Review of Mediation Analysis on Microbiome Data

Haoyi Zheng Ziman Jiang

Department of Biostatistics
University Of Michigan, Ann Arbor

June 10, 2025

Overview

1. Introduction
2. Literature Review
3. Numerical Analysis
4. Conclusion

Significance of Microbiome as a Mediator in Mediation Analysis

Significance of Microbiome as a Mediator in Mediation Analysis

- Microbiome as a mediator between treatment/exposure and health outcomes.
- Understanding microbial mediation helps refine cause-effect theories and develop targeted interventions.
- Fat intake affects BMI; obesity is linked to gut microbiome. Question: Does fat intake affect BMI through the gut microbiome?

Challenges with Microbiome Data

Challenges with Microbiome Data

- High-dimensional, compositional data limit the effectiveness of traditional analysis.
- Sparsity and zero-inflation common in microbiome datasets.
- Few specialized methods exist for microbiome mediation analysis.

Project Overview

Project Overview

- Review and evaluate methods in microbiome mediation analysis.
- Design numerical experiments to evaluate method performance.
- Provide guidelines for future microbiome studies.

Basic Frameworks in Mediation Analysis

Basic Frameworks in Mediation Analysis

1. **Structural Equation Model (SEM):**
   - **Test Formula:** Indirect Effect $= \alpha\beta$
   - **Hypothesis:** Test if the indirect effect $\alpha\beta \neq 0$.
2. **Counterfactual Framework:**
   - **Test Formula:**
     NIE $= Y(X = 1, M(X = 1)) - Y(X = 1, M(X = 0))$
   - **Hypothesis:** Test if the natural indirect effect $NIE \neq 0$.
   - **Advantages:** Provides a natural definition of effects without relying on controlling variables; robust in nonlinear models.
   - **Trend:** Most recent methods in mediation analysis are based on the counterfactual framework.

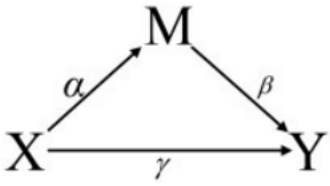
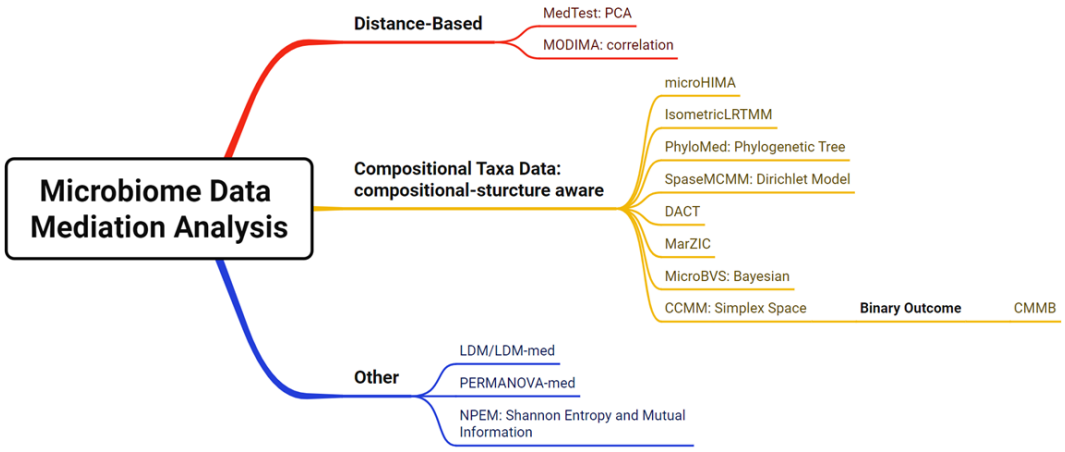


Figure: Mediation Analysis Framework

Method Classification Overview



Review of Mediation Analysis Methods

Table 1: Comparison of different mediation analysis methods for microbiome data. Abbreviations: SEM = Structural Equation Model; ilr = isometric logratio transformation; alr = additive logratio transformation; LRT = log-ratio transformation; LM = linear regression model; Taxa = using raw taxonomic units directly as mediator without applying compositional data transformations. Pseudocount refers to the technique of adding a small constant (usually 0.5) to zero counts to avoid mathematical issues in log-ratio transformations or other compositional analyses.

