Author’s response to reviews

Title: Sparse Generalized Linear Model with L0 Approximation for Feature Selection and Prediction with Big Omics Data

Authors:
Zhenqiu Liu (liuzx@cshs.org)
Fengzhu Sun (fsun@dornsife.usc.edu)
Dermot McGovern (dermot.mcgovern@cshs.org)

Version: 1 Date: 17 Nov 2017

Author’s response to reviews:

11/17/2017

BioData Mining

Dear Editor:

We have attempted to address the reviewer’s concerns for our paper entitled “Sparse Generalized Linear Model with L0 Approximation for Feature Selection and Prediction with Big Omics Data”. The revised manuscript is attached. We want to thank the reviewer and Associate Editor for their constructive comments. The point-to-point responses to reviewer’s comments are below. Thank you.

We look forward to hearing from you soon.

Best regards,

Zhenqiu Liu

Author’s responses to reviewer’s comments on manuscript (BIDM-D-17-00015) entitled “Sparse Generalized Linear Model with L0 Approximation for Feature Selection and Prediction with Big Omics Data”

We appreciate your constructive comments on our paper. Our point-to-point responses are given below, where the numbering corresponds to the paragraph numbering in your report.
Response to Reviewer 1:

1. Algorithm: the initialization of eta is not clear. If eta = beta = 0 as the author stated, the solution will be stuck in beta=0.

We have set the initial values with a nonzero random vector with very small values on the top of page 7. Many thanks. By the way, except for linear regression, the initialization of eta = beta = 0 seems to be worked for other glm models including Poisson and logistic regression. Thanks.

2. Algorithm: based on the reweighting algorithm, is it possible that the algorithm will be stuck in a local minimum? Will it be sensitive to a different starting point of eta?

Theoretically, the approximation for L0 is non-convex, it is possible to stick in a local optima with some initial values. However, our proposed algorithm approximates L0 with sequential convex optimizations, it is pretty robust with randomized initializations from our simulations on the bottom of first paragraph of page 12.

3. Algorithm: it is not obvious to the reviewer how the sparsity is achieved using a ridge-type regression in each step. It seems that a final step of thresholding is needed to set these very small coefficients to 0's. Please state it explicitly.

Great point. We have added a couple of sentences to discuss the sparsity of the model and state the threshold explicitly at the end of first paragraph on page 8. Thanks.

4. Lemma: the lemma has not been rigorously proved in the reviewer's opinion: the proof is more of an intuitive justification of the proposed iterative algorithm.

Thanks for pointing out that. We have changed the Lemma to Algorithm Justification.

5. Simulations: the selection of lambda should be based on the same criteria (AIC and BIC, or CV) when comparing different methods (table 2).

We redid the simulation with BIC criteria (log(N)) for all model in Table 2, and demonstrated that L1, SCAD and MC+ need a larger penalty than BIC to achieve the similar sparsity in the supplementary Table 1.

6. Simulations: I would suggest comparing to L1 regression using glmnet, since most applications are still using L1 penalization due to the super efficiency of glmnet.

We added the glmnet with the efficient path algorithm for L1. We demonstrated that L1, SCAD and MC+ did not perform well with either BIC or cross-validation, and pointed out that many
popular packages including the commercial MATLAB usually choose a larger lambda one standard deviation above the minimum test error with cross-validation, which is arbitrary and leads to larger bias. To overcome such bias in parameter estimation, some packages re-estimate the parameters with the selected features and standard GLM model. Unlike these methods, our proposed method performed better without any postprocessing. The discussion was on the top paragraph of page 12.

7. Simulations: ANSF and PTM criteria are good. The evaluation will be strengthened if some measure of false positive control (e.g. empirical FDR) is also reported since one of the aims is to select the gene signature.

FDR was reported both on Table 2 and Table 3. Thanks.

8. Application: the authors may consider reporting the coefficients (magnitude) for the selected features. This will also guide the design of the simulations.

As suggested, the coefficients of the selected features are reported in Table 2 of the supplement file.

9. Application: the evaluation (based on AUCs) may be too optimistic since the filtering (P < 0.01) seems to have been conducted before the model training. If the AUCs are reported, I will suggest that screening/filtering and tuning parameter selection are both performed on the training data sets and evaluation on the test data set (which has not been used for feature screening and tuning parameter selection).

We redid it with the screening for training data only. The results did not change much. If I understand correctly, screening first will exclude the discarded genes in model selection. It is biased, because those genes filtered out will never be included in the model, and the search space becomes smaller, but the test AUC will probably be fine. Thanks.

10. Application: it is interesting to compare to other methods. For example, compare to the simple univariate testing procedure (coupled by FDR control), and the L1 penalization procedure to demonstrate the benefit of the proposed procedure.

We reported the results of prediction with the same number of top genes from statistical test On Table 2 of the supplementary file, and also briefly discussed the results at the end of the first paragraph on page 14. Thanks.

Minor revision:
1. P3 L21-28. It is not clear how these applications are related to the proposed method. Need to rewrite.

As suggested, we rewrote the second paragraph on page 3 to make the link of the proposed method and application more smoothly. Many thanks.

2. "Debulking" should be defined since readers may not be familiar with it.

As suggested. Suboptimal debulking (cytoreduction) is defined at the beginning of the second paragraph on page 3.