# *Moment Matters*: Mean and Variance Causal Graph Discovery from Heteroscedastic Observational Data

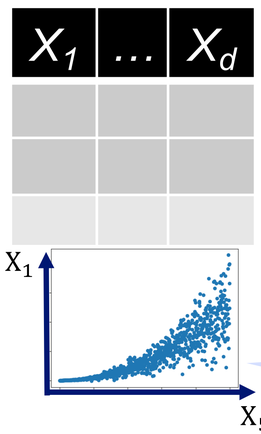
Yoichi Chikahara[1].

[1]Communication Science Laboratories. NTT, Inc.

This work was supported by JST ACT-X (JPMJAX23CF).

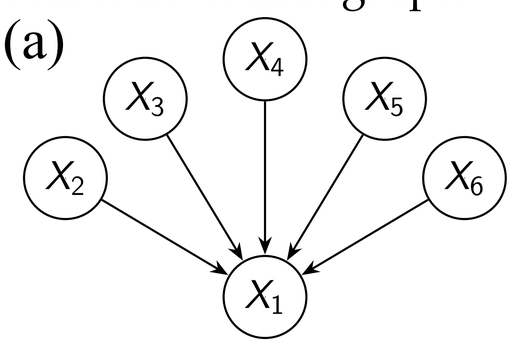
## Standard Graphs vs. Mean & Variance Graphs: Which ones are more interpretable?

Observational data $D$

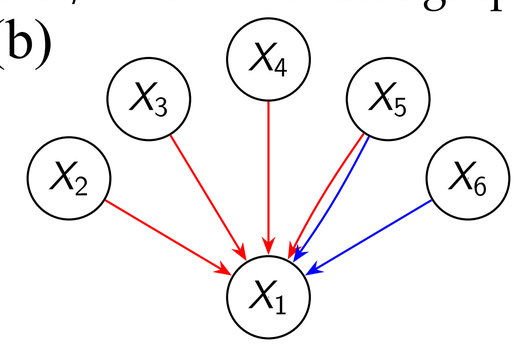



Heteroscedastic data

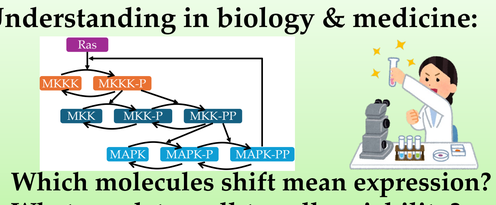
Standard causal graph (a)




Mean / Variance causal graphs (b)




**Understanding in biology & medicine:**



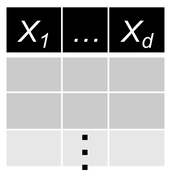

- **Which molecules shift mean expression?**
- **What regulates cell-to-cell variability?**

***Latent* discrimination detection in decision-making:**

- **Risk aversion by rejecting subgroups with high outcome variance is usual, but is it really *fair*?**

## Problem Setup: Bayesian Causal Discovery

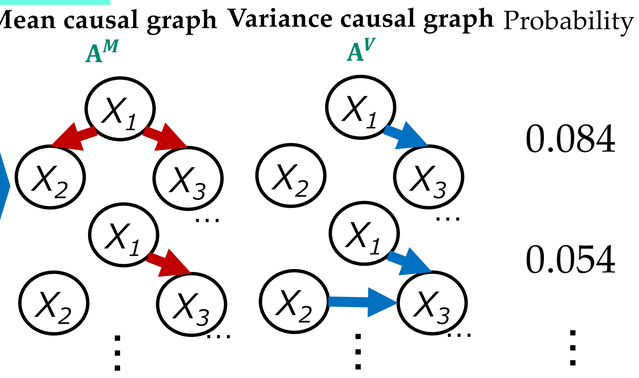
**Input** Observational data $D$




**Input** Prior knowledge about true causal graph structure

| Edge of $\mathbf{G}^M$ | Edge of $\mathbf{G}^V$ | Order |
|---|---|---|
| $X_1 \to X_3$ | $X_1 \to X_3$ | $\pi(X_1) < \pi(X_2)$ |
| ⋮ | ⋮ | ⋮ |

**Output** *Posterior over causal graphs* $\mathrm{P}(\mathbf{A}^M, \mathbf{A}^V | D)$




- **Advantage**: Can evaluate probability on structural feature of causal graphs

$\mathbb{E}_{\mathrm{P}(\mathbf{A}^M, \mathbf{A}^V | D)}[f(\mathbf{A}^M, \mathbf{A}^V)]$; e.g., $f(\mathbf{A}^M, \mathbf{A}^V) = \mathbb{I}(X_i \to X_j \text{ in } \mathbf{A}^M)$ or $\mathbb{I}(X_i \rightsquigarrow X_j \text{ in } \mathbf{A}^V)$ (path)

