Journal of Data Science 0 (0), 1–15
??? 2025
DOI: 10.6339/25-JDS1192
Computing in Data Science

HIMA: An R Package for High-Dimensional Mediation Analysis

Haixiang Zhang[1,†], Yinan Zheng[2,†], Lifang Hou[2], and Lei Liu[3,*]

[1]*School of Mathematics and KL-AAGDM, Tianjin University, Tianjin 300350, China*
[2]*Department of Preventive Medicine, Northwestern University, Chicago IL 60611, USA*
[3]*Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA*

Abstract

Mediation analysis plays an important role in many research fields, yet it is very challenging to perform estimation and hypothesis testing for high-dimensional mediation effects. We develop a user-friendly R package *HIMA* for high-dimensional mediation analysis with varying mediator and outcome specifications. The *HIMA* package is a comprehensive tool that accommodates various types of high-dimensional mediation models. This paper offers an overview of the functions within *HIMA* and demonstrates the practical utility of *HIMA* through simulated datasets. The *HIMA* package is publicly available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=HIMA.

Keywords *DNA methylation; mediator selection; penalized estimate; quantile regression*

1 Introduction

Mediation analysis is a statistical methodology employed to investigate the underlying mechanisms through which an independent variable influences a dependent variable through an intermediary variable (mediator). This approach typically involves evaluating indirect, direct, and total effects within a specified statistical framework (Baron and Kenny, 1986). A key objective of mediation analysis is to test the mediation hypothesis, which suggests that the impact of the independent variable on the dependent variable is transmitted through one or more mediating variables. By investigating these pathways, we can achieve a more comprehensive understanding of the causal relationships among the variables. For instance, mediation analysis can enhance the refinement of interventions to improve their efficacy in clinical trials. In observational studies, mediation analysis can serve to pinpoint intervention targets and elucidate the underlying mechanisms of diseases. In recent years substantial research efforts have been dedicated to developing methodologies for mediation analysis. For example, Shen et al. (2014) provided an approach for estimating quantile mediation effects. Sun et al. (2021) introduced a Bayesian framework for conducting mediation analysis. Zhang and Li (2023) considered the statistical mediation analysis on large-scale datasets. Zhang (2025) introduced an efficient adjusted joint significance test, along with a Sobel-type confidence interval for assessing the mediation effect.

With the technological advancement in data collection, high-dimensional mediation analysis has become an indispensable tool in omic studies. Works include linear mediation models for continuous outcomes (Zhang et al., 2016; James et al., 2022; Perera et al., 2022), logistic mediation models for binary outcomes (Wu et al., 2018), mediation models for censored survival

*Corresponding author. Email: lei.liu@wustl.edu.
†Co-first authors.

Received August 30, 2024; Accepted June 23, 2025

outcomes (Luo et al., 2020; An and Zhang, 2023), compositional mediation analysis for microbiome data as mediators (Zhang et al., 2018; Wang et al., 2020; Zhang et al., 2021a; 2021b), quantile mediation analysis (Zhang et al., 2024) for quantiles of continuous outcomes. For more information, please refer to Zhang et al. (2022).

The field of mediation analysis offers various R packages. For example, *BayesGmed* provides Bayesian causal mediation analysis (Yimer et al., 2023); *mma* supports mediation analysis with multiple mediators (Yu and Li, 2017). In view of the growing interest and research activity, it is desirable to develop a user-friendly and comprehensive R package on high-dimensional mediation analysis. In this paper, we introduce the comprehensive R package *HIMA* based on several published papers, including Zhang et al. (2016; 2021a; 2021b; 2021), Perera et al. (2022), Zhang et al. (2024) and Bai et al. (2024). From a practical application perspective, we have developed a function `hima()` to implement various statistical methodologies for estimating and testing high-dimensional mediation effects. The `hima()` function is capable of automatically selecting the suitable method based on the data types of the outcome and mediator variables. Note that the `hima()` function is designed to be extensible, allowing future features and methods to seamlessly integrate within its framework, ensuring a consistent and user-friendly experience.

The rest of the article is organized as follows: In Section 2, we briefly review several methods for high-dimensional mediation analysis. In Section 3, we present detailed arguments for the `hima()` function. In Section 4, we illustrate the applications of the package to various models, including linear mediation models with continuous outcomes, logistic mediation models with binary outcomes, Cox mediation models for censored survival outcomes, compositional mediation analysis for microbiome data, and quantile mediation analysis. Section 5 concludes with a discussion.

