BaCaDI: Bayesian Causal Discovery with Unknown Interventions

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath,
Bernhard Schölkopf, Andreas Krause

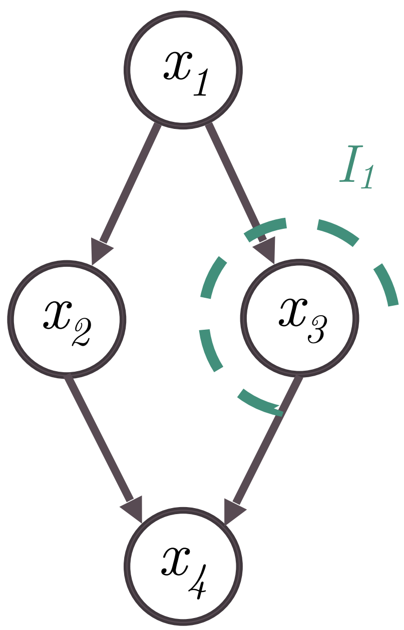





ETH zürich

Learning &
Adaptive Systems

Toy example: Causal Bayesian Networks



e.g. hard intervention

- **Factorized likelihood**

$$p(x_1, \ldots, x_4 \mid \mathbf{G}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1) \cdot p(x_4 \mid x_2, x_3)$$

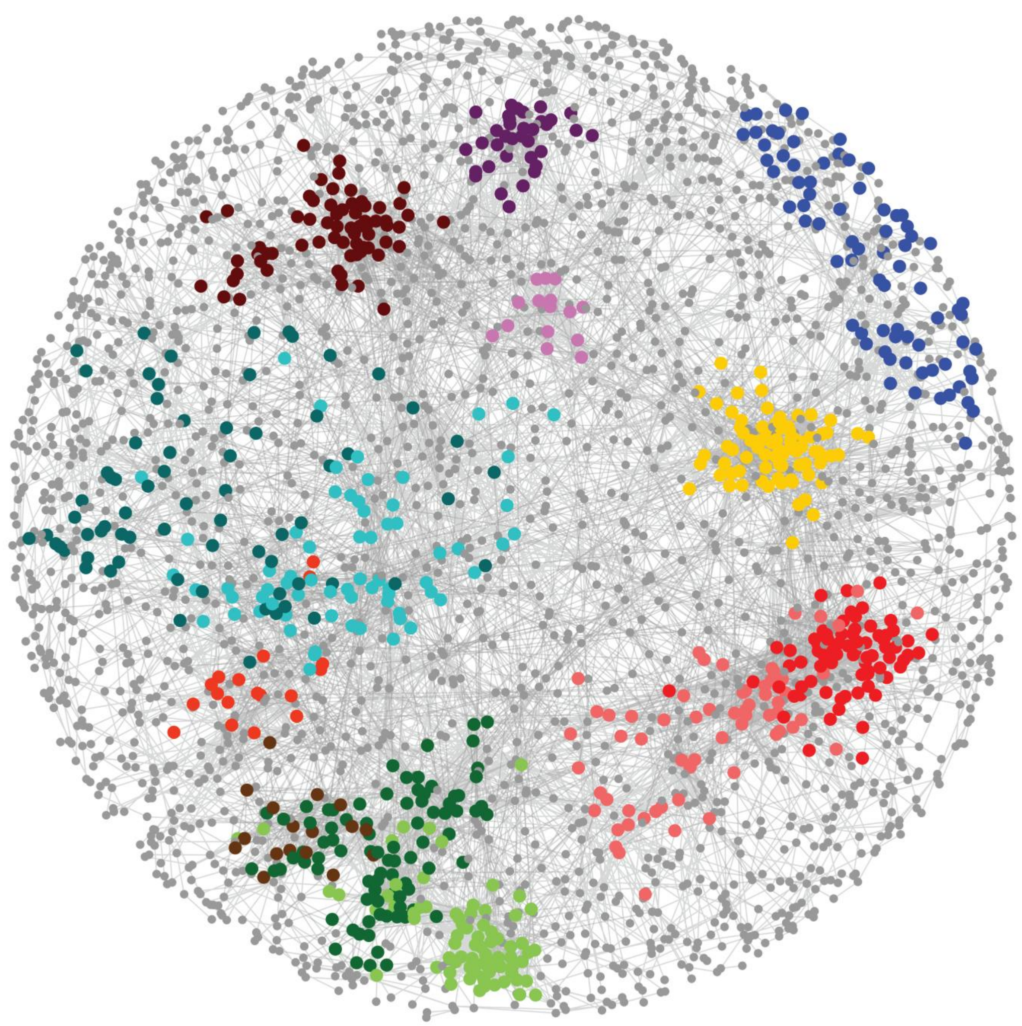
- **... with local switching after intervention**

