LaMME

To: Erricos Kontoghiorghes in charge of

From: Chiquet, Grandvalet, Ambroise

Subject: Response to CSDA review CSDA-D-17-00692 (e199)

Date: February 13, 2018

Dear editor,
We are grateful to the reviewers for their careful reading of our paper entitled

"Sparsity by Worst-case Penalties"

and their enriching remarks. This letter aims to provide a point-by-point response to their comments and suggestions.

Referee 1

* Reviewer 1 *** The algorithm starts from a sparse initial guess, i.e., the active set A = . I think it may be useful to discuss briefly how the computational time changes for an experimental setup, in which the choice of the initial guess is randomized.

**** Rponse (dans la lettre) :Yves: Yves: L'algorithme est efficace du fait que nous partons de rien. L'efficacit de l'algo repose sur le fait que nous ne rsolvons pas de gros systme et une solution randomise serait couteuse

Julien: Si l'on part d'une solution proche de l'optimale effectivement on rsout peu de systmes...

Yves: Une partie de la rponse est dans le warm start que nous avons utilis dans le protocole....

**** Rponse (ajouter dans le texte) :Yves: Dire que c'est important de partir de zro

*** In section 5, I would suggest to switch the results on simulated settings and on real-world data, such that the latter can be commented in light of the properties and conclusions obtained with synthetic data.

**** Rponse (faisable) :Christophe:

OK on fait

*** In the simulated data section, I would suggest to add the F-measure between the support of the true coefficient vector and the estimated one , as a performance measure to evaluate the selection properties of the different algorithms (see Section IV in Gasso et al., 2009).

**** La F mesure n'est pas super lisible, :Julien: donc on ne rajoute pas mais on met les FP et FN sur la figure 8

**** Rajouter la citation en citant la F-Mesure :Christophe:

@articlegasso2009recovering, title=Recovering sparse signals with a certain family of non-convex penalties and DC programming, author=Gasso, Gilles and Rakotomamonjy, Alain and Canu, Stéphane, journal=IEEE Transactions on Signal Processing, volume=57, number=12, pages=4686–4698, year=2009, publisher=IEEE

*** In line 244, :Christophe: :DONE: the reference should not be in parenthesis.

*** Figure 6 could be improved, :Julien: as the axes labels are not easily readable and colors are not distinguishable. *** In line 377, :Yves: **** Quadrupen, Spam_lars sont des mthodes du second ordre et ne sont pas sensible aux corrlation... au I would suggest to explain the reason why quadrupen, SPAMS-LARS and lars are not sensitive to the level of correlation between features. *** In the caption of Figure 8, specify the x-axis. :Julien:

**** C'est un log-10 (λ_1)

Gasso G., Rakotomamonjy A., Canu S. (2009). Recovering sparse signals with a certain family of non convex penalties and DC programming. IEEE Transactions on Signal Processing, 57(12), 4686-4698.

Referee 2

The authors proposed a new optimization algorithm to solve the sparse linear regression with a norm penalty term, including group-lasso and elastic-net. The proposed method is both accurate in estimation and extremely computationally fast, as shown in extensive

simulation studies. It is very good that the authors also provided the accompanying R-package (quadrupen) on CRAN.

The paper is interesting in computational aspects and fit the theme of the journal. However, I think a few concerns should be addressed before publication is possible. ** Major comments: *** 1. As the authors also admitted in the paper, the proposed algorithm is highly similar to the active-set methods, except the procedures taken in Step 3. I understand that the authors argue that quadrupen is more flexible in handling a wider range of penalties and look at the problem from a different perspective. However, either methodological innovations or numerical improvement should be proved to make the contribution enough to consider it as an alternative algorithm.

**** Rponse dans le texte Step 3 speeds up the algorithm .... that the main advantage....

*** 2. In the simulation studies, under the same tuning parameter, a more conservative selection of the quadrupen was observed in the QTL example comparing to that of glmnet, but not in the examples showed in Figure 8. I was wondering if this is an universal phenomenon? If it is the authors should provide some theoretical insights on why does this happen and maybe run some more monte carlo simulations to verify it, since basically they are solving the same underlying optimization problem.

