Package 'SparseMCMM'

June 19, 2023

Type Package

Title SparseMCMM: Estimating and testing the microbial causal mediation effect with the high-dimensional and compositional microbiome data

Version 2.0.0

Description Sparse Microbial Causal Mediation Model (SparseMCMM) is designed for the high dimensional and compositional microbiome data. SparseMCMM utilizes the linear log-contrast regression and Dirichlet regression to quantify the causal direct effect of the treatment and the causal mediation effect of the microbiome on the outcome under the counterfactual framework while addressing the compositional structure of microbiome data. Further it implements regularization techniques to handle the high-dimensional microbial mediators and identify the signature causal microbes. Furthermore, a splitting strategy (Rinaldo et al; 2019) is incorporated to account for the biases introduced by the regularization techniques employed. SparseMCMM is particularly effective in examining the mediation effect of the microbiome within a standard three-factor (treatment, microbiome, and outcome) causal study design (Wang et al. 2020). Moverover, the analytic procedure of SparseMCMM can be harnessed to explore the influences of the microbiome on health disparities.This is depicted in an extension of the model, SparseMCMM_HD, as elucidated in Wang et al. (2023). We also discuss the differences and relevance between SparseMCMM and SparseMCMM_HD (Wang et al; 2023). It's noteworthy that the mathematical expressions of the Residual Disparity Measure (RDM), Manipulable Disparity Measure (MDM), and Overall Disparity Measure (ODM), proposed by SparseMCMM_HD, align precisely with the formulas for the Causal Direct Effect of treatment (DE), the Mediation Effect through the microbiome (ME), and the Total Effect (TE) on the outcome, as formulated in our SparseMCMM. To simplify the discussion, we will refer to these as DE, ME, and TE henceforth.

Date 2023-06-19

License Artistic-2.0

Imports Compositional, stats, nloptr

Suggests testthat (>= 3.0.0)

Encoding UTF-8

LazyData true

URL https://github.com/chanw0/SparseMCMM

BugReports http://github.com/chanw0/SparseMCMM/issues

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Depends R (>= 4.3.0)

Config/testthat/edition 3

R topics documented:

`SparseMCMM-package` *SparseMCMM: Estimating and testing the microbial causal mediation effect with the high-dimensional and compositional microbiome data*

Description

Sparse Microbial Causal Mediation Model (SparseMCMM) is designed for the high dimensional and compositional microbiome data. SparseMCMM utilizes the linear log-contrast regression and Dirichlet regression to quantify the causal direct effect of the treatment and the causal mediation effect of the microbiome on the outcome under the counterfactual framework while addressing the compositional structure of microbiome data. Further it implements regularization techniques to handle the high-dimensional microbial mediators and identify the signature causal microbes. Furthermore, a splitting strategy (Rinaldo et al; 2019) is incorporated to account for the biases introduced by the regularization techniques employed.

SparseMCMM is particularly effective in examining the mediation effect of the microbiome within a standard three-factor (treatment, microbiome, and outcome) causal study design (Wang et al. 2020). MOverover, the analytic procedure of SparseMCMM can be harnessed to explore the influences of the microbiome on health disparities.This is depicted in an extension of the model, SparseMCMM_HD, as elucidated in Wang et al. (2023). We also discuss the differences and relevance between SparseMCMM and SparseMCMM_HD (Wang et al; 2023). It's noteworthy that the mathematical expressions of the Residual Disparity Measure (RDM), Manipulable Disparity Measure (MDM), and Overall Disparity Measure (ODM), proposed by SparseMCMM_HD, align precisely with the formulas for the Causal Direct Effect of treatment (DE), the Mediation Effect through the microbiome (ME), and the Total Effect (TE) on the outcome, as formulated in our SparseMCMM. To simplify the discussion, we will refer to these as DE, ME, and TE henceforth.

SparseMCMM consists of three components:

Component 1: Report the estimates and if applicable standard errors of DE, ME and TE respectively based on the split strategy.

Component 2: Report the point and if applicable 95% confidence interval estimates of ME_j for each microbe.

Component 3: Report the overall mediation test results: OME and CME tests.

Consequentially, SparseMCMM provides a clear and sensible causal path analysis among an exposure, compositional microbiome and outcome of interest.

Details

Package:	SparseMCMM
Type:	Package
Version:	2.0
Date:	2023-06-19
License:	Artistic-2.0

Author(s)

Chan Wang and Huilin Li.

Maintainer: Chan Wang <Chan.Wang@nyulangone.org> and Huilin Li <Huilin.Li@nyulangone.org>

References

Wang C, Hu J, Blaser MJ, and Li H (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics. 36(2):347-355.

Wang C, Ahn J, Tarpey T, Stella SY, Hayes RB and Li H (2023). A microbial causal mediation analytic tool for health disparity and applications in body mass index.

`SimulatedData` *Simulated Data A simulated datalist for test and illustration. There are 100 subjects (50 cases and 50 controls) and 10 taxa.*

Description

Simulated Data A simulated datalist for test and illustration. There are 100 subjects (50 cases and 50 controls) and 10 taxa.

