'écrit (voir par exemple [20] ou [16]) :

$$\hat{\theta}^{(\lambda)} = \underset{\theta \in \Delta_X \subset \mathbb{R}^n}{\arg\max} \underbrace{\frac{1}{2} \|y\|_2^2 - \frac{\lambda^2}{2} \left\|\theta - \frac{y}{\lambda}\right\|_2^2}_{=D_\lambda(\theta)}. \tag{2}$$

En particulier, la solution duale $\hat{\theta}^{(\lambda)}$ est unique, contrairement à la solution primale $\hat{\beta}^{(\lambda)}$.

2.1 Tests de sphère

Reprenant une terminologie déjà introduite pour d'autres règles de sélection *safe rules* [15, 16], nous appelons *tests de sphère* les tests qui utilisent des boules comme région de sécurité. Pour un test de sphère, on choisit une boule contenant $\hat{\theta}^{(\lambda)}$ de centre c et de rayon r, c'est-à-dire $\mathcal{C} = B(c, r)$. La règle de sélection correspondante est définie par :

$$\text{Si } |x_j^\top c| + r\|x_j\| < 1, \text{ alors } \hat{\beta}_j^{(\lambda)} = 0. \tag{3}$$

Notez que pour un centre donné, plus le rayon est petit, plus la stratégie de sélection sera discriminante. Ainsi, le but principal des *safe rules* est de trouver des boules avec un petit rayon de manière à éliminer autant de variables x_j que possible.

2.2 Règles de sélection dynamiques

Pour approcher une solution $\hat{\beta}^{(\lambda)}$ du problème Lasso, des algorithmes itératifs sont souvent utilisés. Notons $\beta_k \in \mathbb{R}^p$ l'estimée courante après k itérations de l'algorithme itératif de notre choix. Les règles de sélection dynamiques cherchent à découvrir des régions de sécurité qui se contractent quand k augmente. Pour les construire, nous avons besoin de points duaux admissibles : $\theta_k \in \Delta_X$. Suivant [15] (voir aussi [18]), nous pouvons en générer par une transformation simple des résidus courants $\rho_k = y - X\beta_k$, en définissant θ_k comme

$$\theta_k = \alpha_k \rho_k, \tag{4}$$

$$\text{où } \alpha_k = \min\Big[\max\left(\frac{y^\top \rho_k}{\lambda \|\rho_k\|^2}, \frac{-1}{\|X^\top \rho_k\|_\infty}\right), \frac{1}{\|X^\top \rho_k\|_\infty}\Big].$$

Un tel θ_k, admissible pour le dual, est proportionnel à ρ_k et est le point le plus proche de y/λ dans Δ_X. On choisit un tel point car la solution optimale duale $\hat{\theta}^{(\lambda)}$ est la projection de y/λ sur Δ_X, et que $\hat{\theta}^{(\lambda)}$ est proportionnel à $y - X\hat{\beta}^{(\lambda)}$.

Notre règle de sélection dynamique consiste à choisir le centre $c = \theta_k$ dans (3). Nous avons prouvé qu'un rayon égal à $r = \sqrt{2\left(P_\lambda(\beta_k) - D_\lambda(\theta_k)\right)}/\lambda$ donne une *safe rule*. Notez que la quantité $P_\lambda(\beta_k) - D_\lambda(\theta_k)$ est simplement le saut de dualité obtenu pour le point primal β_k et le point dual θ_k.

Remarque 1. On peut raffiner le test de sphère en un test de dôme. Malheureusement, sa formulation est trop lourde pour être donnée ici ; *cf.* [16] pour plus de détails.

Remarque 2. Notez que si $\lim_{k\to+\infty} \beta_k = \hat{\beta}^{(\lambda)}$ (convergence du primal), alors nous pouvons montrer que $\lim_{k\to+\infty} \theta_k = \hat{\theta}^{(\lambda)}$ (convergence du dual) et que la convergence du primal n'est pas altérée par les règles de sélection sûres. Éliminer les coefficients inutiles ne peut que faire décroître la distance à une solution primale.