**** Pour rpondre cela nous montrons les FP et TN et pour s'apercevoir que c'est toujours le cas.... **** Ajouter une phrase commentaire sur la figure 8 :Yves:

glmnet garde les petits coefficients qui ne coutent pas cher dans l'objectif

*** 3. For solving the elastic-net problem, the authors compared the proposed method to other two optimization strategies. It seems that the proposed method only gains obvious efficiency in the regime where 1 and 2 are very small. In the case if we choose parameters probably (according the theoretical rate or simply by cross- validation), what is the advantage of using quadrupen?

**** Prciser ce que l'on fait sur la figure 6 ...

**** Pas seulement sur les petites valeurs, c'est vrai que c'est l qu'il y le plus de diffrence mais pour des valeurs de lambda1 lambda2

**** Mais c'est justement sur les petites valeurs que c'est coteux.

** Minor comments: *** 1. The presentation of the figures in the paper is not very straightforward to readers. For example, I found Figure 1,2,3,6 kind of hard to understand by reading the captions. More basic introductions should be made in the captions or in the context describing the figures. Moreover, please pay attention to the size of fonts in the labs, titles and legends across the figures. Currently they are quite differing.

**** Plus gros caption :Yves:

*** 2. In the section 5.2.3 and 5.2.4, the authors benchmarked the methods by Lasso, letting 2 = 0. Is it possible to consider the more general elastic-net problem (say for a fixed 2) for at least some of the packages, especially in measuring the model selection accuracy? I understand that the authors have pointed out that they have different rules in determining tuning parameters.

**** Rpondre dans la lettre et le papier on ne parcours pas les młmes solutions de la młme manire.

Fixer lambda puis alpha pour glmnet Faire bouger lambda1 et lambda2 pour nous

Referee 3

The article proposes a novel algorithm, which supposedly solves the elastic- net problem and approximates the solutions to LASSO and l,1 version of group LASSO. The idea seems interesting but based on the current version of the article it is difficult to assess the properties and correctness of the pro- posed algorithm. Its description is very sketchy. Also, the paper lacks the results on its convergency. Additionally, many parts of the paper suffer from the lack of precision and contain wrong statements or unjustified claims. The list of detailed remarks is included below. *** 1. Equations (4) and the one above (4) give the wrong impression concerning the simplicity of the solution of problem (3). Namely, the solution of the inner problem (maximizing over or minimizing over) will depend on the value of the second parameter, which makes the outer optimization problem rather difficult.

**** Rponse dans la lettre: c'est comme les quation d'EM o les quations d'update sont simples et le problme globale est complexe :Christophe:

**** On explique la fin du paragraphe: et, defined by the extreme points of the convex polytope B1. This number of 87 points typically increases exponentially in p, but, with the working-set strategy, 88 the number of configuration actually visited typically grows linearly with the 89 number of non-zero coefficients in the solution .

*** 2. It is not true that the quadratic problem (5) is a different formulation of (3).

Under some circumstances it only asymptotically (for) approximates (3). In Section 3 the authors have shown that this asymptotic approximation works for LASSO and for l,1 version of group LASSO. Their derivations strongly rely on the simple form of l1 or l balls.

**** We agree with the reviewer, as stated in the paper line 100

It is hard to verify the authors statement that similar asymptotic approximations would work for other norms, like e.g. the Sorted L-One Norm (see OSCAR or SLOPE (Bogdan et al, AOAS 2015)). If this assertion is indeed true, I suggest to add a Section with a proper mathematical justification.

**** Ajouter OSCAR dans la lettre. :Christophe: Voir oscar.tex

OSCAR The sparsity inducing penalties can be adapted to pursue different goals, such as having equal coefficients. This was first implemented for ordered features with the fused Lasso (Tibshirani et al., 2005), which encourages sparse and locally constant solutions by penalizing the ℓ_1-norm of both the coefficients and their successive differences.

