Dear Editor,

Thank you for giving us the opportunity to revise our manuscript according to reviewers' recommendations.

We would like to acknowledge the reviewers for their positive comments and their insightful suggestions. We have now revised the manuscript accordingly (see our point-by-point responses below).

Thank you again for your consideration.

Sincerely yours

Magali Richard

—---

Review by Pierre Neuvial

This paper presents an R package called hdmax2, which implements and enhances a method called HDMAX2 recently published by some of the authors for high-dimensional mediation analysis (Jumentier et al 2023). The goal of mediation analysis is to quantify the indirect effect of a variable M in the causal relationship between and exposure X and an outcome Y. While it is already not obvious to properly define and to perform mediation analysis in a classical setting, the HDMAX2 method addresses the case of high-dimensional mediation, meaning that the number of potential mediators is much larger than the sample size.

The main contributions of this paper are:

- availability of an implementation of the HDMAX2 method as an R package
- extension of the original method to binary outcomes and to binary, categorical, and multivariate exposures (instead of only continuous variables).
- two case studies that were not described in the Jumentier et al paper

The paper is well written and illustrated: in particular, Figure 1 provides a useful graphical summary of the method. The method and two new case studies are described in detail. I believe this is a very nice contribution, which fits the scope of JOBIM 2024 very well.

My questions are the following:

1) The max-squared test is by construction valid when the two tests are independent. In the context of (high-dimensional) mediation analysis, it seems likely that the two p-values corresponding to a

given potential mediator will be correlated even in absence of actual mediation. In this case, the max-squared p-value could be invalid. This point deserves to be discussed in the manuscript. Adding (in Supplementary Materials) plots showing the distribution of HDMAX2 p-values in both of the use cases considered, similar to Fig S3 in Jumentier et al (2023), would strengthen the manuscript.

Such a plot could also be added to the current vignette (simulated data).

We thank the reviewer for this important point about the validity of the max-squared test when p-values are correlated. We have conducted additional analyses to address this concern:

- *We calculated the correlation between the two p-values in all three use cases presented in the package vignette. Results show very weak correlations between the p-values, supporting the independence assumption. The correlation plot was also added for both use cases and the simulated example in the package vignette.*
- *We have added histograms of HDMAX2 p-values for both use cases and the simulated example in the package vignette. In all three scenarios, these histograms exhibit the expected uniform distribution under the null hypothesis, with a peak near zero corresponding to candidate mediators. This pattern supports the statistical validity of our approach.*
- *The combination of weak p-value correlations and proper uniform distribution of p-values under the null hypothesis provides strong evidence for the validity of the max-squared test in our context.*
- *We added a sentence to highlight the max-squared assumptions in the manuscript (line 149): 'We provide package vignettes to demonstrate the use of the package with simulated data and the two use cases presented in the article. The vignettes include descriptive plots to verify that the assumptions of the max-squared test are met, namely weak correlations among p-values and a uniform distribution of p-values.'*

2) Have the p-values obtained in use case 1 "HER2 and breast cancer" been **a**djusted for multiple testing?

Regarding multiple testing adjustment in the HER2 and breast cancer use case: since we focused on the top 10 mediators rather than testing significance, no multiple testing adjustment was required. The top-ranking approach was used to identify the most promising mediators for further investigation. For sake of clarity, we added in the manuscript the following sentence, line 180: 'Since our focus was on selecting the top 10 mediators rather than testing for statistical significance, no false discovery rate (FDR) control or multiple testing adjustment was required.'

3) A nice feature of the method is that it offers a statistically-grounded way to decide from the data which variables to include as mediators for the second analysis step. However, in the two applications described in the paper, the **final choice seems to have been made somewhat arbitrarily** (top 10 and top 2 scoring mediators, respectively). Can the authors discuss the influence of this choice on the results and their interpretation?

We added in the Description of the package vignette of the Results section (line 152) the following sentences:

‘For each example, the vignettes illustrate (i) the use of {\tt hdmax2::run_AS()}, (ii) the selection of potential mediators, (iii) the application of the {\tt hdmax2::estimate_effect()} function, and (iv) visualization with the {\tt hdmax2::plot()} function. The selection of candidate mediators is inherently tied to the biological question under investigation and the intended purpose of the results (e.g., experimental validation of candidates, public health recommendations, or data mining). Users can adopt various approaches for mediator selection based on their objectives, such as a top-ranking approach, statistical testing, FDR control, multiple testing correction, or p-value aggregation. In this paper and the package vignettes, we present two use cases to demonstrate the functionality of the {\tt hdmax2} package with different types of exposures and outcomes in high-dimensional mediation analysis. For simplicity, we opted for a top-ranking approach in both examples. Finally, the vignettes also illustrate the use of the package’s helper functions.’

4) The choice of the number of latent components is an ubiquitous problem which induces some level of arbitrariness in any data analysis. While one can not expect the authors to solve this problem in general, it would be useful if they could discuss the influence of the choice of the number K of latent components on the results in the two use cases. **Are the results somewhat robust to this choice**?

