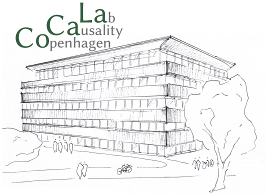
# Causal Change Point Detection

Shimeng Huang, Rikke Nielsen, Jonas Peters, Niklas Pfister

ETH-UCPH-TUM Workshop on Graphical Models

October 11, 2022




UNIVERSITY OF
COPENHAGEN

# Before this project...

## Supervised Learning and Model Analysis with Compositional Data

Shimeng Huang, Elisabeth Ailer, Niki Kilbertus, Niklas Pfister

The compositionality and sparsity of high-throughput sequencing data poses a challenge for regression and classification. However, in microbiome research in particular, conditional modeling is an essential tool to investigate relationships between phenotypes and the microbiome. Existing techniques are often inadequate: they either rely on extensions of the linear log-contrast model (which adjusts for compositionality, but is often unable to capture useful signals), or they are based on black-box machine learning methods (which may capture useful signals, but ignore compositionality in downstream analyses).
We propose KernelBiome, a kernel-based nonparametric regression and classification framework for compositional data. It is tailored to sparse compositional data and is able to incorporate prior knowledge, such as phylogenetic structure. KernelBiome captures complex signals, including in the zero-structure, while automatically adapting model complexity. We demonstrate on par or improved predictive performance compared with state-of-the-art machine learning methods. Additionally, our framework provides two key advantages: (i) We propose two novel quantities to interpret contributions of individual components and prove that they consistently estimate average perturbation effects of the conditional mean, extending the interpretability of linear log-contrast models to nonparametric models. (ii) We show that the connection between kernels and distances aids interpretability and provides a data-driven embedding that can augment further analysis. Finally, we apply the KernelBiome framework to two public microbiome studies and illustrate the proposed model analysis. KernelBiome is available as an open-source Python package at this https URL.

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG); Applications (stat.AP)
Cite as: arXiv:2205.07271 **[stat.ML]**
(or arXiv:2205.07271v1 **[stat.ML]** for this version)
https://doi.org/10.48550/arXiv.2205.07271

# Background & Motivation

- Detecting changes in sequential or time series data has long been of interest

# Background & Motivation

- Detecting changes in sequential or time series data has long been of interest

$$X_i \rightarrow Y_i \qquad\qquad X_i \boldsymbol{\rightarrow} Y_i$$

$$i = 1, \cdots, \tau - 1 \qquad\qquad i = \tau, \cdots, n$$

$X_i$ your desire of ice cream, $Y_i$ your actual consumption of ice cream, and at $\tau$ you found out that lactase pills are a thing!

# Background & Motivation

- Detecting changes in sequential or time series data has long been of interest

$$X_i \longrightarrow Y_i \qquad\qquad X_i \longrightarrow Y_i$$

$$i = 1, \cdots, \tau - 1 \qquad\qquad i = \tau, \cdots, n$$

$X_i$ your desire of ice cream, $Y_i$ your actual consumption of ice cream, and at $\tau$ you found out that lactase pills are a thing!

Given: data $X$ and $Y$

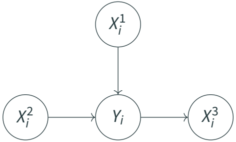
Ideal output: there is a causal change point at $\tau$.

# Background & Motivation

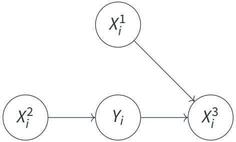
- In economics literature, changes in how $Y$ is affected by others are often referred to as structural changes

# Background & Motivation

- In economics literature, changes in how $Y$ is affected by others are often referred to as structural changes
- Here we consider observing **multivariate sequential data** where the causal structure affecting a particular variable changes

