Peer Community In Genomics

High-dimensional mediation analysis: Unraveling pathways linking external exposures to health outcomes

Guillaume Laval *based on peer reviews by* ***Gaspard Kerner*** *and* ***Pierre Neuvial***

Florence Pittion, Basile Jumentier, Aurélie Nakamura, Johanna Lepeule, Olivier François, Magali Richard (2025) hdmax2, an R package to perform high dimension mediation analysis. HAL, ver. 4, peer-reviewed and recommended by Peer Community in Genomics. https://hal.science/hal-04658960

Submitted: 12 September 2024, Recommended: 10 March 2025

Cite this recommendation as:
Laval, G. (2025) High-dimensional mediation analysis: Unraveling pathways linking external exposures to health outcomes. *Peer Community in Genomics*, 100416. 10.24072/pci.genomics.100416

Published: 10 March 2025

Pittion et al. (2025) introduce an R package called hdmax2, which implements an enhanced version of the "High-Dimensional Mediation Analysis using the Max-Squared" (HDMAX2) method originally proposed by Jumentier et al. (2023) for high-dimensional mediation analysis. The goal of mediation analysis is to quantify the indirect effect of a variable M in the causal relationship between exposure X and outcome Y. The fundamental concept behind HDMAX2 methods is to use a latent factor mixed model to estimate the effects of unobserved confounders and a max-squared test to identify significant mediators. The HDMAX2 method represents a significant advancement in the case of high-dimensional mediation, such as DNA methylation or gene expression analysis, where the number of mediators often far exceeds the sample size.

The main contributions of this article are the implementation of the HDMAX2 method as an R package, and an extension of the original method to binary outcomes and to binary, categorical, and multivariate exposures, as opposed to only continuous variables. The package includes visualization tools, helper functions for mediator selection, and options for handling multivariate exposures. A key strength of the package lies in its versatility. The new package, hdmax2, accommodates a variety of data types. This flexibility makes it a valuable tool for researchers analyzing high-throughput molecular data. Finally to illustrate this flexibility, the authors present two case studies that were not described in the Jumentier et al. (2023) analysis. In the first case study, the authors employed mediation analysis to assess the potential causal role of DNA methylation in the pathway linking the HER2 status of breast cancer (a marker for an aggressive breast cancer subtype) to a survival risk score, which was derived from a six-gene expression signature and is inversely correlated with patient survival.

In the second case study, the authors conducted mediation analysis to explore the role of gene expression in the pathway linking patient gender to the occurrence of multiple sclerosis specific subtypes: clinically isolated syndrome and relapsing-remitting multiple sclerosis. These illustrate the relevance of hdmax2 to study the transcriptome and the methylome.

In conclusion, the hdmax2 R package will be invaluable for handling high-dimensional molecular data in the study of the intricate pathways through which exposures influence health outcomes.

References:

Jumentier B, Barrot C-C, Estavoyer M, Tost J, Heude B, François O, Lepeule J (2023) High-dimensional mediation analysis: A new method applied to maternal smoking, placental DNA methylation, and birth outcomes. Environmental Health Perspectives, 131, 047011. https://doi.org/10.1289/EHP11559

Pittion F, Jumentier B, Nakamura A, Lepeule J, Francois O, Richard M (2025) hdmax2, an R package to perform high dimension mediation analysis. HAL, ver. 4 peer-reviewed and recommended by PCI Genomics https://hal.science/hal-04658960

Reviews

Evaluation round #1

DOI or URL of the preprint: https://hal.science/hal-04658960
Version of the preprint: 3

Authors' reply, 14 February 2025

Download author's reply

Decision by Guillaume Laval, posted 20 December 2024, validated 23 December 2024

This paper presents the R package hdmax2, which implements a method for high-dimensional mediation analysis recently published by some of the authors. The two reviewers are rather positive, making some suggestions and comments that should be easily addressed. A revision that takes account of these points will make for a much stronger paper.

PCI Genomics Managing Board note: This preprint was originally reviewed as part of the JOBIM conference (Les Journées Ouvertes en Biologie, Informatique et Mathématiques), which is why one reviewer refers to this preprint as fitting the scope of JOBIM. This review was originally conducted for the JOBIM conference, but was expanded upon once the preprint was submitted to PCI Genomics.

Reviewed by Pierre Neuvial, 20 December 2024

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

2) Have the p-values obtained in use case 1 ”HER2 and breast cancer” been adjusted for multiple testing?

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?

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?

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.

Minor:

- caption of Fig 3: ”Total number of individuals”
- line 116: ”will directly impact”

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

considerd in the paper
- 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

Reviewed by Gaspard Kerner, 30 November 2024

Download the review