We acknowledge this important point and have added substantial analysis and discussion in the manuscript:

We added a new supplementary figure (Figure 1) showing (i) scree plots for each use case and (ii) upset plots comparing selected mediators with k, k-1, and k+1 for both use cases

We also added explicit discussion of K selection and robustness in the manuscript:

- *For use case 1 (line 177 and 182): 'For the initial step of the HDMAX2 approach, we opted to use K=2 latent factors in the association study.* ***The choice of K was guided by the PCA scree plot (Supplementary Figure 1A).*** *Subsequently, we identified the top 10 potential mediators with the lowest max-squared p-values. Since our focus was on selecting the top 10 mediators rather than testing for statistical significance, no false discovery rate (FDR) control or multiple testing adjustment was required.* ***We observed that the selection of the top 10 mediators is only partially robust to the choice of K=2, suggesting that, depending on the downstream application, users may wish to evaluate multiple values of K and consider the intersection of the resulting mediators (Supplementary Figure 1B-C).'***
- *For use case 2 (line 219): '****We used K=2 latent factors based on the PCA scree plot (Supplementary Figure 1D)****’. And line 244 : ‘****Given that different K values produce slightly different top mediator lists, further investigation may benefit from exploring additional potential mediators (Supplementary Figure 1E-F)****.'*

5) I appreciate that the authors have made available a vignette to analyze simulated data based on a TCGA study, including an example of plot corresponding to Figure 2. However, this vignette does not seem to be finalized (as of December 16, 2024) as it contains the mention: "THIS VIGNETTE IS CURRENTLY UNDER DEVELOPMENT, SO ITS CONTENT IS PROVISIONAL". Moreover, given the focus of the manuscript with respect to the methodological paper already published by the authors

(Jumentier et al, 2023), the authors should also **provide vignettes corresponding to the two use cases** highlighted in the manuscript.

The vignettes have been updated and finalized: we removed the development notice and added comprehensive documentation for both use cases described in the manuscript.

Minor:

- caption of Fig 3: "Total number of individuals"

- line 116: "will directly impact"

Minor corrections have been done.

Review questions:

Title and abstract

- Does the title clearly reflect the content of the article? Yes
- Does the abstract present the main findings of the study? Yes

Introduction

- Are the research questions/hypotheses/predictions clearly presented? Yes
- Does the introduction build on relevant research in the field? Yes

Materials and methods

- **Are the methods and analyses sufficiently detailed to allow replication by other researchers? No: I recommend that the authors provide Rmarkdown vignettes to reproduce their analysis of the two use cases considered in the paper**

These modification have been done, see point (5).

- Are the methods and statistical analyses appropriate and well described? Yes

Results

- In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? Yes
- Are the results described and interpreted correctly? Yes

Discussion

- Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? Yes
- Are the conclusions adequately supported by the results (without overstating the implications of the findings)? Yes

—--

Review by Gaspard Kerner

Summary of the paper:
The paper presents hdmax2, an R package designed to facilitate high-dimensional mediation analysis. The package builds upon the HDMAX2 framework, which integrates latent factor mixed models (LFMM) to estimate unobserved confounders and a max-squared test to identify significant mediators. This package represents a significant step forward in making complex mediation analysis more accessible and robust. A key strength of the package lies in its versatility. hdmax2 accommodates a variety of data types, including univariate and multivariate exposures and binary or continuous outcomes. This flexibility makes it a valuable tool for researchers analyzing high throughput molecular data, such as DNA methylation or gene expression, where the number of mediators often far exceeds the number of samples.

The paper showcases the package through two case studies:
1) Breast Cancer and HER2 Status: Explores the mediating role of DNA methylation in the pathway linking HER2-positive breast cancer status to a survival risk score.
2) Gender and Multiple Sclerosis (MS) Subtypes: Investigates gene expression as a potential mediator in the pathway linking gender to MS subtypes.

The package includes visualization tools, helper functions for mediator selection, and options for handling multivariate exposures. However, the paper focuses on univariate exposure models, leaving the multivariate capabilities underexplored.

My comments are primarily minor and aimed at enhancing clarity and providing additional context in a few areas.

We thank the reviewer for its time and constructive comments.

Minor comment on Abstract:
The abstract summarizes the purpose and contributions of the study, emphasizing the development of a method that addresses statistical challenges in high-dimensional mediation analysis. However, it would benefit from explicitly connecting the features of the package to the case studies presented, particularly in explaining how the results demonstrate its utility.

We have addressed the suggestion about the abstract by adding the following sentence: 'We demonstrate its application through two high-dimensional case studies examining DNA methylation and gene expression as mediators, with binary outcomes and both continuous and binary exposures. These examples illustrate practical aspects of the method, including latent factor selection and mediator identification.'

Minor comment on Materials and Methods

The methodology is described in sufficient detail, including the use of LFMM for estimating latent confounders, the max-squared test for assessing mediators, and the R package's features. Figure 1 nicely illustrates the workarounds of the package. A significant limitation is the lack of discussion on the selection and interpretation of the number of latent factors K. For example, in the breast cancer case study, K=2 is mentioned without further justification, and for the MS case study, K is not mentioned at all.