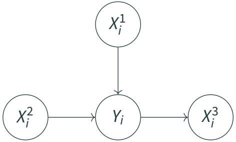



$i = 1, \cdots, \tau$

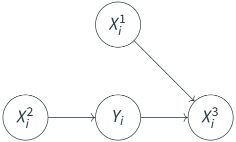



$i = \tau + 1, \cdots, n$

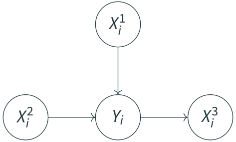
# Examples of interests



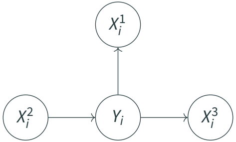

$i = 1, \cdots, \tau$




$i = \tau + 1, \cdots, n$

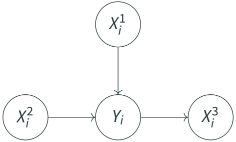



$i = 1, \cdots, \tau$

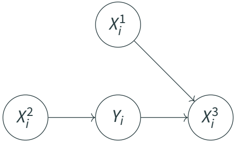



$i = \tau + 1, \cdots, n$

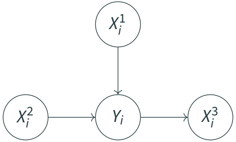
# Examples of interests



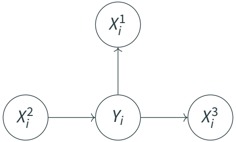

$i = 1, \cdots, \tau$



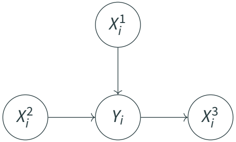

$i = \tau + 1, \cdots, n$



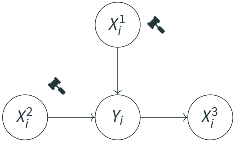

$i = 1, \cdots, \tau$




$i = \tau + 1, \cdots, n$




$i = 1, \cdots, \tau$




$i = \tau + 1, \cdots, n$

# Setup

We observe a sequence of independent $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$ with

$$Y_i = f_i(X_i, \epsilon_i), \quad i = 1, \cdots, n.$$

**Def. Causal Change Point (CCP)**

We call the time points $\tau_1, \cdots, \tau_{J-1}$ the complete set of **causal change points** (CCPs) if for some $J \in \{1, \ldots, n-1\}$ and $\{\tau_0, \ldots, \tau_J\} \subseteq \{1, \ldots, n\}$ with $1 = \tau_0 < \cdots < \tau_J = n$, we have

$$f_i = \sum_{j=1}^{J} f_{\tau_j} \cdot \mathbb{1}_{(\tau_{j-1}, \tau_j]}(i),$$

and $\forall k \in \{1, \cdots, J\}, f_{\tau_k} \neq f_{\tau_{k-1}}$.

# Goals

Two goals:

# Goals

Two goals:

1. **Test existence of change points**
   for a time interval $I = \{t, \cdots, t+m\} \subseteq \{1, \cdots, n\}$

$$\mathcal{H}_0^{\text{CP}}(I) : \nexists k \in I \text{ s.t. } k \text{ a CCP}$$

# Goals

Two goals:

1. **Test existence of change points**
   for a time interval $I = \{t, \cdots, t+m\} \subseteq \{1, \cdots, n\}$

$$\mathcal{H}_0^{\mathrm{CP}}(I) : \nexists k \in I \text{ s.t. } k \text{ a CCP}$$

2. **Estimate the complete set of causal change points**
   this could be achieved based on Goal 1.

# Test existence of change points

**Idea**: Test existence of change points via testing the existence of an invariant set or invariant function.

# Test existence of change points

a. **Invariant sets**

**Def. Invariant set**

For a time interval $I = \{t, \cdots, t+m\}$ with $t, m \in \mathbb{N}$, a set $S \subseteq \{1, ..., d\}$ is called *invariant within $I$* with respect to $(X, Y)$ if for all $i, j \in I$

$$Y_i | X_i^S \stackrel{d}{=} Y_j | X_j^S.$$

For a time interval $I$, we aim to test the hypothesis

$$\mathcal{H}_0^{\text{set}}(I) : \exists S \subseteq \{1, \cdots, d\} \text{ s.t. } S \text{ is invariant within } I.$$

