universität
wien

Digitize!
Computational Social Sciences
Bundesministerium
Bildung, Wissenschaft
und Forschung

Bayesian Causal Structure Learning

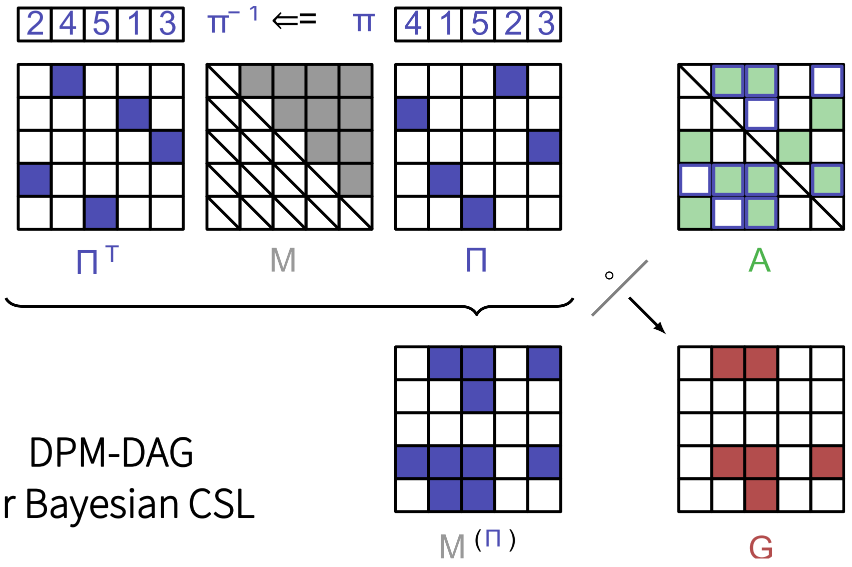
Simon Rittel[1,2], Sebastian Tschiatschek[1]

[1]University of Vienna, Faculty of Computer Science, Vienna, Austria

[2]University of Vienna, UniVie Doctoral School for Computer Science, Vienna, Austria

Simon Rittel

Sebastian Tschiatschek
(supervison)



DPM-DAG
for Bayesian CSL

Outline of this talk

- Intro to Causality
- Causal Structure Learning
- Bayesian Causal Structure Learning
- Differentiable Probabilistically Masked DAG (*DPM-DAG*)

Causality for Machine Learning

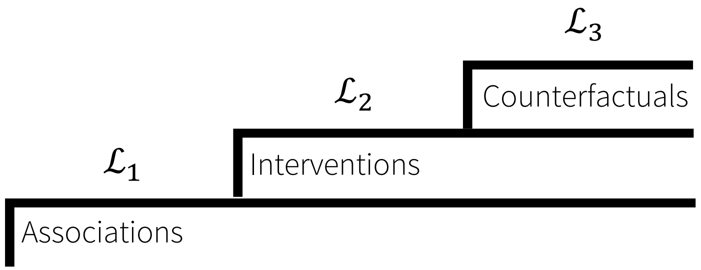
- Fundamental differences

Statistical relations ($\mathcal{L}_1$)

⇆

Causal relations ($\mathcal{L}_{1:3}$)

→



$\mathcal{L}_3$

Counterfactuals

$\mathcal{L}_2$

Interventions

$\mathcal{L}_1$

Associations

- Promising applications

Robustness & generalization

Interpretability & explainability

Fairness

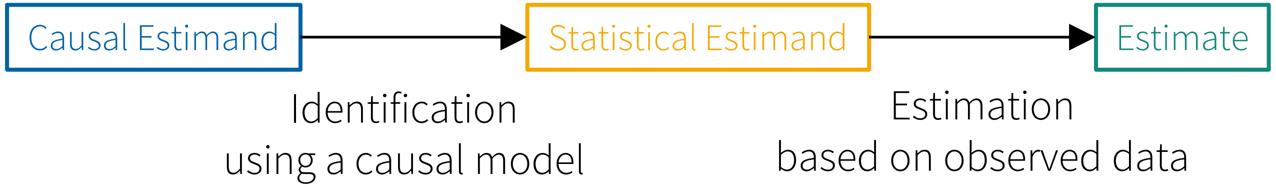
Causal Insights

Causal Model as Basis for Causal Inference

- Indiviudal effect of a treatment T on an outcome Y: $\text{ITE}_i \coloneqq Y_i(T=1) - Y_i(T=0)$

- Average treatment effect: $\text{ATE} \coloneqq \mathbb{E}_i[\text{ITE}_i]$

- Identifiability: Causal effect can be consistently estimated from observed data



Causal Estimand → Statistical Estimand → Estimate

Identification using a causal model

Estimation based on observed data

- Controlling/adjusting for a set of confounders X:

$$\text{CATE} = \mathbb{E}_X\big[\mathbb{E}_Y[Y|T=1,X] - \mathbb{E}_Y[Y|T=0,X]\big] = \frac{1}{|\mathcal{D}_{T=1}|}\sum_{i\,\in\,\mathcal{D}_{T=1}} \hat{\mu}_{Y|T=1,\,X}(X_i) - \frac{1}{|\mathcal{D}_{T=0}|}\sum_{j\,\in\,\mathcal{D}_{T=0}} \hat{\mu}_{Y|T=0,\,X}\big(X_j\big)$$

Recap of Bayesian Networks

- Graphical model:
  One-to-one mapping between nodes $\boldsymbol{V_i} \in \boldsymbol{V}$ of a direct acyclic graph (DAG) $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ and random variables $\boldsymbol{X_i} \in \boldsymbol{X}$

- Local Markov Condition:
  Given the parents $\mathbf{pa}$ of a node $\boldsymbol{V_i}$ in the DAG $\boldsymbol{G}$, the corresponding random variable $\boldsymbol{X_i}$ is independent of all its non-descendants $\mathbf{nd}$.

- Bayesian Network Factorization:
  Given a joint probability distribution $\boldsymbol{P_X}$ and a DAG $\boldsymbol{G}$,
  $\boldsymbol{P_X}$ factorizes according to $\boldsymbol{G}$ if:

$$P(\boldsymbol{X}) \coloneqq P\left(\{X_i\}_{i=1}^{D}\right) = \prod_{i}^{D} P(x_i|\mathbf{pa}(X_i))$$

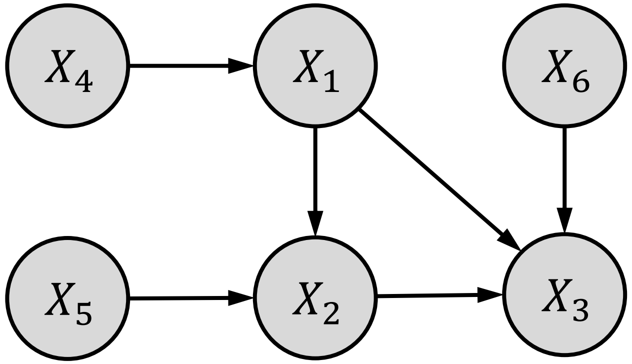
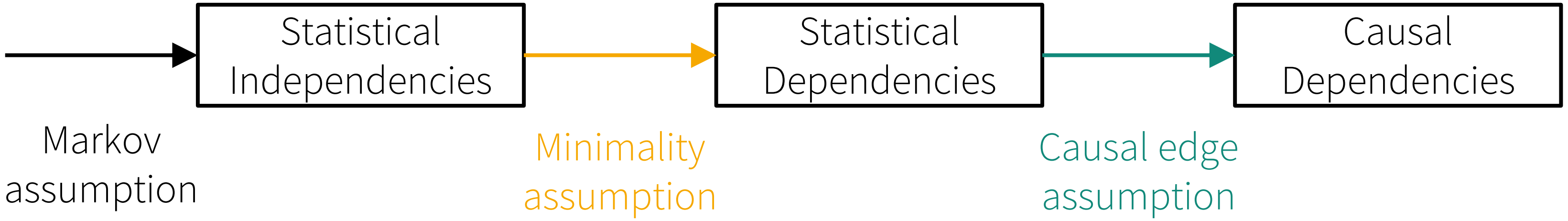


Fig. 1: DAG over six random variables

Causal Graphs

- Minimality assumption:
  1. Local Markov condition $X_i \perp_P \mathrm{nd}(X_i) \mid \mathbf{pa}(X_i)$
     (implies d-separation as *global Markov condition)*
  2. Adjacent nodes in the DAG $\boldsymbol{G}$ are dependent $X_i \sim X_j \; in \; G \implies X_i \perp_P X_j$
     (no additional independences)

