Response to reviews of BIOM2021236M.R1

1 Response to Reviewer 2

Thank you for your very careful second reading and the excellent comments. We respond to each point below.

1. *Population and sample eigenvectors: A key thesis of the revised manuscript that is emphasized multiple times (even in the abstract) is that the new method outperforms optimal rotationally invariant methods in situations where sample eigenvectors are far from population eigenvectors. I think this claim is not suffciiently supported . . . I think either more compelling evidence should be provided, or the claim can be downemphasized and perhaps mentioned once in the manuscript as a possible conjecture.*

   This makes sense. We have downplayed this point and merely mention it as a conjecture; **see the last paragraph of Section 4**.

2. *Gene networks in Figures 4 and 5: I found these networks diffciult to interpret. It would greatly help if the presentation of these figures could be improved. Perhaps one could present less methods, but superimpose resulting edges onto a joint set of nodes?*

   Thanks for the great suggestion, we have adopted your approach; **see Figures 3 and 4 and a discussion of the figures at the end of Section 5**.

3. *counts per million mapped reads: My understanding is that this method of normalizing counts is not recommended for RNA-Seq experiments. For example, DESeq2 uses another method of normalization. Why was this method chosen? Do the results change with other normalization methods?*

   There are indeed a number of normalization methods, but most of the work in this area has been for identifying differentially expressed genes. It is not always clear what normalization to use for more complex analyses, such as the coexpression network study in this paper. In particular, the most recent version of the edgeR user's guide states in Section 2.16 that

   "Inputing RNA-seq counts to clustering or heatmap routines designed for microarray data is not straight-forward, and the best way to do this is still a matter of research. To draw a heatmap of individual RNA-seq samples, we suggest using moderated log-counts-per-million."

   We have followed this CPM approach in our analyses. CPM normalization for co-expression has also been shown to have good performance (Johnson and Krishnan, 2022).

4. *Could the authors provide a bit more detail on splitting into training and testing set? Is time ignored when splitting the samples? Also is the sample covariance (that first centers observations; as in the simulation) used for both test and training samples?*

   We do ignore time, for the sake of sample size. This was also done in the original Saul et al. (2017) paper. We also state that we use the sample covariance matrix, which first centers the observations; **see the second-to-last paragraph of Section 5**.

5. *OracMSG: (method defined on Page 16): Why do support points need to be generated here? Since all the (σ_j, σ_k, r_{jk}) are known for the oracle method, I would think that the oracle rule can be computed directly without any need for optimization.*

   We performed clustering even though the true parameters are all known in order to reduce the number of support points to ease the computational burden. When p is large the number of true support points could be very large, which dramatically increased the computation time. We have now mentioned this; **see Section 4.2**.

6. *Some notation on indexing the population/sample covariance matrix should be made more explicit in the manuscript. In particular, it appears that throughout the text there is the convention on single/double-indexing that $\sigma_j^2 = \sigma_{jj}$ and $s_j^2 = s_{jj}$ that confused me in light of e.g. (2). (Similarly for r_{jk}.)*

   We have now clarified that $\sigma_{jj} = \sigma_j^2$; **see the first paragraph of Section 2.1**.

7. *Page 17, "The corrected version consistently outperformed the uncorrected version.": I think this should hold provably (not only just empirically) by a convex analysis projection argument.*

   Thanks for this insight! However, we do not exactly project our matrix onto the space of positive semi-definite matrices because we want positive-definiteness. Our minimum eigenvalue is c_α, as we describe in (14), and we haven't yet thought through how this would affect the argument.

8. *Page 3, "The article is organized as follows.", something is off with the numbering of Sections in this paragraph.*

   Thank you for pointing this out, we have fixed the numbering.

9. *Page 4, "incentived" typo*

   Fixed, thanks.

10. *Page 4, "about index i ... different index j": Maybe phrase this in terms of different parameters rather than indices?*

    We have followed your suggestion; **see the last paragraph of Section 2.1**.

11. *Page 5, "that minimizes the compound risk." → "that minimizes an estimate of the compound risk"?*

    Fixed, thanks.

12. *Page 5, after equation 3: Maybe explicitly write that β_S, β_I are scalars (i.e., $\beta_S, \beta_I \in \mathbb{R}$)? Has $\boldsymbol{I}$ been explicitly defined as the identity matrix?*

    We have now written that β_S and β_I are scalars and have identified $\boldsymbol{I}$ as the $p \times p$ identity matrix; **see after equation (3)**.

13. *Page 9: "In this case, the optimal separable rule for estimating σ_{jk} borrows from all of the σ_j and r_{jk}" → Perhaps use different indices when referring to the first σ_{jk} and when referring to all parameters?*

    We have now reworded this as "In this case, the optimal separable rule for estimating the jkth entry of $\boldsymbol{\Sigma}$ borrows from all of the σ_j and r_{jk}"; **see after (11)**.

14. *Page 10: "We propose to calculate estimate" → "we propose to estimate"?*

    Fixed, thanks.

15. *Page 18: "The data are from the NCBI" → "The data are available from the NCBI"?*

    Fixed, thanks.

16. *Page 21: "These imposes significant" → "impose"*

    Fixed, thanks.