## Mean-Variance HNM, Mean & Variance Graphs

**Assume each $X_j$'s values are determined by the mean-variance HNM:**

$$X_j = m_j\left(\boldsymbol{X}_{\mathrm{pa}^M(j)}\right) + v_j\left(\boldsymbol{X}_{\mathrm{pa}^V(j)}\right) E_j \quad \text{for } j = 1, \ldots, d,$$

mean function; variance function; zero-mean noise

$$\mathbb{E}[X_j \mid \boldsymbol{X}_{\mathrm{pa}(j)}] = m_j(\boldsymbol{X}_{\mathrm{pa}^M(j)}) \qquad \mathbb{V}[X_j \mid \boldsymbol{X}_{\mathrm{pa}(j)}] = \mathbb{V}[E_j]\left(v_j(\boldsymbol{X}_{\mathrm{pa}^V(j)})\right)^2.$$

**Definition: Mean & Variance Causal Graphs**

Mean causal graph $G^M$ has edge $X_i \to X_j$ if and only if $X_i \in \boldsymbol{X}_{\mathrm{pa}^M(j)}$,
variance causal graph $G^V$ contains edge $X_i \to X_j$ if and only if $X_i \in \boldsymbol{X}_{\mathrm{pa}^V(j)}$.

**Adjacency Matrices:** $\mathbf{A}^M, \mathbf{A}^V \in \{0,1\}^{d \times d}$

## Proposed Method: Variational Approach

**Goal** Approximate the posterior as $\mathrm{P}(\mathbf{A}^M, \mathbf{A}^V \mid \mathcal{D}) \approx \mathrm{P}_\Phi(\mathbf{A}^M, \mathbf{A}^V)$

**Identifiability Conditions** Mean & Variance graphs are DAGs with shared permutation

Adjacency matrix (DAG) = Permutation$^\mathrm{T}$ matrix (Order) Upper-triangular matrix (Edge) Permutation matrix (Order)

$$\mathbf{A}^M = \mathbf{\Pi}^\mathrm{T}\mathbf{U}^M\mathbf{\Pi}$$
$$\mathbf{A}^V = \mathbf{\Pi}^\mathrm{T}\mathbf{U}^V\mathbf{\Pi}$$

※$A_{i,j} = U_{\pi_i,\pi_j}$

**Factorized Variational Distribution** $\mathrm{P}(\mathbf{A}^M, \mathbf{A}^V) = \sum_{\mathbf{U}^M, \mathbf{U}^V, \mathbf{\Pi}} \mathrm{P}(\mathbf{U}^M)\,\mathrm{P}(\mathbf{U}^V)\,\mathrm{P}(\mathbf{\Pi})$

- **Summation is computationally intractable ☹**
- **Sampling operations for binary matrices $\mathbf{U}^M, \mathbf{U}^V, \mathbf{\Pi}$ are NOT differentiable ☹**

**Differentiable sampling** Approximate gradient using differentiable function

$\mathbf{U}^M \sim \mathrm{P}_{\boldsymbol{\phi}^M}(\mathbf{U}^M)$
$\mathbf{U}^V \sim \mathrm{P}_{\boldsymbol{\phi}^V}(\mathbf{U}^V)$
$\mathbf{\Pi} \sim \mathrm{P}_{\boldsymbol{\varphi}}(\mathbf{\Pi})$

**Gumbel-softmax sampling** [Jang+; ICLR2016]:
Idea: Softmax(Unnormalized log prob. + Gumbel variables)
For each $i, j \in \{1, \ldots, d\}$,

$$\hat{U}_{i,j} \in [0,1] = \frac{e^{(\log \phi_{i,j} + G_{i,j,1})/\tau}}{e^{(\log \phi_{i,j} + G_{i,j,0})/\tau} + e^{(-\log \phi_{i,j} + G_{i,j,1})/\tau}}$$

**Gumbel-topK + SoftSort** [Prilo+; ICML2020]:
Idea: SoftSort(Log prob. + Gumbel variables)

$$\hat{\boldsymbol{\pi}} \in \mathbb{R}^d = \mathrm{SoftSort}(\log \boldsymbol{\varphi} + \boldsymbol{G}) \Rightarrow \hat{\mathbf{\Pi}} \in [0,1]^{\mathrm{d} \times \mathrm{d}}$$

Forward: Discrete sampling by taking argmax

Backward: Use $\hat{U}_{i,j} \in [0,1]$ or $\hat{\mathbf{\Pi}} \in [0,1]^{\mathrm{d} \times \mathrm{d}}$ for gradient approximation

