# Lasso concomitant avec répétitions (CLaR): au-delà de la moyenne pour des réalisations multiples en présence d'un bruit hétéroscédastique

Quentin Bertrand[1], Mathurin Massias[1], Alexandre Gramfort[1] & Joseph Salmon[2]

[1] *INRIA, Université Paris-Saclay*
[2] *IMAG, Univ Montpellier, CNRS, Montpellier, France*

**Résumé.**

Les régularisations par normes induisant de la parcimonie sont fréquemment utilisées pour la régression en grande dimension. Une limite des estimateurs ainsi obtenus (le Lasso étant l'exemple canonique) est que le paramètre de régularisation dépend du niveau de bruit, qui varie entre les jeux de données et les expériences. Des estimateurs comme le concomitant Lasso Owen [2007], Sun and Zhang [2012] ou le square-root Lasso Belloni et al. [2011] résolvent cette dépendance en estimant conjointement le niveau de bruit et les coefficients de régression. Cependant, dans de nombreuses applications expérimentales, les données sont obtenues en faisant la moyenne de plusieurs mesures. Cela aide à réduire la variance du bruit, mais diminue considérablement la taille des échantillons, empêchant ainsi une modélisation raffinée du bruit. Dans ce travail, nous proposons un estimateur capable de gérer des structures de bruit complexes (hétéroscédastique) en utilisant l'ensemble des mesures (non moyennées) et une estimation concomitante de la structure du bruit. Le problème d'optimisation qui en résulte reste convexe, ce qui permet d'utiliser des algorithmes efficaces pour le résoudre. Grâce à la théorie du lissage Nesterov [2005], Beck and Teboulle [2012], il est donc possible de recourir à des techniques de descente de coordonnées (qui sont état-de-l'art pour ces approches dans un contexte d'apprentissage statistique en grande dimension) pouvant tirer parti de la parcimonie attendue des solutions. Les avantages pratiques sont illustrés sur des données simulées et sur des applications de neuroimagerie.

**Mots-clés.** Lasso, parcimonie, optimisation convexe, lissage, hétéroscédasticité

**Abstract.**

Sparsity promoting norms are frequently used in high dimensional regression. A limitation of Lasso-type estimators is that the regularization parameter depends on the noise level which varies between datasets and experiments. Estimators such as the concomitant Lasso address this dependence by jointly estimating the noise level and the regression coefficients. As sample sizes are often limited in high dimensional regimes, simplified heteroscedastic models are customary. However, in many experimental applications, data is obtained by averaging multiple measurements. This helps reducing the noise variance, yet it dramatically reduces sample sizes, preventing refined noise modeling. In this work, we

propose an estimator that can cope with complex heteroscedastic noise structures by using non-averaged measurements and a concomitant formulation. The resulting optimization problem is convex, so thanks to smoothing theory, it is amenable to state-of-the-art proximal coordinate descent techniques that can leverage the expected sparsity of the solutions. Practical benefits are demonstrated on simulations and on neuroimaging applications.

**Keywords.** Lasso, sparsity, convex optimization, smoothing, heteroscedasticity

# 1 Estimation concomitante

Dans de nombreuses applications statistiques, le nombre de paramètres $p$ est beaucoup plus grand que le nombre d'observations $n$. Une approche populaire pour résoudre les problèmes de régression linéaire dans de tels scénarios consiste à utiliser des pénalités convexes de type $\ell_1$, popularisées par Tibshirani [1996]. L'utilisation de ces pénalités repose sur un paramètre de régularisation $\lambda$ qui représente un compromis entre la fidélité aux données et la régularisation. Malheureusement, l'analyse statistique révèle que le $\lambda$ optimal devrait être proportionnel au niveau de bruit [Bickel et al., 2009], qui est rarement connu en pratique. Pour résoudre ce problème, il est possible d'estimer conjointement le niveau de bruit et les coefficients de régression. Une telle estimation concomitante [Huber and Dutter, 1974, Huber, 1981] a récemment été adaptée en régression parcimonieuse par Owen [2007] et analysée sous plusieurs noms, tels que square-root Lasso [Belloni et al., 2011] ou scaled Lasso [Sun and Zhang, 2012].