$$p(x_1, \ldots, x_4 \mid \mathbf{G}, I_1) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p^I(x_3) \cdot p(x_4 \mid x_2, x_3)$$

- Can try and learn graph from data — interventions help!

- **... but what if interventions (targets and/or distribution) are unknown?**

Learning causal mechanisms from experimentation & interventions



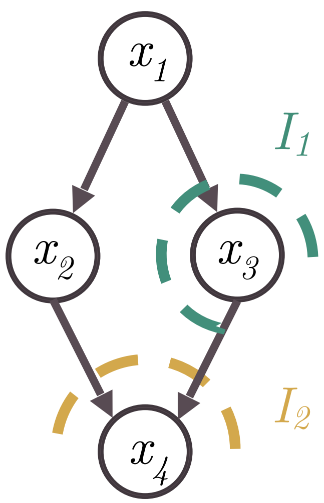
Global genetic landscape of the cell

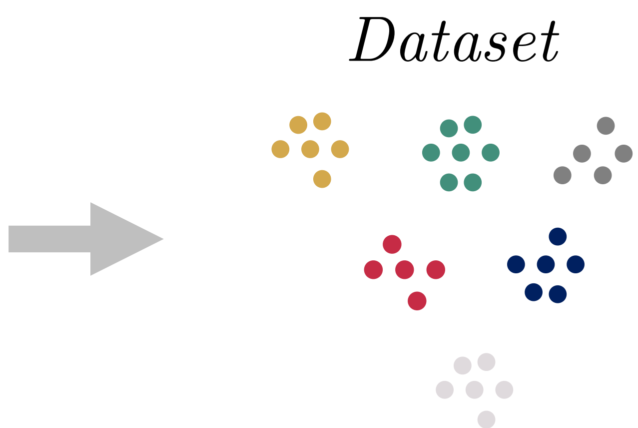
- Collect datasets from *same* underlying causal system under multiple interventions
- Biology: single-cell expression with interventions (drug candidates, gene knockouts)
- *Costly interventions & uncertain (off-target effects, ...)*

Image credit: Raamesh Deshpande

Causal Discovery from multiple contexts

Observe samples from same underlying causal graph (DAG) across different contexts





Context	x_1	x_2	x_3	x_4
●	-1.45	0.89	1.01	0.40
●	-1.21	0.68	0.98	0.36
●	-1.34	0.84	-1.24	-2.22
...	...	...	...	...

$$\mathbf{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$$

$$\mathcal{D}_k = \{\mathbf{x}^{(k,1)}, \ldots, \mathbf{x}^{(k,n_k)}\}$$

with corresponding unknown interventions

$$\mathcal{I} = \{I_1, \ldots, I_M\}$$

Goal: learn $\mathbf{G}, \boldsymbol{\Theta}, \mathcal{I}$ and rigorously account for *uncertainty*

