Response to Reviewer 1

December 14, 2016

We appreciate the helpful feedback from the reviewer. We have addressed your questions and comments. Below we give a point-by-point response to each of the questions posed by the reviewer:

Overall comment

Although the idea of a using a larger set of regularization parameters is interesting, the empirical results are incomplete and including additional scenarios would help strengthen the paper

We have included more details in the empirical results and included a new example (I don't know what though).

Specific Suggestions/comments

1. Can the authors point to examples in the literature where a large set of regularization parameters was used?

   We have added more examples to Section 1.

2. First line on p. 13, "the optimal regularization parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$" should be "the optimal regularization parameters $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, ..., \lambda_M)^\top$"

   We have corrected this typo.

3. The authors note that the models considered in Sections 2.4.2 and 2.4.3 are usually employed with only two regularization parameters but propose using a larger set of regularization parameters. In the corresponding simulations in Sections 3.2 and 3.3, they compare using a larger set of regularization parameters against using two regularization parameters selected using grid search. The grid spaces considered are fairly small, so its not clear if the improved performance for gradient descent is due to the additional regularization parameters or due to the dependence of performance on the grid space. To provide additional insight into whats causing the difference in performance, could the authors also present results using gradient descent, Nelder-mead, and Spearmint for the two parameter case?

   We have added gradient descent, Nelder-mead, and spearmint for the two-parameter case for the examples in Sections 3.2 and 3.3.

4. The authors report average performance and standard errors for the simulations done in Section 3. How many simulation runs were used in each example?

   We have added a sentence regarding the number of simulation runs in the beginning of Section 3.

5. In Section 3, two different starting values were considered for Nelder-mead and gradient descent. How sensitive were the results to the choice of starting values?

We considered four initializations for Nelder-mead and gradient descent for the un-pooled sparse group lasso which resulted in smaller validation error, but not significantly so. Perhaps more importantly, the average validation error attained by gradient descent remained smaller than that compared to Nelder-mead. We would like to keep the paper restricted to two initialization points however since it shows that one does not need very many starting points. In the setting where there are many penalty parameters, the number of possible initialization points grows exponentially; the point of the paper is to show that the computational load does not grow exponentially with the number of penalty parameters (in contrast to grid search). We have added a comment regarding sensitivity of the methods to their starting values in Section 3, as well as an explanation for why we only use two initialization points.

6. The empirical results for gradient descent depend on α, β, and δ. The authors mention ranges considered for these parameters in Section 2.5. How were these parameters ultimately selected in the evaluations in Sections 3 and 4?

   Since the results were not sensitive to the choice of α,β, and δ, we have set the values to be the same in all simulations. We have specified their values in Section 2.5.

7. In Sections 3.2 and 3.3, the authors created a training, validation, and test set, but in Section 3.1 they only consider a training and validation set. Why was a test set not considered in Section 3.1?

   The goal of Section 3.1 was just to illustrate that gradient descent was able to find a validation loss that was nearly the same as grid search. To streamline the paper and to make the empirical results more compelling, we have included a test set to Section 3.1.

8. Table 1 should note that standard errors are given in parentheses.

   This has been corrected.

9. For the simulations in Section 3.3, why did the authors set n = 60 in the first case and n = 90 in the other two cases?

   We have changed the training size for all simulations in Section 3.3 to $n = 90$ for consistency.

10. g is undefined in Table 4

    We had meant for g to denote the true number of groups, but this was already given in equation (?). We have removed it.

11. In Table 4, could the authors provide intuition for why there is a large difference in validation error and test error for gradient descent?

    We have added a paragraph in Section 3.3 (check this again later) to provide some intuition for the observed behavior.