- Strict causal edge assumption:
  Every parent is a direct cause of all its children, i.e. the children are affected by changes in their parents



Causal Structure Learning (*CSL*)

- Functional Causal Model: indexed tuple of
  - endogenous variables $\boldsymbol{X}$,
  - exogeneous noise variables $\boldsymbol{\epsilon}$ with distribution $P_{\boldsymbol{\epsilon}}$,
  - deterministic functions $\boldsymbol{g}$, s. t. $X_i \coloneqq g_i(\mathbf{pa}_{\boldsymbol{G}}(X_i), \epsilon_i)$
- Assumptions:
  - Acyclic causal relations
    → Direct Acyclic Graph (*DAG*) $\boldsymbol{G}$
  - Causal sufficiency
    → no latent confounders and mutually independent noise $\boldsymbol{\epsilon}$

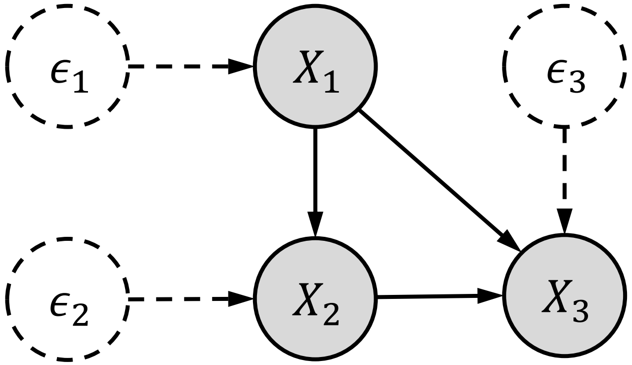


Fig. 2: Causal DAG induced by an acyclic FCM over three observed random variables

Interventions by the do-Operator $(\mathcal{L}_1)$

- Definition
  Hard intervention $\mathbf{do}(X_i = x)$ replaces structural function g_i by the assignment $X_i = x$

- Truncated Factorization
  $$P(\boldsymbol{X}|\mathrm{do}(X_i = x) \coloneqq \delta(X_i = x) \prod_{j \neq i} P\big(x_j|\mathbf{pa}(X_j)\big)$$

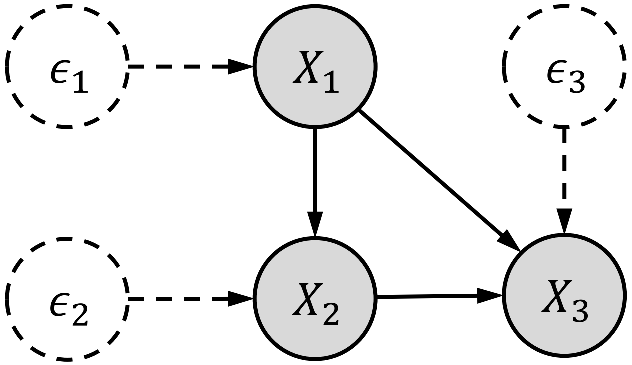


Fig. 2: Causal DAG induced by an acyclic FCM over three observed random variables

Interventions by the do-Operator $(\mathcal{L}_1)$

- Definition
  Hard intervention $\mathbf{do}(X_i = x)$ replaces structural function g_i by the assignment $X_i = x$

- Truncated Factorization
  $$P(\boldsymbol{X}|\mathrm{do}(X_i = x) \coloneqq \delta(X_i = x) \prod_{j \neq i} P\big(x_j|\mathbf{pa}(X_j)\big)$$

- Example:
  - $P(X_1, X_3) = \int P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)\, dX_2$

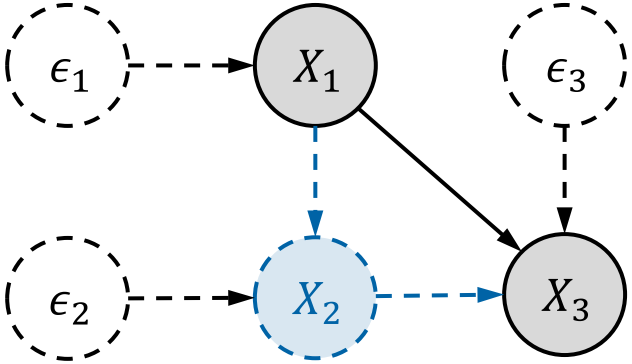


Fig. 2: Causal DAG induced by an acyclic FCM over three observed random variables

Interventions by the do-Operator $(\mathcal{L}_1)$

- Definition
  Hard intervention $\mathbf{do}(X_i = x)$ replaces structural function g_i by the assignment $X_i = x$

- Truncated Factorization
  $$P(\boldsymbol{X}|\mathrm{do}(X_i = x) \coloneqq \delta(X_i = x) \prod_{j \neq i} P\big(x_j|\mathbf{pa}(X_j)\big)$$

- Example:
  - $P(X_1, X_3) = \int P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)\, dX_2$
  - $P(X_1, X_3|do(X_2 = x_2)) = P(X_3|X_1, X_2 = x_2)P(X_1)$

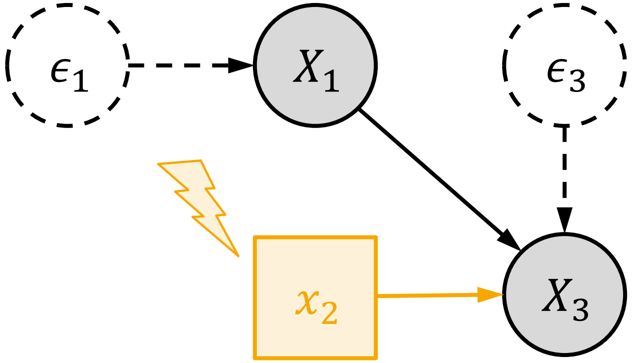


Fig. 2: Causal DAG induced by an acyclic FCM over three observed random variables

Interventions by the do-Operator $(\mathcal{L}_1)$

- Definition
  Hard intervention $\mathbf{do}(X_i = x)$ replaces structural function g_i by the assignment $X_i = x$

- Truncated Factorization
  $$P(\boldsymbol{X}|\mathrm{do}(X_i = x) \coloneqq \delta(X_i = x) \prod_{j \neq i} P\big(X_j|\mathbf{pa}(X_j)\big)$$

- Example:
  - $P(X_1, X_3) = \int P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)\, dX_2$
  - $P(X_1, X_3|do(X_2 = x_2)) = P(X_3|X_1, X_2 = x_2)P(X_1)$
  - $P(X_1, X_3|\, X_2 = x_2) = \frac{P(X_3|X_1, X_2=x_2)P(X_2=x_2|X_1)P(X_1)}{P(X_2=x_2)}$

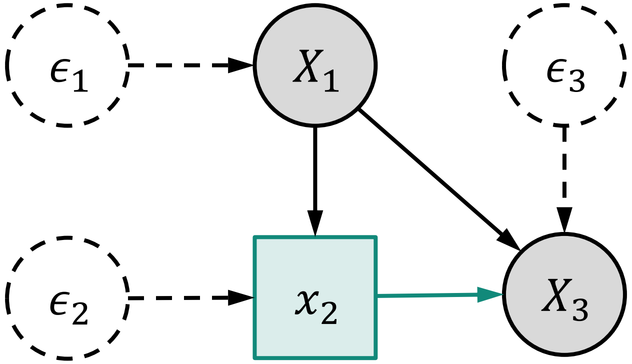


Fig. 2: Causal DAG induced by an acyclic FCM over three observed random variables

The Three Layer Causal Hierarchy by Pearl

Level	Typical Quantity	Typical Activity	Typical Questions
1. Association	$P(Y \vert X = x)$	Seeing	What is? How does observing X change my belief in Y?
2. Intervention	$P(Y \vert do(X = x))$	Doing/Intervening	What if I do X?
3. Counterfactuals	$P(Y_x \vert do(X = x'), y')$	Imagining, Retrospection	Why? Was it X that caused Y?