Bayesian Causal Discovery from multiple contexts

- **Bayesian Inference of CBNs:**

$$p(\mathbf{G}, \boldsymbol{\Theta} \mid \mathcal{D}) \propto p(\mathbf{G}) p(\boldsymbol{\Theta} \mid \mathbf{G}) p(\mathcal{D} \mid \mathbf{G}, \boldsymbol{\Theta}) \longrightarrow \mathbb{E}_{p(\mathbf{G}, \boldsymbol{\Theta} \mid \mathcal{D})}[f(\mathbf{G}, \boldsymbol{\Theta})]$$

posterior *prior terms* *likelihood*

- **Known interventions?**

insert $\mathcal{I}$

- **Unknown interventions?**

$$p(\mathbf{G}, \boldsymbol{\Theta}, \mathcal{I} \mid \mathbf{D}) \propto p(\mathbf{G}) p(\boldsymbol{\Theta} \mid \mathbf{G}) \prod_{k=1}^{M} p\big(I_k^{\text{tar}}\big) p\big(\boldsymbol{\Theta}_{I_k} \mid I_k^{\text{tar}}\big) p(\mathcal{D}_k \mid \mathbf{G}, \boldsymbol{\Theta}, I_k)$$

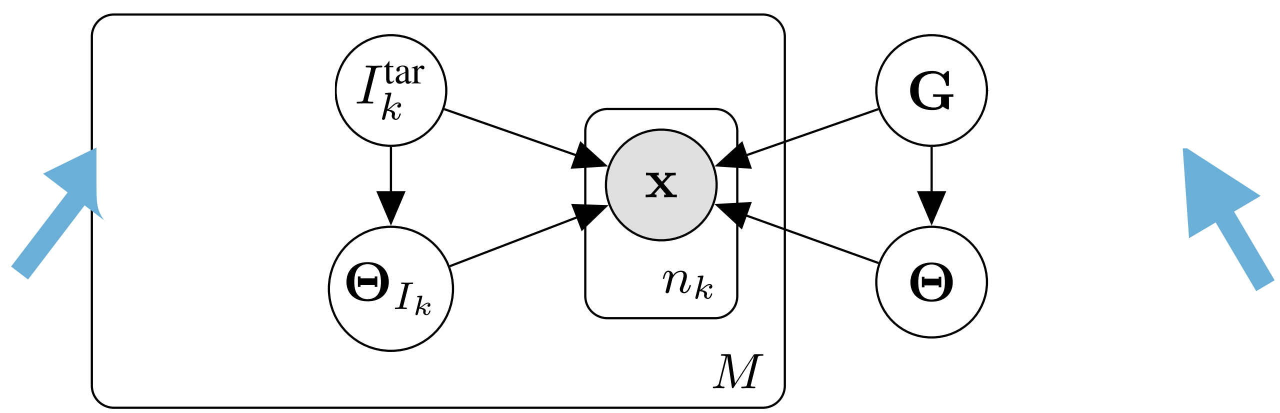
posterior *graph prior* *interv. prior* *interv. likelihood*

- **Intractable!**

BaCaDI: A Differentiable Generative Model over CBNs and Interventions

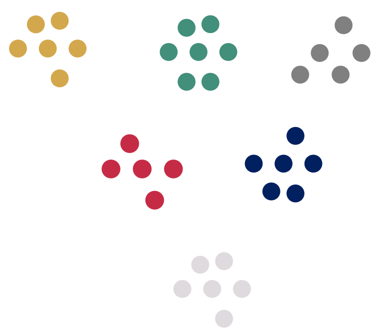
Without loss of generality, use latent variables

$$p(\mathbf{G}, \mathbf{\Theta}, \mathcal{I} \mid \mathbf{D}) \longrightarrow p(\underline{\mathbf{Z}}, \mathbf{\Theta}, \underline{\mathbf{\Gamma}}, \mathbf{\Theta}_I \mid \mathbf{D})$$