# Test existence of change points

a. **Invariant sets**

For a time interval $I$, we aim to test the hypothesis

$$\mathcal{H}_0^{\text{set}}(I) : \exists S \subseteq \{1, \cdots, d\} \text{ s.t. } S \text{ is invariant within } I.$$

# Test existence of change points

a. **Invariant sets**

For a time interval $I$, we aim to test the hypothesis

$$\mathcal{H}_0^{\text{set}}(I) : \exists S \subseteq \{1, \cdots, d\} \text{ s.t. } S \text{ is invariant within } I.$$

To achieve this, for each $S \in \mathcal{P}(\{1, \cdots, d\})$ we test

$$\mathcal{H}_0^{S}(I) : S \text{ is invariant within } I,$$

e.g. Chow test (Chow, 1960), and we reject $\mathcal{H}_0^{\text{set}}(I)$ if we reject $\mathcal{H}_0^{S}(I)$ for all $S \in \mathcal{P}(\{1, \cdots, d\})$.

In the linear Gaussian case, this is essentially as in seqICP (Pfister et al., 2019).

# Test existence of change points

a. **Invariant sets**
b. **Invariant functions**

**Def. Invariant function**

For a time interval $I = \{t, \cdots, t+m\}$ with $t, m \in \mathbb{N}$, a function $f : \mathcal{X} \to \mathbb{R}$ is called *invariant within* $I$ with respect to $(X, Y)$ if for all $i, j \in I$

$$Y_i - f(X_i) \stackrel{d}{=} Y_j - f(X_j).$$

For a time interval $I$, we are interested in testing the hypothesis

$$\mathcal{H}_0^{\text{fun}}(I) : \exists\, f \in \mathcal{F} \text{ s.t. } f \text{ is invariant within } I.$$

# Test existence of change points

a. **Invariant sets**
b. **Invariant functions**

   For a time interval $I$, we are interested in testing the hypothesis

$$\mathcal{H}_0^{\text{fun}}(I) : \exists\, f \in \mathcal{F} \text{ s.t. } f \text{ is invariant within } I.$$

   In this case, we could test

$$\mathcal{H}_0^{\text{iv}}(I) : \forall i \in I \; (\exists\, f \text{ s.t. } Y_i - f(X_i) \text{ is independent of the indices}).$$

   This may be related to exploiting the exclusion restriction of an instrumental variable e.g., by using the Anderson-Rubin test (Anderson and Rubin, 1949).

# Test existence of change points

**Idea**: Test existence of change points via testing the existence of an invariant set or invariant function.

a. **Invariant sets** $\mathcal{H}_0^{\text{set}}(I) : \exists S \subseteq \{1, \cdots, d\}$ s.t. $S$ is invariant within $I$
b. **Invariant functions** $\mathcal{H}_0^{\text{fun}}(I) : \exists f \in \mathcal{F}$ s.t. $f$ is invariant within $I$

**Note**: Recall

$$\mathcal{H}_0^{\text{CP}}(I) : \nexists k \in I \text{ s.t. } k \text{ a CCP},$$

if $\mathcal{H}_0^{\text{set}}(I)$ or $\mathcal{H}_0^{\text{fun}}(I)$ is tested at the correct level, then $\mathcal{H}_0^{\text{CP}}(I)$ is tested at the correct level.

# Estimate the complete set of causal change points

One way to achieve Goal 2 is to utilize the approaches for Goal 1.

# Estimate the complete set of causal change points

One way to achieve Goal 2 is to utilize the approaches for Goal 1.

i. Search for candidates of CCP

# Estimate the complete set of causal change points

One way to achieve Goal 2 is to utilize the approaches for Goal 1.

i. Search for candidates of CCP

**Conj. Any CCP induces a change in $Y|X$**

Under certain faithfulness conditions, if $k \in \{2, \cdots, n\}$ is a CCP, then

$$Y_k|X_k \overset{d}{\neq} Y_{k-1}|X_{k-1}.$$

# Estimate the complete set of causal change points

One way to achieve Goal 2 is to utilize the approaches for Goal 1.