3 Mise en œuvre et résultats

3.1 Détails d'implémentation

Notre méthode ainsi que les méthodes de l'état de l'art pour la sélection de variables ont été implémentées à partir du code de descente par coordonnées de Scikit-Learn [21]. Le code est écrit en Python et Cython afin de générer du code C bas niveau, offrant ainsi d'excellentes performances. En pratique, nous mettons à jour la sélection de variables dynamique tous les 10 passages sur les variables restantes. Les itérations sont arrêtées lorsque le saut de dualité est inférieur à une tolérance fixée par l'utilisateur.

3.2 Données utilisées

Les données utilisées pour démontrer l'impact applicatif de ce travail sont les données publiques fournies par le logiciel MNE [22]. Ce sont des données combinant MEG et EEG et obtenues sur une machine Neuromag Vectorview avec 306 capteurs MEG et 60 électrodes EEG au sein du Martinos Center au Massachusetts General Hospital à Boston. Les données sont échantillonnées à 600 Hz. L'expérience est une stimulation auditive et visuelle avec des stimuli auditifs délivrés de façon monaurale dans l'oreille gauche ou droite et des stimuli visuels présents dans l'hémichamp gauche ou droite. Les stimuli furent présentés de façon randomisée avec un temps moyen entre deux stimuli de 750 ms. Les données sont décrites avec plus de détails dans [22].

L'algorithme est mis en œuvre après avoir discrétisé l'espace des sources avec un écart moyen de 6 ou 7 mm sur la surface corticale. Ceci conduit à travailler avec un p environ égal à 22,000. Après suppression des capteurs MEG et EEG dis-fonctionnant durant l'expérience, 360 capteurs sont utilisés. On a donc ici $n = 360$. La localisation de sources est effectuée dans la section suivante sur le pic d'activité cérébrale induit par les stimuli auditifs dans l'oreille gauche (environ 100 ms après stimulation).

3.3 Résultats numériques

La Figure 1,(a) présente la proportion de sources ignorées par les différentes règles de sélection sur le dataset M/EEG décrit ci-dessus. La proportion est présentée en fonction du nombre d'itérations K dans la descente par coordonnées. Comme la règle SAFE de [15] n'est pas dynamique, pour un λ donné la proportion de variables ignorées ne dépend pas de K. Les règles proposées par [18] sont plus efficaces mais elles ne bénéficient pas vraiment du caractère dynamique de l'approche. Notre méthode GAP SAFE permet d'ignorer beaucoup plus de variables, en particulier quand le paramètre de régularisation λ devient faible, cas particulièrement pertinent en pratique. De plus, même pour les faibles valeurs de λ (notez l'échelle logarithmique) où aucune variable n'est ignorée au début de l'optimisation, la règle GAP SAFE réussit à supprimer de plus en plus de variables lorsque K augmente. Enfin, la figure démontre que la technique GAP SAFE dôme n'apporte presque aucune amélioration supplémentaire par rapport à la sphère.

La principale motivation pour les méthodes de sélection telles que présentées ici est de réduire le temps de calcul. En effet, le temps pour calculer les règles ne doit pas être plus long que le temps gagné lors de la descente par coordonnées. Nous avons ainsi comparé le temps nécessaire pour calculer un chemin de régularisation du Lasso complet en précisant différentes valeurs pour la tolérance du critère d'arrêt. La Figure 1(b) présente les résultats. Pour des valeurs faibles de la tolérance sur le saut de dualité, ce qui est nécessaire pour estimer le bon support, on obtient des gains en temps de calcul d'un facteur 7 sur cet exemple. C'est une amélioration significative en pratique. Bien évidemment les résultats en terme de localisation sont identiques à ceux obtenu sans sélection dynamique comme dans [6].