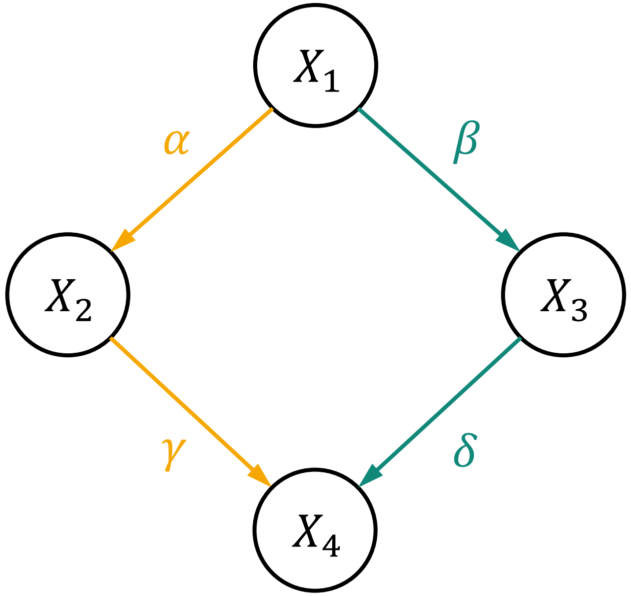
Typical Assumptions for Independence-based CSL

- Markov assumption:

$$X \perp_G Y \mid Z \quad \Longrightarrow \quad X \perp_P Y \mid Z$$

- Faithfulness:

$$X \perp_G Y \mid Z \quad \Longleftarrow \quad X \perp_P Y \mid Z$$



$$X_4 \coloneqq \gamma X_2 + \delta X_3 = \underbrace{(\alpha\gamma + \beta\delta)}_{\neq 0} X_1$$

Typical Assumptions for Independence-based CSL

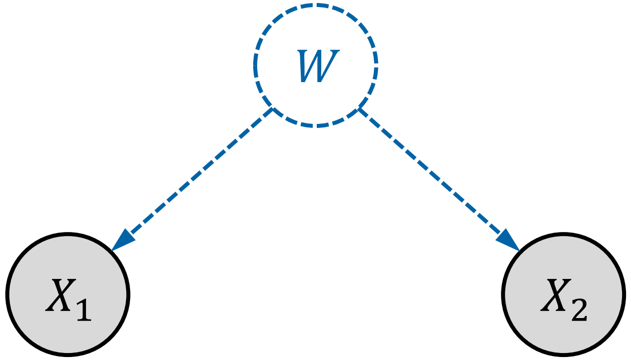
- Markov assumption:

$$X \perp_G Y \mid Z \quad \Longrightarrow \quad X \perp_P Y \mid Z$$

- Faithfulness:

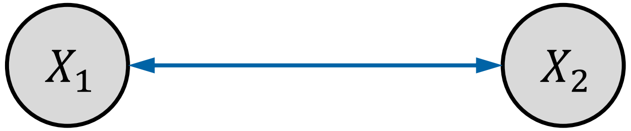
$$X \perp_G Y \mid Z \quad \Longleftarrow \quad X \perp_P Y \mid Z$$

- Causal sufficiency:
  No unobserved confounders.



$X_1 \perp X_2 \mid W$

True graph



$X_1 \not\perp X_2$

Projected graph

Typical Assumptions for Independence-based CSL

- Markov assumption:

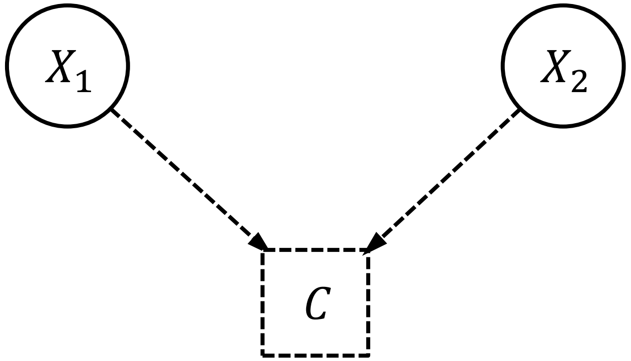
  $$X \perp_G Y \mid Z \implies X \perp_P Y \mid Z$$

- Faithfulness:

  $$X \perp_G Y \mid Z \impliedby X \perp_P Y \mid Z$$

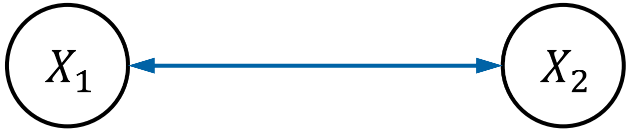
- Causal sufficiency:
  No unobserved common causes

- No selection bias:
  No conditioning on unobserved colliders



$X_1 \perp X_2$

True graph



$X_1 \not\perp X_2 \mid C$

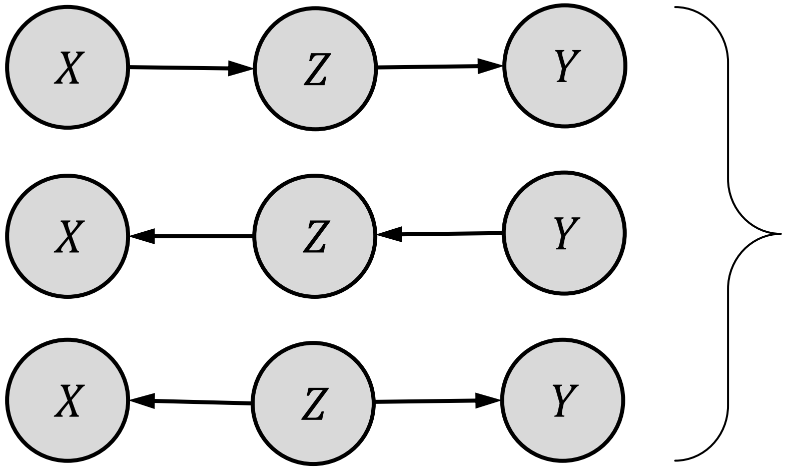
Projected graph

Review of Unshielded 3-node Structures

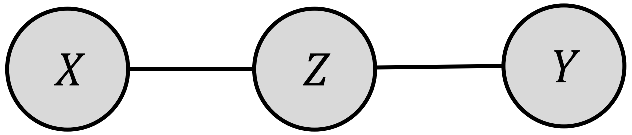
Causal trail

Evidential trail

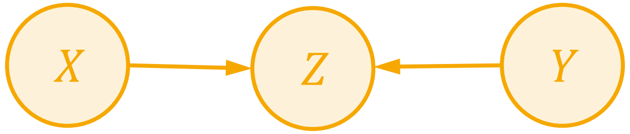
Common cause



$X \not\perp Y$
$X \perp Y \mid Z$



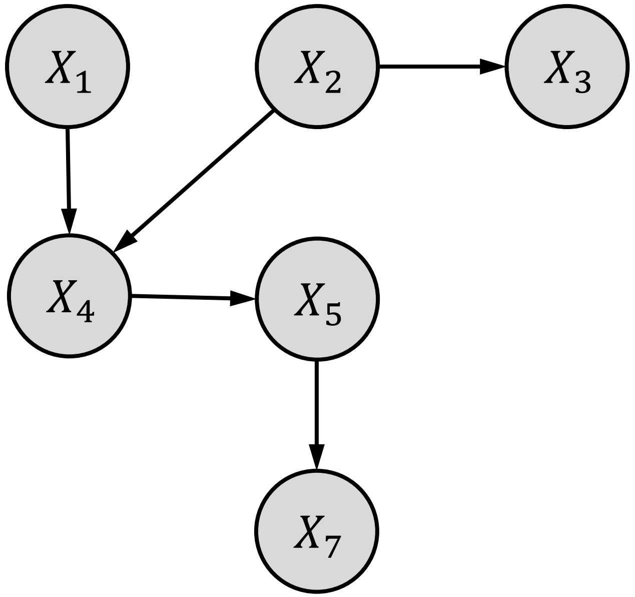
Common effect



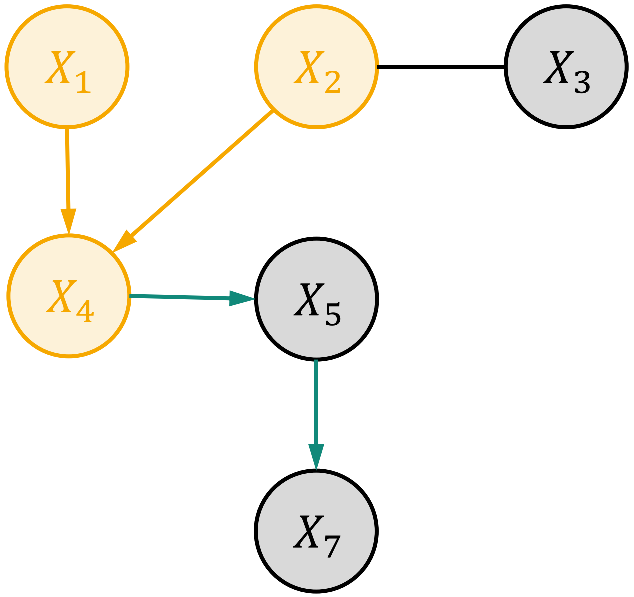
$X \perp Y$
$X \not\perp Y \mid Z$

Sketch of the PC-algorithm

1) Start with a complete undirected graph
2) Eliminate edges between variables that are (conditionally) independent
3) Add arrow marks at colliders in identified v-structure
4) Propagate arrows such that no additional v-structures are formed that were not detected



