Causal Discovery on Gut Microbial Data for Disease Risk Prediction

Mariana Paco Mendivil [1] Candus Shi [2] Nicole Zhang [3] Mentor: Dr. Biwei Huang [4] Mentor: Dr. Jelena Bradic [5]

[1]mpacomendivil@ucsd.edu [2]c6shi@ucsd.edu [3]nwzhang@ucsd.edu [4]bih007@ucsd.edu [5]jbradic@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Background

- Association vs. Causation
- Causal Discovery Algorithms in Gut Microbiome Studies Previous research has explored causal discovery in gut microbiome studies, notably using algorithms like PC-stable to construct causal networks
  and implement do-calculus for estimating microbe-microbe and microbe-outcome causal effects. More recent advancements include CD-NOD,
  specifically designed for heterogeneous data, which is particularly valuable for gut microbiome research where samples often come from different studies.
  These algorithms aim to identify true causal relationships by performing conditional independence tests and orienting edges using established rules, while accounting for the unique challenges of microbiome data analysis.

Research Questions

1. Microbe-Microbe: How do the microbe-microbe interaction networks between the healthy and diseased participants differ?
2. Microbe-Disease: What microbes have a causal relationship to disease status?
3. Prediction: Is it possible to predict disease status with the current composition of the dataset given causal representation learning techniques?
   How do they differ with the microbes learned in question 2?

Data

To answer the questions above, we apply our framework to gut microbial data that investigated T2D and an individual participant data meta-analysis dataset that investigated PCOS.

- T2D: For T2D, we use the NIH Human Microbiome Project (HMP2) dataset, filtered to healthy visits with 16S sequencing. Includes 153 insulin-sensitive (IS) and 178 insulin-resistant (IR) samples.
- PCOS: For PCOS, we use a dataset aggregated from 14 different clinical studies across Asia and Europe, filtered to individual-level samples with 16S sequencing. Includes 435 healthy controls (HC) and 513 PCOS patients.

Causal Discovery

Causal discovery is all about recovering the true causal structure of system given observed data. One way to model this causal structure is through a graph One of the oldest and most widely-used general-purpose causal discovery algorithms is PC. It follows these key steps:

1. Start with a fully connected graph (every variable connected to every other).
2. Remove edges based on statistical independence tests.
3. Identify v-structures (patterns like $X \to Y \leftarrow Z$) to infer causal directions.
4. Apply Meek's rules to orient additional edges while preserving consistency.

The result is a CPDAG (Completed Partially Directed Acyclic Graph), which represents a set of causal structures consistent with the observed data.

Why Use PC?

- Works for different data types (as long as independence tests match the data distribution).
- Efficient for large datasets.
- Assumes **no hidden confounders**, **causal Markov condition**, and **faithfulness**.

Methods

In this study, we use causal discovery algorithms and compare them with predictive modeling to explore the causal relationships between the gut microbiome and two diseases: T2D and PCOS. Due to the high-dimensionality of the data and small sample sizes, we first select features through sparse estimation methods and sure-screening to reduce the number of microbes.

1. Filter out rare OTUs. Remove microbes where all samples have less than 1% relative abundance.
2. Feature selection and sure screening. For the microbe-microbe network, we use two methods, SparCC and graphical lasso to reduce the number of edges between pairs of microbes. For the microbe-disease network, we use logistic lasso regression to reduce the number of features that are not helpful in predicting disease.
3. Causal discovery algorithms. For the microbe-microbe network, we implement PC-stable with a max depth of 2. For the microbe-outcome network, we implement CD-NOD where the covariates correspond to the heterogeneity index.
4. Variational autoencoder. xxx.

T2D

Et rutrum ex euismod vel. Pellentesque ultricies, velit in fermentum vestibulum, lectus nisi pretium nibh, sit amet aliquam lectus augue vel velit. Suspendisse rhoncus massa porttitor augue feugiat molestie. Sed molestie ut orci nec malesuada. Sed ultricies feugiat est fringilla posuere.

