**Work remaining**

Finalize code and write appendix?

**Author response**

We thank the reviewers for making the effort for detailed comments that led to significant improvements of the paper. A few common issues raised by the referees that we have addressed were about (a) the need for demonstration of real-world applicability using a real data example, and open-sourcing the codes, (b) motivating the problem and analysis more and contextualizing theoretical conditions more, (c) discussing the tuning parameter selection method and how it ensures correct estimates to be used for testing. Below are our detailed responses to comments by individual referees. Substantial changes have been highlighted in blue in the main paper.

**Referee 1**

This paper studies the joint estimation and inference problem for data integration based on multiple and multi-layered Gaussian graphical models. Conceptually, the proposed method, coined as the Joint Multiple Multi-Layer Estimation (JMMLE), aims to estimate the vertical (multi-layer) and horizontal (multiple graphs) dependence structures by borrowing similarity information within each layer and the hierarchical links. Methodically, the JMMLE is implemented as a combination of the neighborhood selection and group lasso type regressions. A debiasing procedure is developed based on the \ell\_{1}/\ell\_{2} penalized regressions. Asymptotic distributions for the model parameters and validity of testing procedures are established. Below are some of my major concerns.

Model assumptions: The framework proposed by the authors requires the knowledge of the group sparsity patterns on the precision matrices of every layer, as well as the group sparsity patterns on the regression coefficient matrices. How practical that information can be obtained in real applications (such as in the Omics data integration problem)? In addition, since the proposed framework is has a hierarchical dependency (besides the similarity structure), if the group sparsity patterns of the coefficient matrices and the precision matrices from the lower layer are given, how is the group sparsity pattern of the upper layer affected given that information? In such a multi-task learning/estimation problem, it seems more natural to only assume similarity patterns on the lower layer such that the features in the upper layer (due to the regression model structure imposed by the authors in Section 2.1) is automatically influenced by the correlation/dependence structures in the multiple response variables. Overall, I think the model framework in this paper is over-demanding.

*The group sparsity assumptions make our framework more general, meaning that they are certainly not necessary, but rather can be plugged in as and when more information is available. We clarify this in Remark 2 in page 7 now.*

Tuning parameters (section 2.2.2): There are three tuning parameters in JMMLE. The tuning parameter selection is done by the BIC and its high-dimensional version (HBIC), which in my opinion is not a convincing way to promote the sparsity in estimating the parameters in the high-dimensional setting. First, I don’t quite understand the definition of the HBIC given by the authors (e.g., why the sample size information is on the scale of \log\log{n} and why the model complexity is on the scale of \log(pq)?). The authors should either give a derivation of their criteria or give a reference for those formula. More seriously, it is unclear whether or not the BIC-tuned parameters obey the conditions in the theory developed in Section 2.3 for the JMMLE estimators.

*The use of HBIC is motivated by existing high-dimensional literature (Kim Et al, JMLR, 13:1037-1057, 2012; Wang et al, AoS, 41-2505-2536, 2013) where the tuning parameter selected as the minimizer of HBIC asymptotically identifies the oracle estimator in ultra-high dimensional penalized problems. We have clarified this in section 2.2.2 now (pg 8). We further tie back these choices to testing in subsection 3.4.*

Simulation: In Section 4.1, in terms of the overall performance of MCC, JMMLE is not as good as the separate estimation methods for each layers, at least in the higher sparsity settings. (Table 2, column MCC, settings (30,60,10), (200,200,150),…) In addition, the number of simulations(=50) is relatively small.

*This seems to be the result of JMMLE producing slightly sparser (=lower TPR than separate) models in these settings. However the lower RF values mean that JMMLE estimates are much more accurate. We now indicate this in Section 4.1. We have now increased the number of replications from 50 to 150 for each setting.*

Real data: Since the paper is motivated from the data integration problem of Omics data, a real application should be added to demonstrate the usefulness of the proposed method.