True graph



PCDAG :
skeleton + v-structures

Known Identifiable Causal Models

- Linear Gaussian model with equal or known variance (Loh & Bühlmann 2014) $Y \coloneqq aX + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(\mu, \sigma)$

- Linear non-Gaussian model (LiNGAM) (Shimizu et al.. 2006) $Y \coloneqq aX + \epsilon \quad \text{with} \quad \epsilon \nsim \mathcal{N}(\mu, \sigma)$

- Nonlinear additive noise model (ANL) (Hoyer et al., 2008) $Y \coloneqq f(X) + \epsilon \quad \text{where } f \text{ is nonlinear}$

- Post-nonlinear causal model (PNL) (Zhang & Hyvärinen, 2009) $Y \coloneqq g(f(X) + \epsilon) \quad \text{where } g \text{ is nonlinear \& invertible}$

Greedy Equivalence Search (*GES*)

1) Initialization by an empty graph

2) Forward equivalence search
   Add the edge that most increases the score and maps the resulting graph then to its MEC

3) Backward equivalence search
   Remove the edge that will most increase the score until no further edges can be removed

- Score equivalence:
  Graphs of the same MEC are assigned the same score

- Locally consistent scoring criterion:
  Score prefers edge additions that remove incorrect dependencies and edge deletions that remove incorrect dependencies

- Decomposable score function:

$$S(\boldsymbol{G}, \boldsymbol{X}) = \sum_{d=1}^{D} S(X_i, \boldsymbol{pa_G}(X_i))$$

Continuous Relaxation of the Discrete Graph Structure & Acyclicity

- Converting the combinatorical optimization problem into a continuous program

$$\min_{\boldsymbol{G}\in\{\mathbf{0},\mathbf{1}\}^{\boldsymbol{D}\times\boldsymbol{D}}} \boldsymbol{S}(\boldsymbol{G}) \quad \text{subject to } \boldsymbol{G} \in \boldsymbol{G_{acyclic}} \qquad \Leftrightarrow \qquad \min_{\boldsymbol{G}\in[\mathbf{0},\mathbf{1}]^{\boldsymbol{D}\times\boldsymbol{D}}} \boldsymbol{S}(\boldsymbol{G}) \quad \text{subject to } h(\boldsymbol{G}) = \mathbf{0}$$

- Differentiable DAG-Constraint $h\left(\boldsymbol{G}_{\text{acyclic}}\right) = 0\,,\quad h\left(\boldsymbol{G}_{\text{cyclic}}\right) > 0$
  - $h_1(\boldsymbol{G}) = \operatorname{tr}\left(e^{\boldsymbol{G}\circ\boldsymbol{G}}\right) - D$ (Zheng et al. , 2018)
  - $h_2(\boldsymbol{G}) = \operatorname{tr}\left(\left(\boldsymbol{I} + \frac{\mathbf{1}}{\boldsymbol{D}}(\boldsymbol{G}\circ\boldsymbol{G})\right)^D\right)$ (Yu et al., 2019)
  - $h_3(\boldsymbol{G}) = \log\det(s\boldsymbol{I} - \boldsymbol{G}\circ\boldsymbol{G}) + D\log s$ (Bello et al. , 2023)

Research Areas in Causal Structure Learning

- Relaxing assumptions
  - No assumed causal sufficiency : FCI algorithm (Spirtes et al., 2001)
  - No assumed acyclicity CCD algorithm (Richardson, 1996)
  - Neither of both: SAT-based causal discovery (Hyttinen et. al., 2013)

Research Areas in Causal Structure Learning

- Relaxing assumptions
- Improving computational scalability
    - Limiting the number of potential parents: PNS-algorithm (Bühlmann et al., 2014)
    - Omitting some CI test: RFCI-algorithm (Colombo et al, 2012)
    - Considering only one edge change at a time: GES-algorithm (Chickering, 2002)
    - Continuous relaxation of the binary adjacency matrix: NOTEARS-algorithm (Zheng et al., 2018)

Research Areas in Causal Structure Learning

- Relaxing assumptions
- Improving computational scalability
- Increasing robustness:
  - Additional CI-tests: Order-independent PC/FCI (Colombo & Maathuis, 2014)

Research Areas in Causal Structure Learning (non-exhaustive)

- Relaxing assumptions
- Improving computational scalability
- Increasing robustness
- Identifiable functional models
- Focus only on local structure relevant for downstream task
- **Modeling uncertainty in the prediction**
- Combining with interventional data

Independence-based CSL

- Based on Conditional Independence (CI) tests
- Additional assumption of faithfulness
- Iterative restriction of CI test to avoid all pairwise combinations
- Point estimate as output
- Sound in the large sample limit

Bayesian CSL

$$p(\boldsymbol{G}, \boldsymbol{\Theta} | \boldsymbol{X}) \propto p(\boldsymbol{G}) p(\boldsymbol{\Theta} | \boldsymbol{G}) p(\boldsymbol{X} | \boldsymbol{G}, \boldsymbol{\Theta})$$

- Quantifying the uncertainty in the posterior
- Incorporation of probabilistic domain knowledge via prior
- Sound in the large sample limit

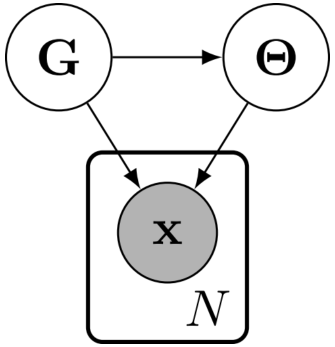


Fig. 2: Generative model

Generative Model

$$p(\boldsymbol{G}, \boldsymbol{\Theta}, \boldsymbol{X}) = p(\boldsymbol{G})p(\boldsymbol{\Theta}|\boldsymbol{G})p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta})$$

$$p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta}) = \prod_{n=1}^{N} \prod_{d=1}^{D} p\left(X_d^{(n)} \middle| \mathrm{pa}_{\boldsymbol{G}}\left(X_d^{(n)}\right), \boldsymbol{\Theta}\right)$$

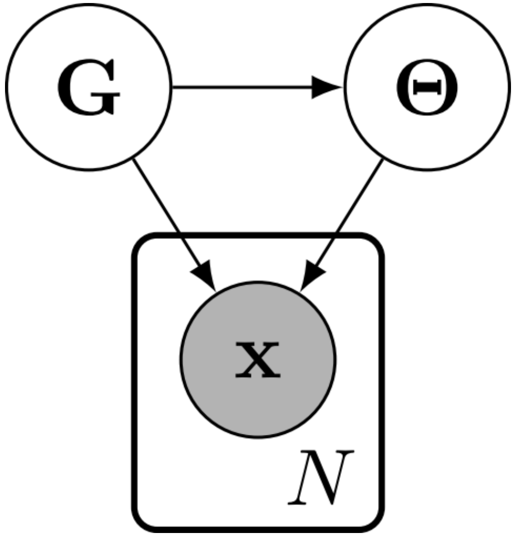


Fig. 2: Generative model

Marginalized Generative Model

$$\mathrm{p}(\boldsymbol{G}, \boldsymbol{X}) = p(\boldsymbol{G}) \int p(\boldsymbol{\Theta}|\boldsymbol{G}) p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta}) \boldsymbol{d\Theta}$$

$$\leq p(\boldsymbol{G}) \boldsymbol{p}_{\boldsymbol{\Theta}^*}(\boldsymbol{X}|\boldsymbol{G})$$

$$\text{where } \boldsymbol{\Theta}^* \coloneqq \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{X}|\boldsymbol{G}, \boldsymbol{\Theta})$$

