Response to reviews of BIOM2021236M

1 Response to the Associate Editor

Thank you for your very valuable comments. We respond to each point below.

1. *As a journal, Biometrics presents papers that focus on the development of new methods and results of use in biosciences. As of now, the application seems more an afterthought, the use of a convenient (easily available) dataset to fit the scope of the journal. I believe that the authors should motivate their approach from the gene expression data or other bio-related applications starting from the Introduction.*

   We have revised our introduction and have also completely redone our data analysis to study a question about gene network inference that arose during some of our previous collaborative work. We were involved in an analysis of bulk RNA-seq expression from mouse brain in a study of the neurogenomic response to social intrusion (Saul et al., 2017). We have now used some of these data to study gene networks in different regions of the mouse brain; **see Section 5**.

2. *It is not really intuitively clear why the vectorization works, despite the approach not taking into consideration the structure of the covariance matrix. Also, the manuscript proposes a series of ad-hoc expedients to implement the method. See the clustering-based exemplar algorithm and replacing non-positive eigenvalues with a chosen positive value c in the positive-definite correction. These expedients do not seem to be completely justified by the proposed separable compound-decision theory framework, and their effects on the estimation are unclear. A referee also commented on the impossibility to follow closely the implementation of the EM algorithm. In short, while I appreciate that the theoretical justification for the asymptotic efficiency of the proposed vectorization is not clear, the current implementation reads more like a clever algorithm than a well-founded statistical method. The authors should justify all the steps of the algorithm and provide an assessment of their contribution to its overall performance.*

   The success of our vectorization perspective is indeed surprising. In our revision we have made several attempts to contextualize this finding.

   (a) In **Section 2.1** we now note that measuring the success of covariance matrix estimation using Frobenius risk does not incentivize positive-definiteness. In other words, there can exist estimators of $\mathbf{\Sigma}$ that have low Frobenius risk but which are not positive-definite. So the fact that vectorization works is in part a consequence of the risk function.

(b) In **Section 2.2** we note that several existing shrinkage methods can already be interpreted as vectorized methods. The sample covariance matrix, the linear shrinkage method of Ledoit and Wolf (2004), and the adaptive thresholding method of Cai and Liu (2011). The adapative thresholding method is also not guaranteed to be positive-definite.

(c) In **Section 4.4** we now study one possible reason why vectorization may work better than existing rotationally invariant procedures that shrink only eigenvalues but use sample eigenvectors in their estimators. Since sample eigenvectors can be poor estimators of population eigenvectors in high-dimensions, we hypothesized that our vectorized approach would be better in simulation settings where the sample and population eigenvectors were more different. This appears to be the case; **see Figure 3**.

We agree that projecting our final estimator on to the space of positive-definite matrices is not warranted by the compound decision formalism. Indeed, our uncorrected estimator still works very well in our simulation settings, and furthermore the (uncorrected) oracle vectorized estimator can be dramatically better than the oracle rotationally invariant estimator; **see Figure 2**. Nevertheless, it does appear that in finite samples, positive-definiteness correction can improve performance. We mention in the **Discussion** that a non-separable class of estimators that can incorporate this constraint is worth exploring in the future.

In our new **Section 3.3** we have also clarified our EM updates and have added details about iimplementation.

In short, in our paper we explore a new perspective on covariance matrix estimation, a vectorization perspective founded on a compound decision theoretic interpretation of the Frobenius risk. We define a new class of estimators, derive the optimal vectorized estimator within this class, and estimate the optimal estimator using approaches from the compound decision literature. Finally, we argue that these estimators work well for covariance matrices whose sample and population eigenvectors are very different.

3. *I agree with one of the referees that the title should be re-thought, since compound decision theory or compound decisions appear quite relevant to the approach, more than the empirical Bayes aspect of the paper.*

   We have changed our title following your suggestions.

4. *on page 4, it appears to me that the choice of the risk function $\hat{R}$ does not correspond to the Frobenius risk. The rationale for choosing this risk function should be provided. It somehow appears the choice is dictated by Ledoit et al (2004) and Proposition 1, and not by independent first principles. If so, it should be explicitly stated before.*

   We have clarified in **Section 2.2** that the choice $\hat{R}$ follows directly from first principles.

   (a) In (4) we restrict ourselves to the class of linear vectorized estimators of $\mathbf{\Sigma}$.

   (b) In Proposition 1 we show that $\hat{R}$ is an asymptotically unbiased estimate of the Frobenius risk for any fixed $\mathbf{\Sigma}$ and any fixed linear estimator.