2 Methods

In this section, we review the following high-dimensional linear mediation model (Figure 1):

$$Y = c + \gamma X + \beta_1 M_1 + \cdots + \beta_p M_p + \boldsymbol{\eta}'\mathbf{Z} + \epsilon, \tag{1}$$

$$M_k = c_k + \alpha_k X + \boldsymbol{\zeta}_k'\mathbf{Z} + e_k, \quad k = 1, \ldots, p, \tag{2}$$

where X is the treatment (or exposure), Y is the outcome, $\mathbf{Z} = (Z_1, \ldots, Z_q)'$ is a vector of confounding variables or covariates, M_k's are potential mediators; γ is the "direct effect" of X on Y, after adjusting for all mediators and covariates. Furthermore, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a vector of parameters relating the treatment to p mediating variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of parameters relating the mediators to the dependent variable adjusting for the effects of the treatment and covariates. $\boldsymbol{\zeta}_k$'s and $\boldsymbol{\eta}$ are the parameters of covariates. In addition, c and c_k's are the intercept terms; ϵ and e_k's are error terms. The "indirect effect" in the path $X \to M_k \to Y$ is $\alpha_k\beta_k$, $k = 1, \ldots, p$. Let $S_0 = \{k : \alpha_k\beta_k \neq 0, k = 1, \ldots, p\}$ be the index set of significant mediators.

The observed data are represented as $(X_i, Y_i, \mathbf{Z}_i, \mathbf{M}_i)$, where $\mathbf{M}_i = (M_{i1}, \ldots, M_{ip})'$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})'$, for $i = 1, \ldots, n$. It is important to note that the dimension of mediators p is much larger than the sample size n. Consequently, many traditional regression methods fail in Equation (1). To address this issue, we have proposed a novel statistical inference procedure for high-dimensional mediation effects in Zhang et al. (2016), which is briefly summarized as the "three-step approach":

Step 1. (Mediator screening). First, mediators are scaled with mean zero and variance one. Perform a series of marginal regression models for p mediators:

$$Y = c + \gamma X + \beta_k M_k + \boldsymbol{\eta}'\mathbf{Z} + \epsilon, \quad k = 1, \ldots, p, \tag{3}$$

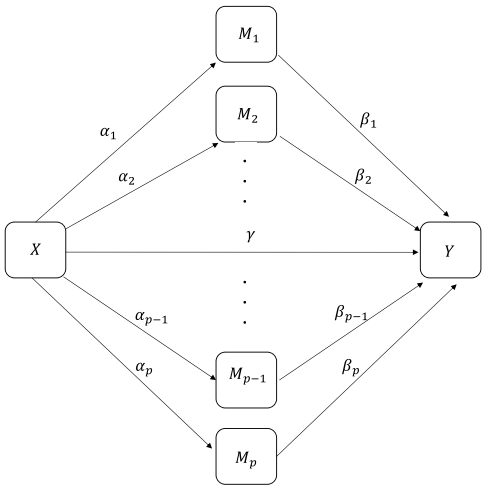

Figure 1: A scenario of high-dimensional mediation model, where X is an treatment, M_k's are mediators, and Y is the outcome.

where the ordinary least squares estimators of β_k's in (3) are denoted as $\breve{\beta}_k$, $k = 1, \ldots, p$. Along the lines of Fan and Lv (2008), we can identify a subset $\Omega_1 = \{1 \leqslant k \leqslant p$: M_k is among the top d mediators with the largest marginal effects $|\breve{\beta}_k|$ for the response $Y\}$, where d is an integer specifying the number of top mediators. e.g., $d = [2n/\log(n)]$.

Step 2. (Penalized estimate). Conduct variable selection for the mediators $\{M_k\}_{k\in\Omega_1}$ by minimizing the penalty-based criterion,

$$Q^{ols} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - c - \gamma X_i - \sum_{k\in\Omega_1}\beta_k M_{ik} - \boldsymbol{\eta}'\mathbf{Z}_i\right)^2 + \sum_{k\in\Omega_1} pen(\beta_k), \tag{4}$$

where $pen(\cdot)$ is a penalty function, such as MCP (Zhang, 2010).