*Following the referee’s suggestion, we have now added a real data analysis section (Section 5), with external references corroborating the findings. (tba in paper)*

Notations: I am not sure if the notation system in the current version is optimal. In many places, the notation is not conventional and quite confusing (I have to guess the meanings!) For examples, the authors use \hat to denote both deterministic and random quantities in Section 2.3; the constants c\_{1}, c\_{4} in Theorem 2 (and subsequent theorems) are meaningless in that c\_{1}=12\*c\_{1} unless the explicit values of c\_{1} is given. I can go this list on and on… So I strongly recommend the authors improve the notation and be concise in writings.

*We agree with the reviewer that the notations are sometime confusing and superfluous. We have overhauled the notations throughout the paper to minimize confusion.*

Technicality: Theoretical properties derived for JMMLE estimators seem to be consequences of standard concentration inequalities (in particular, given that the Gaussian distribution assumptions on the data) and the debiasing techniques. I don’t see the essential technical difficulty in this paper. If I were incorrect, then please highlight your difference/contributions from existing results.

*For the estimation results, the presence of an optional grouping structure in B and generalization to K>1 adds add technical complexity to the proofs. More importantly, we work in presence of minimal assumptions, steering clear of conditions used in the previous papers, like Incoherence and Uniform Irrepresentability that are often difficult to verify in practice. Similarly for testing, while we do work with debiased estimates and prove parallel results, we do not use additional conditions as required by Zhang/Zhang 2014 and Mitra/Zhang 2016, and prove the results for \*generic\* parameter estimates satisfying standard convergence rates (conditions T1-T3). We added notes to this generalizability in the introduction (1.1 second paragraph), section 2 (Remark 5) and section 3 (Remark 6).*

Minor points:

— page 6: the authors call the estimator \hat{\Omega}\_{x}^{k} as the \*graphical lasso\* in estimating the precision matrices of the X-layer. I think this is not precise since there is no \ell\_{1} penalty involved in the estimator. Rather, it is just an MLE of \Omega\_{x}^{k} on a restricted support.

*The reviewer is correct, we have changed this now.*

— page 12: in line -10, simularities -> similarities.

— page 13: in the definition of the debiased estimator \hat{\beta}\_{j}^{deb}, there is a missing hat in the residual of initial estimate \beta^{(\init)}.

*We have corrected these typos.*

**Referee 2**

In this paper, the authors have developed a hierarchical model based approach for data integration. Hierarchical approach for data integration is common in the recent advancement of multi-omics and multi-platform data sets. Here the authors have used multi-layered Gaussian graphical model structure to do combine modeling and inference from different conditions or subtypes and different types of molecular information. The formulation involves decomposing the multi-layer problem into a set of two layers. Then using GGM framework to model the outcomes from each pair of layers. The relationship between and within the layers are established using the nodes in the layers as the nodes of GGM. The sparsity is developed using elementwise penalization of the precision matrices. An efficient algorithm is developed to fix the complex multi-layer model across all data platforms. Overall, the paper is well written with underlying theoretical justification and elicitation through simulations. However, I have following comments regarding the article.

1. Why there is no real life applications with the model? There are several multi-omics studies/data sets that are available for practical use and demonstration? Why no attempt is made on that. In a sterile simulation setting the model will most of the time work, where as in reality it will breakdown. That is not always bad, it often reveals any shortcoming/restriction of the model and thus makes the user aware of that. At least one real life application is sorely needed.

*Following the referee’s suggestion, we have now added a real data analysis section (Section 5), with external references corroborating the findings. (tba in paper)*

2. Regarding the group penalty, did the authors considered graph laplacian structure? Diagonal dominance can be created several ways it’s not fully clear in this situation how Diagonal dominance enforce the group selection. Does it depends on the partial correlation among the nodes?

