Reviewer’s report

Title: A biplot correlation range for group-wise metabolite selection in mass spectrometry

Version: 0 Date: 10 Aug 2018

Reviewer: Robert Flight

Reviewer's report:

Manuscript Number: BIDM-D-18-00020

Full Title: A biplot correlation range for group-wise metabolite selection in mass spectrometry

Reviewer Name: Robert M Flight

### Summary Statement

In this manuscript, the authors create a novel method (BCR) based on observations of 'biplots' that attempts to find variables that contribute to the separation of classes of samples. By the use of this procedure, they hope to avoid some of the pitfalls inherent in multiple testing correction of results from standard statistical tests (such as the use of Benjamini-Hochberg to control the false discovery rate), as well as finding related variables contributing to the separation of sample classes. The authors provide background on biplots, an explanation of their method, and then test their method against other methods using a synthetic dataset that should have some of the properties of the biological data of interest, as well as biological dataset, where only their BCR method returns any results.

I think this manuscript is potentially an important contribution, because it is known that single *biomarkers* are not robust across variations in individuals and rarely are robust across replicate studies, and groups of biomarkers should provide better indications of pathology.

I do think the BCR method is conceptually interesting, and potentially solves a very real problem in the analysis of various high-feature, biological -omics datasets, however, I do have some concerns about the submitted manuscript.

Before bringing up points of concern, I will briefly summarize my understanding of the BCR methodology from my reading of the manuscript:

- Perform dimensionality reduction on two class data

- Take the first two dimensions ordered by importance / variance explained, etc

- For each class of points
- create a multivariate normal distribution of the scores, and then define a 95% 2D CI for that class of samples

- find variables that contribute to increasing scores for those samples in the 95% CI for each of the two dimensions.

- filter to the top 5% (default) of variables in those two directions

- filter out variables that have weak correlation to the response variable by statistical evaluation of correlation, removing those either with p-value $\geq 0.1$ or 0.1 FDR (unclear from the manuscript)

Had the MatLab code been provided to the reviewers by the authors, it would be much easier to evaluate if my summary is correct.

## Major Points of Concern

1. Throughout the manuscript, the term "statistical approach" is used to describe the BCR methodology. However, there are no typically defined "statistical tests" defined that give the end-user an idea of the "confidence" that the variables returned are truly associated with the condition of interest. There is also no test to verify that the method does not return false positives when there are truly no variables associated with the response variable, as would be the case from completely random data. I understand that the "properties" of the loading vectors are "statistical" due to the fact that they can encode the covariance of two variables, however I don't think that justifies the use of the word "statistical approach" as used in the manuscript. The only way that I could see the continued use of "statistical approach" in this manuscript, is if the covariance of the loading vectors to the vectors defining the 95% confidence interval of the scores in the biplot were used. According to the description given in the manuscript, this is NOT what is being done (see my description of the algorithm above).

2. Does the BCR procedure produce valid results if the two most important dimensions do not actually separate the groups? Relatedly, does BCR return false positives in the case when there are no variables that separate the classes of samples?

3. For the analysis of the simulated data, O-PLS is used as the dimensionality reduction method. Based on my own understanding of O-PLS, O-PLS was developed to remove sources of variance *orthogonal* to the underlying sources of variance (batch vs biological variance for example), so that they can be removed before considering the rest of the data and hopefully uncover the *biological* contributions to the variance. I do not think it is appropriate to use it in this context.

4. The type of statistical test used as the basis for the FDR results is never explicitly identified. I am assuming that a t-test was used to generate the uncorrected statistics, and then the p-values were FDR corrected using the two different methods (FDR1 and FDR2).

5. The correlation analysis at the end of the BCR procedure may be appropriate for non-categorical data, but I don't think it is appropriate for categorical data of the nature discussed. Logistical regression would be more appropriate. It is also unclear which values are used in the correlation test; the raw values for that variable, or the loadings for that variable? It is not clear from the description.
6. The plots of the paired variables X1,X5, X9,X10, X28,X29,X30 are helpful, but it would be easier to see how the variables contribute to separation or not through plots of their values in each sample class, with the examples of some that contribute to separation in tandem. It would also be important to address the question of whether some variables contribute to separation in higher dimensions from the dimensionality reduction??

Is there a good reason why the loadings for variables X9-X30 are so much smaller than X1-X8, as observed in Fig 3??

7. The analysis of the simulated data does not address which of the variables are selected by each method. I would expect that X1-X8 are found in every case by each of the methods evaluated. However, which other variables are also found by each of the methods would illuminate the advantages of BCR over the others, vs the current argument of "more variables found" on average.

8. Which variables are being removed by the correlation test is not addressed. I would expect that many of the noise variables are removed, but are some of the other "separating variables" also removed, and does this explain why BCR never finds all 30 variables?

9. Related to the above two points, Fig 3 does demonstrate that the related variables X1-X4 are in one direction, and X5-X8 are in the other direction, implying that the inter-variable correlation is captured. However, given that the idea of BCR is inspired by 'biplots', and that an analyst using BCR *does not* have to use 'biplots' directly when using BCR, what is it about the BCR method that captures which variables are related / correlated? I believe this *is* due most of the dimensionality reduction methods capturing the covariance of variables in the data, however, it is implied that FDR and STOCSYO cannot capture these relationships. That may be so, but if X1-X4 and X5-X8 are captured repeatedly by either the FDR or STOCSYO methods, then what advantage does BCR have over those other methods? There also does not appear to be anything in the BCR method that inherently groups the variables together in reporting the variables, negating the primary claim of the method by the authors. This cannot be evaluated without reporting *which* variables are reported by each method.

10. In contrast to what is indicated from the guidelines to authors page, none of the software for generating simulated data, the MatLab package to run BCR, the simulated data, or the biological data analyzed are available. My understanding of the guidelines to authors (https://biodatamining.biomedcentral.com/submission-guidelines/preparing-your-manuscript) implies that *all* of these should be available, and in no way constitute protected data or source code (the metabolomics data are not from human clinical samples, for example). See "Data and Materials" on https://biodatamining.biomedcentral.com/submission-guidelines/preparing-your-manuscript. I would expect software to be archived on figshare or zenodo or data dryad, with the simulated data, and the metabolomics data should be in metabolights or metabolomics workbench, or if it does not meet their standards, could be hosted on figshare, zenodo, or data dryad.

## Minor Points

1. Introduction, Pg 1, Line 31: "current improvements show that >12,000 metabolic features can be extracted [11]", suggest changing this to read: "current improvements in data processing have demonstrated that ...."

2. Reference 23 does not seem to have anything about post-analysis, based on the version I can access
from this URL: https://pdfs.semanticscholar.org/045f/6250ad3a3e1c5337f4859c4464617926cd5c.pdf

3. In the methodology, the explanation of how PCA works is rather extensive, however, the full explanation of PCA is overwhelming, and a full explanation does not seem to be necessary for understanding the BCR method. In general, BCR requires some kind of loadings, representing how much a variable contributes to that new dimension, and scores, representing the score of the sample in that new dimension.

A graphical representation of the decomposition, and then how BCR chooses the variables in that lower dimensional representation would greatly help illustrate what is going on. The `biplot` figure does help, but it is disconnected from a representation of the scores and loadings.

4. It is surprising that none of the methods had false positives. Given that neither of the FDR methods control family wise error rate (unlike bonferroni correction), I would have expected some of the noise variables to have shown up some of the time, given that they control the expected number of false positives only.

Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

Do you want to get recognition for reviewing this manuscript? Add a record of this review to Publons to track and showcase your reviewing expertise across the world’s journals. Signing up is quick, easy and free!

Yes