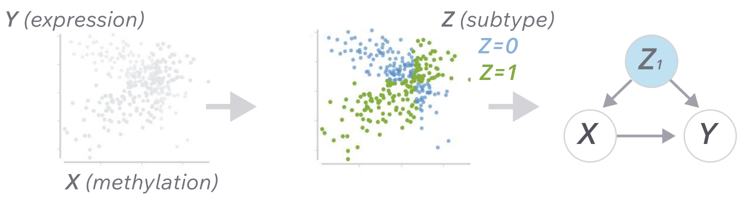
Causal Mixture Models: Characterization and Discovery

Sarah Mameche, Janis Kalofolias, Jilles Vreeken
CISPA Helmholtz Center for Information Security

Summary We address mixtures of populations with different generating processes using causal models. For inference, we integrate conditional mixture modelling into score-based causal DAG search algorithms.

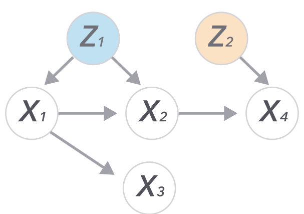
Motivating Example I
Disease Heterogeneity



Example methylation (X)-expression(Y) mechanism in colon adenocarcinoma cancer patients *(The Cancer Genome Atlas (TCGA); Chang et al., 2020)*
Latent Factor different cancer subtypes
Challenge I heterogeneous generating process of $Y \mid X$, shown as colored points/latent variable Z
Previous Work assumes heterogeneity results from specific interventions *(Kumar et al., 2024)*

Motivating Example II
Treatment Resistance

Example resistance of patients to different antibiotics (X)
Latent Factors differences due to multiple factors, e.g., prior exposure (Z_1) vs. batch effects (Z_2)



Challenge II potentially multiple independent latent variables with unknown points of influence
Previous Work assumes global environments *(Huang et al., 2020)* or global latent-class confounder *(Mazaheri et al., 2024)*

Goals
- discover latent sources of heterogeneity (Z)
- discover observed and latent causal structure (DAG)

Model
Causal Mixture Model (CMM)

Graphical Causal Model given by a directed acyclic graph (DAG) over a set of observed variables $\mathbf{X}$ and a set of latent categorical variables $\mathbf{Z}$, where each Z_i has K_i values.

Structural Causal Model where each variable given its causes is modelled through a mixture of regressions (MLR),

$$X_j = f(\mathbf{Pa}_j, b_j) + N_j$$

with $f(x, b_j(z)) = \beta_{jz}^T x + \beta_{jz}^{(0)} x$ and $N_j \sim \mathcal{N}(0, \sigma^2)$ with $N_j \perp\!\!\!\perp \mathbf{Pa}_j$.

Assumption **Causal Markov Condition** *We assume that the marginal distribution over $\mathbf{X}$ is given as*

$$p_{\mathbf{X}}(x) = \prod_{X_j \in \mathbf{X}} p^{\mathrm{MLR}}_{X|\mathbf{Pa}_j}(x, \mathbf{pa}_j; \mathbf{B}, \gamma, \sigma^2)$$

$$\sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\|\beta_k^T \mathbf{pa}_j - x\|}{2\sigma^2}\right)$$

Guarantees
Consistent Scoring

Definition **BIC, latent-aware** *Given samples $\mathcal{D} = \{x_1, \ldots, x_r\}$ and a hypothesis $\mathcal{H}$, the Bayesian Information Criterion (BIC) is*

$$\mathrm{BIC}(\mathcal{H}) := -2 \log p_{\mathbf{X}}(\mathcal{D} \mid \hat{\theta}) + d \log r$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta_K} p^{\mathrm{MLR}}_{X|\mathbf{Y}}(x, \mathbf{y}; \theta)$$