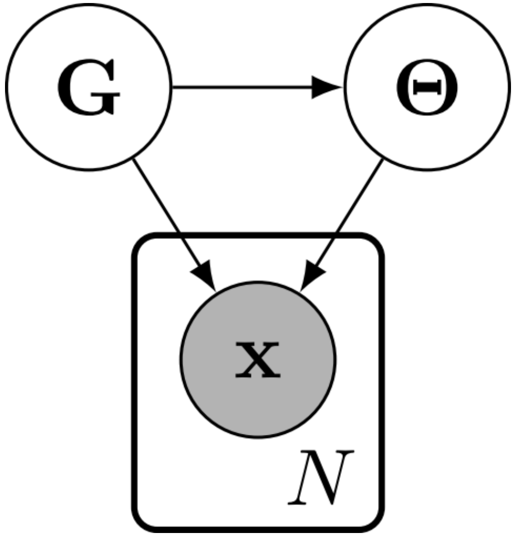


Fig. 2: Generative model

Graph Posterior

$$p_{\boldsymbol{\Theta}^*}(\boldsymbol{G}|\boldsymbol{X}) = \frac{p_{\boldsymbol{\Theta}^*}(\boldsymbol{G},\boldsymbol{X})}{p(\boldsymbol{X})} \propto p_{\boldsymbol{\Theta}^*}(\boldsymbol{G},\boldsymbol{X})$$

$$\text{where } \boldsymbol{\Theta}^* \coloneqq \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{X}|\boldsymbol{G},\boldsymbol{\Theta})$$

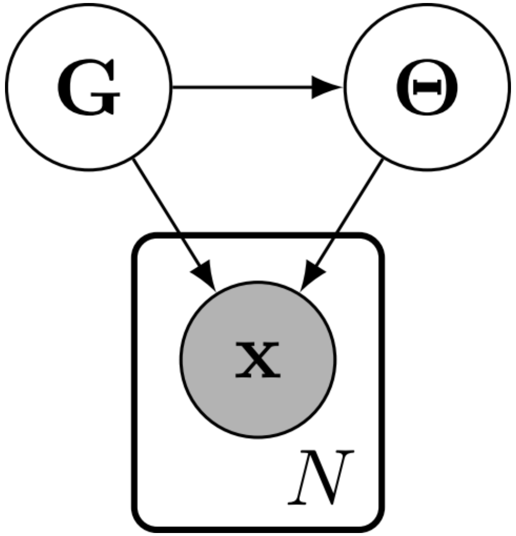


Fig. 2: Generative model

Probabilistic Graph: REINFORCE estimator [1]

- Independent Bernoulli distributed RV models each edge

$$G_{ij} \sim \mathrm{Bern}(\phi_{ij})$$

- Score function gradient estimator for its parameters

$$\eta = \boldsymbol{\nabla}_{\boldsymbol{\phi}}\, \mathbb{E}_{p_{\Theta^*(X)}}[f(\boldsymbol{X})] = \boldsymbol{\nabla}_{\boldsymbol{\phi}} \int p_{\Theta^*(X)} f(\boldsymbol{X})\mathbf{d}\boldsymbol{X} = \int f(\boldsymbol{X}) \boldsymbol{\nabla}_{\boldsymbol{\phi}} p_{\Theta^*(X)} \mathbf{d}\boldsymbol{X} = \int f(\boldsymbol{X}) p_{\Theta^*(X)}\, \boldsymbol{\nabla}_{\boldsymbol{\phi}} \log p_{\Theta^*(X)}\, \mathbf{d}\boldsymbol{X} =$$

$$= \mathbb{E}_{p_{\Theta^*(X)}}\left[f(\boldsymbol{X}) \boldsymbol{\nabla}_{\boldsymbol{\phi}} \log p_{\Theta^*(\boldsymbol{X})}\right]$$

$$\hat{\eta}_N = \frac{1}{N} \sum_{n=1}^{N} f\left(\widehat{\boldsymbol{X}}^{(n)}\right) \boldsymbol{\nabla}_{\boldsymbol{\phi}} \log p_{\Theta^*\left(\widehat{\boldsymbol{X}}^{(n)}\right)} \qquad \text{where} \quad \widehat{\boldsymbol{X}}^{(n)} \sim p_{\Theta^*}(\boldsymbol{X})$$

[1] Rezende et al. , ‘MC Gradient Estimation in Machine Learning’, in the Journal of Machine Learning Research, (2020)

Probabilistic Graph: Pathwise Gradient Estimator [2]

- Independent perturbed Gumbel distributed RV models each edge

$$G_i \sim \mathrm{Gumbel}(0, 1)\ , \qquad \phi_i + G_i \sim \mathrm{Gumbel}(\phi_i, 1)$$

- Perturbed Gumbel-Softmax samples

$$\arg\max_{i \in \mathbb{I}} (\phi_i + G_i) \sim \frac{\exp(\phi_i)}{\sum_{j \in \mathbb{I}} \exp\left(\phi_j\right)}$$

- $\mathbf{Softmax}$ as continuous, differentiable relaxation of the $\mathbf{arg\,max}$ operator (equivalence for $\boldsymbol{\tau} \to \mathbf{0}$)

$$Z_i = \frac{\exp\left((\phi_i + G_i)/\tau\right)}{\sum_{j \in \mathbb{I}} \exp\left((\phi_i + G_i)/\tau\right)}$$

[2] Maddison et al. , ‘The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables’, in Proceedings of the International Conference on Learning Representations, (2017)

Probabilistic Graph: Pathwise Gradient Estimator

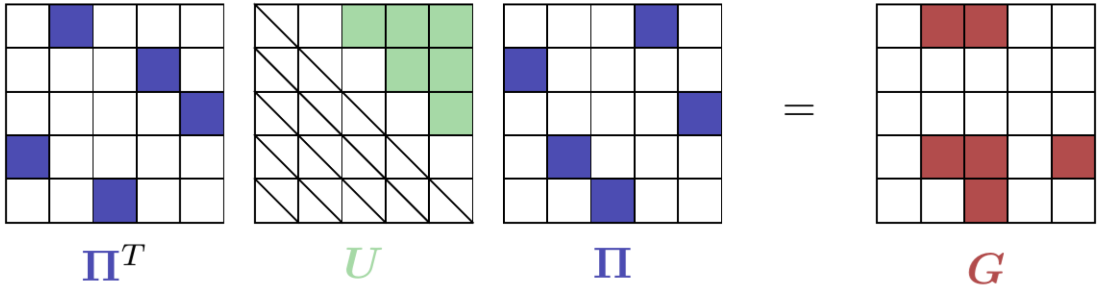
- Straight-through estimator
  discrete samples ($\arg\max$) in the forward pass and continuous samples ($\mathit{s}\mathrm{oftmax}$) in the backward pass

- Logistic samples for binary RV
  $$[G_1 + \phi_1 > G_0 + \phi_0] = [\,\underbrace{G_1 - G_0}_{\doteq L} + \underbrace{\phi_1 - \phi_0}_{=:\,\phi} > 0\,]\,, \qquad \text{where } L \sim \mathrm{Logistic}(0{,}1)$$

- Sigmoid as 2-dim version of Softmax
  $$Z_i = \left(1 + \exp\left(-\frac{L_i + \phi_i}{\tau}\right)\right)^{-1}$$

Enforcing Acyclicity

1) Permuted upper triangular matrix [3]



$$p(\boldsymbol{G}) = \sum_{\boldsymbol{\Pi} \in \mathcal{P}_D(\boldsymbol{G})} p(\boldsymbol{G}, \boldsymbol{\Pi})$$

Mean-field approximation: $p(\boldsymbol{G}, \boldsymbol{\Pi}) = p(\boldsymbol{U})p(\boldsymbol{\Pi})$

2) Differentiable acyclicity constraint [4]

$$h(\boldsymbol{G}_{\text{cyclic}}) > 0 \quad , \quad h(\boldsymbol{G}_{\text{acyclic}}) = 0$$

$$p(\boldsymbol{G}) \propto e^{-\lambda h(\boldsymbol{G})}$$

$$p(\boldsymbol{G}_{\text{cyclic}}) \xrightarrow[\lambda \to \infty]{} 0 \quad , \quad p(\boldsymbol{G}_{\text{acyclic}}) \xrightarrow[\lambda \to \infty]{} \frac{1}{|\mathbb{G}_{\boldsymbol{acyclic}}|}$$