Step 3. (Joint significance test). Let $\Omega_2 = \{k : \hat{\beta}_k \neq 0, k = 1, \ldots, p\}$ be the index set of the mediators survived in Step 2, where $\hat{\beta}_k$'s are the penalized estimates in (4). We consider the multiple testing problem:

$$H_{0k} : \alpha_k\beta_k = 0 \leftrightarrow H_{1k} : \alpha_k\beta_k \neq 0, \quad k \in \Omega_2. \tag{5}$$

The idea of joint significance (JS) test is that H_{0k} is rejected when both $\alpha_k = 0$ and $\beta_k = 0$ are simultaneously rejected. To control the family wise error rate (FWER), Zhang et al. (2016) proposed Bonferroni corrected p-values: $P_{corr,k} = \max(P_{corr,1k}, P_{corr,2k})$, where $P_{corr,1k} = \min(P_{raw,1k} \cdot |\Omega_2|, 1)$ and $P_{corr,2k} = \min(P_{raw,2k} \cdot |\Omega_2|, 1)$ with $P_{raw,1k} = 2\{1 - \Phi(|\hat{\beta}_k|/\hat{\sigma}_{1k})\}$, and $P_{raw,2k} = 2\{1 - \Phi(|\hat{\alpha}_k|/\hat{\sigma}_{2k})\}$. Here $\hat{\alpha}_k$ is the ordinary least squares estimator for α_k and $\hat{\sigma}_{2k}$ is the corresponding estimated standard error; $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, and $\hat{\sigma}_{1k}$ is the estimate of standard error for $\hat{\beta}_k$; $|\Omega_2|$ is the cardinality, i.e., the number of elements in the set Ω_2. An estimated index set of significant mediators is $\hat{S} = \{k : P_{corr,k} < 0.05, k \in \Omega_2\}$.

A comparable methodology can be applied to high-dimensional mediation models with a binary outcome. In addition, Zhang et al. (2021), Zhang et al. (2021a) and Zhang et al. (2024) investigated the survival mediation model, the compositional mediation model for microbiome mediators, and the quantile mediation model, respectively.

3 Arguments for `hima()`

Before installing *HIMA*, please ensure that the R package *qvalue* (Storey et al., 2024) is installed through Bioconductor:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("qvalue")
```

The R package *HIMA* is publicly available from the Comprehensive R Archive Network at https://CRAN.R-project.org/package=HIMA. To install from CRAN:

```
install.packages("HIMA")
```

The arguments and outputs of the function `hima()` are presented in Tables 1 and 2, respectively. To use the `hima()` function, ensure that your data is prepared according to the following guidelines:

3.1 Formula Argument (`formula`)

Define the model formula to specify the relationship between the outcome, treatment, and covariates. Ensure the following:

- **General Form**: Use the format `Outcome ~ Exposure + Covariates`. Note that the `Exposure` variable represents the exposure of interest and it has to be listed as the first independent variable in the formula. Covariates are optional.
- **Survival Data**: For survival outcomes, use the format `Surv(Time, Status) ~ Exposure + Covariates`. See data examples `SurvivalData$PhenoData` for more details.

3.2 Phenotype Data (`data.pheno`)

The `data.pheno` object should be a `data.frame` or `matrix` containing the phenotype information for the analysis. Key requirements include:

- **Rows:** Represent samples.
- **Columns:** Include variables such as the outcome, treatment, and optional covariates.
- **Formula Consistency:** Ensure that all variable names specified in the `formula` argument (e.g., `Outcome`, `Treatment`, and `Covariates`), exactly match the column names in the `data.pheno`.

3.3 Mediator Data (`data.M`)

The `data.M` object should be a `data.frame` or `matrix` containing high-dimensional mediators. Key requirements include:

- **Rows:** Represent samples, aligned with the rows in `data.pheno`.
- **Columns:** Represent mediators (e.g., CpGs, genes, or other molecular features).
- **Mediator Type:** Specify the type of mediators in the `mediator.type` argument. Supported types include (more types will be added in the future):
  - "`gaussian`" for continuous mediators (default, e.g., DNA methylation data).
  - "`compositional`" for microbiome or other compositional data.

Table 1: Overview of the arguments in function `hima()`.

Arguments	Description
`formula`	An object of class formula representing the overall effect model to be fitted, specified as `outcome` ~ `exposure` + `covariates`. The "exposure" variable (the variable of interest) must be listed first on the right-hand side of the formula; For survival outcomes, use the format `Surv(time, status)` ~ `exposure` + `covariates`
`data.pheno`	A data frame containing the exposure, outcome, and covariates specified in the formula. Variable names in data.pheno must match those in the formula. When `scale` = `TRUE`, the exposure and covariates will be scaled with mean zero and variance one (the outcome retains its original scale)
`data.M`	A data.frame or matrix of high-dimensional mediators, with rows representing samples and columns representing mediator variables. When `scale` = `TRUE`, `data.M` will be scaled with mean zero and variance
`mediator.type`	A character string indicating the data type of the high-dimensional mediators (`data.M`). Options are: "`gaussian`" (default): for continuous mediators; "`compositional`": for compositional data mediators (e.g., microbiome data)
`penalty`	A character string specifying the penalty method to apply in the model. Options are: "`DBlasso`": De-biased LASSO; "`MCP`": Minimax Concave Penalty. Note: Survival HIMA and microbiome HIMA can only be performed with "`DBlasso`". Quantile HIMA and efficient HIMA can only apply "`MCP`"
`quantile`	Logical. Indicates whether to use quantile HIMA (hima_quantile). Default is `FALSE`. If `TRUE`, specify the desired quantile(s) using the `tau` parameter; otherwise, the default `tau` = 0.5 (i.e., median) is used
`efficient`	Logical. Indicates whether to use efficient HIMA (hima_efficient). Default is `FALSE`. Applicable only for classic HIMA with a continuous outcome and `mediator.type` = "`gaussian`"
`scale`	logical. Determines whether the exposure (variable of interest), mediators, and covariate(s) (if included) are standardized to a mean of zero and a variance of one. No scaling will be applied to Outcome
`sigcut`	Numeric. The significance cutoff for selecting mediators. Default is 0.05
`contrast`	A named list of contrasts to be applied to factor variables in the covariates (cannot be the variable of interest)
`subset`	An optional vector specifying a subset of observations to use in the analysis
`verbose`	Logical. Determines whether the function displays progress messages. Default is `FALSE`

3.4 Data Scaling

If `scale` is set to `TRUE`, the exposure, mediators, and covariate(s) (if included) are standardized to a mean of zero and a variance of one. No scaling will be applied to outcome.

Table 2: The details on outputs from function `hima()`.

Values	Description
`ID`	Mediator ID or name
`alpha`	Coefficient estimates of treatment (X) → mediators (M) (adjusted for covariates)
`beta`	Coefficient estimates of mediators (M) → outcome (Y) (adjusted for covariates and treatment)
`alpha*beta`	The estimated indirect (mediation) effect of exposure on outcome through each mediator
`rimp`	The proportion of each mediator's mediation effect relative to the sum of the absolute mediation effects of all significant mediators
`p-value`	The joint p-value assessing the significance of each mediator's indirect effect, calculated based on the corresponding statistical approach
`tau`	The quantile level of the outcome (applicable only when using the quantile mediation model)

4 Illustrations

4.1 Linear Mediation Models with Continuous Outcomes

In this section, we provide guidance on how to use the function `hima()` for analyzing high-dimensional linear mediation models with continuous outcomes (Zhang et al., 2016). In that study, the authors applied the method to the Normative Aging Study (NAS), a longitudinal cohort initiated in 1963 (Bell et al., 1966), to investigate how DNA methylation markers mediate the relationship between smoking and lung function. The analysis included 290 male participants without lung disease, with DNA methylation profiled using the Illumina 450K BeadChip, yielding 484,548 CpG sites as potential mediators. Lung function was assessed using four continuous outcomes: FEV1, FVC, FEV1/FVC, and MMEF. Since the original application dataset is not publicly available, we illustrate the use of the method with a simulated example provided in *HIMA*.

Step A (Data Preparation): The high-dimensional mediators (corresponding to `data.M`) are structured as a data.frame or matrix, with rows denoting samples and columns representing mediator variables. The exposure, outcome, and covariates (corresponding to `data.pheno`) are organized as a data.frame, where the variable names in `data.pheno` must correspond exactly to those specified in the formula.

```
library(HIMA)
data(ContinuousOutcome)
pheno_data <- ContinuousOutcome$PhenoData
mediator_data <- ContinuousOutcome$Mediator
head(pheno_data)
```

```
  Treatment   Outcome Sex Age
1         1 3.9535874   0  19
2         1 4.9808737   0  25
```

```
3          1 6.3215341   1  56
4          0 3.9490133   0  56
5          0 0.2820298   0  50
6          1 2.0730989   1  37
```

```
str(mediator_data)
```

```
 num [1:500, 1:100] 0.337 -0.318 2.834 2.096 0.717 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:500] "S1" "S2" "S3" "S4" ...
  ..$ : chr [1:100] "M1" "M2" "M3" "M4" ...
```

Step B (Function Execution): The function `hima()` can be executed as follows:

```
e1 <- hima(Outcome ~ Treatment + Sex + Age,
      data.pheno = pheno_data,
      data.M = mediator_data,
      mediator.type = "gaussian",
      penalty = "MCP",
      scale = FALSE
      )
```

The implementation of the `hima()` function for the method described in Perera et al. (2022) follows the same procedure as in Step B, with the exception of the configuration where `penalty` = "`DBlasso`". The `hima()` function for the method in Bai et al. (2024) follows Step B, except when `efficient` = `TRUE`.

Step C (Outputs of Significant Mediators): Based on the execution of the function `hima()`, the significant mediators are identified, with detailed information presented in Table 2. For example, it is evident from the results that our method accurately identifies the three significant mediators, denoted as M_1, M_2, and M_3.

