Referee report on CSDA-D-17-00692

Sparsity by Worst-Case Penalties

by Y. Grandvalet, J. Chiquet and C. Ambroise

The article proposes a novel algorithm, which supposedly solves the elastic-net problem and approximates the solutions to LASSO and $l_{\infty,1}$ 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 proposed 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 γ or minimizing over β) will depend on the value of the second parameter, which makes the "outer" optimization problem rather difficult.

2. It is not true that the quadratic problem (5) is a different formulation of (3). Under some circumstances it only asymptotically (for $\eta \to \infty$) approximates (3). In Section 3 the authors have shown that this asymptotic approximation works for LASSO and for $l_{\infty,1}$ version of group LASSO. Their derivations strongly rely on the simple form of l_1 or l_∞ balls. 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.

3. I suggest to mathematically formalize the Section 2.4 on the Geometrical Interpretation. Specifically, it is not clear at all that the solution belongs to the intersection of all the balls centered at $\gamma \in B_\star^\eta$. Note, that for small c and large η, this intersection would be an empty set. Also, please, note that $\hat{\beta}$ depends on γ, thus maximizing

$$||\hat{\beta}(\gamma) - \gamma||$$

does not seem to be a simple task.

On the other hand the statement that for each β, $\hat{\gamma}(\beta)$ belongs to the set of extreme points does not seem to require a special justification.

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 $l_{\infty,1}$ version of group LASSO. Where are crosses in graph number 4 in Figure 2 ?

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

5. It is not clear at all if the first two steps of the algorithm converge. Note that $\hat{\gamma}(\beta)$ is selected as the extreme point, which is most distant to β. On the other hand the quadratic penalty used to estimate $\hat{\beta}(\hat{\gamma})$ privileges β close to $\hat{\gamma}$. For large λ, when the penalty dominates the first term in the objective function, $\hat{\beta}(\hat{\gamma})$ will be very close to $\hat{\gamma}$ and $\hat{\gamma}$ will not longer be coherent (most distant) to $\hat{\beta}$.

   I would welcome some mathematical result illustrating the convergence of the first two steps of the algorithm.

6. It would be good to explain what is the meaning of g_j. Did not we select γ as the most distant extreme point ? Why do we calculate g_j by minimizing over all points from the dual ball ?

7. I do not quite see how to formulate a stopping criterion based on Proposition 1. In the algorithm the worst case gradients are computed for γ from the ball, while the Proposition 1 requires γ to be in the closed complement of the ball.

8. Real Data Analysis - comparison of LASSO and *quadrupen.*

   There is no information which η was used here. It is quite natural that *glmnet* and *quadrupen* give different results since *quadrupen* is only the approximation to LASSO. 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 η 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 monotonic, 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.

   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.