[3] Charpentier et al. , 'Differentiable DAG sampling', in Proceedings of the International Conference of Learning Representations, (2022)

[4] Lorch et al. , 'DiBS: Differentiable Bayesian structure learning', in Advances in Neural Information Processing Systems, volume 34, pp. 24111-2413, (2021)

Variational Posterior Not Constrained to DAGs

$$D_{\mathrm{KL}}\left(p_{\boldsymbol{\phi}}(\boldsymbol{G}) | p_{\lambda}(\boldsymbol{G})\right)$$

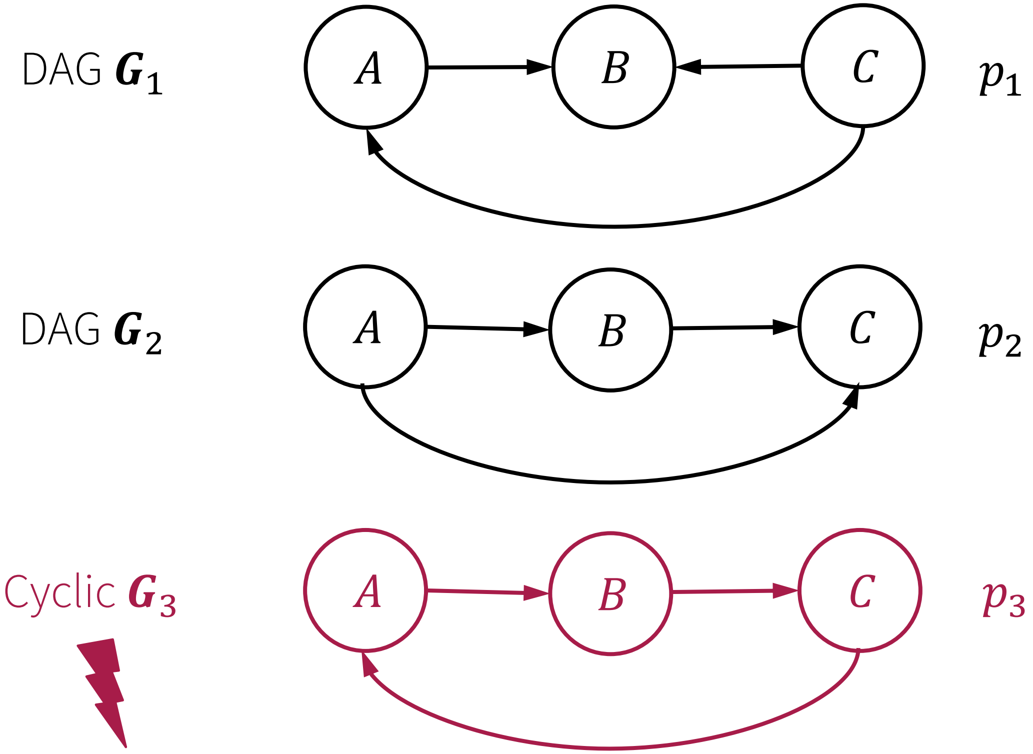
KL-Divergence

Posterior [5]

Prior

$$p_{\boldsymbol{\phi}}(\boldsymbol{G}) = \prod_{i \neq j} p_{\phi_{ij}}(\boldsymbol{G}_{ij})$$

$$p_{\lambda}(\boldsymbol{G}) \propto \exp(-\lambda h(\boldsymbol{G}))$$



[5] Geffner et al. , ‘Deep end-to-end causal inference’, in NeurIPS Workshop on Causality for Real-world Impact, (2022)

Incorporating Probabilistic Knowledge in a Gibbs Prior

- Number of expected causes for every node [3]
  - Erdös-Renyi graphs

$$p(\boldsymbol{G}) \propto p^{\|\boldsymbol{G}\|_1}(1-p)^{E-\|\boldsymbol{G}\|_1}$$

  - Scale-free graphs

$$p(\boldsymbol{G}) \propto \prod_{i=1}^{D}\left(1+\left\|G_i^{\mathrm{T}}\right\|_1\right)^{-3}$$

- Additional sparsity regularization [4]

$$p(\boldsymbol{G}) \propto \beta\|\boldsymbol{G}\|_F^2$$

- Prior over a single edge $\boldsymbol{p_{ij}}$

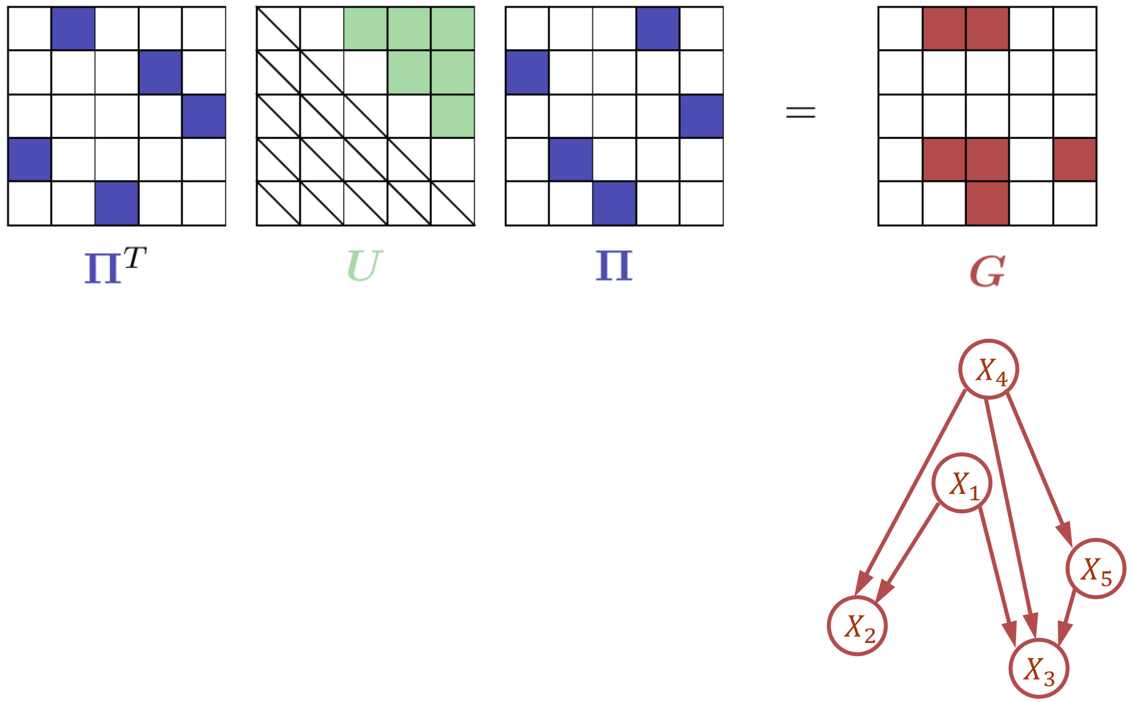
$$p(\boldsymbol{G}) \propto \left(q_{ij}\boldsymbol{G}_{ij} + \left(1-q_{ij}\right)\left(1-\boldsymbol{G}_{ij}\right)\right)$$

$$p\left(\boldsymbol{G} \in \mathbb{G}_{ij}\right) = \frac{q_{ij}}{p_{ij} + \left(1-p_{ij}\right)} = q_{ij} \coloneqq \frac{p_{ij}}{\left|\mathbb{G}_{ij}\right|}$$

[4] Lorch et al. , ‘DiBS: Differentiable Bayesian structure learning’, in Advances in Neural Information Processing Systems, volume 34, pp. 24111-2413, (2021)

[5] Geffner et al. , ‘Deep end-to-end causal inference’, in NeurIPS Workshop on Causality for Real-world Impact, (2022)

Differentiable Probabilistic DAG (*DP-DAG*) [3]