	MedTest	**MODIMA**	**CCMM**	**SparseMCMM**	**IsometricLRTMM**	**microHIMA**
Framework	Distance-based	Distance-based	Counterfactual	Counterfactual	SEM	SEM
Mediator	PCoA	Distance matrix	alr	alr	ilr	ilr
Overall ME	✓	✓	✓	✓	×	×
Mediator-wise ME	×	×	✓	×	✓	✓
Zero counts	-	-	Pseudocount	Pseudocount	Pseudocount	Pseudocount
Regularization	×	×	×	L1 penalty	de-biased Lasso	de-biased Lasso
Model for mediator	LM	LM	Compositional algebra	Dirichlet regression	LM	Log-linear regression
Model for response	LM	LM	Linear log-contrast	Linear log-contrast	LM	LM

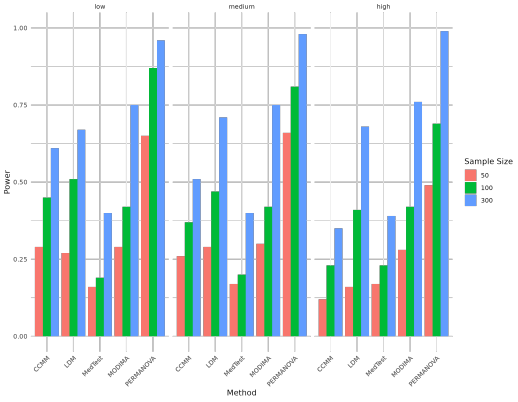
	MedZIM	**MicroBVS**	**PhyloMed**	**LDM**	**PERMANOVA-med**	**NPEM**
Framework	Counterfactual	Counterfactual	Phylogenetic tree	Inverse regression	Distance-based	Information-theoretic
Mediator	Taxa	Taxa	LRT	Taxa	Distance matrix	Taxa
Overall ME	×	✓	×	✓	✓	×
Mediator-wise ME	✓	✓	✓	✓	×	✓
Zero counts	-	Pseudocount	Pseudocount	-	-	-
Regularization	×	Bayesian priors	×	✓	×	×
Model for mediator	LM	Log-linear regression	Log-linear regression	Inverse regression	Distance-based	Information-based
Model for response	LM	LM	Log-linear regression	Inverse regression	Distance-based	Information-based

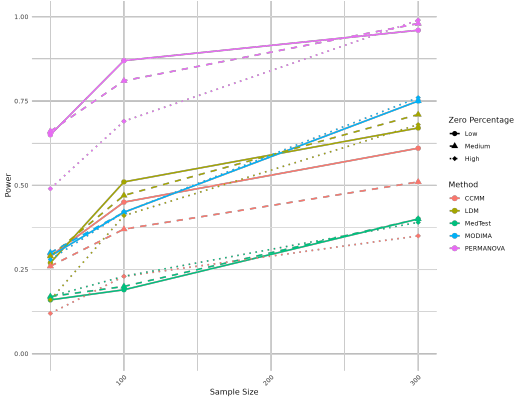
Method Comparison for Mediation Analysis

Method	Applies Overall ME	Applies Individual ME	Reason for Exclusion or Limitation
MedTest	Yes		
MODIMA	Yes		
CCMM	Yes	Yes	
LDM	Yes	Yes	
PERMANOVA	Yes		
SparseMCMM	No		Excluded due to computational time limitations.
MicroBVS	No	Yes	Applies Individual ME but excluded for Overall ME due to computational time limitations.
IsometricLRTMM		No	Excluded because no implementation code is available.
PhyloMed		No	Excluded because it requires phylogenetic tree information.
MedZIM		Yes	
npEM		Yes	

Simulation Scenario: Sample Size Effect

- Zero-Percentage Levels: Low (30%), Medium (50%), High (60%).
- Data input for all analyses was transformed into compositional data (SparsDOSSA2 provides absolute abundance).
- The focus is on testing the overall mediation effect.
- Due to computational efficiency limitations, results from SparseMCMM and MicroBVS are not shown.