Dans ces derniers travaux, la modélisation du bruit repose sur un seul paramètre, sa variance. Cependant, dans divers contextes appliqués, il est d'usage d'agréger des données de natures différentes ou provenant de sources différentes pour augmenter le nombre d'observations. Cela conduit souvent à de l'hétéroscédasticité: les données sont contaminées par des niveaux de bruit non uniformes (différents selon les caractéristiques ou les échantillons). C'est le cas des données magnéto-électroencéphalographiques (M/EEG), où les signaux observés proviennent de trois types de capteurs: gradiomètres, magnétomètres et électrodes, ce qui conduit à des amplitudes et des covariances très différentes pour le bruit. Des tentatives de traiter l'hétéroscédasticité dans ce contexte ont été analysées par Daye et al. [2012], Wagener and Dette [2012], Kolar and Sharpnack [2012], Dalalyan et al. [2013]. De plus, des algorithmes rapides reposant sur les techniques de lissage de la communauté d'optimisation [Nesterov, 2005, Beck and Teboulle, 2012] ont été étendus à la régression hétéroscédastique dans un contexte multi-tâches, par le lasso concomitant généralisé lisse (SGCL, Massias et al. [2018]). Le SGCL est conçu pour estimer conjointement les coefficients de régression et la *matrice de co-standard déviation* du bruit [1]. Cependant, dans certaines applications, telles que les données M/EEG, le nombre de

[1] c'est-à-dire la racine carrée de la matrice de covariance du bruit

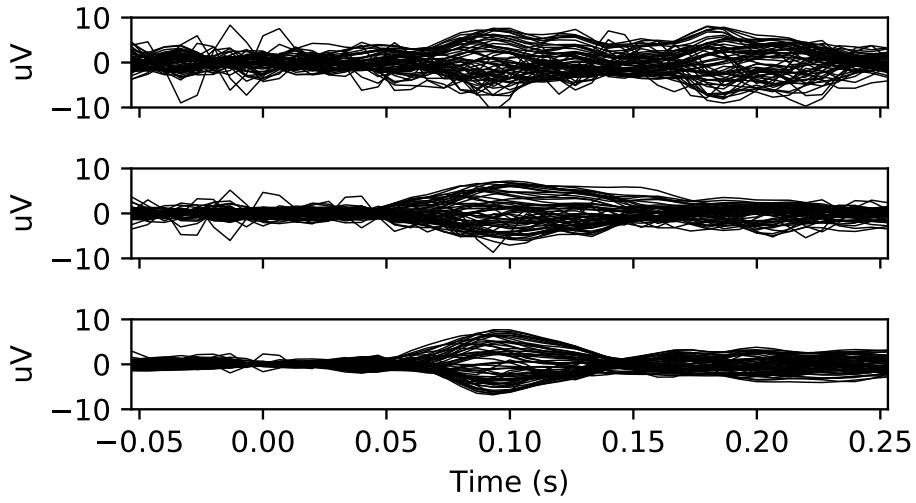


Figure 1: Amplitude $\bar{Y}$ de $n = 59$ signaux EEG, moyennés sur les $r = 5$ (haut), $r = 10$ (milieu) et $r = 50$ (bas) répétitions. Au fur et à mesure que le nombre de répétitions augmente, le bruit diminue et les mesures deviennent lisses, révélant la réponse du cerveau autour de 0.1 s.

paramètres dans la matrice de co-standard déviation ($\approx 10^4$) est généralement égal au nombre d'observations, ce qui rend statistiquement impossible une estimation précise.