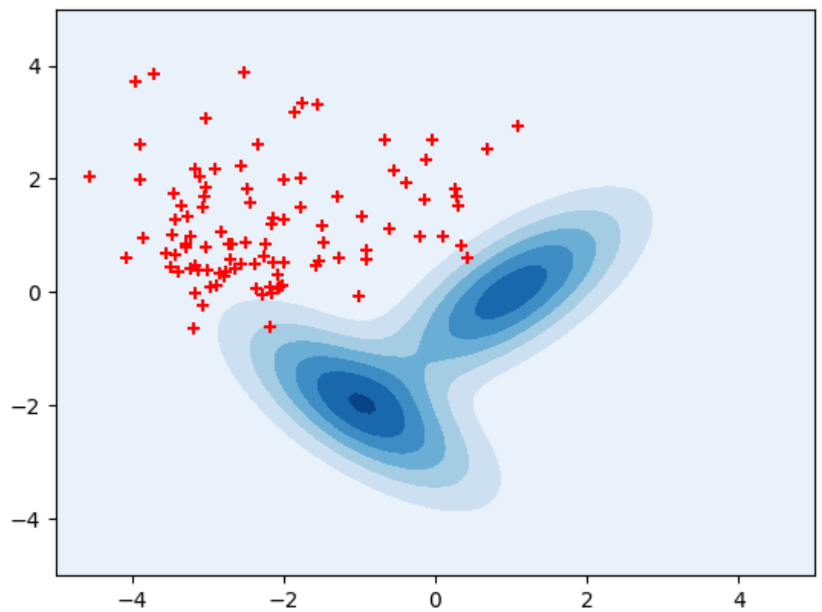
[*DiBS*, Lorch et al. 2021]

Joint Variational Inference using SVGD



Learn posterior $\rightarrow$ $p(\mathbf{Z}, \mathbf{\Theta}, \mathbf{\Gamma}, \mathbf{\Theta}_I \mid \mathbf{D})$

Implementation:
Use score and apply
particle transform **SVGD**



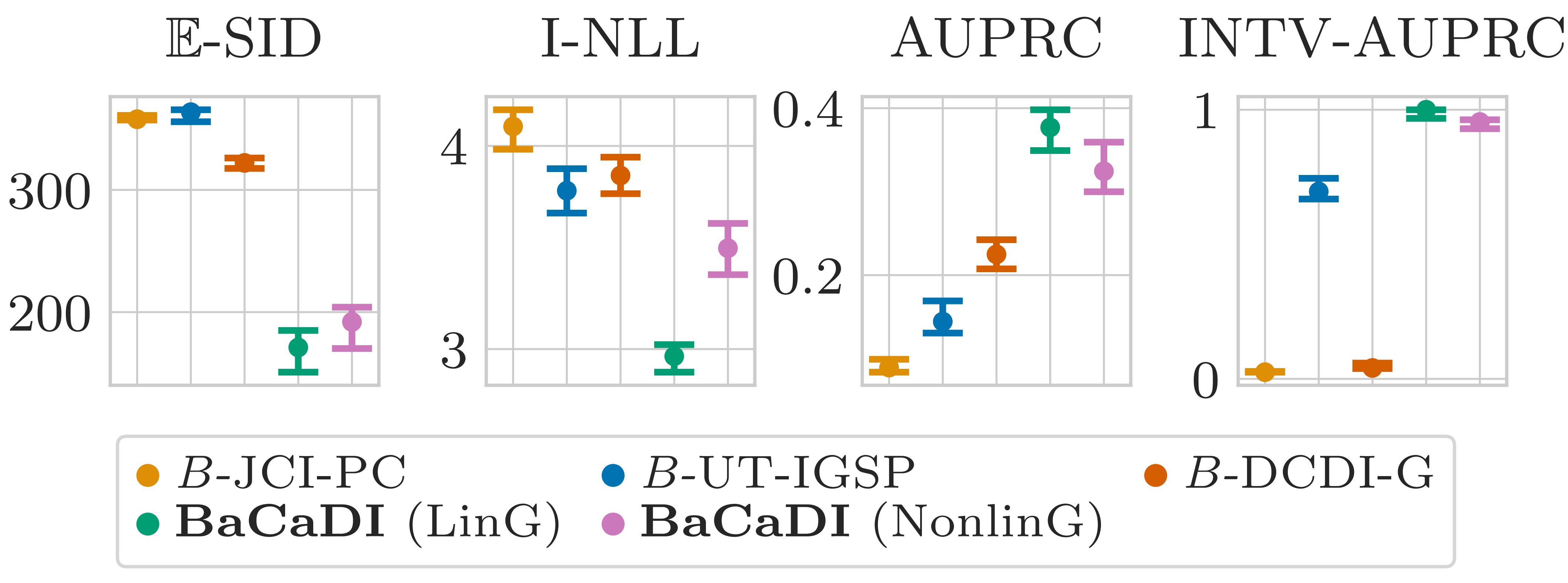
- ***Note:* joint inference**
  - Inference of the interventions and causal Bayesian network is interdependent
  - Estimate epistemic uncertainty jointly