(c) A natural procedure is to minimize the empirical risk $\hat{R}$ over the class of linear estimators.

(d) In Proposition 2 we show that the estimator obtained after empirical risk minimization is very closely related to the Ledoit and Wolf (2004) estimator.

5. *Figures 1 and 2 do not allow to appreciate uncertainty. Perhaps a table would be better.*

   We now give exact numerical results for all simulations in our **Supporting Information**.

6. *The manuscript requires careful proofreading and improvements. in the structure of the presentation and the visualization of the results. Non-exhaustive examples are reported below:*

   - p.1 "To overcome these issues, various methods have been developed to estimate high-dimensional covariance matrix" → matrices

   This has been fixed.

   - p. 1 "about the structure of population covariance matrix" → of the population covariance matrix

   This has been fixed.

   - p. 2 "Other common structured methods assume that the covariance matrix is banding" → banded?

   This has been edited to "banded".

   - p. 2 "yet can still outperform the sample covariance matrix" What does this mean?

   It means that these unstructured model have been shown to have lower estimation risk comparing to sample covariance matrix. To make it clear, we have changed to "yet still have less estimation error comparing to the sample covariance matrix".

   - p. 3 "is squared error loss"→ is the squared error loss

   This has been fixed.

   - p. 4 "Now consider the problem of estimating the vectorized $\mathbf{\Sigma}$ under risk 1" → risk (1) Equations are typically referenced between brackets. The same issue appears in many other points where equations are referenced throughout the manuscript.

   This has been fixed.

   - p. 8 "A major advantage of nonparametric estiation of the prior"→ estimation

   This has been fixed.

   - page 8, "Figure 1 shows that different K have similar estimation accuracy compared to the exemplar algorithm, while Table 1 shows that they can be significantly faster" However, at this point Figure 1 refers to models that have not yet been presented. Also, the reference to Table 1 anticipates point made in Section 3.2

   This has been fixed.

- p. 16 "Finally, we have so far assumed that our data multivariate normal" → are normal

This has been fixed.

- In Equations (1) and (2) the arguments of the risk functions are θ and δ. However, on page 5 the arguments of $\hat{R}$ are β_S, and β_I.

We now use $\hat{R}_L$ to denote the risk estimate for the class of linear decision rules. We also make the distinction between $\hat{R}_L$ and R more explicit in **Proposition 1**.

2 Response to Referee 1

Thank you for your positive comments, and your careful reading and extremely helpful suggestions. We respond to them below.

2.1 Response to referee section "CorShrink"

Thank you so much for bringing our attention to Dey and Stephens (2018). It is indeed extremely closely related to our work. We now discuss connections to their work in several places.

1. In the last paragraphs of **Section 3.2** we note that the main difference between our approaches is that Dey and Stephens (2018) do not shrink the variances while we do, as our $\hat{t}_d$ estimates σ_j using its supposed posterior expectation assuming a prior distribution of $\hat{G}_d$. Furthermore, our Proposition 3 shows that our shrinkage approach can be interpreted as viewing the triples $(\sigma_j, \sigma_k, r_{jk})$ as if they were drawn from the trivariate prior G_{od}, which allows for possible dependencies between r_{jk} and the σ_j. For example, model 2 in our simulations in Section 4 considers a covariance matrix where entries with larger standard deviations also tend to have larger correlations. In contrast to the method of Dey and Stephens (2018), our approach can learn this structure if it exists and take advantage of it for more accurate estimation. We demonstrate this in Model 2 in our simulations.

2. We have added CorShrink to our simulation studies for estimating both covariance matrices (**Figure 1**) and correlation matrices (**Table 3 in the Supporting Information**). We found that for estimating covariance matrices, our approach could outperform CorShrink in many cases, likely because it was able to accurately shrink the standard deviations. In the **Supporting Information** we also compare MSG and MSGCor to CorShrink for estimating correlation instead of covariance matrices. We find that in that case, CorShrink performs slightly better than MSG and MSGCor except in our Model 2, likely because the additional flexibility that our method trades off lower bias for higher variance.

