Response to the AoAS review

Manuscript: AOAS 1906-007

Title: A compositional model to assess expression changes from single-cell RNA-Seq data

Authors: Ma, Korthauer, Kendziorski, and Newton

Date: 11th August 2020

1 Overview

We thank the editor, associate editor, and both referees for the critical assessment of our manuscript, and we are delighted that a suitably revised version may be considered appropriate for publication in AoAS. We prepared a number of additional numerical experiments and have added a new subsection (3.5 Diagnostics) in order to clarify the workings of the methodology related to the useful points raised in review. The editor has also noted how essential is a real data application. Our motivating example (Figure 1) concerns a problem of stem-cell biology. While the present paper is not the primary report on this data set, the case is especially interesting because the gene-level effects are so subtle that they allude existing approaches and yet these effects are confirmed in separate experiments. In the following we describe our response to the review and how the manuscript has changed. We very much appreciate the opportunity to further improve the presentation of our work.

We indicate editor/reviewer comments in italics and our responses in non-italic, blue font. In the revised manuscript, changes are indicated in a non-italic, red font.

2 Response to Editor

To summarize some of the key points:

1. *Questions about modeling assumptions and their impact: There are several questions about the modeling assumptions made that were raised by reviewers in terms of aptness, flexibility, and sensitivity. These include the choice of clustering method, assumption of same number of clusters in both conditions, common independent shape parameters, model fit of mixture especially considering the abundance of zeros. These assumptions should be better justified, sensitivity assessed and where possible/appropriate the model allowed to accommodate the extra flexibility.*

On modeling assumptions and impact:

The reviewers raise important questions and we have done a number of additional computations to address them. In summary, our modeling choices respond to characteristics of scRNA-seq data and to opportunities for improving power through clustering. In some cases there are straightforward alternatives, such as on the distance function provided to the clustering algorithm, and in such cases we provide both a reasoned argument for the choice taken as well as numerical experiments investigating alternative choices. In other cases we are guided by computational feasibility and analytic tractability. For example, the use of negative binomials and the constant-shape assumption is somewhat pragmatic. Mixtures of these negative binomials provide a very flexible model class and have widely demonstrated utility for sequence-count data; and by using constant shape we are able to plug directly into an existing Bioconductor package (EBSeq) in order to produce needed probabilities. It would be onerous to develop a non-constant-shape method, but of course we can (and do) investigate the operating characteristics of the standing method when the shapes in the generative process are not constant, for example. Overall, the expanded numerical experiments reported in the revision offer evidence of satisfactory operating characteristics of scDDboost in the motivating stem-cell example as well as in a range of scenarios we might reasonably expect to see in practice.

(a) *choice of clustering method*: scDDboost gains a power advantage over other methods by leveraging subtype information that is revealed through cell-level clustering. Such clustering utilizes each cell's entire expression profile. The choice of clustering method does affect the reported output, but several facts are relevant:

   i. The type I error rate of scDDboost is controlled in any case, as evidenced by permutation calculations and simulation studies. Different clustering methods have different capacity to reveal the underlying cellular subtypes; inaccuracies here may diffuse the shrinkage effects (information sharing) between genes, but numerical experiments show that these inaccuracies do not inflate the false discovery rate (FDR). The negative control permutation study (Figure 7) was repeated under the SC3 clustering method and showed no excess in discovered gene calls (Supplementary Material, Section 3.5). Further, we include both SC3 and the default clustering in the simulation and also find control of FDR in this case where there is a positive non-null fraction (Figure 5; Supplementary Figure S15.)

   ii. The scDDboost software is modular. The default code uses a specific distance matrix balancing robustness and efficiency, but other cell-cell dis-

tance matrices may be provided on input. Furthermore, by averaging over results from randomized distance matrices, scDDboost gains stability and is not reliant, for example, on a single partitioning of the cells into estimated subytypes.

iii. We report numerical experiments using both the default clustering and the SC3 clustering, with a comparison in the revision, Section 3.5.

(b) *same number of clusters per condition*: The proposed method assumes that an overall number of cellular subtypes may be utilized by tissue at various proportions in the two cellular conditions. It does not assume the same number of clusters appear in each condition, but rather that the number K of allowable subtypes is the same. This is not restrictive in the sense that the frequency of a given subtype may approach zero in one condition, and the union of subtypes appearing in the two conditions serves to inform the total number K. The point is discussed in the revision, pages 6-7.

(c) *common shape parameters*: The theoretical formalism does not require the assumption of constant shape per gene, but we adopt this assumption in numerical studies in order to utilize existing codes for the mean pattern probabilities. Initial calculations (Figure 7; original and revision) demonstrate at one level that the type-I error rate is not adversely affected, in the sense that the FDR computed from multiple random permutations is well controlled. We performed additional numerical experiments on the impact of the constant-shape assumption. Section 3.5 (revision) reports that a formal hypothesis test of the constant-shape assumption often fails, but that the empirical effect of this failure is not substantial. In a number of examples, most of the genes showing non-constant shape that are DD by scDDboost are also called (or nearly called) by other assessments (MAST, DESeq2, scDD, t-test), (Supplementary Table S4; Figure S14). Importantly, the splatter tool that we use to generate data in the simulation study does not respect the constant shape assumption, and yet scDDboost demonstrates favorable sampling properties in this study.