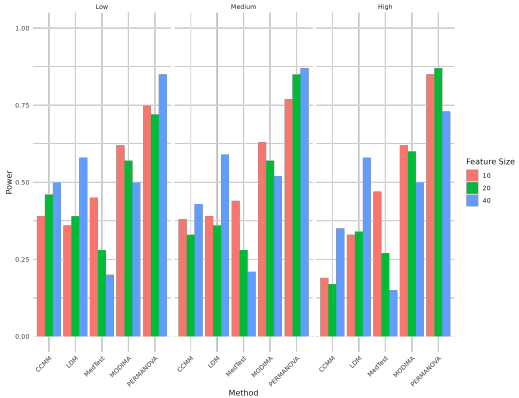


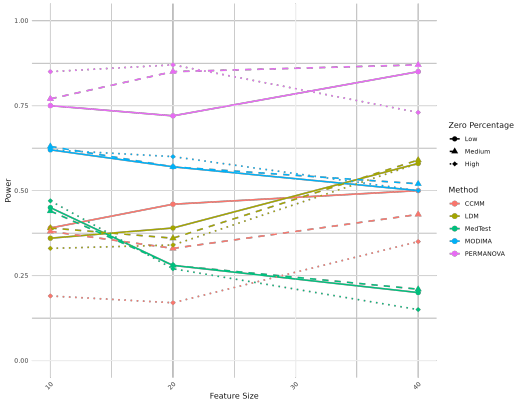


Simulation Scenario: Feature Size Control in Method 2

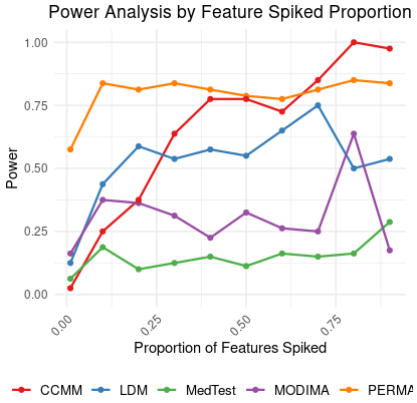
Overview: The figures below demonstrate the impact of controlling feature size on mediation analysis using Method 2:

- Method 2 involves subsampling based on library size to control data sparsity.
- Analyzing how varying the feature size affects the detection of mediation effects.





Simulation Scenario



- **Sample Size**: Fixed at 100.
- **Feature Size**: Fixed at 50.
- **Adjusted Parameter**: The percentage of effective features among all features is varied.
- The goal is to evaluate the impact of effective feature proportion on method performance under this controlled scenario.

Power Analysis Results

Method	Power
CCMM	0.53
LDM	0.55
MODIMA	0.10
MedTest	0.05
PERMANOVA	0.68

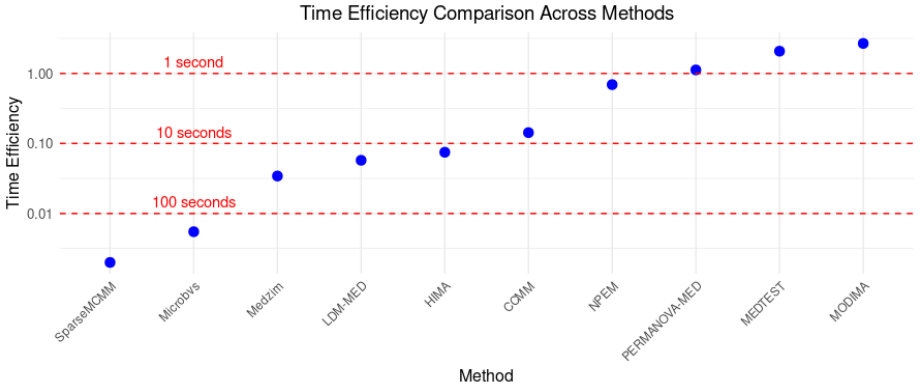
Experiment Settings:

- **Feature size:** 100
- **Sample size:** 50
- **Effective features:** 40%

Execution Results:

- **100 datasets were generated.**
- **No errors occurred in any method.**
- **LDM:** Runtime increased from 20s to 40s.
- **CCMM:** Runtime increased from 10s to 300s.
- CCMM showed "NOT CONVERGE" warnings but generated results.

Time Efficiency of 10 Methods

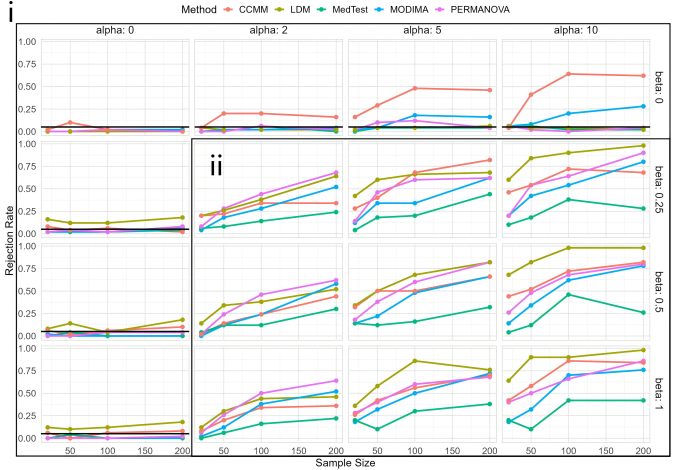


- **Scenario:** The comparison is based on a sample size of 100 and a feature size of 50.
- **Time Efficiency:** Calculated as $\log_{10}(1/\text{Time})$, where higher values indicate better efficiency.
- Methods are evaluated based on runtime efficiency under controlled simulation settings.

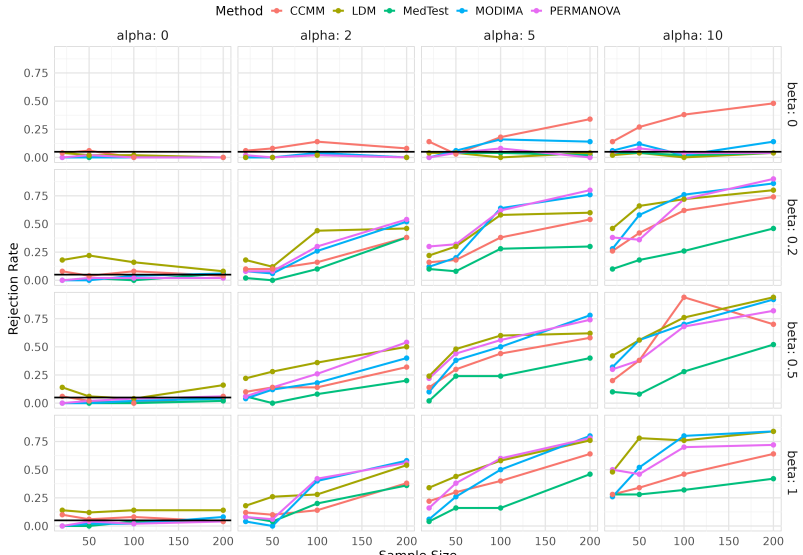
Simulation Scenario: Pathway Strength Control

Overview: The figure below demonstrates the effect of controlling pathway strengths in mediation analysis:

- α pathway represents the strength of the relationship between the treatment and the mediator.
- β pathway represents the strength of the relationship between the mediator and the outcome.