2.2 Response to referee section "Separable rules, symmetry and the diagonal"

1. *While reading the paper I was a bit confused by the fact that (R1) treats diagonal entries and off-diagonal in the same way. It would feel more natural to me to define a separable estimator as in (R1) for $j \neq k$ ... In the Multivariate Gaussian case this is not an issue, since $X_{\cdot j} \neq X_{\cdot k}$ for $j \neq k$ with probability 1. But the proposed approach would also apply with minor modifications to e.g., discrete measurements (with known likelihood) in which case $X_{\cdot j} \neq X_{\cdot k}$ could happen with non-zero probability. Furthermore, while this issue is ignored for most of the paper, diagonal and off-diagonal entries are treated in a different way at the end of Section 2.3 (in the context of the clustering based algorithm). I believe it would be valuable to make the above distinction more explicit!*

   Thank you for this excellent suggestion. We completely agree and have now redefined our class of separable estimators to be

$$\mathcal{S} = \{\boldsymbol{\delta} : \delta_{kj} = \delta_{jk} = t(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}),\ 1 \leq k < j \leq p, \quad \delta_{jj} = \widetilde{t}(\boldsymbol{X}_{\cdot j}),\ j = 1, \ldots, p)\},$$

   so that the on- and off-diagonal entries are treated separately. In **Section 3.1** we derive the optimal estimator in this class. This estimator turns out to be the same as the estimator we actually implemented in our previous draft, but redefining $\mathcal{S}$ makes this much more rigorous.

2. *A related question I had concerns the symmetry of the estimator. Does the estimator satisfy the symmetry property $\hat{\sigma}_{jk} = \hat{\sigma}_{kj}$? Similarly, the reader may wonder whether the estimation procedure enforces symmetry in the prior G in (6) with respect to σ_j, σ_k.*

   In our new implementation we now explicitly enforce symmetry. We construct the support points of our estimated discrete prior to be symmetric and we also force the estimated masses at these points to be symmetric as well. See **Section 3.3** for details.

3. *The paper emphasizes the Multivariate Gaussian problem. In that case, the sample covariance is sufficient, and so, it would seem more natural to me to define a separable rule, as one that satisfies,*

$$\hat{\sigma}_{jk} = \bar{t}(X_{\cdot j}^T X_{\cdot k}), \quad \bar{t} : \mathbb{R} \to \mathbb{R}_+$$

   instead of (R1). I am wondering why the authors chose to use (R1) instead.

   We chose our definition of a separable estimator because it is more general than $\bar{t}(\boldsymbol{X}_{\cdot j}^\top \boldsymbol{X}_{\cdot k})$. We actually initially tried this class of estimators but found that its performance was worse than the estimators we proposed in the manuscript. Intuitively, the reason is that the statistic $X_{\cdot j}^\top X_{\cdot j}$ is not sufficient for the parameter σ_{jk}.

4. *If I understand correctly, one of the benefits of estimator (3) is that it works without specific assumptions on the noise distribution. Instead the approach here appears to more heavily rely on a known likelihood (and specifically a multivariate Gaussian likelihood). It would be great if some simulations could be conducted under misspecification,*

wherein the data is not multivariate Gaussian, but the approach is applied assuming multivariate Gaussian data.

We now study the performance of our proposed method when the multivariate Gaussian assumption is violated. **Tables 4 and 5 in the Supporting Information** show that our procedures can still provide relatively accurate estimates for uniform or negative binomial data.

2.3 Response to referee section "Compound decision theory or empirical Bayes?"

1. *The result following equation (6) could be presented separately as a Theorem (or Proposition). It should be highlighted that part of the statement is that the estimator in (R1) that optimizes (R2) depends on all the σ_j and ρ_{jk}.*

   We have followed your suggestion and added **Proposition 3** along with a discussion after it highlighting the fact that the optimal estimator depends on all the σ_j and ρ_{jk}.

2. *The term "compound decision theory" or "compound decisions" etc. deserves to be in the title!*

   Thank you for the suggestion! We have revised the title accordingly.

3. *On the other hand, I would downemphasize the empirical Bayes aspect of the paper (and my preference would be to even remove it from the title). To me, an empirical Bayes analysis (especially since here the goal is to perform well, in a frequentist sense, in terms of (R2)) has the following flavor: "Let us mimic what a Bayesian would do, when we do not know the true prior, and in a way that we can get frequentist guarantees as well". However, I do not think a Bayesian would start approaching this problem by first positing (6) and then using (8)2. Instead a more natural Bayesian approach would be to set a prior for the full covariance matrix Σ, e.g., as in Berger et al. [2020] and references therein. I would restrict the discussion of empirical Bayes to just saying that model (6) appears formally when optimizing over separable rules (R1). To approximate the unknown optimal rule, one follows an empirical Bayesian "Generalized Maximum Likelihood" approach with respect to the pairwise composite likelihood.*

   This is an excellent point and we have done exactly as you suggested. We now only refer to empirical Bayes ideas as a method of estimating the optimal compound decision rule.