Usage

```
SimulatedData
```

Format

A data list with three components:

Treatment A treatment vector for 100 subjects. 0 represents the control group and 1 represents the case group.

otu.com A 100 x 10 numeric matrix containing compositional microbiome data. Each row is a subject, and each column is a taxon. The row sum equals 1.

outcome A outcome vector for 100 subjects.

SparseMCMM *A main function in SparseMCMM framework*

Description

This function provides estimates of DE, ME, TE, and component-wise MEs. Additionally, it calculates the statistical significances of OME and CME using a permutation procedure grounded on Models (1) and (2).

Usage

```
SparseMCMM(Treatment, otu.com, outcome, n.split=10,
                    dirichlet.penalty=seq(0,1,0.1),
                    lm.penalty1=seq(0,1,0.1), lm.penalty2=seq(0,2,0.2),
               low.bound1=NULL, up.bound1=NULL, low.bound2=NULL, up.bound2=NULL,
                    num.per=NULL)
```

Arguments

Treatment	A numeric vector of the binary treatment (takes the value 1 if it is assigned to the treatment group and takes the value 0 if assigned to the control group) with length = sample size (n).
otu.com	A n*p numeric matrix containing compositional microbiome data. Each row represents a subject, and each column represents a taxon (given the rank, for example, the genus rank) or an OTU. The row sum equals 1.
outcome	A numeric vector of the continuous outcome with length = sample size (n).
n.split	An integer value. The number of repetitions regarding the split strategy. See Details for a more comprehensive discussion on the split strategy. Default=10.
dirichlet.penalty	A numeric vector that includes potential tuning parameters employed during the estimation process in relation to Model 2. Default=seq(0,1,0.1).
lm.penalty1	A numeric vector that includes potential tuning parameters employed during the estimation process in relation to Model 1. Default=seq(0,1,0.1).
lm.penalty2	A numeric vector that includes potential tuning parameters employed during the estimation process in relation to Model 1. Default=seq(0,2,0.2).
low.bound1	A numeric vector representing the lower bounds of the controls employed during the estimation process in relation to Model 1. Default=NULL, there is no lower bound.
up.bound1	A numeric vector representing the upper bounds of the controls employed during the estimation process in relation to Model 1. Default=NULL, there is no upper bound.
low.bound2	A numeric vector representing the lower bounds of the controls employed during the estimation process in relation to Model 2. Default=NULL, there is no lower bound.
up.bound2	A numeric vector representing the upper bounds of the controls employed during the estimation process in relation to Model 2. Default=NULL, there is no upper bound.
num.per	An integer value, the number of permuations. statistical significances of tests TME and CME are calculated based on these permutations. Default=NULL, No calculation for hypothesis test.

Details

Within the SparseMCMM framework, regularization techniques are employed to carry out variable selection, aiding in the identification of signature causal microbes. To account for the biases introduced by the regularization techniques employed, we further implement splitting strategy (Rinaldo et al; 2019), which can handle arbitrary penalties and provide asymptotically validated inference. Specifically, we randomly divide the dataset into two equal halves: the first half is utilized for variable selection, while the second half is dedicated to parameter estimation. The estimates of DE, ME, TE, and component-wise MEs were then calculated. We repeated this data splitting procedure multiple times (n.split times) to ensure robustness and accuracy in our estimations and inference.

If n.split > 1, the 'Esitmated Causal Effects' component is a 2 x 3 matrix. The first row contains the average estimates of DE, ME, and TE based on n.split times of repetitions. Correspondingly, the second row provides the standard errors associated with these averages. If n.split=1, the 'Estimated Causal Effects' component is a vector of length 3. This vector contains the estimates for DE, ME, and TE, which are calculated based on a single, random data split.

If n.split > 1, the 'Estimated component-wise Mediation Effects' component is a 3 x p matrix. The first row contains the average estimates of compoment-wise MEs for all p taxa based on n.split times of repetitions. Correspondingly, the second and third rows provide the lower and upper 95% confidence interval associated with these averages. If n.split=1, the 'EEstimated component-wise Mediation Effects' component is a vector of length p. This vector contains the estimates for compoment-wise MEs which are calculated based on a single, random data split.

Value

A list which contains three elements:

Esitmated Causal Effects Estimates of direct effect (DE), mediation effect (ME) and total effect (TE). See `Details` for a more comprehensive discussion.

Estimated component-wise Mediation Effects Estimates of component-wise MEs for all mediators. See `Details` for a more comprehensive discussion.

Test P-values for tests OME and CME if num.per is not NULL, otherwise, no this item.

Author(s)

Chan Wang and Huilin Li.

References

Wang C, Hu J, Blaser MJ, and Li H (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics. 36(2):347-355.

Wang C, Ahn J, Tarpey T, Stella SY, Hayes RB and Li H (2023). A microbial causal mediation analytic tool for health disparity and applications in body mass index.

Rinaldo A, Wasserman L, G'Sell M (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference.

Examples

```
# library(SparseMCMM)
#
# ##### Simulation data
# Treatment=SimulatedData$Treatment;
# otu.com=SimulatedData$otu.com
# outcome=SimulatedData$outcome
```

```
#
# ##### SparseMCMM function
# SparseMCMM(Treatment,otu.com,outcome,n.split=1,
#            num.per=10)
```

Index