Regarding the selection of K, we have thoroughly addressed this concern (also corresponding to point (4) of reviewer 1):

We added a new supplementary figure (Figure 1) showing (i) scree plots for each use case and (ii) upset plots comparing selected mediators with k, k-1, and k+1 for both use cases

We also added explicit discussion of K selection and robustness in the manuscript:

- ○ *For use case 1 (line 177 and 182): 'For the initial step of the HDMAX2 approach, we opted to use K=2 latent factors in the association study.* ***The choice of K was guided by the PCA scree plot (Supplementary Figure 1A).*** *Subsequently, we identified the top 10 potential mediators with the lowest max-squared p-values. Since our focus was on selecting the top 10 mediators rather than testing for statistical significance, no false discovery rate (FDR) control or multiple testing adjustment was required.* ***We observed that the selection of the top 10 mediators is only partially robust to the choice of K=2, suggesting that, depending on the downstream application, users may wish to evaluate multiple values of K and consider the intersection of the resulting mediators (Supplementary Figure 1B-C).'***
- ○ *For use case 2 (line 219): '****We used K=2 latent factors based on the PCA scree plot (Supplementary Figure 1D)****'. And line 244 : '****Given that different K values produce slightly different top mediator lists, further investigation may benefit from exploring additional potential mediators (Supplementary Figure 1E-F).'***

Minor comment on Case study 1:

The breast cancer case study presents both the total effect (adjusted for age) and the mediation results. However, the total effect of 0.30 is introduced in a somewhat disconnected manner. If it is part of a preliminary analysis, this should be clarified, and its relevance to the HDMAX2 results (e.g., comparison to indirect effects in Figure 2C) should be explicitly discussed.

Regarding the total effect discussion in Case study 1, we have clarified its context and relevance:

Line 176: 'In a preliminary analysis, after adjusting for the confounding effect of age, we found that the total effect of HER2-positive status resulted in a 0.30 higher risk score (t-test, p=0.007, sd = 0.11).'

Line 190: 'This indicates that the mediated effect is detrimental to patient survival, resulting in the observed total effect of 0.23 (sd = 0.11) on the risk score, similar to our preliminary observation.'

Minor comment on Case study 2:
It is worth questioning why the authors chose this case study. While the initial motivation to assess the relationship between gender and MS is clear and compelling, and the discussion of negative results can be valuable, the authors acknowledge that the lack of significant findings is likely due to the small sample size. They proceed to discuss the top-ranked, albeit not statistically significant, results and provide reasoning for their potential biological relevance. If the authors believe that limited statistical power might be a common challenge when applying the hdmax2 model, it would be beneficial to include a power analysis in the paper to better address such scenarios

While we acknowledge the importance of statistical power, conducting a formal power analysis for high-dimensional mediation is complex due to several factors:

- *Our experience with epigenetic mediation studies typically identifies around ten mediators, which aligns with theoretical expectations (as EWAS studies identify ~100 candidates, and √100 ≈ 10). To achieve great power in mediation studies, sample sizes would need to be squared compared to traditional association studies, making such studies prohibitively expensive.*
- *A study-specific power analysis would require simulation parameters closely matching the observed data structure, which is challenging to establish reliably.*

We have added a paragraph discussing this aspects in the discussion (line 260):
‘To take sample size in consideration in high-dimensional mediation studies, we suggest not relying solely on statistical significance, instead we recommend using FDR control with flexible thresholds. For small candidate lists (like our top 10), higher FDR thresholds (e.g., 30%) may be acceptable, allowing for individual examination of candidates while acknowledging potential false discoveries. While limited sample sizes affect statistical power, results can still be meaningful when analyzed with appropriate FDR thresholds and biological validation.’

Additional minor comments:
1) Is this sentence accurate: “Upon observing a significant decrease in CIS-MS occurrence among women (see Fig 3A), we sought to investigate this phenomenon further”? (line 187-188) (now line 207) Did the authors intend to refer to RR-MS instead of CIS-MS?

Yes, this was an error. We have corrected the text to refer to RR-MS instead of CIS-MS.

2) The statement, “Remarkably, most of the top 10 identified mediators were associated with genes known to be involved in breast cancer biology, thus supporting the biological relevance of our approach,” is likely accurate. However, it appears to rely on “PubMed hits,” defined as the number of outputs from the search "(Breast cancer) AND (‘Gene Symbol‘)." I recommend that the authors either explicitly mention in the text the methodology used to link genes and breast cancer or refine their search methodology to provide stronger evidence supporting this claim.

We have clarified our PubMed search methodology by adding explicit details about our search strategy (line 200): This association relies on `PubMed hits' defined as the number of outputs from the search (Breast cancer) AND (Gene Symbol), it represents a preliminary assessment of biological relevance that should further be investigated by experimental approaches.