[Liu and Wang 2016]

Experiments: SERGIO

SERGIO: a single-cell expression simulator guided by gene regulatory networks

Interventions: gene knockouts



[Dibaeinia and Sinha 2020]

References

- Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data, 2020.
- Payam Dibaeinia and Saurabh Sinha. SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.
- Nir Friedman and Daphne Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. Machine Learning, 50(1):95–125, January 2003.
- Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In Advances in Neural Information Processing, 2016.
- Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian Structure Learning. Advances in Neural Information Processing Systems, 34:24111–24123, 2021.
- Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint Causal Inference from Multiple Contexts. 2016. doi: 10.48550/ARXIV.1611.10351. URL https://arxiv.org/abs/1611.10351.
- Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. Neural computation, 27(3):771–799, 2015.
- Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-Based Causal Structure Learning with Unknown Intervention Targets, 2020.

Thank you!

- Paper: https://arxiv.org/pdf/2206.01665.pdf
- Contact: ahaegele@ethz.ch
- Code: https://github.com/haeggee/bacadi

BaCaDI: Bayesian Causal Discovery with Unknown Interventions

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath,
Bernhard Schölkopf, Andreas Krause

Appendix

BaCaDI: Bayesian Causal Discovery with Unknown Interventions

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath,
Bernhard Schölkopf, Andreas Krause

ETH zürich

Learning &
Adaptive Systems

Experimental setup

- Sample random graphs (ER-2, SF-2) with 20 nodes and random parameters for local conditionals
- Random (hard) interventions on one target, bounded away from 0 with small noise
- 100 observational samples, 10 samples per interventional context
- Baselines — bootstrapped versions of:

  JCI-PC [Mooij et al. 2016]

  UT-IGSP [Squires et al. 2020]

  DCDI-G [Brouillard et al. 2020]
- Metrics: Expected SID [Peters et al. 2015], held-out interventional NLL,
  AUPRC for edge prediction, AUPRC for intervention targets

Proposition 1

Proposition 1 *Under the extended generative model in Eq. 5 and Figure 1, it holds that*

$$\mathbb{E}_{p(\mathbf{G},\mathbf{\Theta},\mathcal{I}|\mathbf{D})}[f(\mathbf{G},\mathbf{\Theta},\mathcal{I})] = \tag{6}$$
$$\mathbb{E}_{p(\mathbf{Z},\mathbf{\Theta},\mathbf{\Gamma},\mathbf{\Theta}_\mathcal{I}|\mathbf{D})}\left[\frac{\mathbb{E}_{p(\mathbf{G}|\mathbf{Z})}\mathbb{E}_{p(\mathcal{I}^{\text{tar}}|\mathbf{\Gamma})}[f(\mathbf{G},\mathbf{\Theta},\mathcal{I})\cdot\mathbf{\Psi}]}{\mathbb{E}_{p(\mathbf{G}|\mathbf{Z})}\mathbb{E}_{p(\mathcal{I}^{\text{tar}}|\mathbf{\Gamma})}[\mathbf{\Psi}]}\right]$$