```
summary(e1)
```

```
Summary of HIMA results:
----------------------------------
Number of significant mediators: 3

Top mediators (sorted by p-value):
  ID      alpha       beta alpha*beta     rimp      p-value
1 M3 -0.9803014 -0.7443280  0.7296657 39.74247 2.049845e-12
2 M2  0.8627820  0.7738197  0.6676377 36.36401 1.203775e-09
3 M1  0.4507821  0.9731558  0.4386812 23.89351 1.109704e-03
```

4.2 Logistic Mediation Models with Binary Outcomes

In this section, we investigate the application of the `hima()` function to high-dimensional mediation models with binary outcomes, which involves the following steps:

Step A (Data Preparation): The high-dimensional mediators (corresponding to `data.M`) are structured as a data.frame or matrix, with rows denoting samples and columns representing mediator variables. The exposure, outcome, and covariates (corresponding to `data.pheno`) are organized as a data.frame, where the variable names in `data.pheno` must correspond exactly to those specified in the formula.

```
library(HIMA)
data(BinaryOutcome)
pheno_data <- BinaryOutcome$PhenoData
mediator_data <- BinaryOutcome$Mediator
head(BinaryOutcome$PhenoData)
```

```
  Treatment Disease Sex Age
1         1       0   0  50
2         0       1   1  55
3         0       1   1  28
4         0       0   0  55
5         0       0   0  62
6         0       0   0  33
```

```
str(mediator_data)
```

```
 num [1:500, 1:100] 0.038618 2.496059 -0.000867 0.860103 1.301687 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:500] "S1" "S2" "S3" "S4" ...
  ..$ : chr [1:100] "M1" "M2" "M3" "M4" ...
```

Step B (Function Execution): The function `hima()` can be executed as follows:

```
e2 <- hima(Disease ~ Treatment + Sex + Age,
      data.pheno = pheno_data,
      data.M = mediator_data,
      mediator.type = "gaussian",
      penalty = "MCP",
      scale = FALSE
      )
```

Step C (Outputs of Significant Mediators): Based on the execution of the function `hima()`, the significant mediators are identified, with detailed information presented in Table 2. For example, it is evident from the results that our method accurately identifies the three significant mediators, denoted as M_1, M_2, and M_3.

```
summary(e2)
```

```
Summary of HIMA results:
---------------------------------
```

```
Number of significant mediators: 3

Top mediators (sorted by p-value):
  ID      alpha       beta alpha*beta     rimp      p-value
1 M3 -0.8669016 -0.8793065  0.7622722 39.82902 1.610871e-09
2 M1  0.7948203  0.6095748  0.4845024 25.31544 8.570595e-09
3 M2  0.7731996  0.8627613  0.6670867 34.85554 2.984815e-08
```

4.3 Survival Mediation Models with Censored Outcomes

In this section, we provide a guidance to demonstrate how the function `hima()` operates for high-dimensional survival mediation models with censored outcomes. The details regarding the simulated datasets can be found in Zhang et al. (2021). In that study, they applied the high-dimensional mediation method for survival to The Cancer Genome Atlas (TCGA) lung cancer cohort to identify DNA methylation markers mediating the effect of smoking on lung cancer survival (Tomczak et al., 2015). A total of 593 patients with complete clinical and epigenetic data were included in the study. The survival outcome is the number of days from diagnosis to death, with a 59% censoring rate. A total of 379,330 DNA methylation markers measured by the Illumina 450K array are considered as potential mediators. The exposure variable is pack-years of smoking, and the outcome is survival time. Due to data availability restrictions, we will provide a simulated dataset available in *HIMA* to illustrate our method.

Step A (Data Preparation): The high-dimensional mediators (corresponding to `data.M`) are structured as a data.frame or matrix, with rows denoting samples and columns representing mediator variables. We recommend the following layout of `data.pheno`: the first column represents the exposure, the second column corresponds to the event indicator status (`TRUE` denotes non-censored observations), the third column represents the observed failure time, and any subsequent columns represent adjusted covariates, where the variable names in `data.pheno` must correspond exactly to those specified in the formula.