4 Conclusion

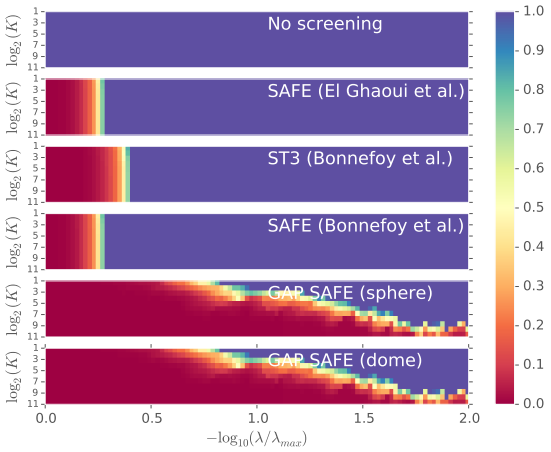
La contribution de ce travail est la présentation d'une nouvelle règle de sélection de variables pour les problèmes inverses sous contrainte de parcimonie ℓ_1. Dans le cadre de la MEG et l'EEG cette approche qui utilise astucieusement des résultats de dualité convexe pour le Lasso permet d'obtenir des gains en temps de calcul d'un facteur 7 sur données réelles. Les travaux futures porteront sur la généralisation de ces idées aux pénalités structurées de type group-Lasso afin de prendre en compte l'orientation des sources et leur dynamiques temporelles [6].

Références

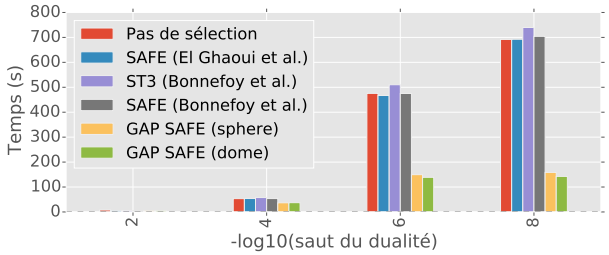
[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61 (electronic), 1998.

[3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.



(a) Proportion de variables ignorées en fonction de λ et du nombre d'iterations K.



(b) Temps mis pour converger en utilisant différentes règles de sélection.

FIGURE 1 – Résultats sur données MEG et EEG ($n = 360, p = 22494$).

[4] J. Mairal, "Sparse coding for machine learning, image processing and computer vision," Ph.D. dissertation, École normale supérieure de Cachan, 2010.

[5] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "TIGRESS : Trustful Inference of Gene REgulation using Stability Selection." *BMC systems biology*, vol. 6, no. 1, p. 145, 2012.

[6] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods," *Physics in Medicine and Biology*, vol. 57, no. 7, pp. 1937–1961, 2012.

[7] K. Matsuura and Y. Okabe, "Selective minimum-norm solution of the biomagnetic inverse problem." *IEEE Trans. Biomed. Eng.*, vol. 42, no. 6, pp. 608–615, June 1995.

[8] W. Ou, M. Hämaläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver," *NeuroImage*, vol. 44, no. 3, pp. 932–946, Feb. 2009.

[9] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[10] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[11] D. Strohmeier, J. Haueisen, and A. Gramfort, "Improved meg/eeg source localization with reweighted mixed-norms," in *Proc. IEEE PRNI 2014*, June 2014, pp. 1–4.

[12] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, 2000.

[13] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004, with discussion, and a rejoinder by the authors.

[14] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.

[15] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," *J. Pacific Optim.*, vol. 8, no. 4, pp. 667–698, 2012.

[16] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening tests for lasso problems," *arXiv preprint arXiv :1405.4897*, 2014.

[17] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *J. Roy. Statist. Soc. Ser. B*, vol. 74, no. 2, pp. 245–266, 2012.

[18] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "A dynamic screening principle for the lasso," in *EUSIPCO*, 2014.

[19] ——, "Dynamic Screening : Accelerating First-Order Algorithms for the Lasso and Group-Lasso," *ArXiv e-prints*, 2014.

[20] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, 2007.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn : Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[22] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446 – 460, Feb. 2014.