Lorsque les observations sont contaminées par un fort bruit et que le rapport signal sur bruit (SNR) est trop bas, à condition que les mesures puissent être répétées, une idée naturelle est de les moyenner. En effet, sous l'hypothèse que le signal d'intérêt est corrompu par des réalisations indépendantes d'un bruit additif, moyenner les différentes mesures permet de diviser la variance du bruit par le nombre de répétitions. Cet effet est illustré dans Figure 1 pour des données d'électroencéphalographie (EEG). Dans ce cas, 5, 10 et 50 répétitions de la réponse électrique du cerveau à un stimulus sont observées, et révèle progressivement une réponse cérébrale environ 100 ms après stimulation à l'instant 0. C'est généralement ce type de données moyennées qui est utilisé directement dans les estimateurs statistiques (obtenus ensuite sous forme de solution d'un problème d'optimisation), alors que les observations individuelles pourraient être utilisées pour mieux caractériser le bruit et améliorer l'estimation statistique Gramfort et al. [2013], Ou et al. [2009].

Dans ce travail, nous proposons l'estimateur Concomitant Lasso with Repetitions (CLaR), conçu pour exploiter toutes les mesures disponibles collectées lors de répétitions d'expériences. La formulation concomitante proposée dans le problème d'optimisation sous-jacent présente deux avantages importants: premièrement, la covariance de bruit est un paramètre explicite du modèle, sur lequel il est facile d'ajouter des contraintes structurelles (par exemple, la diagonalité par blocs) et deuxièmement, la théorie du lissage

conduit à une fonction de coût pouvant être minimisée à l'aide de techniques efficaces de descente par coordonnées proximales.

Le modèle multi-répétitions est:

$$\forall l \in [r], \quad Y^{(l)} = X\mathrm{B}^* + S^*\mathrm{E}^{(l)} \in \mathbb{R}^{n\times q} \frac{1}{2nqr} \sum_1^r \left\|Y^{(l)} - X\mathrm{B}\right\|^2_{S^{-1}} + \frac{1}{2n}\operatorname{Tr}(S) \quad , \tag{1}$$

où $\mathrm{B}^*$ et $S^*$ sont les mêmes pour les $r$ répétitions $Y^{(1)}, \ldots, Y^{(r)}$, les $\mathrm{E}^{(l)}$ sont indépendants et composés chacun d'entrées i.i.d. gaussiennes centrées réduites.

CLaR estime les paramètres du précédent modèle par[2]

$$(\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{\mathrm{S}}^{\mathrm{CLaR}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ S \succeq \underline{\sigma}}}{\arg\min} \frac{1}{2nqr} \sum_1^r \left\|Y^{(l)} - X\mathrm{B}\right\|^2_{S^{-1}} + \frac{1}{2n}\operatorname{Tr}(S) + \lambda \left\|\mathrm{B}\right\|_{2,1} \quad , \tag{2}$$

où $\lambda > 0$ contrôle la parcimonie de $\hat{\mathrm{B}}^{\mathrm{CLaR}}$ et $\underline{\sigma} > 0$ contrôle la plus petite valeur propre de $\hat{\mathrm{S}}^{\mathrm{CLaR}}$.

En estimant conjointement les coefficients de régression et la structure de bruit, cet estimateur montre des améliorations de l'identification de support par rapport aux estimateurs utilisant des données moyennées ou supposant un bruit homoscédastique.

# 2 Estimation du support

Nous décrivons ici quelques éléments numériques validant les propriétés énoncées précédemment.

Plus précisément nous illustrons la capacité de notre estimateur à récupérer le support des solutions. Pour cela on part de simulations avec $n = 150$ observations, $p = 500$ caractéristiques (ou co-variables), $q = 100$ tâches (dimension temporelle dans le cas M/EEG). La matrice des co-variables $X$ est aléatoire avec les colonnes corrélées selon une structure de Toeplitz de paramètre $\rho_X = 0.6$ (corrélation entre $X_{:i}$ et $X_{:j}$ qui vaut $\rho_X^{|i-j|}$), et ses colonnes sont normalisée pour avoir une norme euclidienne valant 1. Le vrai coefficient $\mathrm{B}^*$ a 30 lignes non zéros dont les entrées sont indépendantes, gaussiennes, centrées réduites. $S^*$ (la matrice de co-standard déviation du bruit) est une matrice de Toeplitz avec le paramètre $\rho_S$. Le SNR est fixe et constant pour toutes les répétitions

$$\mathrm{SNR} := \|X\mathrm{B}^*\| / \|X\mathrm{B}^* - Y^{(l)}\| \ . \tag{3}$$

Pour Figures 2 and 3, on fournit des courbe ROC, c'est-à-dire le taux de vrais positifs, contre le taux de faux positifs. Pour les quatre estimateurs, la courbe ROC est obtenue en faisant varier la valeur du paramètre de régularisation $\lambda$ sur une grille géométrique de 160 points, en partant de $\lambda_{\max}$ (spécifique à chaque algorithme) à $\lambda_{\min}$, le dernier étant également spécifique et choisi pour obtenir des taux de faux positifs suffisamment importants.

