response_major_revision.Rmd

Shiqiang Jin

First author comment

- 1) In Algorithm 3, the step 5 should be explained more clearly. Here, I am not sure if:
  - 1.1 we draw a single additional element from N_+ with some probability P and do not draw any additional element with probability 1-P or
  - 1.2 we draw a single additional parameter from N_+ and the probability of selecting a specific element depends on the marginal likelihoods of models with additional elements and alpha.

Probably, the latter case is true (because the former case does not make too much sense) but the authors should explain it more clearly in the paper.

Answer:

- 2) The authors argue that they "establish model selection consistency for the proposed method in the high dimensional setting". However, there is no proof that Algorithm 3 selects the best subset of predictors as $T \to \infty$. Contrary, I can think of the situation where the search procedure does not have a chance to select the best subset of predictors. Imagine that we have 6 potential predictors $x_1, x_2, x_3, x_4, x_5, x_6$ where x_1, x_2, x_3 are all strongly correlated with the dependent variable y and each predictor in $\{x_4, x_5, x_6\}$ is only weakly correlated with y, but $x_4 + x_5 + x_6 = y$. We want to select the best 3 predictors and we start the search algorithm by selecting x_1, x_2, and x_3 in the first step. Under certain circumstances, the procedure will never have a chance to compute the marginal likelihood of the model containing x_4, x_5, x_6 as predictors. If the new algorithm allowed for selecting the subset of "worse" predictors at the end of its stochastic part the problem would disappear. The inability to select the specific predictors may be a huge problem when a large number of predictors is considered and we want to identify the most significant predictors, as shown in the real data application. The standard stochastic search algorithms, including the MC3 algorithm, always find the best subset as $T \to \infty$.

Answer: The stochastic step is a MCMC process, obtaining the limiting distribution as $T \to \infty$. Then the mode of it is the best subset. For the second question, we need to run a simulation with this setting to see what's gonna happen. On the other hand, in the real data scenario, this situation is rare and does not possibly happen.

- 3) The idea of a combined deterministic and stochastic procedure can probably be extended to frequentist model selection approaches. In the frequentist model selection algorithms, the Akaike, Schwarz, Mallows, or cross-validation criteria are computed and they all use similar formulas to m_+ and m_-. The authors could discuss extension of their algorithm to frequentist model selection.

Answer: Yes, the hybrid search of the proposed method can be easily extended to the frequentist method. The modification is replacement of model selection criterion. But the algorithm that makes use of the Sherman-Morrison formula to reduce the computational burden cannot be extended to frequentist methods.

- 4) Some statisticians argue that mechanical model selection techniques have their disadvantages and the appropriate misspecification tests should be applied to select the best predictors in the model (e.g., Spanos, 2010). I would like to see how the misspecification testing can be incorporated in the model selection algorithm.

Answer: Adding this technique may imporve the algorithm, but it also may lose our focus of the proposed method – hybrid search and the use of the Sherman-Morrison formula.

- 5) The main advantage of the new algorithm is the reduction of the computational burden associated with searching for the best subset of predictors. The authors should provide some statistics presenting the improvement in the number of steps and/or the time required to find the optimal subset in comparison to other methods.

Answer: yes, maybe we need to run the simulation again and record the time it cost.

- 6) Some editorial comments: the expressions "MCP", "LASSO", and "SCAD" should be explained, the expression "setting that p>n" should be "setting where $p > n$" (page 4). Answer: yes, change it to setting where $p > n$.

Reviewer #2:

I have two general comments on this very interesting contribution and suggest to resubmit the paper containing some extended analysis of proposed algorithm.

- 1) Extension of the simulation study:
  - why y has unit variance? How interpret prior parameters a_sigma and b_sigma in this case? How results change in the case of estimation for multivariate gaussian Data Generating Process with different relation of the variance of the error term and the variances of the processes generating regressors?

Answer:

i) y can be other variance values. We think if we conduct a fair comparion with other methods (MCP, lasso), the value of variance of y is not important. Besides, when we do the real data analysis, usually we studentize the data first. This make y becomes unit variance.

ii) the hyperparameters a_sigma and b_sigma are set to make the distribution of σ^2 become flat. $a_\sigma = b_\sigma = 1$

iii) I don't completely understand the question. Is it a hetergeneous variance?

- How algorithm works in heteroscedastic or generalized regression environment. I would encourage Authors to conduct more extended simulation study focused not only on the simple linear regression, but on some generalized cases with heteroscedasticity or correlation in the vector of error terms.

The hybrid method can be extended to the GLM but not the Sherman-Morrison formula. Since two auther comment about this part, should we extend our hybrid idea to the GLM cases? Or we can comment that our 2nd and 3rd projects are targeting these issues.

- Some details concerning posterior analysis about inference on variance of the error term should be presented and discussed. Perhaps there is a linkage between ranks of analysed algorithms and inference about parameter of the stochastic structure.

- 2) Extension of the empirical study: Why empirical part is not conducted for the whole dataset? The progress in search methods developed for model selection (or model averaging) should avoid the step that Authors call the pre-screening procedure. What is the purpose of selecting much smaller set of regressors, that are marginally correlated with endogenous variable. Does final results are invariant with respect to this step?

Answer: I remember at first we try to use the whole dataset, but my computer's memory (8 GB) is too small to store the huge matrix. Maybe we can use beocat to run the whole dataset and compare it with result with pre-screening procedure.