**Objective Function**

$$\max_{\Phi,\Theta} \; \mathbb{E}_{\mathbf{A}^M, \mathbf{A}^V \sim \mathrm{P}_\Phi}\left[\log \mathrm{P}_\Theta(\mathcal{D} \mid \mathbf{A}^M, \mathbf{A}^V)\right] - \lambda\Omega_{\Phi,\Theta},$$

$$\log \mathrm{P}_\Theta(\mathcal{D} | \mathbf{A}^M, \mathbf{A}^V) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{(x_{i,j} - m_j(\boldsymbol{x}_{i,\mathrm{pa}^M(j)}; \boldsymbol{\theta}_j^M))^2}{2(v_j(\boldsymbol{x}_{i,\mathrm{pa}^V(j)}; \boldsymbol{\theta}_j^V))^2} + \log v_j(\boldsymbol{x}_{i,\mathrm{pa}^V(j)}; \boldsymbol{\theta}_j^V)\right)$$

$$\Omega_{\Phi,\Theta} = \lambda_\Phi \mathrm{KL}\left(\mathrm{P}_\Phi(\mathbf{U}^M, \mathbf{U}^V) \,||\, \mathrm{P}(\mathbf{U}^M, \mathbf{U}^V)\right) + \lambda_{\Theta^M}\sum_{j=1}^{d}\|\boldsymbol{\theta}_j^M\|_2^2 + \lambda_{\Theta^V}\sum_{j=1}^{d}\|\boldsymbol{\theta}_j^V\|_2^2,$$

**Algorithm 1** Parameter Learning for Mean and Variance Causal Graphs

**Require:** Observational dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$; parameters $\Phi, \Theta$; node-ordering constraints $\mathcal{O}$
Initialize all parameters $\Phi = \{\boldsymbol{\phi}^M, \boldsymbol{\phi}^V, \boldsymbol{\psi}\}$ and $\Theta = \{(\boldsymbol{\theta}_j^M, \boldsymbol{\theta}_j^V)_{j=1}^d\}$
**while** Stopping criterion not met **do**
  **while** Updates for parameters $\Theta \setminus \{(\boldsymbol{\theta}_j^V)_{j=1}^d\}$ and $\Phi \setminus \{\boldsymbol{\phi}^V\}$ not converged **do**
    ① Sample mean and variance DAGs $\mathbf{A}^M, \mathbf{A}^V \sim \mathrm{P}_\Phi(\mathbf{A}^M, \mathbf{A}^V)$
    ② Compute ELBO-based objective function in Eq. (8)
    ③ Update parameters $\Theta \setminus \{(\boldsymbol{\theta}_j^V)_{j=1}^d\}$ and $\Phi \setminus \{\boldsymbol{\phi}^V\}$ using scaled gradient
    **if** node-ordering constraints $\mathcal{O}$ are provided **then**
      Project updated parameters $\psi$ onto the feasible set by Eq. (11)
    **end if**
  **end while**
  **while** Updates for variance-related parameters $\{(\boldsymbol{\theta}_j^V)_{j=1}^d\}$ and $\{\boldsymbol{\phi}^V\}$ not converged **do**
    ① Sample mean and variance DAGs $\mathbf{A}^M, \mathbf{A}^V \sim \mathrm{P}_\Phi(\mathbf{A}^M, \mathbf{A}^V)$
    ② Compute ELBO-based objective function in Eq. (8)
    ③ Update variance-related parameters $\{(\boldsymbol{\theta}_j^V)_{j=1}^d\}$ and $\{\boldsymbol{\phi}^V\}$ using standard gradient
  **end while**
**end while**
**return** Learned parameters $\Phi$ and $\Theta$

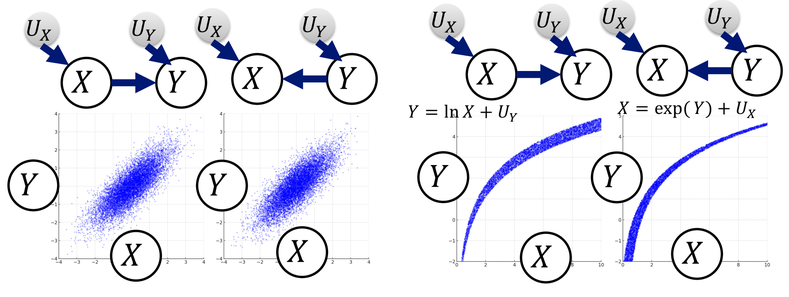
**Order knowledge** $\pi(X_1) < \pi(X_2)$ ⋮

$\mathcal{I}(\boldsymbol{\psi}) = \{\psi_i + c_{i,j} \le \psi_j \mid (i,j) \in \mathcal{S}\}$

