Revision

Manuscript: AOAS1906-007

Date: 2020-04-07

Editor

Thank you for submitting your manuscript to The Annals of Applied Statistics. Two reviewers, and associate editor, and I have reviewed it carefully. Reports are given in the links below.

. . .

To summarize some of the key points:

1. `Questions about modeling assumptions and their impact: There are several questions about the modeli`

our response

2. `Comparison to other methods in real data and simulation:`

   The comparison with existing methods needs more details, with more detailed discussion of results and better description of simulation construction as well as description of more and less difficult scenarios. Also, the comparison should be extended to the real data analysis and results discussed, highlighting what if anything was detected by the authors approach that would not have been found with alternatives.

our response

Reviewer 1

Question about number of clusters at each condition

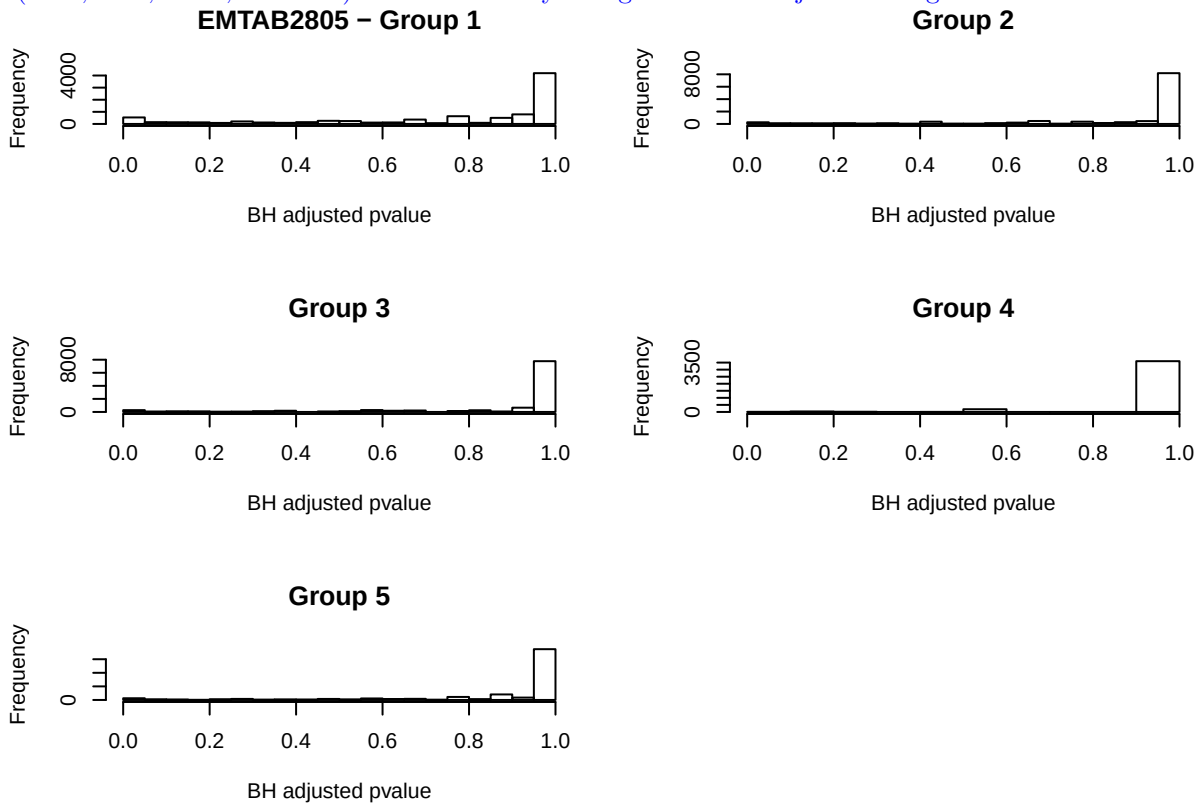
(1) One key assumption they made is that the number of cell clusters in both conditions is the same (K). This does not allow the scenarios where different conditions may have different cell subtypes (e.g., normal versus controls). Of course, if the cell types are different, then all genes should have different distributions. Should not they first decide whether this is the case?