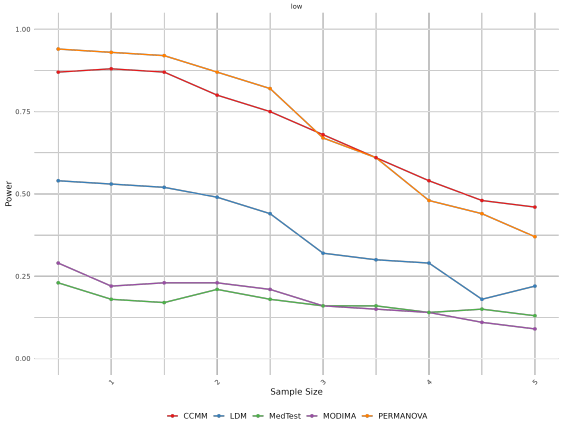
Simulation Scenario: Pathway Strength Control



Simulation Scenario: Noise Control

Overview: The impact of controlling noise levels on mediation analysis:

- Increasing noise levels can obscure the mediation effect, making it harder to detect.
- This analysis examines how different levels of noise affect the robustness of mediation detection.



Testing Individual Mediation Effects

- **Recall:** Sensitivity of the method in correctly identifying true mediation effects.
- **Precision:** Accuracy in identifying true mediation effects among all detected effects.
- **FPR (False Positive Rate):** Rate at which non-mediation effects are incorrectly identified as mediation effects.
- **FDP (False Discovery Proportion):** Proportion of falsely identified mediation effects among all detected effects.

Effect Size: 5					Effect Size: 10				
Method	**Recall**	**Precision**	**FPR**	**FDP**	**Method**	**Recall**	**Precision**	**FPR**	**FDP**
ccmm	0.128	0.285	0.0075	0.715	ccmm	0.080	0.224	0.0111	0.776
hima	0.010	0.400	0.0000	0.600	hima	0.020	0.533	0.0009	0.467
ldm	0.100	0.381	0.0194	0.619	ldm	0.083	0.173	0.0158	0.827
medzim	0.130	0.419	0.0131	0.581	medzim	0.055	0.114	0.0171	0.886
microbvs	0.350	0.183	0.6250	0.817	microbvs	0.347	0.206	0.5350	0.794
npem	0.000	0.000	0.0000	N/A	npem	0.003	1.000	0.0000	0.000
Effect Size: 15					**Effect Size: 20**				
Method	**Recall**	**Precision**	**FPR**	**FDP**	**Method**	**Recall**	**Precision**	**FPR**	**FDP**
ccmm	0.107	0.198	0.0174	0.802	ccmm	0.140	0.210	0.0211	0.790
hima	0.013	0.250	0.0015	0.750	hima	0.020	0.174	0.0038	0.826
ldm	0.135	0.213	0.0200	0.787	ldm	0.138	0.242	0.0172	0.758
medzim	0.035	0.053	0.0248	0.947	medzim	0.030	0.042	0.0276	0.958
microbvs	0.360	0.217	0.0521	0.783	microbvs	0.375	0.202	0.0592	0.798
npem	0.003	0.125	0.0007	0.875	npem	0.030	0.333	0.0024	0.667

Conclusion

- **Techniques in Microbiome Research:** Most mediation analysis methods in microbiome research have effectively addressed high-dimensional data challenges. Many of these methods leverage the compositional nature of microbiome data using log-ratio transformations like log-contrast.
- **Challenges with Zero-Inflated Data:** Most methods do not adequately handle the zero-inflated nature of microbiome data. Current approaches often fail to capture the sparsity and the excess zeros, leading to reduced model power and accuracy.
- **Testing Individual Mediation Effects (ME):** While some existing methods can test individual ME, they tend to be either too stringent or too lenient in their criteria. Moreover, there is a lack of approaches that can simultaneously test both overall and individual mediation effects in a balanced manner.
- **Future Directions:** New methods need to be more aware of zero-inflation and the compositional nature of microbiome data. It is crucial to develop approaches that can simultaneously observe both overall and individual mediation effects while directly integrating these features to improve accuracy and robustness in mediation analysis.

References

The End