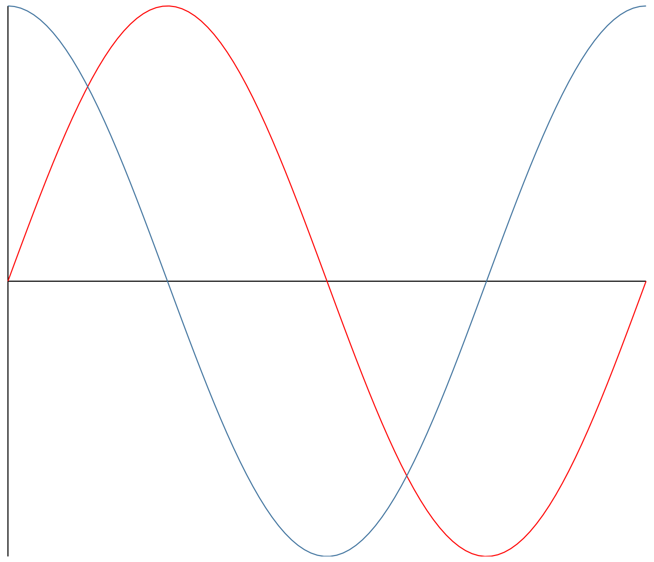

Figure 1. Another figure caption.

PCOS

Nulla eget sem quam. Ut aliquam volutpat nisi vestibulum convallis. Nunc a lectus et eros facilisis hendrerit eu non urna. Interdum et malesuada fames ac ante *ipsum primis* in faucibus. Etiam sit amet velit eget sem euismod tristique. Praesent enim erat, porta vel mattis sed, pharetra sed ipsum. Morbi commodo condimentum massa, *tempus venenatis* massa hendrerit quis. Maecenas sed porta est. Praesent mollis interdum lectus, sit amet sollicitudin risus tincidunt non.

Etiam sit amet tempus lorem, aliquet condimentum velit. Donec et nibh consequat, sagittis ex eget, dictum orci. Etiam quis semper ante. Ut eu mauris purus. Proin nec consectetur ligula. Mauris pretium molestie ullamcorper. Integer nisi neque, aliquet et odio non, sagittis porta justo.

- Sed consequat id ante vel efficitur. Praesent congue massa sed est scelerisque, elementum mollis augue iaculis.
  - In sed est finibus, vulputate nunc gravida, pulvinar lorem. In maximus nunc dolor, sed auctor eros porttitor quis.
  - Fusce ornare dignissim nisi. Nam sit amet risus vel lacus tempor tincidunt eu a arcu.
  - Donec rhoncus vestibulum erat, quis aliquam leo gravida egestas.
- Sed luctus, elit sit amet dictum maximus, diam dolor faucibus purus, sed lobortis justo erat id turpis.
- Pellentesque facilisis dolor in leo bibendum congue. Maecenas congue finibus justo, vitae eleifend urna facilisis at.

Variational Autoencoder

A different kind of highlighted block.

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

Interdum et malesuada fames $\{1, 4, 9, \ldots\}$ ac ante ipsum primis in faucibus. Cras eleifend dolor eu nulla suscipit suscipit. Sed lobortis non felis id vulputate.

A heading inside a block

Praesent consectetur mi $x^2 + y^2$ metus, nec vestibulum justo viverra nec. Proin eget nulla pretium, egestas magna aliquam, mollis neque. Vivamus dictum $\mathbf{u}^\intercal\mathbf{v}$ sagittis odio, vel porta erat congue sed. Maecenas ut dolor quis arcu auctor porttitor.

Another heading inside a block

Sed augue erat, scelerisque a purus ultricies, placerat porttitor neque. Donec $P(y \mid x)$ fermentum consectetur $\nabla_x P(y \mid x)$ sapien sagittis egestas. Duis eget leo euismod nunc viverra imperdiet nec id justo.

Conclusion/Future Work

Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Phasellus libero enim, gravida sed erat sit amet, scelerisque congue diam. Fusce dapibus dui ut augue pulvinar iaculis.

First column	Second column	Third column	Fourth
Foo	13.37	384,394	α
Bar	2.17	1,392	β
Baz	3.14	83,742	δ
Qux	7.59	974	γ

Table 1. A table caption.

Donec quis posuere ligula. Nunc feugiat elit a mi malesuada consequat. Sed imperdiet augue ac nibh aliquet tristique. Aenean eu tortor vulputate, eleifend lorem in, dictum urna. Proin auctor ante in augue tincidunt tempor. Proin pellentesque vulputate odio, ac gravida nulla posuere efficitur. Aenean at velit vel dolor blandit molestie. Mauris laoreet commodo quam, non luctus nibh ullamcorper in. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos.

Nulla varius finibus volutpat. Mauris molestie lorem tincidunt, iaculis libero at, gravida ante. Phasellus at felis eu neque suscipit suscipit. Integer ullamcorper, dui nec pretium ornare, urna dolor consequat libero, in feugiat elit lorem euismod lacus. Pellentesque sit amet dolor mollis, auctor urna non, tempus sem.

References

Please visit our website at https://nzhang20.github.io/Causal-Discovery-on-Gut-Microbial-Data-for-Disease-Risk-Prediction/