*The diagonal dominance assumption mainly ensures restricted eigenvalue type assumptions on model quantities, and tightening the quantity Q\_0 that affects finite sample error rates in theorem 2. We discuss this after theorem 2.*

3. When working with multiple conditions, problem arises when same node is shared or found to be present in multiple cases. Across X1, X2, X3. I am curious if so how the model explore such behaviors?

*At present our framework is applicable for non-overlapping group structures. Expansion to non-overlapping groups is indeed a promising next step, and we discuss this in the caveats (Section 6.2).*

4. The tuning parameter selection section need more detail description. For high dimension, can some consistency property of the BIC be established in that regard? Can you compute the full solution path of your tuning parameter? Based on your model is it possible to explore the choice of tuning parameter by calculating the profile marginal likelihood?

*We agree with the reviewer that tuning parameters warrant more discussion, and have added more material (2.2.2 and 3.4) to address this.*

5. No clear description on the computation complexity, scalability, and convergence properties of the algorithm. No mention of the software availability. For practicality, reason a short section in the appendix need to be created with clear description on how to use the developed software. This makes me wonder about the reproducibility of the claims.

6. Without real life justification, this paper is just a collection of some attractive results. I strongly urge the authors to focus more on the story telling and make the claim accessible/reproducible through sharing their codes. This paper need some serious content revision in narration. The paper overall is novel and has merit but lacks in real life based illustration and claim justification.

*Following the referee’s suggestion, we have now added a real data analysis section (Section 5), with external references corroborating the findings. (tba in paper). We have also formalized our source code and are attaching them in the supplementary material (tba).*

**Referee 3**

The paper combines ideas from Ma and Michailidis (2016) and Lin et al. (2016a) to jointly estimate grouped hierarchical models. The work is a technical extension of existing results. Although the pieces are put together in a straightforward way, the work might be of interest to practitioners whose data falls into the framework. In the current form, the paper is not suitable for publication in JMLR. Below, I will provide more detailed comments that might improve the quality of the manuscript.

There is no real world example in the paper that might benefit from the framework. Without this, I am not convinced that we need a complicated model as described in the paper. The authors should illustrate how the current methodology helps in solving a problem that could not have been solved with existing approaches.

*Following the referee’s suggestion, we have now added a real data analysis section (Section 5), with comparisons to current methods and external references corroborating the findings. (tba in paper)*

Writing is not clear. Especially the introduction. It is not clear what the problem is, nor what the major contribution of the paper is. In my opinion, too much is assumed of the reader --- what are hierarchical models, grouped graphical models, how are graphical models used for the problem at hand. I would suggest the authors to find a colleague that is not working in the area of graphical models to see if he/she would understand the paper. Even after the introduction, often notation is used before it is defined, which makes reading difficult. For example, page 5, second paragraph, "We assume known structured sparsity patterns..." What does the structured sparsity mean here?

*We agree with the reviewer that the beginning of the paper can use better narration and motivation. Keeping this in mind, we have reorganized the introduction, and added more exposition to section 2.*

Condition 2 does not seem to be a condition. It is a definition of diagonal dominance.

*We agree with the reviewer. We have now labeled them as Definition 1 and 2 to navigate this confusion.*

In Theorem 1, are there any requirements on the quantity R(p,q,n)? For example, can this diverge with p,q,n?

*The referee raises a great point. While R(p,q,n) does not need to obey any explicit bounds for Algorithm 1 to have a limit point, having a tighter O(\sqrt{log(pq)}/n) bound ensures that the limit point lies close to the population parameters with high enough probability. We address this after stating Theorem 1.*

Assumption E1 on diagonal dominance seems strong. In the literature on estimation of graphical models it is not commonly used. Why is it needed? Can it be relaxed?

*The diagonal dominance assumption ensures restricted eigenvalue type assumptions on model quantities, and in result controls the quantity Q\_0 that tightens finite sample error rates in theorem 2. We discuss this after theorem 2, and also briefly in the discussion (tbd).*

What is \Psi in equation 2.15 and 2.16?