## Identifiability Conditions

**Unidentifiable = Different models can yield the same observational data distribution**

Unidentifiable : **Linear** + Gaussian Noise | Identifiable : **Nonlinear** + Additive Noise




**Mean & variance causal graphs are identifiable (= uniquely determined) from observational data under the following sufficient conditions:**

**Theorem 3.5:** *Under Assumptions 3.1, 3.2, 3.3, and 3.4, mean and variance causal graphs $G^M$ and $G^V$ are identifiable from observational distribution $\mathrm{P}(\boldsymbol{X})$ if for $j = 1, \ldots, d$, (A) $m_j$ is a nonlinear function, (B) $v_j$ is a piecewise function,* ***but not a constant function****, and (C) $E_j$ is a Gaussian noise.*

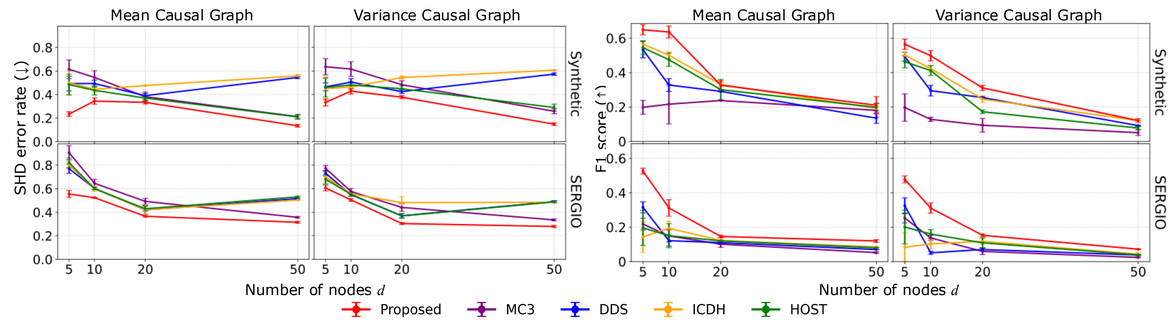
**Assumption 3.1** (Causal sufficiency)**:** *Exogenous noises satisfy $E_i \perp E_j$ for any $i, j \in \{1, \ldots, d\}$.*

**Assumption 3.2** (Causal minimality)**:** $\mathrm{P}(\boldsymbol{X})$ *satisfies the causal minimality condition with respect to moment-agnostic causal graph $G$.*

**Assumption 3.3:** *Mean and variance graphs $G^M$ and $G^V$ are directed acyclic graphs (DAGs).*

**Assumption 3.4** (Shared permutation condition)**:** *There exists an identical permutation (a.k.a., topological ordering) of mean and variance causal graphs $G^M$ and $G^V$.*

## Experimental Results




| | $n = 100$ | | $n = 200$ | | $n = 853$ | |
|---|---|---|---|---|---|---|
| | SHD (↓) | F1 (↑) | SHD (↓) | F1 (↑) | SHD (↓) | F1 (↑) |
| MC3 | 22.6 ± 1.1 | 0.19 ± 0.04 | 20.3 ± 1.3 | 0.20 ± 0.07 | 18.9 ± 0.7 | 0.14 ± 0.05 |
| DDS | 17.6 ± 0.7 | 0.20 ± 0.03 | 16.6 ± 0.6 | 0.20 ± 0.02 | 15.0 ± 0.7 | 0.28 ± 0.02 |
| ICDH | 20.0 ± 1.0 | **0.21 ± 0.01** | 17.5 ± 0.5 | 0.22 ± 0.01 | 14.5 ± 0.5 | 0.27 ± 0.03 |
| HOST | 16.1 ± 0.5 | 0.19 ± 0.01 | 15.0 ± 0.3 | **0.33 ± 0.02** | **13.5 ± 0.8** | **0.38 ± 0.03** |
| PROPOSED | **16.0 ± 0.3** | 0.20 ± 0.02 | **14.9 ± 0.4** | 0.31 ± 0.02 | 13.7 ± 0.5 | 0.36 ± 0.02 |
| PROPOSED +25% | 14.9 ± 0.4 | 0.34 ± 0.03 | 14.8 ± 0.5 | 0.35 ± 0.02 | 13.4 ± 0.4 | 0.37 ± 0.02 |
| PROPOSED +50% | **13.2 ± 0.5** | **0.36 ± 0.02** | **13.2 ± 0.3** | **0.36 ± 0.02** | **13.1 ± 0.2** | **0.45 ± 0.03** |

**Table 4: Estimated posterior probability of MEK → ERK in inferred variance causal graph structure**

| | $n = 100$ | $n = 200$ | $n = 853$ |
|---|---|---|---|
| PROPOSED | 0.585 ± 0.009 | 0.592 ± 0.008 | 0.620 ± 0.019 |