i. Search for candidates of CCP
ii. Test each candidate by testing its surrounding intervals

# Testing a candidate (based on invariant sets)

**Algorithm 1** Test a potential causal change point

**Require:** $(X_1, Y_1), \cdots, (X_n, Y_n)$, a candidate $k$, confidence level $\alpha$

1: Construct interval sets $I_1 = \{1, \cdots, k-1\}$ and $I_2 = \{k, \cdots, n\}$
2: **for** $S \in \mathcal{P}(\{1, \cdots, d\})$ **do**
3: Compute the test statistic $T(X, Y, I_1, I_2, S)$ and $p$-value $p_S$
4: **end for**
5: Let $p = \max\{p_S : S \in \mathcal{P}(\{1, \cdots, d\})\}$
6: **return** $p < \alpha$

# Preliminary results

**Data generating process**:

1. 15 random DAGs with 6 nodes where node $Y$ has 2 parents and 1 child.
2. total number of time points ($n$): $180 \times 2^{\{1,2,3,4,5\}}$
3. 1 CCP at $1/3$ of and 1 distributional shift in the covariates at $2/3$ of the total number of observations
4. 30 repetition for each DAG & sample size combination

**Test**: Chow test (Chow, 1960)

# Preliminary results

**Data generating process**:

1. 15 random DAGs with 6 nodes where node $Y$ has 2 parents and 1 child.
2. total number of time points ($n$): $180 \times 2^{\{1,2,3,4,5\}}$
3. 1 CCP at $1/3$ of and 1 distributional shift in the covariates at $2/3$ of the total number of observations
4. 30 repetition for each DAG & sample size combination

**Test**: Chow test (Chow, 1960)

For this experiment, we only test the *oracle candidates*, meaning at the true CCP and the time of distributional shift.

# Preliminary results

The following compares the naive approach using Chow test and algorithm 1.

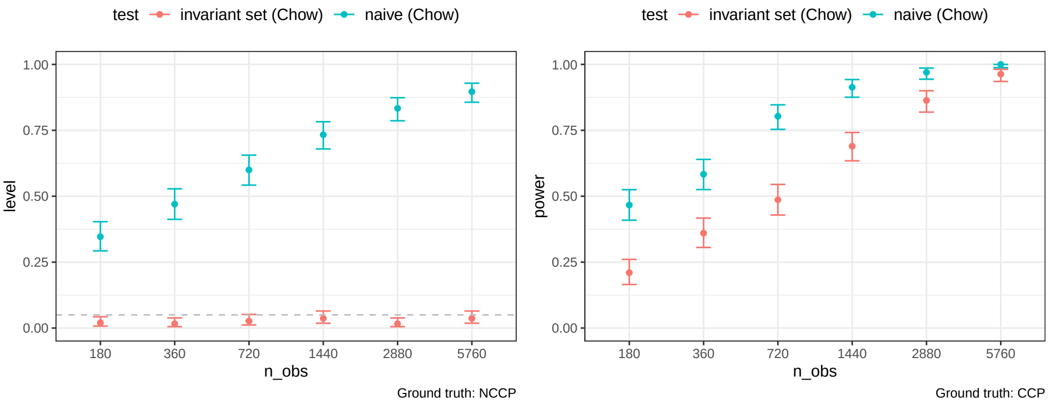



**Figure 1:** Estimated level with binomial CI **Figure 2:** Estimated power with binomial CI

# Summary

- Goals:
  1. Test existence of CCP
  2. Estimate location of CCP

# Summary

- Goals:
  1. Test existence of CCP
  2. Estimate location of CCP
- Approaches:
  a. Invariant sets
  b. Invariant function

# Open questions

- So far we used the oracle candidates. How to efficiently and unbiasedly search for candidates?
- The invariant function approach could be more computationally efficient and more general. What tests can be used to explore the exclusion restriction criteria in the IV setup? E.g. time does not affect $X$ linearly; in an under-identified situation?

# References

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of mathematical statistics*, 20(1):46–63, 1949.

G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605, 1960.

N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.