Even when there is no ordering between features, equality can be desired for interpretability purposes. OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) has been conceived in this siprit, to infer clusters of variables in a supervised setting (Bondell and

Reich, 2008). It is based on a penalizer encouraging the sparsity of the regression coefficients and the equality of the non-zero entries. By this means, correlated predictors that have a similar effect on the response form "predictive clusters" represented by a single coefficient.

$$\mathcal{H}_{\boldsymbol{\beta}^*}^{\text{oscar}} = \{\} \quad .$$

The dual assumption is that the ℓ_∞-norm of $\boldsymbol{\gamma}$ should be controlled, say:

$$\mathcal{D}_{\boldsymbol{\gamma}}^{\text{Oscar}} = \{\}$$

,

where η_γ is defined from η_β and **conv** denotes convex hull, so that Problem (**??**) reads:

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\gamma}\in\{-\eta_\gamma,\eta_\gamma\}^p} \left\{ \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \right\} \\ \Leftrightarrow &\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + 2\lambda\eta_\gamma \|\boldsymbol{\beta}\|_1 + \lambda \|\boldsymbol{\beta}\|_2^2 \quad , \end{aligned}$$

The lagrangian formulation of OSCAR as a constrained optimization can be expressed as

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \sum_{j=1}^{p} \left(c(j-1)+1\right) |\beta|_{(j)},$$

with $|\beta|_{(1)} \leq |\beta|_{(2)} \leq \cdots \leq |\beta|_{(p)}$. The penalty term can be expressed in a form close to our adverse quadratic penalty. Let us consider the adverse vector domain to be

$$\mathcal{D}_{\boldsymbol{\gamma}} = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^p | \boldsymbol{\gamma} = \begin{pmatrix} \alpha_1 1 \\ \alpha_2(c+1) \\ \alpha_3(2c+1) \\ \vdots \\ \alpha_p(p-1)c+1 \end{pmatrix}, \; c \in \mathbb{R}^+, \|(\alpha_1, \cdots, \alpha_p)\|_\infty \leq \eta_\alpha \right\}$$

and the permutation matrix

$$P_{\boldsymbol{\beta}} = \left\{ \mathbb{I}_{(\text{rank}(|\beta|_{(i)}=j)} \right\}_{i=1\cdots p, j=1\cdots p} .$$

We can reformulate the previous lagrangian as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\gamma}\in\mathcal{D}_{\boldsymbol{\gamma}}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2 + \lambda \|\boldsymbol{\beta} + P_{\boldsymbol{\beta}} P_{\boldsymbol{\beta}} \boldsymbol{\gamma}\|_2 \quad .$$

The rewriting of the initial problem allows to see that the very same optimization adaptive constraint algorithm used for the elastic net can be used to solve the OSCAR problem (see Table **??**).

*** 3. I suggest to mathematically formalize the Section 2.4 on the Geomet- rical Interpretation. Specifically, it is not clear at all that the solution belongs to the intersection of all the balls centered at B.

**** On ne comprend pas ce qu'il ne comprend pas... rponse dtaill de Yves

Note, that for small c and large , this intersection would be an empty set.

**** Oui mais on ne l'a pas pos comme cela... car c'est crit sous forme lagrangienne et c'est celle l qui nous intresse La vue gomtrique est intressante pour l'interprtation....

Also, please, note that depends on , thus maximizing ——() —— 1 does not seem to be a simple task.

**** Dans la lettre: :Christophe: We agree that it is not a simple task, this is why we adopt an alternate optimization point of view... It is an EM algo