```
library(HIMA)
data(SurvivalData)
pheno_data <- SurvivalData$PhenoData
mediator_data <- SurvivalData$Mediator
head(SurvivalData$PhenoData)
```

```
  Treatment Status       Time Sex Age
1         1   TRUE 0.01100768   0  32
2         1   TRUE 0.03030355   0  58
3         1   TRUE 0.05730512   1  28
4         0  FALSE 0.67173725   1  40
5         1   TRUE 0.01536887   0  41
6         1  FALSE 1.22992066   0  23
```

```
str(mediator_data)
```

```
 num [1:300, 1:100] 0.86 2.749 -0.578 0.176 0.803 ...
```

```
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:300] "S1" "S2" "S3" "S4" ...
  ..$ : chr [1:100] "M1" "M2" "M3" "M4" ...
```

Step B (Function Execution): The function `hima()` can be executed as follows:

```
e3 <- hima(Surv(Time,Status) ~ Treatment + Sex + Age,
      data.pheno = pheno_data,
      data.M = mediator_data,
      mediator.type = "gaussian",
      penalty = "DBlasso",
      scale = FALSE
      )
```

Step C (Outputs of Significant Mediators): Based on the execution of the function `hima()`, the significant mediators are identified, with detailed information presented in Table 2. For example, it is evident from the results that our method accurately identifies the three significant mediators, denoted as M_1, M_2, and M_3.

```
summary(e3)
```

```
Summary of HIMA results:
---------------------------------
Number of significant mediators: 3

Top mediators (sorted by p-value):
  ID      alpha       beta alpha*beta     rimp      p-value
1 M1  0.8972695  0.7862471  0.7054755 36.44005 3.019807e-14
2 M2  0.8510845  0.8201913  0.6980521 36.05660 1.115108e-12
3 M3 -0.6988647 -0.7618957  0.5324620 27.50335 2.586902e-10
```

4.4 Compositional Mediation Model for Microbiome Mediators

In this section, we demonstrate the use of the `hima()` function for high-dimensional mediation analysis with compositional microbiome data. The real dataset, described in Zhang et al. (2021a), comes from a murine study investigating how gut microbiota mediate the effect of antibiotic treatment on body weight. The study included 36 male mice, with microbial DNA profiled using the MO BIO PowerSoil kit and processed by the QIIME pipeline, yielding 149 genera. After filtering rare taxa, 36 taxa were retained. Zero counts were replaced with 0.5, and data were transformed to compositions. The exposure is subtherapeutic antibiotic treatment (X = 1 vs. 0), and the outcome is body weight on day 116. A simulated dataset from *HIMA* is used for illustration.

Step A (Data Preparation): (a) The data frame `data.pheno`: the first column represents the treatment, the second column represents the outcome and any subsequent columns represent adjusted covariates. (b) The data frame `data.M`: a data.frame or matrix of high-dimensional compositional matrix (mediators), i.e., the sum of each row in `data.M` is 1. In practical applications, there may be many zeros in the microbiome count data matrix M_{count}, which could

be replaced with 0.5 before being transformed into the compositional matrix `data.M`. Ensure that all variable names specified in the formula argument, such as `Outcome`, `Treatment`, and `Covariates`, exactly match the column names in the `data.pheno` dataset.

```
library(HIMA)
data(MicrobiomeData)
pheno_data <- MicrobiomeData$PhenoData
mediator_data <- MicrobiomeData$Mediator
head(MicrobiomeData$PhenoData)
```

```
  Treatment   Outcome Sex Age
1         0 4.8584340   0  59
2         1 2.8420944   0  33
3         0 1.7659530   1  63
4         0 3.8308817   1  47
5         1 1.5642848   0  21
6         0 0.1988363   1  23
```

```
str(mediator_data)
```

```
 num [1:300, 1:100] 0.0339 0.5481 0.0712 0.0186 0.1501 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:300] "S1" "S2" "S3" "S4" ...
  ..$ : chr [1:100] "M1" "M2" "M3" "M4" ...
```

Step B (Function Execution): The function `hima()` can be executed as follows:

```
e4 <- hima(Outcome ~ Treatment + Sex + Age,
      data.pheno = pheno_data,
      data.M = mediator_data,
      mediator.type = "compositional",
      penalty = "DBlasso",
      scale = FALSE
      )
```

Step C (Outputs of Significant Mediators): Based on the execution of the function `hima()`, the significant mediators are identified, with detailed information presented in Table 2. For example, it is evident from the results that our method accurately identifies the three significant mediators, denoted as M_1, M_2, and M_3.