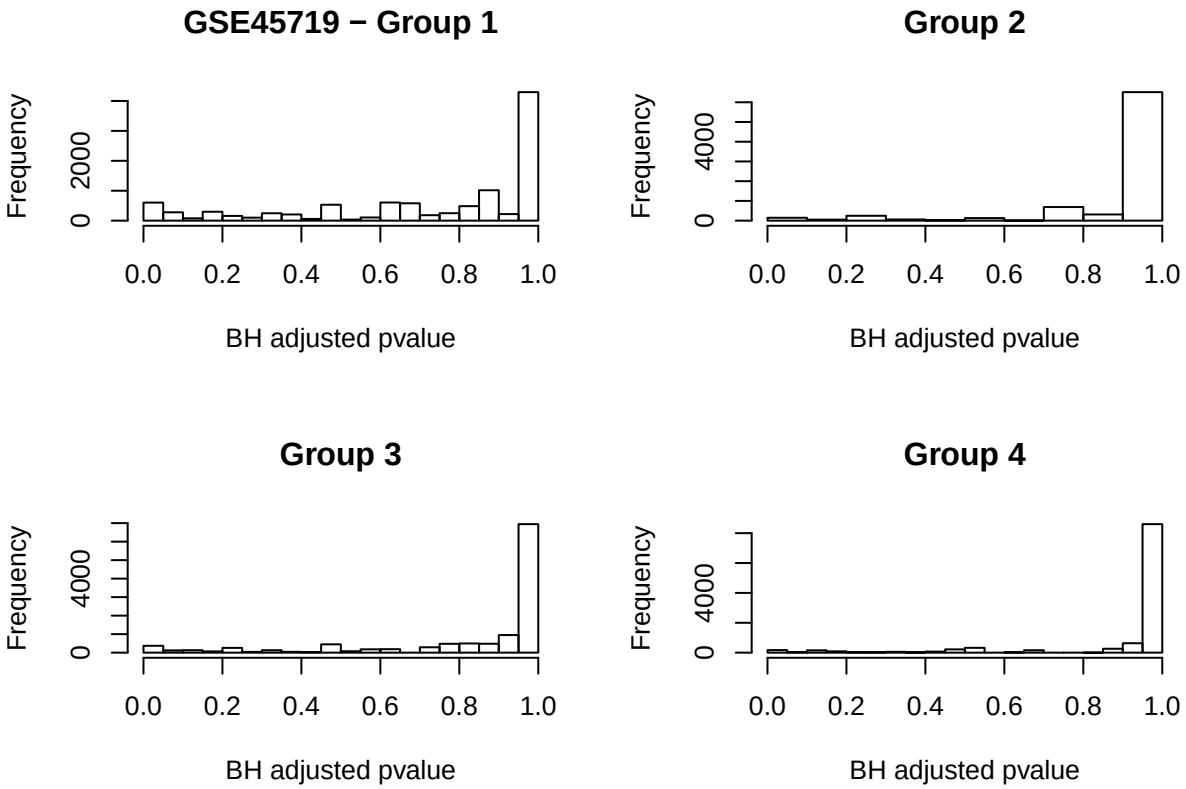
It is not an assumption for number of clusters to be same across two conditions. Our method pooled cells from both conditions first and identified a global clustering for all the cells, the number of clusters (K) is for the global clustering. These two conditions can have different number of clusters given that one condition can have zero proportion of cells from a cluster. For example, $K = 3$, there are 3 clusters (labelled as A, B, C) globally. Condition 1 having 3 types, 60% from A, 30% from B and 10% from C, while condition 2 only have 2 types 50% from B and 50% from C. It is not necessary that different cell types will induce different distributions over the whole genome. Actually the major proportion of genes should still have equivalent distribution. Which is key observation we proposed in the paper and we propose an empirical Bayesian framework with a specifically constructed prior to handle the scenario when both means and proportions changed. **add a sentence to clarify on page 6 middle; also note language on page 7 to deal with this case**

Questions about fitting data with mixture of NB

(2) Single cell RNA-seq data often have lots of zeros. I was wondering how well the mixture of multinomial distributions really fit the data. Some plots that show the model fits would be useful.