1) $\boldsymbol{\Pi} \sim p_{\psi}(\boldsymbol{\Pi})$ Gumbel-Softsort

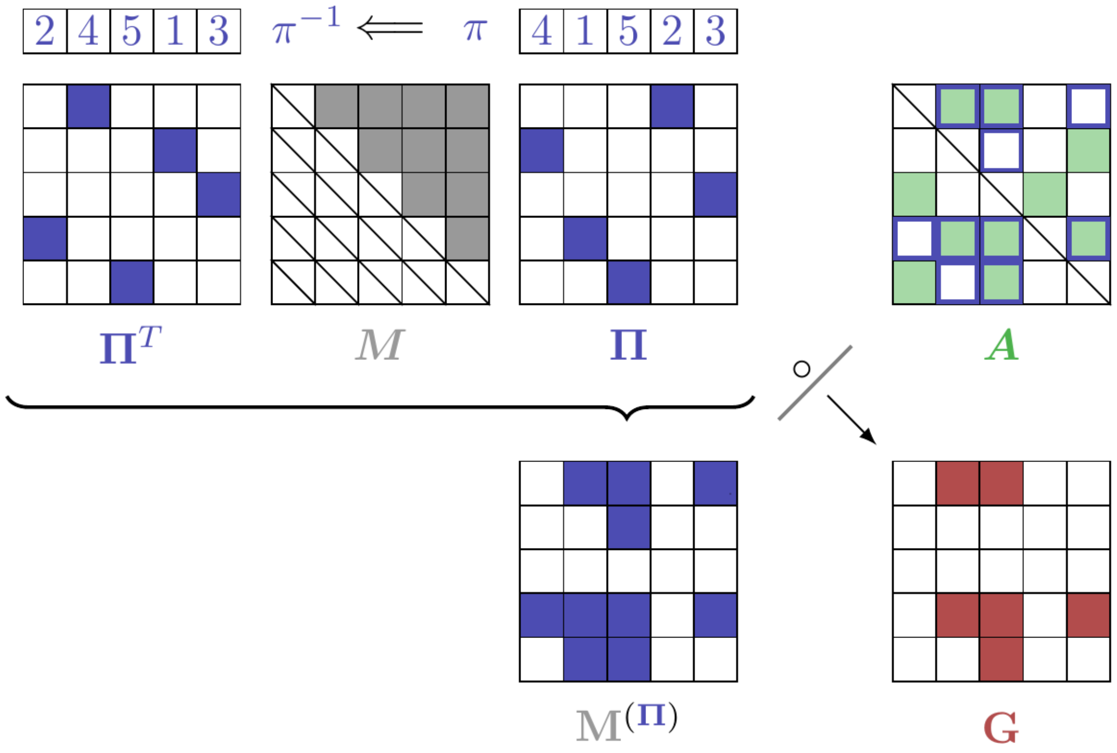
2) $\mathbf{U} \sim p_{\phi}(\boldsymbol{U})$ Gumbel-Softmax

3) $\mathbf{G} = \mathbf{U}^{(\boldsymbol{\Pi})} = \boldsymbol{\Pi}^T \mathbf{U}\, \boldsymbol{\Pi}$

$$p_{\psi,\phi}(\boldsymbol{G},\boldsymbol{\Pi}) = p_{\psi}(\boldsymbol{\Pi}) \prod_{i \neq j} \boldsymbol{p}_{\boldsymbol{\phi}_{ij}}\left(\mathbf{U}_{ij}^{(\boldsymbol{\Pi})}\right)\left[\mathbf{U}_{ij}^{(\boldsymbol{\Pi})} = \mathbf{G}_{\mathrm{ij}}\right]$$

[3] Charpentier et al. , 'Differentiable DAG sampling', in in Proceedings of the *International Conference of Learning Representations*, (2022)

Differentiable Probabilistically Masked DAG (*DPM-DAG*) [6]



1) $\mathbf{\Pi} \sim p_{\boldsymbol{\psi}}(\boldsymbol{\Pi})$ Gumbel-Softsort

2) $\mathbf{M}^{(\mathbf{\Pi})} = \mathbf{\Pi}^T M \, \mathbf{\Pi}$

3) $\mathbf{A} \sim p_{\boldsymbol{\phi}}(\boldsymbol{A})$ Gumbel-Softmax

4) $\mathbf{G} = \mathbf{M}^{(\mathbf{\Pi})} \circ \mathbf{A}$

$$p_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{G},\boldsymbol{\Pi}) = p_{\boldsymbol{\psi}}(\mathbf{\Pi}) \prod_{i \prec j \text{ in } \mathbf{\Pi}} p_{\boldsymbol{\phi}}(\boldsymbol{A}_{ij} = \boldsymbol{G}_{ij}) \prod_{j \prec i \text{ in } \mathbf{\Pi}} [0 = \boldsymbol{G}_{ij}]$$

[6] Rittel and Tschiatschek, 'Specifying Prior Beliefs over DAGs in Deep Bayesian Causal Structure Learning', in Proceedings of the European Conference on Artificial Intelligence, (2023)

Prior specification

- Gumbel-SoftSort is equal in distribution to the Plackett-Luce distribution

$$\arg\max_{i\in\mathbb{I}\setminus\mathbb{S}}(\psi_i + g_i) \sim p\left(\frac{\exp(\psi_i)}{\sum_{j\in\mathbb{I}\setminus\mathbb{S}}\exp(\psi_j)}\right) \implies p^{(\mathrm{PL})}(i \prec j) = p(\mathbf{M}_{ij}^{(\mathbf{\Pi})} = 1) = \frac{w_i}{w_i + w_j}$$

- Prior over permutation

$$D_{\mathrm{KL}}\left(p_{\boldsymbol{\psi}}(\mathbf{\Pi})\middle|p_{\boldsymbol{\omega}}(\mathbf{\Pi})\right) \approx \sum_{i}^{D} w_i(\log w_i - \log \omega_i)$$

- Prior over unmasked part of *A*

$$D_{\mathrm{KL}}\left(p_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{G}|\mathbf{\Pi})\middle|p_{\boldsymbol{\gamma}}(\boldsymbol{G}|\mathbf{\Pi})\right) = \sum_{\mathbf{\Pi}}\sum_{i\prec j \text{ in } \mathbf{\Pi}} a_{ij}\frac{\log a_{ij}}{\log \gamma_{ij}} + (1 - a_{ij})\frac{\log\left(1 - a_{ij}\right)}{\log\left(1 - \gamma_{ij}\right)}$$

Variational Loss for Bayesian CSL

- Maximizing evidence lower bound (*ELBO*) $\max\limits_{\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\Theta}} \mathcal{L}$

- DP-DAG $\mathcal{L} = \mathbb{E}_{\mathbf{G}\sim \mathrm{p}_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{G})}[\log p_{\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{G})] - \beta \underbrace{D_{\mathrm{KL}}\big(p_{\boldsymbol{\phi}}(\boldsymbol{A})\big|\big(p_{\boldsymbol{\phi}}(\mathbf{A})\big|p(\mathbf{A})\big)\big)}_{\prod_{i\neq j} D_{\mathrm{KL}}\left(p_{\boldsymbol{\phi_{ij}}}\left(\boldsymbol{A}_{ij}\right)\middle|p\right)}$

- DPM-DAG $\mathcal{L} = \mathbb{E}_{\mathbf{G}\sim \mathrm{p}_{\boldsymbol{\psi},\boldsymbol{\phi}}(\boldsymbol{G})}[\log p_{\boldsymbol{\Theta}}(\boldsymbol{X}|\boldsymbol{G})] - D_{\mathrm{KL}}\big(p_{\boldsymbol{\phi}}(\boldsymbol{G}|\boldsymbol{\Pi})\big|p_{\boldsymbol{\gamma}}(\boldsymbol{G}|\boldsymbol{\Pi})\big) - D_{\mathrm{KL}}\big(p_{\boldsymbol{\psi}}(\boldsymbol{\Pi})\big|p_{\boldsymbol{\omega}}(\boldsymbol{\Pi})\big)$

