1 Comments to the editor

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.*

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.*

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 agree that compound decision theory should be more emphasized than empirical Bayes in our paper. Following your suggestions, we have changed the title from "A Empirical Bayes Approach to Covariance Matrix Estimation" to "A Compound Decision Approach to Covariance Matrix Estimation".

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 squared and scaled the Frobenius norm $\|A\|_F = \sqrt{\sum_{j,k=1}^{p} a_{jk}^2}$ to $\frac{1}{p^2}\sum_{j,k=1}^{p} a_{jk}^2$ in order to apply fundamental theorem of compound decisions, for which the risk function should be scaled by the number of parameters because the empirical distribution has mass probability of $\frac{1}{d}$ where d is the number of parameters.
The transformation of Frobenius norm will not change the value of $(\widehat{\beta}_S, \widehat{\beta}_I)$ since it is determined by minimizing $RE(\beta_S, \beta_I)$.

5. *Figures 1 and 2 do not allow to appreciate uncertainty. Perhaps a table would be better.*
   We add the tables of the medians and uncertainty of Frobenius errors in supporting information. The uncertainty of Frobenius error is measured by the difference of 75% quantile and 25% quantile in 200 replicates. We use this range to measure the uncertainty because the errors of most methods are not following normal distribution.

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

   It has been fixed.

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

   It has been fixed.

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

   It 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. For example, in [2], the numerical result shows that their linear shrinkage method has much lower risk than 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

It 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.*

We used to use \ref to reference equation, we have changed it to \eqref so that they are referenced between brackets.

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

It 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

We have moved the setting of the experiment to the start of Section 3 Numerical Results.

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

It 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 have changed the notation $\hat{R}$ to RE which means risk estimator and this differentiates it from R.

2 Comments to Referee1

Thank you so much for the detailed and inspiring suggestions on our paper. We learnt a lot through them. We agree with your comments and address them in the updated draft version. Please find the following responses for more details.

2.1 comments to section "CorShrink"

After carefully reviewed [1], we think it is very related to our paper and it is valuable to make the comparison.

As mentioned in Section 1, our method is quite similar to [1]. Both of us used empirical Bayes estimator and nonparametric estimation of prior distribution. The difference is that our method aim to estimate covariance matrix which requires to also shrink standard deviations.

We also compared our two methods by numerical experiments which are shown in supporting information. We designed two simulations to compare the performance on correlation matrix estimation and covariance matrix estimation. In the first simulation, we compare covariance matrix estimation. For corShrink, $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\widehat{SD})\widehat{\boldsymbol{R}}\text{diag}(\widehat{SD})$, where $\widehat{SD}$ are sample standard deviations. In the second simulation, $\widehat{\boldsymbol{R}} = \text{diag}(1/\widehat{SD})\widehat{\boldsymbol{\Sigma}}\text{diag}(1/\widehat{SD})$, where $\widehat{SD}$ are standard deviations derived from $\widehat{\boldsymbol{\Sigma}}$. For each setting, we take n as 20 and 100.

2.2 comments to section "Separable rules, symmetry and the diagonal"

$$\widehat{\Sigma} = (\widehat{\sigma}_{jk})_{1\le j,k\le p}, \quad \widehat{\sigma}_{jk} = t(X_{\cdot,j}, X_{\cdot,k}), \quad t : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{1}$$

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 $j \neq k$, and as follows on the diagonal:*

$$\hat{\sigma}_{jj} = \tilde{t}(X_{\cdot j}), \quad \tilde{t} : \mathbb{R}^n \to \mathbb{R}$$

We have changed the rule to

$$\mathcal{S} = \{\boldsymbol{\delta} : \delta_{kj} = \delta_{jk} = t(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}),\ 1 \le k < j \le p, \quad \delta_{jj} = \widetilde{t}(\boldsymbol{X}_{\cdot j}),\ j = 1, \ldots, p)\}, \tag{2}$$

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.

We agree that it is not very precise to use $I(\boldsymbol{X}_{\cdot j} = \boldsymbol{X}_{\cdot k})$ to differentiate

on-diagonal entries and off-diagonal entries. So we changed separable rule to the above class of rules.

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!

2. *A related question I had concerns the symmetry of the estimator. Does the estimator satisfy the symmetry property $\hat{\sigma}_{jk} = \hat{\sigma}_{kj}$?*

   Yes, we only calculate estimates of lower triangular matrix of Σ, i.e., $\hat{\sigma}_{jk}$ where $j \geq k$. Then the upper triangular matrix estimation is calculated by the symmetry of covariance matrix $\sigma_{kj} = \sigma_{jk}$. Correspondingly, we should fix the Bayesian model in Section2.2 from

$$\boldsymbol{A} \mid \boldsymbol{\eta} \sim f(\cdot \mid \boldsymbol{\eta}), \quad \boldsymbol{\eta} \sim G_p(a, b, \gamma) = \frac{1}{p^2} \sum_{j,k=1}^{p} I(\sigma_j \leq a, \sigma_k \leq b, r_{jk} \leq \gamma). \tag{3}$$

   to

$$(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k}) \mid \boldsymbol{\eta} \sim f(\cdot \mid \boldsymbol{\eta}), \quad \boldsymbol{\eta} \sim G_{nd}(a, b, \gamma) = \frac{2}{p(p-1)} \sum_{1 \leq k < j \leq p} I(\sigma_j \leq a, \sigma_k \leq b, r_{jk} \leq \gamma). \tag{4}$$

   and

$$\boldsymbol{X}_{\cdot j} \mid a \sim \widetilde{f}(\cdot \mid a), \quad a \sim G_d(a) = \frac{1}{p} \sum_{j=1}^{p} I(\sigma_j \leq a). \tag{5}$$