We are using mixture of negative binomial to model the counts instead of multinomial distributions. Mixture of NB are flexible to approximate a lot of distributions. Specifically, NB can approximate constant 0 by arbitrary accuracy. As we know the density of $P(X = 0) = 1 - p$ and $P(X > 0) = p$ given $X \sim \text{NB}(1, p)$. Further, scRNA-seq data tends to be overdispersed so that many people are using NB to model it. (cite DESeq2, cite Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression). Also, the empirical data (EMTAB2805, GSE45719) showed that more than 80% of genes their nonzero counts are overdispersed. In addition, given our estimated group labels, we test whether there is a strong evidence to reject fitting NB model at each group. Given the cells from same group, we excluded those genes with more than 70% of 0s, as we know NB can approximate 0 well. For the remaining genes, we do the goodness of fit tests following the procedure proposed in the paper "Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation(Guosheng Yin,Yanyuan Ma)" (2013, EJS, vol 7, 412-427.). We found very few genes would reject the negative binomial model





Questions not sure

(3). For the transcriptional bursting data analysis,
some comparisons with other methods such as MAST, DESEQ2, scDD would be useful, as they did for synthetic data sets.

Reviewer 2

Questions not sure

1.How does the model behave when both the proportions and the means are different? The calculation p7 (before the key issue) should be detailed. What hypotheses on $f_{g,k}$ allow this result ?

When both the proportions and the means are different, the distributions of the two conditions can be the same or different. Namely, for those subgroups having the same mean, even the proportions changed as long as their aggregation remained the same across conditions, it will not lead to differential distributions. Theorem 1 gives the sufficient and necessary condition for what kind of change of means and proportions in order to have the same distribution.
We assume the only parameter differs $f_{g,k}$ for k is the mean. That is for the negative binomial model, we have subgroup dependent mean parameters and shared shape parameter.

2.It is assumed that the shape parameter is constant and independent to the population classes. Is it a realistic hypothesis ? What is it impact on the results ? Is it possible to relax this hypothesis to consider a shape parameter depending on k ? A discussion on this hypothesis should be added in the discussion.

For the fixed shape parameter of NB among groups, we performed tests on each gene.
The testing procedure is following: the null hypothesis is that all groups having the same shape parameter. Assuming we have K groups. Denote the distribution of transcripts at group i, gene g as $f_{g,i}$. Under the null,

$f_{g,i} \sim \text{NB}(r, q_i)$. We pool the data and use the joint likelihood to obtain the MLE and condifence interval of r.
Under alternative, $f_{g,i} \sim \text{NB}(r_i, q_i)$. We obtain the MLE of the group specific r_i. If there is at least one MLE of r_i being outside of the 95we perform two sample t-test on whole genome with FDR correction, 544 out of tnose 555 genes have adjusted p value smaller than 0.05. We conclude that for the majority of genes, it is ok to assume a shared shape parameter across groups and even when the model is inappropriate, we are likely not to introduce higher false discovery rate. We also run the same procedure on GSE45719. We found 1447 (3%) genes having been flagged and 977 of them have estimated posterior of being DD > 0.95. Similarly, we perform t-test with FDR correction to found that 892 among those 977 genes also have adjusted p value $<$ 0.05. I would say overall the proportion of genes may be affected by the same shape parameter assumption is low and when the same shape parameter assumptipon is not appropriate, it is unlikely we lose the control of false discovery rate. It is possible to extend with group dependent shape parameter, but for the shape parameter, the density involves gamma function of it, which make us difficult to find an explicit form of the posterior of shape parameters. **maybe add a couple of sentences to the concluding remarks as an area for further experiment, with results noted and added to supplement**