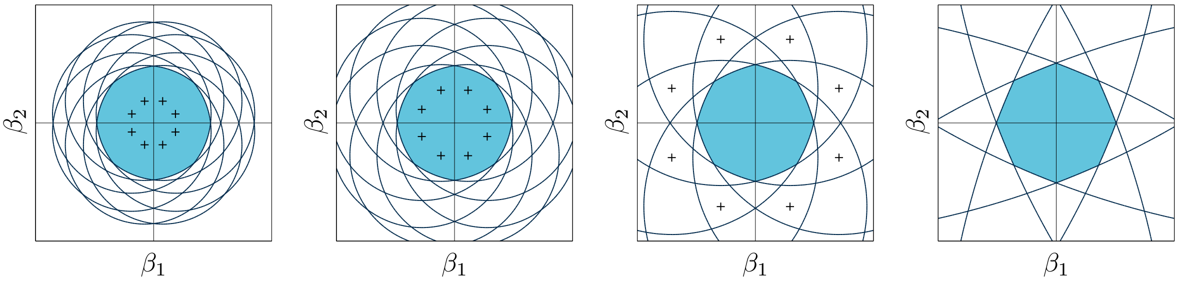

Figure 1: Admissible sets (patches) for the OSCAR, defined by the intersection of the Euclidean balls whose centers are represented by crosses and boundaries are displayed in black.

On the other hand the statement that for each , () belongs to the set of extreme points does not seem to require a special justification.

**** Dans l'article : changer les captions des figures 1 et 2 :Yves:

Figures 1 and 2 require an extended description.

It is not clear what is represented by different colors. Also, according to the description of Figure 1, it contains the graphs for Elastic Net, l and OSCAR, while according to the description in the text the first two graphs contain LASSO, and l,1 version of group LASSO.

**** Dans la figure 1 on reprsente des versions L2 ifes :Yves:

Where are crosses in graph number 4 in Figure 2 ? **** A l'extrieur de la figure :Yves:

*** 4. The formulation (5) is asymptotically equivalent to LASSO when . It would be good to provide some results (at least empirical) to show how the accuracy of this approximation depends on . Should good depend on p, the correlation structure or the sparsity of the signal ?

**** Rponse dans la lettre: :Yves: LA formulation 5 permet de mlanger L2 et autre pnalit lorsqu'il n'y pas de L2 : We always use formulation 3 to solve the problem

le lambda qui tend vers l'infini c'est une faon de voir les

*** 5. It is not clear at all if the first two steps of the algorithm converge. Note that () is selected as the extreme point, which is most distant to . On the other hand the quadratic penalty used to estimate () privileges close to . For large , when the penalty dominates the first term in the objective function, () will be very close to and will not longer be coherent (most distant) to . I would welcome some mathematical result illustrating the convergence of the first two steps of the algorithm.

**** Prouver la convergence de l'algo

*** 6. It would be good to explain what is the meaning of gj.

Did not we select as the most distant extreme point ? **** Yes but in the active set

Why do we calculate gj by minimizing over all points from the dual ball ?

**** We need to consider gj to add a variable in the active set

*** 7. I do not quite see how to formulate a stopping criterion based on Proposition 1. **** upper bound est le terme de droite :Yves:

In the algorithm the worst case gradients are computed for from the ball, while the Proposition 1 requires to be in the closed complement of the ball.

**** :Yves: dans la lettre

**** Expliquer que l'algo produit une squence de gamma non admissible et c'est pour ces solutions que nous cherchons avoir une borne sur l'optimum du critre...

*** 8. Real Data Analysis - comparison of LASSO and quadrupen. There is no information which was used here.

**** We do not use eta see above :Christophe:

It is quite natural that glmnet and quadrupen give different results since quadrupen is only the approximation to LASSO.

**** Not true, it is the same problem thanks to formulation 3

How do these two algorithms compare in execution time for this relatively large data set ?

****

*** 9. The synthetic data examples are rather small (p = 100). It would be good to see the results when p can reach several thousands (say 5000). Which was used for quadrupen ?

****

The authors use LARS as the benchmark. Please, note that LARS is only an approximation to LASSO. For example LARS paths are mono- tonic, which is not true about the LASSO path (here variables may appear and disappear many times along the path). Thus, in general, LARS does not solve the LASSO optimization problem and should not be considered as the true optimum.

**** We use LAR the package implements the lasso :Julien:

Instead of comparing D(method), which is never negative, even if a given method is better than LARS, I suggest to directly compare the values of the objective function.

****

References

Bondell, H. D., Reich, B. J., 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64 (1), 115–123.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1), 91–108.