   , as well as the proposed estimator from

$$\widehat{G}_p = \arg\max_{G \in \mathcal{G}} \prod_{j,k=1}^{p} \int f(\boldsymbol{A}_{jk} \mid \boldsymbol{\eta}) dG(\boldsymbol{\eta}), \tag{6}$$

   to

$$\widehat{G}_{nd} = \arg\max_{G_{nd} \in \mathcal{G}} \prod_{1 \leq k < j \leq p} \int f(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k} \mid \boldsymbol{\eta}) dG_{nd}(\boldsymbol{\eta}) \prod_{j=1}^{p} \int \widetilde{f}(\boldsymbol{X}_{\cdot j} \mid a) dG_d(a), \tag{7}$$

   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 correspondence, we modified our sample exemplar algorithm, we first clusters sample points (s_j, s_k, r_{jk}) of lower triangular part of sample covariance matrix and get the K centroids (a, b, γ), then find their symmetric points (b, a, γ). We also updated EM algorithm to guarantee each (a, b, γ) and (b, a, γ) has the same weight.

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.

$X_{\cdot j}^T X_{\cdot k}$ is sufficient statistics of the single entry σ_{jk}. In fact, for $(\boldsymbol{X}_{\cdot,j}, \boldsymbol{X}_{\cdot,k})$, the sufficient statistics for η is $\boldsymbol{S}_{jk} = \begin{pmatrix} s_{jj} & s_{jk} \\ s_{kj} & s_{kk} \end{pmatrix}$. We also tried density based on single entries s_{jk} before but the performance is not as good as using $(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot k})$.

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.*

This is a good point. To measure the performance of our method on nonGaussian distribution, we did some simulations on negative binomial distribution. We first generate independent distributed Y and then let $X = LY$, where L is the Cholesky decomposition of Σ. The simulation results are presented in supporting information.

2.3 comments to 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 rewrote the equation after (6) as a separate proposition to show the equivalence between the Frobenius risk and Bayes risk.

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

   Yes, we agree that "compound decision" is more deserved to be highlighted in the paper so we replaced "nonparametric empirical Bayes" with "compound decision" in the title.

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.3*

2.4 comments 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 put our code of methods and simulation in supporting information.

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?

In our implementation, we used EM algorithm to estimate the prior distribution. We added the related details to Section 3 Numerical Results. The maximum number of iterations is set as 200. Iteration early stops when the criterion $\frac{L(\boldsymbol{w}^{(k+1)})-L(\boldsymbol{w}^{(k)})}{L(\boldsymbol{w}^{(k)})} \leq 10^{-4}$ is met. We also tried interior point, but it always failed to find the optimal solution. The possible reason is the density $f(\cdot|a,b,\gamma)$ has very different scales, from 10^{-22} to 2 and this makes interior point method difficult to solve.

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)*

   The sentence "However, they modify only the sample eigenvalues and not the sample eigenvectors." is not very appropriate so we changed it to "However, as rotation-equivariant estimators (Stein, 1975, 1986), shrinkage estimators modify only the sample eigenvalues and keep the sample eigenvectors.".

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

   We added the proof in the appendix.

   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 $\hat{\Delta}_{jk}^2$)? You're right. We realized that the previous defined $\hat{\Delta}_{jk}^2$ is not unbiased estimator of Frobenius risk of $\boldsymbol{\delta}$ (1). So we changed the text to describe it as an "approximately unbiased risk estimator" since the factor $\frac{n-1}{n}$ is closed to 1. It is also shown in appendix B.

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

   In the updated manuscript, $\boldsymbol{S}$ is defined as $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$ as it is first mentioned in page 4, section 2.1.

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

   No. We did not implement it in terms of other norms.

   Else perhaps remove this sentence.

   It has been removed.

3 Comments to Referee2

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.*
   In section 3.3, methods compared and section 3.4, data analysis, we used clustering-based exemplar algorithm, where we find K clusters of all off-diagonal sample points. K is set to be p to balance accuracy and speed according to result of section 3.2.

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 did the simulation for $n = 100$ and $p = 1000$ using the different covariance matrix models in Section 3.1. Please see the supporting information for the simulation results.

3. *Continuing on the numerical experiments, models 1 to 5 impose various structures on* $\mathbf{\Sigma}$. *It will be interesting to see the performance of the proposed method if a spiked covariance structure is imposed on* $\mathbf{\Sigma}$.
   We added the spiked model as the sixth matrix models in Section 3.1. Spiked covariance matrix is generated by $\Sigma = \boldsymbol{U}\mathrm{diag}(\boldsymbol{l})\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is a randomly generated orthogonal unit matrix, $\boldsymbol{l} = \{4, 3, 2, 1, \ldots, 1\}$.

4. *Typo on page 6: "The density of* $f(\cdot \mid \boldsymbol{\eta}_{jk})$ *of ...".*
   Thanks for poiting this out. We've fixed this sentence as "$f(\cdot \mid \boldsymbol{\eta}_{jk})$, the density of $\boldsymbol{A}_{jk}$, depends on $\boldsymbol{\eta}_{jk} = (\sigma_j, \sigma_k, r_{jk})^\top$, ...".

References

[1] Dey, K. K., & Stephens, M. (2018). CorShrink: Empirical Bayes shrinkage estimation of correlations, with applications. bioRxiv, 368316.

[2] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2), 365-411.