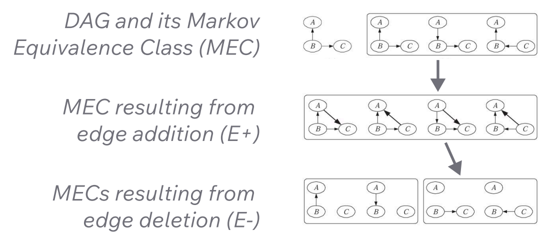
Lemma **Non-Gaussianity of Direct Effect** *(informal)* *Given an effect Y with a set of causes $\mathbf{X}$ and Z, under mild assumptions, the distribution of $Y \mid \mathbf{X}$ does not degenerate to a Gaussian.*

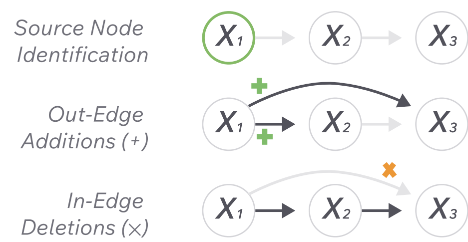
Theorem **Consistency** *(informal)* *The latent-aware BIC is a consistent scoring criterion.*

Background
Score-Based Causal Discovery

I. Greedy Equivalence Search (GES) searches over equivalence classes of DAGs to optimise a scoring criterion *(Chickering, 2002)*

II. TOPIC identifies source nodes and constructs the DAG in topological order *(Xu et al, 2025)*



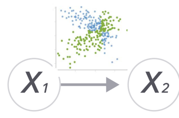


Algorithms
CMM Discovery

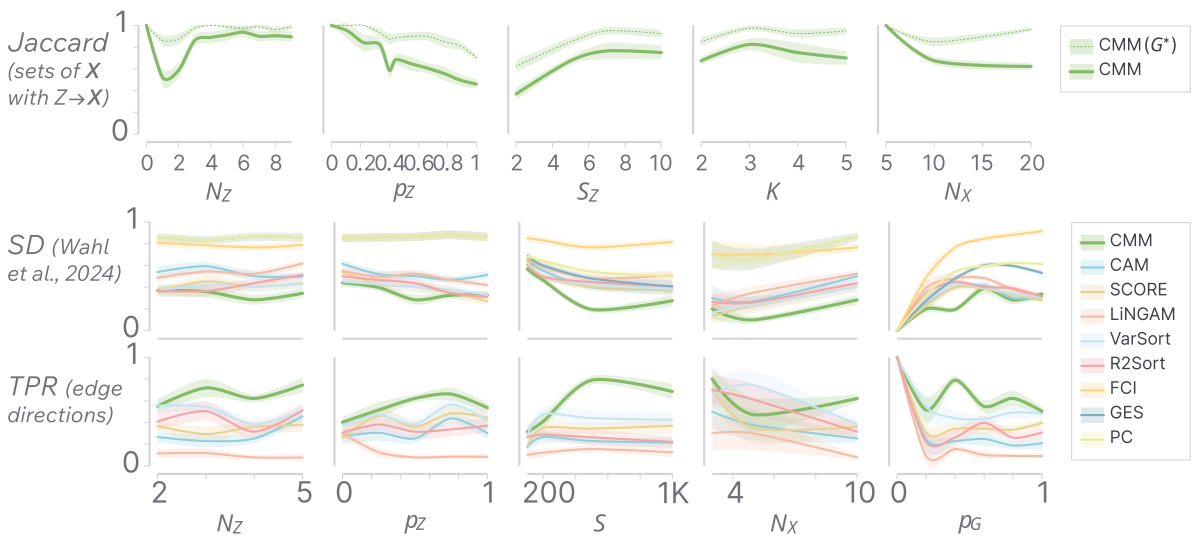
I. CMM (GES) w/ latent-aware BIC

II. CMM (TOPIC) w/ latent-aware BIC

Scoring for each variable (X_2) given its direct predecessors (X_1) in a given graph, we infer an MLR using EM and use the latent-aware BIC to pick K_i



Evaluation
CMM Discovery in Synthetic Data



Ablation Studies
Nonlinear Mixtures

Real-World Benchmark
Flow Cytometry Data

github.com/srhmm/cmm