[2]on utilise la notation: $S \succeq \underline{\sigma}$ si $S - \underline{\sigma}\,\mathrm{Id}$ est une matrice symétrique semi-définie positive.

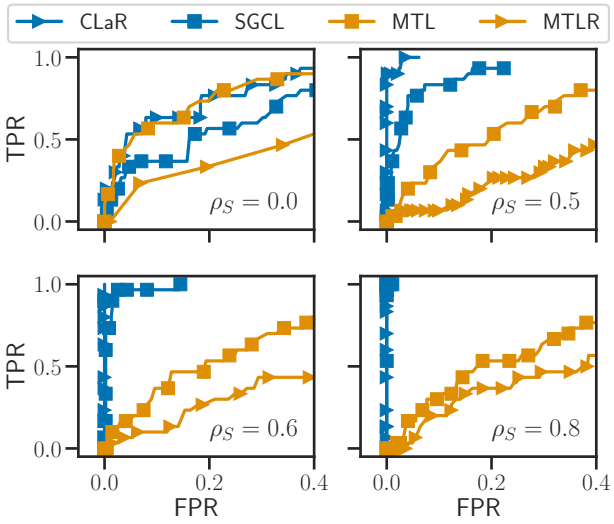


Figure 2: Courbes ROC pour les estimateurs CLaR, SGCL, MTL et MTLR avec $\rho_X = 0.6$, SNR $= 0.03$, $r = 50$ pour différentes valeurs de $\rho_S$.

# References


A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, 2013.

J. Daye, J. Chen, and H. Li. High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, 68(1):316–326, 2012.

A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013.

P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

P. J. Huber and R. Dutter. Numerical solution of robust regression problems. In *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*, pages 165–172. Physica Verlag, Vienna, 1974.

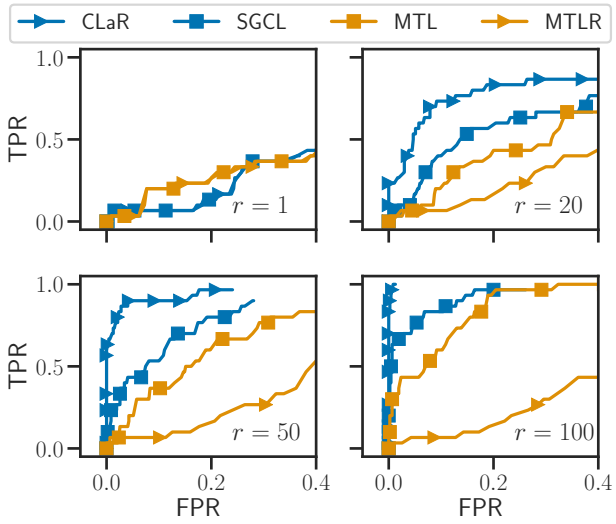


Figure 3: Courbes ROC pour les estimateurs CLaR, SGCL, MTL et MTLR avec $\rho_X = 0.6$, $\rho_S = 0.4$, SNR $= 0.03$, pour différentes valeurs de $r$.

M. Kolar and J. Sharpnack. Variance function estimation in high-dimensions. In *ICML*, pages 1447–1454, 2012.

M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *AISTATS*, volume 84, pages 998–1007, 2018.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1): 127–152, 2005.

W. Ou, M. Hämaläinen, and P. Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, Feb 2009.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

J. Wagener and H. Dette. Bridge estimators and the adaptive Lasso under heteroscedasticity. *Math. Methods Statist.*, 21:109–126, 2012.