Question about sensitivity to clustering

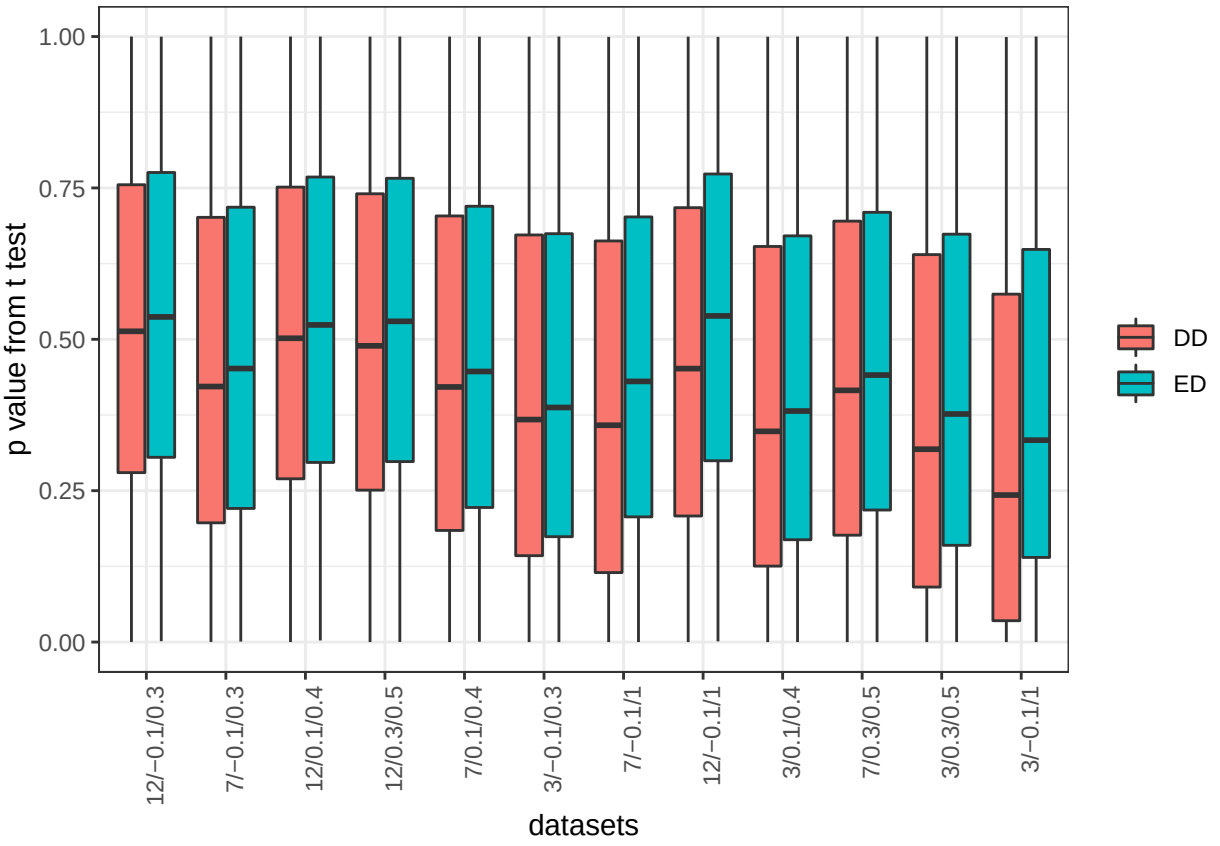
3.For the clustering, the authors should better justify their choice. The results seem to be strongly related to the clustering method.

We have tried different clustering method, sc3 on 3 datasets, (EBTAB2805, GSE45719, GSE79102). the correlation coefficients of the local fdr are 0.92,0.94,0.96. Our method is not very sensitive to the clustering methods.

Questions for simulations

Finally for the simulation part, the data generation should be explained and it would be nice to have criteria evaluating the difficulty of the simulated datasets. For example $K = 12$ seems very difficult, and $K = 7$ also (On Figure S7, some ROC curves are closed to the first bisector. Moreover on this plot, configurations are not precised).