(d) *model fit, considering 0's*: Though certainly restrictive compared to a non-parametric model, finite mixtures of negative binomials represent a relatively flexible model class for gene-level data. For example, having one mixture component with mean near 0 can represent genes with substantial zero counts. We investigate model fit more thoroughly in the revision, Section 3.5. For several data sets we cluster cells and then deploy a formal test of the negative-binomial model within clusters (where the unmixed negative binomial is presumed to hold). We report that relatively few genes show evidence against NB within

subtype, and that even for these there is general agreement between scDDboost and other methods on inference about differential distribution. Additional details of this analysis are included in Supplementary Material Table S3 and Figure S12. As suggested in the review, we include additional plots showing the goodness-of-fit of the overall model in the stem-cell example (Figure 10).

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.*

   On comparison to other methods

   (a) *real data analysis*: We emphasize that the stem-cell example in Figure 1 (page 4) is precisely a case where the proposed method delivers at a conventional FDR a non-empty list of differentially distributed transcripts where other methods do not. Because the biology of this case happens to be well understood, we have external evidence that the additional discoveries are not false calls. The revision reports additional diagnostics (Section 3.5) relevant to this stem-cell example.

   (b) *other methods/better description of simulation*: Comparisons are central to our evaluation of scDDboost and the revision aims to make these comparative findings more evident. Beyond the empirical study, in which multiple methods are compared on a dozen published data sets, the primary comparisons derive from a set of simulations. The revised Section 3.1 is completely re-written to better explain the simulation settings, and further details are included in a revised Supplementary Material Section 3.1. We appreciate the opportunity to expand on this important point and regret not being clear enough in the original. We have also taken the opportunity to boost the replication size, going from $n = 2$ replicate data sets per parameter setting in the original to $n = 10$ replicates in the revised report, thereby reducing further the Monte Carlo errors. We address the question of difficulty of the 12 distinct simulation settings (e.g. Supp. Figures S6, S7) and also the question of reporting output via ROC (Supp. Figure S9).

   It is important to recognize (and we do so in the revision) that simulations are drawn from a well-established generative model (splatter) that is different from the more descriptive statistical model underlying the scDDboost computations. We are not simply reporting the baseline check that a model-based

method (scDDboost) can do better than structurally simpler methods when the model is true. These computations are not so interesting, frankly (we've done them to check the code, but we don't report them in the manuscript). Rather, splatter provides a level playing field, because it encodes a more mechanistic model reflective of many aspects of scRNA-seq data, and one that would surely be difficult to invert for the sake of inference. For example, the splatter model does not induce a negative-binomial mixture involving constant shape parameters, as we adopt in scDDboost. That we find good operating characteristics indicates some level of robustness to model mis-specification.

3 Response to Reviewers

3.1 Reviewer 1

This paper proposes hierarchical models for identifying genes that shows different distributions between two conditions based on single cell RNA-seq data. The key of the proposed methods is to allow subtypes of cells in each condition- these subtypes obtained by first based existing clustering methods for cells based on gene expression profiles. Overall, the paper was well motivated and presented with statistical rigorousness. I have the following comments:

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?*

   We address this in response to the editor's comment (1b) and at pages 6-7 in the revised manuscript. We disagree with the reviewer's point, *"Of course, if the cell types are different then all genes should have different distributions."* It is a central point of differential expression analysis that many genes have distributions that are the same between cellular conditions being compared. Some genes must have differential expression, but it is not necessary that all genes have an altered distribution. scDDboost methodology leverages this fact to score genes for changes in expression. The methodology accounts for mixing of cellular subtypes in potentially different frequencies in the two conditions, and computes the posterior probability that differential distributions are present at any given gene.

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 emphasize model fit more thoroughly in the revision, as noted in the response to the editor's comment 1d, and have added Figure 10 to present model fits in the main example.

Mixtures of negative binomials (scDDboost is not mixing multinomials) are able to express quite flexible marginal shapes, including the allowance for substantial mass at zero if need be. In Section 3.5 of the revision we report on formal goodness-of-fit tests of the negative binomial. In summary, while some lack of fit is present, it has a limited effect on inference summaries. We used the bootstrap test from Yin and Ma, 2013 at each gene and for each estimated subtype. The majority of cases pass the goodness-of-fit test, but many do not, and present data not consistent with an unmixed negative binomial within subtype. The empirical effect of this lack-of-fit is relatively minor. The numbers of DD genes uniquely called on non-NB genes are small on several data sets (Supplementary Table S3) and there is empirical support for these DD calls (Supplementary Figure S12).

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.*

   We appreciate the reviewer's comment. The language in the original manuscript was not clear. In fact, the primary comparison methods (MAST, scDD, DESEQ2) are considered in the bursting computations (Figure 9). We've edited the language in the discussion around Figure 9.