```
summary(e4)
```

```
Summary of HIMA results:
----------------------------------
Number of significant mediators: 3
```

```
Top mediators (sorted by p-value):
  ID     alpha      beta alpha*beta     rimp p-value
1 M1 0.8161109  3.910193   3.191151 50.11182       0
2 M2 0.8171704 -2.250231  -1.838822 28.87570       0
3 M3 0.7920133 -1.689476  -1.338088 21.01248       0
```

4.5 Quantile Mediation Model

In this section, we provide a guidance to demonstrate how the function `hima()` operates for high-dimensional quantile mediation model. The details regarding the simulated datasets can be found in (Zhang et al., 2024). In that study, they applied the model to a birth cohort of 954 mother–newborn pairs from the Boston Birth Cohort, examining how DNA methylation mediates the impact of maternal smoking on quantiles of infant birthweight (Pearson et al., 2022). DNA methylation in each newborn was profiled using the Illumina Infinium MethylationEPIC BeadChip, yielding data for 865,859 CpG sites. Due to data availability constraints, we illustrate our method using a simulated dataset in *HIMA*.

Step A (Data Preparation): The high-dimensional mediators (corresponding to `data.M`) are structured as a data.frame or matrix, with rows denoting samples and columns representing mediator variables. Ensure that all variable names specified in the formula argument, such as `Outcome`, `Treatment`, and `Covariates`, exactly match the column names in the `data.pheno` dataset.

```
library(HIMA)
data(QuantileData)
pheno_data <- QuantileData$PhenoData
mediator_data <- QuantileData$Mediator
head(QuantileData$PhenoData)
```

```
   Treatment   Outcome Sex Age
1  1.2660838  3.303784   1  37
2  0.2494073  1.119954   1  46
3 -1.4294165 -2.194543   0  58
4 -1.2095963 -1.410079   0  63
5  2.9513044 12.654794   0  62
6  0.2592669  4.935507   0  27
```

```
str(mediator_data)
```

```
 num [1:500, 1:100] 1.656 0.924 1.317 -0.418 2.726 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:500] "S1" "S2" "S3" "S4" ...
  ..$ : chr [1:100] "M1" "M2" "M3" "M4" ...
```

Step B (Function Execution): The function `hima()` can be executed as follows:

```
e5 <- hima(Outcome ~ Treatment + Sex + Age,
      data.pheno = pheno_data,
      data.M = mediator_data,
      mediator.type = "gaussian",
      penalty = "MCP",
      quantile = TRUE,
      scale = FALSE,
      tau = 0.5
      )
```

Step C (Outputs of Significant Mediators): Based on the execution of the function `hima()`, the significant mediators are identified, with detailed information presented in Table 2. For example, it is evident from the results that our method accurately identifies the three significant mediators, denoted as M_1, M_2, and M_3.

```
summary(e5)
```

```
Summary of HIMA results:
----------------------------------
Number of significant mediators: 3