2.4 Response to section "Other remarks"

1. *It would be of great service to the community if the authors could provide code implementing the method and reproducing the results!*

   We now note at the **end of the Introduction** that our code is available on Github and give the link.

2. *Related to the above point, based on the description in the manuscript it would not be possible for me to fully reproduce the results. For example, does the actual implementation use the EM algorithm or an interior point convex solver? If the EM algorithm is used, how many steps are taken or what is the termination criterion? And why is the EM algorithm used (since for these empirical Bayes problem it typically converges extremely slowly compared to interior point solvers, as carefully demonstrated by Koenker and Mizera [2014])? On the other hand, if an interior point solver is used, which one?*

   We have revised our description of our implementation in **Section 3.3** to make our EM algorithm more clear. We aso note that instead of EM, we attempted to maximize the composite likelihood (12) using the interior point solver MOSEK, as advocated by Koenker and Mizera (2014), but ran into computational difficulties because for even moderate sample sizes n, the values of the likelihoods we needed to calculate could be extremely small at certain support points. These small values caused problems for the R wrapper for MOSEK. Our implementation of the EM algorithm was more robust to these small density values.

3. *First paragraph of page 3, "they modify only the sample eigenvalues and not the sample eigenvectors": The situation here is a bit more nuanced, since e.g., sometimes using the sample eigenvectors is provably the optimal thing to do (e.g., if one constrains themselves to rotationally invariant estimators as in Bun et al. [2016] and references therein)*

   Thank you for this reference. We now explicitly distinguish our proposed method from rotationally invariant ones and note in our **Introduction** that while rotationally invariate estimators perform extremely well, Bun et al. (2016) showed that the eigenvectors of any rotationally invariant estimate must be the same as those of the sample covariance matrix. This can be problematic because sample eigenvectors are not consistent when the dimension and the sample size increase at the same rate (Mestre, 2008). This suggests that unstructured but non-rotationally invariant covariance matrix estimators may be worth exploring, as these can modify both the sample eigenvectors as well as the eigenvalues. Our contribution is to propose and study one such class of non-rotationally invariant estimators.

   Incidentally, our new simulations in **Figure 3** suggest that the simulation settings where our estimators can outperform existing methods are exactly those where the sample and population eigenvectors differ most.

4. *Page 5, "it is straightforward to show that": The "straightforward" could be replaced by a short proof in the appendix.*

   We have rewritten this as **Proposition 1** and have added a proof to the Supporting Information.

5. *Also could it be that the risk estimate is not exactly unbiased (because of dividing by n and not $n-1$ in the definition of $\hat{\Delta}^2_{jk}$)?*

   You are absolutely right, thank you for catching this. We have changed our text to mention that the risk estimate is only asymptotically unbaised, and we show this in

the Supporting Information.

6. *Where is the sample covariance $\boldsymbol{S}$ defined?*

   We now define $\boldsymbol{S}$ in the first paragraph of **Section 2.2**.

7. *Section 2.5, "Projections in terms of other matrix norms are also possible". Is this actually being done in the implementation?*

   We did not implement this and so have removed this comment from the manuscript.

8. *Else perhaps remove this sentence.*

   This has been removed.

3 Response to Referee 2

Thank you for your positive comments. We respond to them below.

1. *In the numerical experiments and the real data analysis, what was the size of the grid that was used? I may have missed it but I did not see it mentioned in the caption for figure 2.*

   We have now added a separate **Section 4.3** that specifies the choice of grid size for our proposed method.

2. *In the numerical experiments, the dimension p goes as far as 200. I believe it will be useful to consider a setting where $n = 100$ and $p = 1000$.*

   We have added a simulation with $p = 1000$ in the **Supporting Information** and in **Section 4.4** note that while our methods perform well in our Models 1 through 4, other methods are better in Models 5 and 6 for large p. This is likely due to the fact that our class of estimators may not be be optimal in these settings because the population eigenvectors in these models appear relatively easy to estimate; **see the last paragraph of Section 4.4 for details**.

3. *Continuing on the numerical experiments, models 1 to 5 impose various structures on $\boldsymbol{\Sigma}$. It will be interesting to see the performance of the proposed method if a spiked covariance structure is imposed on $\boldsymbol{\Sigma}$.*

   We have added a spiked covariance matrix as model 6 in our new simulations. As mentioned in our response above, we found that models 5 and 6 constitute settings where our methods may not work as well as existing approaches.

4. *Typo on page 6: "The density of $f(\cdot \mid \boldsymbol{\eta}_{jk})$ of ...".*

   This has been fixed.