Max\_k \Lambda\_min \Sigma\_x^k (given in lemma 2 statement)

*Thanks for pointing out this mistake. We have now added this to Theorem 2.*

In Theorem 2, there is an assumption on l1 convergence of \hat B to B\_0, which does not depend on sparsity of B\_0. How can one estimate B\_0 so quickly? Usual bounds require scale as s\sqrt(log p / n) where s is the number of non-zero elements and s &#8594; &#8734;.

In Theorem 2, is Q\_0 a constant or it depends on minimum eigenvalue of Sigma^k ?

In equation 2.17, if Q\_0 = O(1), why do you keep it inside O(...) ?

In Theorem 3, C\_&#952; is assumed O(1). This is again strange given known rates for sparse estimation. One would expect sqrt(s log p / n).

*We apologize for the lack of clarity. We think this confusion was due to us assigning hats to both deterministic quantities and estimates, as pointed out by another referee. The bounds the current referee refers to are bounds on fixed estimate values of B (or Theta) in an iteration based on which Theta (or B) is estimated in the next iteration. In an effort to keep our framework as general as possible, we did not impose any explicit sparsity conditions on these fixed quantities. Since C\_\beta (or C\_\Theta) depends only on the population parameter B\_0 (or Theta\_0), when that parameter is actually sparse their corresponding sparsity parameters can be a part of C\_\beta (or C\_\Theta). When we do obtain the actual estimates (2.15 - 2.17 in Thm 2 and 2.18 – 2.21 in Thm 3), their finite sample bounds scale with the corresponding sparsity parameters s\_j and B at the rate the referee points out. We have now clarified this in a remark after thm 3 (Remark 3). Finally, the O(1) is indeed superfluous when indicating constants, we have removed them.*

Some comments would be useful on how does joint estimation of multiple layers help compared to separate estimation. More generally, it would be useful to provide comments on assumption and importance of results.

*We have added comments on the assumptions throughout. The computational benefits of joint vs. separate estimation are discussed in Section 6.1.*

Please comment on assumption T1 - T3 and explain what procedure would be able to achieve these rates of convergence and under what additional conditions.

*These are common finite sample bounds in the high-dimensional literature. We discuss them after stating T1-T3 now.*

For Theorem 5 and Section 3.1, what are the gains from considering multiple output regression, rather than applying Zhang and Zhang (2014) node by node. Do we get efficiency or weaker side conditions? In a simulation, can one quantify gains empirically?

In simulation section, tables include SD reported to three digits, but only 50 replications are done.

*We have now increased the number of replications from 50 to 150 for each setting. We have also added testing results for separate estimation as well as the separate lasso-based debiasing of Zhang/Zhang 2014 (Table 5).*

Empirical size of the test is quite far from the nominal. This is a problem.

*This is possibly an offshoot of the higher sample size requirements in testing than estimates, as nominal sizes for JMMLE estimates tend to go down when p,q are kept constant and n is increased. We have mentioned this in section 4.2 now.*

BIC is suggested for selecting tuning parameters. It would be useful to discuss how does one obtain degrees of freedom for the complex model studied here. Can it be shown that the tuning parameters obtained in this way lead to correct tests?

In practice correct selection of tuning parameters for testing in high-dimensional problems is crucial for obtaining honest tests and confidence intervals with correct coverage. For the inferential procedure to be valid one should prove how to select tuning parameters, since even changing constants in front of penalty parameters will change resulting confidence intervals.

*The use of HBIC is motivated by existing high-dimensional literature (Kim Et al, JMLR, 13:1037-1057, 2012; Wang et al, AoS, 41-2505-2536, 2013) where the tuning parameter selected as the minimizer of HBIC asymptotically identifies the oracle estimator in ultra-high dimensional penalized problems. We have clarified this in section 2.2.2 now (pg 8). We further tie back these choices to testing in subsection 3.4, and discuss them in caveats (Section 6.2).*