Top mediators (sorted by p-value):
  ID      alpha       beta alpha*beta     rimp p-value tau
1 M1  0.6546957  0.8922243  0.5841354 29.89891       0 0.5
2 M2  0.7861428  0.8769273  0.6893900 35.28636       0 0.5
3 M3 -0.9391433 -0.7242513  0.6801757 34.81473       0 0.5
```

Note that `tau` is the quantile level of outcome (default is 0.5). A vector of `tau` is also accepted. The details of outputs with `hima()` are presented in Table 2. The results show that our method accurately identifies the three significant mediators. Additionally, the estimators for both α_k and β_k are consistent.

5 Conclusions

For practical applications, we have introduced a comprehensive R package *HIMA* for high-dimensional mediation analysis. In this paper, we have provided details on the usages and arguments of functions in HIMA. Moreover, simulated datasets were used as illustrative examples for evaluating the performance of *HIMA*. The future developments of *HIMA* will involve the incorporation of additional functionalities, including the use of longitudinal biomarkers, multi-modal (e.g., omics and imaging) data as mediators, categorical variables and Poisson count outcomes to enhance high-dimensional mediation analysis.

Supplementary Material

In the Supplementary Material, we provide the R code implementations corresponding to the illustrations presented in Section 4.

Acknowledgments

The authors would like to thank the Editor and an Associate Editor for their constructive and insightful comments that greatly improved the manuscript.

References

An M, Zhang H (2023). High-dimensional mediation analysis for time-to-event outcomes with additive hazards model. *Mathematics*, 11: 1–11.

Bai X, Zheng Y, Hou L, Zheng C, Liu L, Zhang H (2024). An efficient testing procedure for high-dimensional mediators with FDR control. *Statistics in Biosciences*. https://doi.org/10.1007/s12561-024-09447-4

Baron RM, Kenny DA (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6): 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bell B, Rose CL, Damon A (1966). The veterans administration longitudinal study of healthy aging. *The Gerontologist*, 6: 179–184. https://doi.org/10.1093/geront/6.4.179

Fan J, Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 70: 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

James Y, Dai JLS, LeBlanc M (2022). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117: 198–213. https://doi.org/10.1080/01621459.2020.1765785

Luo C, Fa B, Yan Y, Wang Y, Zhou Y, Zhang Y, et al. (2020). High-dimensional mediation analysis in survival models. *PLoS Computational Biology*, 16(4): e1007768. https://doi.org/10.1371/journal.pcbi.1007768

Pearson C, Bartell T, Wang G, et al. (2022). Boston birth cohort profile: Rationale and study design. *Precision Nutrition*, 1: 1–12.

Perera C, Zhang H, Zheng Y, Hou L, Qu A, Zheng C, et al. (2022). HIMA2: High-dimensional mediation analysis and its application in epigenome-wide DNA methylation data. *BMC Bioinformatics*, 23: 1–14. https://doi.org/10.1186/s12859-021-04477-x

Shen E, Chou CP, Pentz MA, Berhane K (2014). Quantile mediation models: A comparison of methods for assessing mediation across the outcome distribution. *Multivariate Behavioral Research*, 49: 471–485. https://doi.org/10.1080/00273171.2014.904221

Storey JD, Bass AJ, Dabney A, Robinson D (2024). qvalue: Q-value estimation for false discovery rate control. R package version 2.38.0.

Sun R, Zhou X, Song X (2021). Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling*, 28: 778–790. https://doi.org/10.1080/10705511.2020.1863154

Tomczak K, Czerwiska P, Wiznerowicz M (2015). Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology*, 19: 68–77.

Wang C, Hu J, Blaser MJ, Li H (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*, 36: 347–355. https://doi.org/10.1093/bioinformatics/btz565

Wu D, Yang H, Winham SJ, Natanzon Y, Koestler DC, Luo T, et al. (2018). Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer. *Journal of Human*

Genetics, 63: 339–348. https://doi.org/10.1038/s10038-017-0385-8

Yimer BB, Lunt M, Beasley M, Macfarlane GJ, McBeth J (2023). Bayesgmed: An R-package for Bayesian causal mediation analysis. *PLoS ONE*, 18: 1–14.

Yu Q, Li B (2017). mma: An R package for mediation analysis with multiple mediators. *Journal of Open Research Software*, 5: 1–11. https://doi.org/10.5334/jors.160

Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894–942.

Zhang H (2025). Efficient adjusted joint significance test and Sobel-type confidence interval for mediation effect. *Structural Equation Modeling*, 32: 93–104. https://doi.org/10.1080/10705511.2024.2392139

Zhang H, Chen J, Feng Y, Wang C, Li H, Liu L (2021a). Mediation effect selection in high-dimensional and compositional microbiome data. *Statistics in Medicine*, 40: 885–896. https://doi.org/10.1002/sim.8808

Zhang H, Chen J, Li Z, Liu L (2021b). Testing for mediation effect with application to human microbiome data. *Statistics in Biosciences*, 13: 313–328. https://doi.org/10.1007/s12561-019-09253-3

Zhang H, Hong X, Zheng Y, Hou L, Zheng C, Wang X, et al. (2024). High-dimensional quantile mediation analysis with application to a birth cohort study of mother-newborn pairs. *Bioinformatics*, 40: 1–8.

Zhang H, Hou L, Liu L (2022). A review of high-dimensional mediation analyses in DNA methylation studies. In: *Epigenome-Wide Association Studies: Methods and Protocols* (W Guan, ed.), 2432.

Zhang H, Li X (2023). A framework for mediation analysis with massive data. *Statistics and Computing*, 33: 1–16. https://doi.org/10.1007/s11222-022-10178-z

Zhang H, Zheng Y, Hou L, Zheng C, Liu L (2021). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*, 37: 3815–3821. https://doi.org/10.1093/bioinformatics/btab564

Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20): 3150–3154. https://doi.org/10.1093/bioinformatics/btw351

Zhang J, Wei Z, Chen J (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*, 34(11): 1875–1883. https://doi.org/10.1093/bioinformatics/bty014