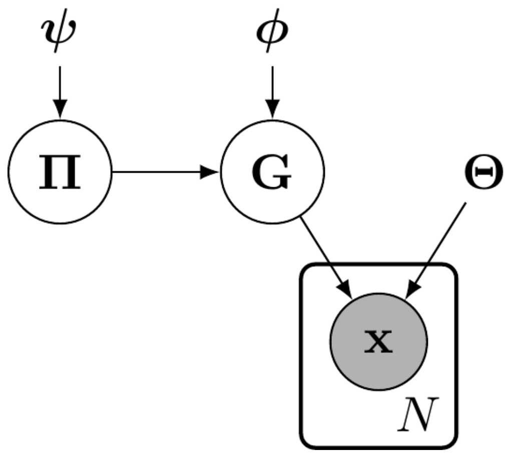
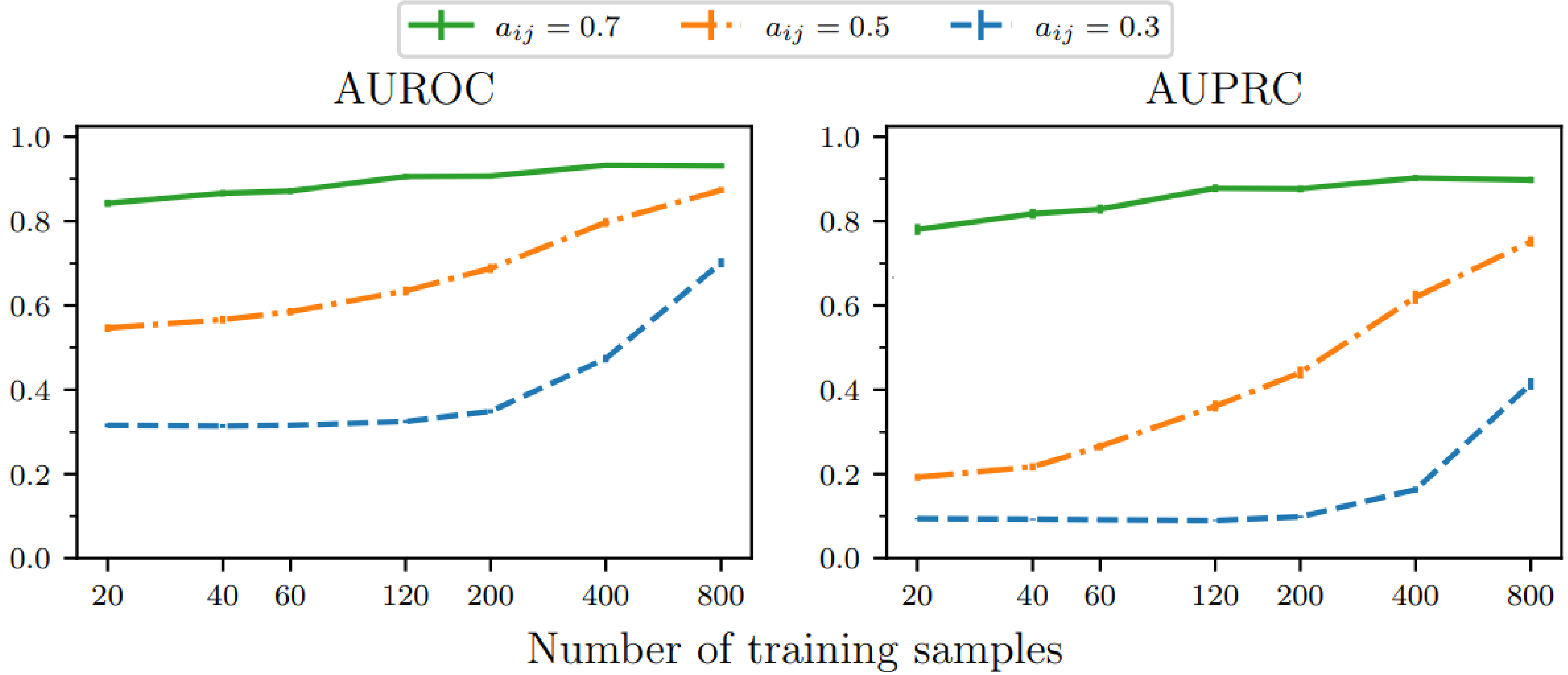


Fig. 3: Generative model of DPM-DAG

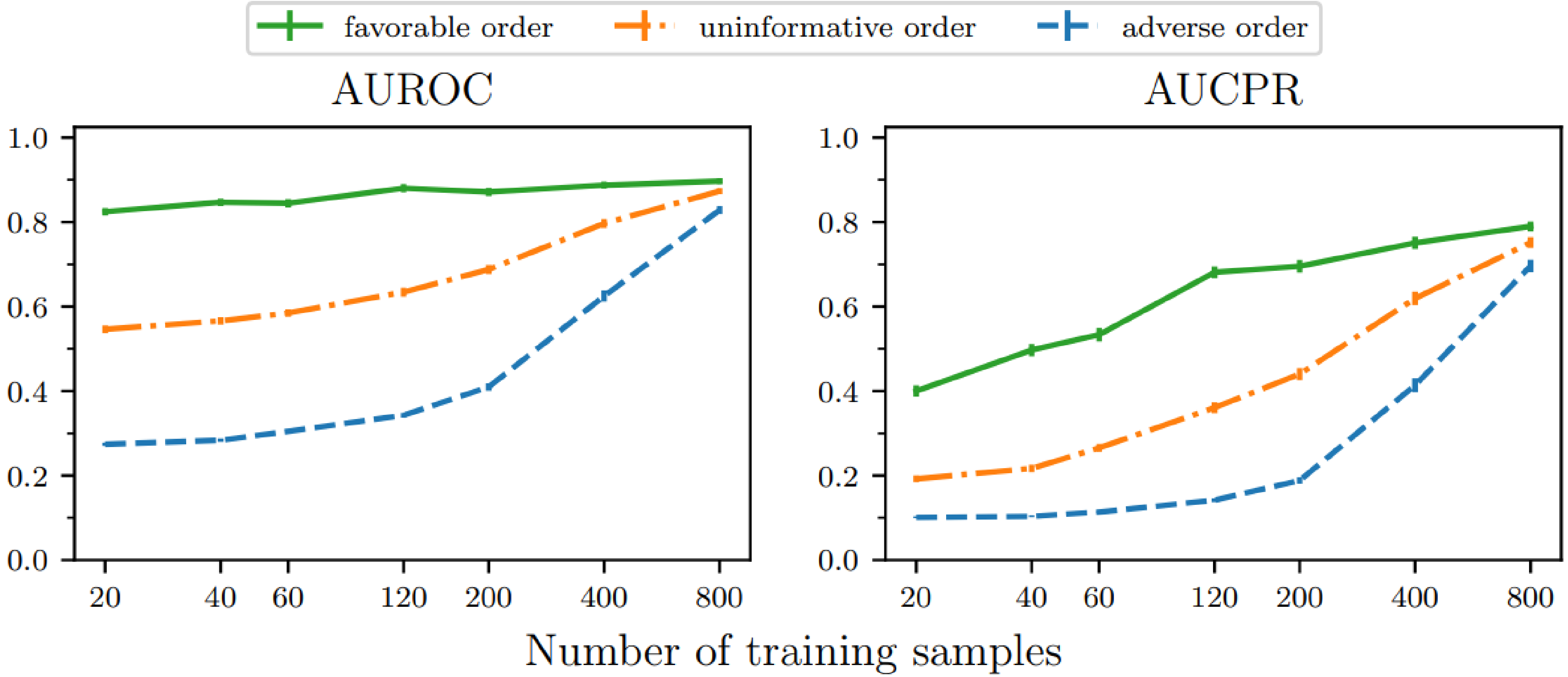
Influence of the prior over unmasked edges p_γ on AUROC (↑) & AUCPR (↑)



Probabilistic knowledge of true causal graph $\boldsymbol{G}^*$

- For $G_{ij}^* = 1$:
  $p(A_{ij} = 1) \coloneqq a_{ij}$
- For $G_{ij}^* = 0$:
  $p(A_{ij} = 1) \coloneqq 1 - a_{ij}$

Influence of the prior over the order p_ω on AUROC (↑) & AUCPR (↑)



- Favorable order: Decreasing permutation weights $\{\boldsymbol{w}_i\}_1^D$ according to a total order admitting $\boldsymbol{G}^*$
- Uninformative order: same permutation weight $\boldsymbol{w}_i$ for each $\boldsymbol{X}_i$
- Adverse order: reversed favorable order

Conclusion

- Introduction to CSL and Bayesian models for it
- Probability distribution over DAGs that enables differentiable sampling (DPM-DAG)
- Edge-wise priors in Bayesian CSL can speed up convergence w.r.t sample efficiency
- Using DPM-DAG for both models allows to reuse the posterior as the next prior

universität wien

Digitize!
Computational Social Sciences
Bundesministerium Bildung, Wissenschaft und Forschung

Thank you very much for your Attention & Interest

Invited Talk on Bayesian Causal Structure Learning