3.2 Reviewer 2

This manuscript deals with a new type of molecular data. Instead of measuring gene expression at tissue level with the RNA-seq technology, the single-cell technology allows to measure gene expression at a cell level (scRNA-seq data), which is more resolutive [sic].

The first analysis performed on RNA-seq experiment is the differential analysis aiming at finding gene altered in their expression when two conditions are compared. The goal of this manuscript is to formulate the same question with scRNA-seq data.

The question and the model proposed to answer are interesting. Nevertheless some explanations and informations [sic] missing.

Please find below my major concerns.

The biological objective of the method is not clearly explained in the introduction. It is indicated only at the end of page 6 with the definitions of differentially distributed and equivalently distributed. I think that it should be announced in the introduction as well as the fact that they work with a mixture of negative binomial distributions. I would also like to have more explanations about the link between the concept « differentially distributed »

and the differential analysis. For me, it is not very clear that the concept « differentially distributed » answers the questions of the biologists.

Thanks for the suggestion. We adopted the phrase *differential distribution* from Korthauer et al. 2016, Genome Biology, because it emphasizes the wide variety of ways that expression distributions may change between conditions. In the revision we emphasize the issue right away, on page 2, and provide justification using sensitivity. And we announce in that introductory section that we will utilize mixtures of negative binomials to accomplish the task.

The authors should also discuss other possible methods to answer this biological question, as well as tests based on kernels.

We appreciate the reviewer's suggestion and note that we compare scDDboost to a number of the widely cited procedures for assessing the significance of expression changes, including purpose-built MAST, DESEQ2, scDD, as well as more generic methods like t-test and the kernel-based KS-test, all of which have limited ability to leverage common features between tests.

Concerning the model,

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 ?*

   Regarding the second point on the calculation before the key issue; we revised the text to clarify how the mixing leads with differential proportions leads to equal margins. We state the condition in terms of component distributions (instead of only using means). To clarify further we use the specific case in Figure 2, and work through the mixing computations in the revised Supplementary Material, Section 2.1. As to the first point, the compositional model does not entail a change in component means between conditions, as discussed on pages 6/7 and indicated in response to the first reviewer's first point and the editor's comment (b).

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.*

   The reviewer raises an excellent question about the constant-shape assumption used by scDDboost. We discuss this in response to the editor (1c). Although the theoretical formalism does not require constant shape, it is convenient to take this assumption in order to utilize EBSeq code for the mean pattern probabilities. In the revision, Section 3.5, we report additional diagnostic calculations aimed at as-

sessing the constant-shape assumption. We developed a likelihood ratio test of common shape and applied it in several examples, finding that a substantial fraction (about 15% in the cases analyzed) of genes reject a constant-shape hypothesis (at 5% FDR). We go on to investigate the impact, finding that in spite of the lack of fit the DD inferences derived from scDDboost are reliable. We are also supported by the simulation study which demonstrates good FDR control in various settings; notably, the constant-shape assumption is not satisfied in the simulation settings either (Supplementary Material 3.5). It is challenging to formulate a more general solution at present, but we suggest this as an interesting topic for future investigation.

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

   See the response to the editor, 1a. The clustering method does affect the output, for sure. The default method in scDDboost demonstrates good operating characteristics in the settings considered, but the code allows for other choices as long as the user can supply a cell-cell distance matrix. Numerical experiments confirm that the FDR is controlled regardless of the clustering method; differences appear to relate to power. The default method combines a robust method with a correlation-based approach that has demonstrated good properties in single-cell data (Kim et al. (2018b)). The reviewer is correct that clustering methodology affects the derived inferences. Two aspects of the proposed system mitigate risks: (1) the averaging over randomized clusterings reduces sensitivity to effects of individual cells, and (2) the code itself is structured to allow various distances to be invoked. This second point makes the system modular and amenable to improvements in clustering technology as they become available. An exhaustive study of distance functions is beyond the present scope, but numerical experiments show broad agreement between methods (Supplementary Material Table S5).

On 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).*

We regret not being more clear in the original submission and we completely re-write the simulation component in the revised Section 3.1. See also our response to Editor point (2b). We address questions about the difficulty of different scenarios in 3.1 and also in Supplementary Figures S6 and S7 which consider both multivariate and bulk-univariate comparisons. The ROC curves have been redrawn (Supp Figure S9). We also include a more challenging scenario involving $K = 15$ subtypes.

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 ?

These are useful comments. We increased the replication size to 10 data sets per case and include ROC curves as well as the original format in Figs 4 and 5. This reduces Monte Carlo error and provides a more clear view of the error rates. We note that the splatter simulation underlying Figures 4 and 5 encodes different modeling assumptions than any of the tested methods, and so we do not expect, for example, to hit the 5% FDR mark exactly using a model-based local FDR approach. We are encouraged by the FDR control that is evident from these and other numerical experiments. We find some model violations, but they are not so egregious as to disable the hypothesis-testing capacity of the tool in the cases considered. The evidence is that the information provided by genomic clustering improves the operation of gene-level tests, but certainly more research is needed to understand this phenomenon more fully.