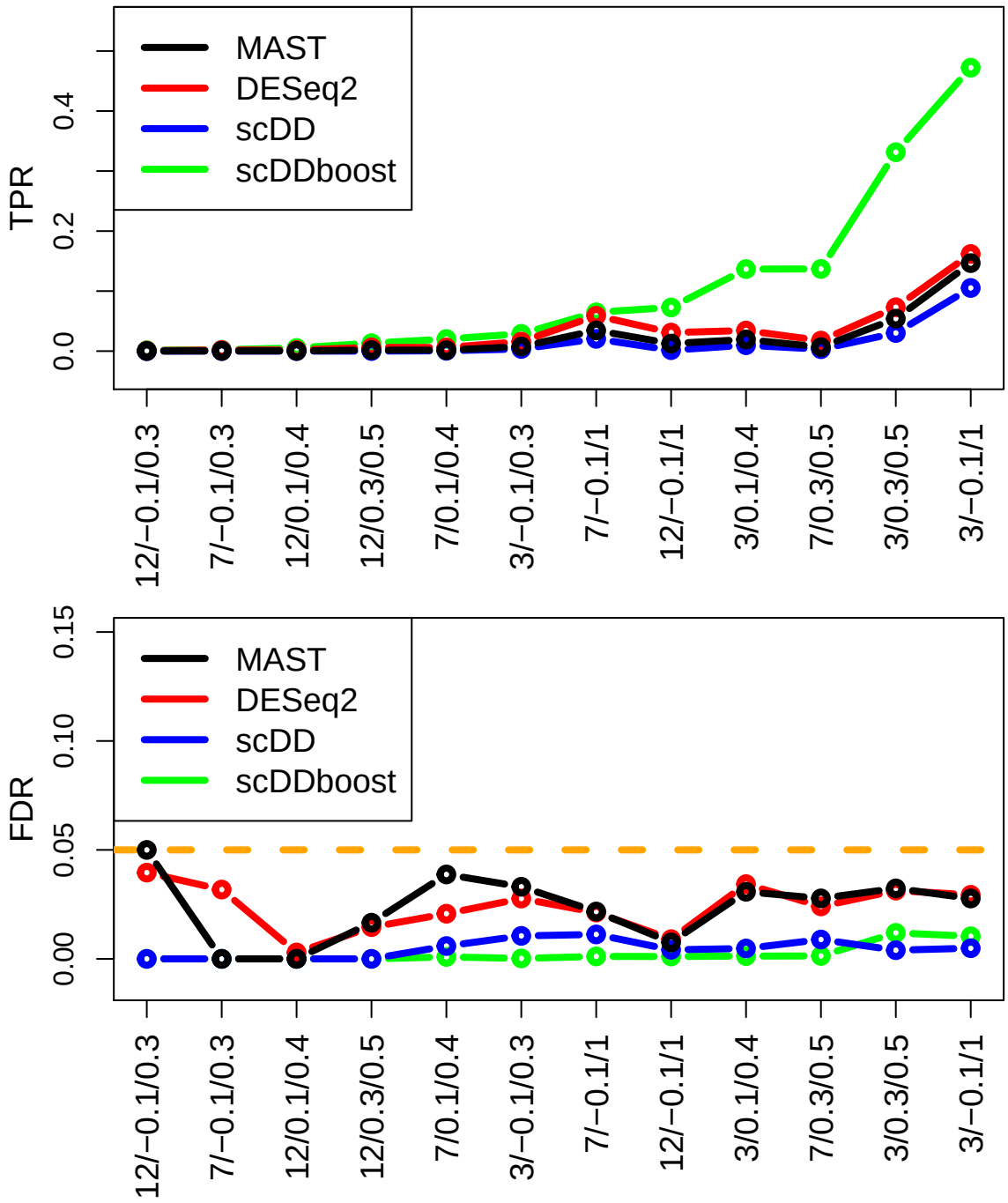
More detailed information of simulation has been placed in the supplementary material, those pca plots showed that groups are collapsed together when there are more groups, which make all the methods difficult to detect DD genes. In addition, we have fixed number of cells(400) so more number of groups will decrease the number of cells per group which makes it harder to detect the change between groups also mixing of more groups weaken the overall signal between conditions. Further to show the difficulty of each simulation settings, we do a t-test on each gene and present a boxplot for every simulation settings. It is easier to detect DD genes if they having small p values compare to those of ED genes, we ordered the boxplot by the same order of the simulation graph in the paper (Fig 4)



Questions for simulations

The results on Figures 4 and 5 are questionable. ScDDboost has the best TPR and a FDR closed to 0, whereas this latter should be controled at 5%. DESeq2 seems to better control the trade-off. Can the authors comment this remark ? In general, I am very surprised by the very small number of replicate datasets per scenario. Is it possible to increase it and to use boxplots and ROC curves to summary the results instead of one figure for the TPR and one figure for the FDR ?

We have done more replicates for each simulation setting, now we have 10 replicates under each settings. And we have similar results as before.
Our method is not a trade-off for more power by inflating the FDR. The splatter simulation having the subgroup structure for cells, which make our model appropriate to be applied


TPR
0.0
0.2
0.4
MAST
DESeq2
scDD
scDDboost
12/−0.1/0.3
7/−0.1/0.3
12/0.1/0.4
12/0.3/0.5
7/0.1/0.4
3/−0.1/0.3
7/−0.1/1
12/−0.1/1
3/0.1/0.4
7/0.3/0.5
3/0.3/0.5
3/−0.1/1
FDR
0.00
0.05
0.10
0.15
MAST
DESeq2
scDD
scDDboost
12/−0.1/0.3
7/−0.1/0.3
12/0.1/0.4
12/0.3/0.5
7/0.1/0.4
3/−0.1/0.3
7/−0.1/1
12/−0.1/1
3/0.1/0.4
7/0.3/0.5
3/0.3/0.5
3/−0.1/1