with weighting $\mathbf{\Psi} = p(\mathbf{\Theta}|\mathbf{G})p(\mathbf{\Theta}_\mathcal{I}|\mathcal{I}^{\text{tar}})p(\mathbf{D}|\mathbf{G},\mathcal{I},\mathbf{\Theta})$ *and* $p(\mathbf{D}|\mathbf{G},\mathcal{I},\mathbf{\Theta}) = \prod_{k=1}^{M} p(\mathcal{D}_k|\mathbf{G},\mathbf{\Theta},I_k^{\text{tar}},\mathbf{\Theta}_{I_k})$.

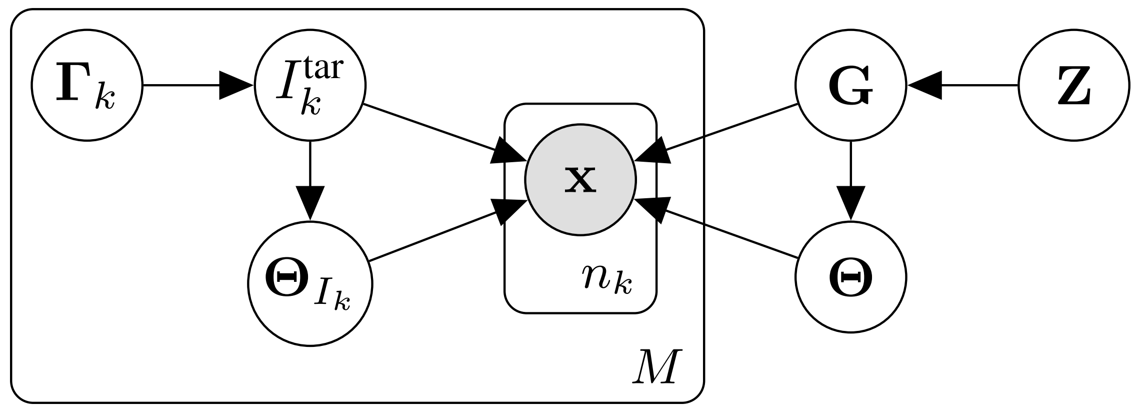
Proposition 2

Proposition 2 *As $\alpha \to \infty$ and $\beta \to \infty$ the posterior expectation in Prop. 1 converges to the simpler expression*

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{G},\mathbf{\Theta},\mathcal{I}^{\text{tar}},\mathbf{\Theta}_{\mathcal{I}}|\mathbf{D})}[f(\mathbf{G},\mathbf{\Theta},\mathcal{I}^{\text{tar}},\mathbf{\Theta}_{\mathcal{I}})] \\ \to\ &\mathbb{E}_{p(\mathbf{Z},\mathbf{\Theta},\mathbf{\Gamma},\mathbf{\Theta}_{\mathcal{I}}|\mathbf{D})}[f(\mathbf{G}_\infty(\mathbf{Z}),\mathbf{\Theta},\mathcal{I}^{\text{tar}}_\infty(\mathbf{\Gamma}),\mathbf{\Theta}_{\mathcal{I}})] \end{aligned} \tag{13}$$

with $\mathbf{G}_\infty(\mathbf{Z})_{i,j} = \mathbf{1}_{\mathbf{u}_i^\top \mathbf{v}_j > 0}$ *and* $\mathcal{I}^{\text{tar}}_\infty(\mathbf{\Gamma})_{k,i} = \mathbf{1}_{\gamma_{i,k} > 0}$. *In the limit, the marginal posterior over discrete structures* $p(\mathbf{G},\mathcal{I}^{\text{tar}}|\mathbf{D})$ *is supported on* $\{\mathbf{G} | \mathbf{G} \in \{0,1\}^{d\times d} \wedge$ $\mathbf{G}$ *is acyclic*$\} \times \{0,1\}^{M\times d}$ *and, thus, a valid probability mass function over DAGs and intervention targets.*

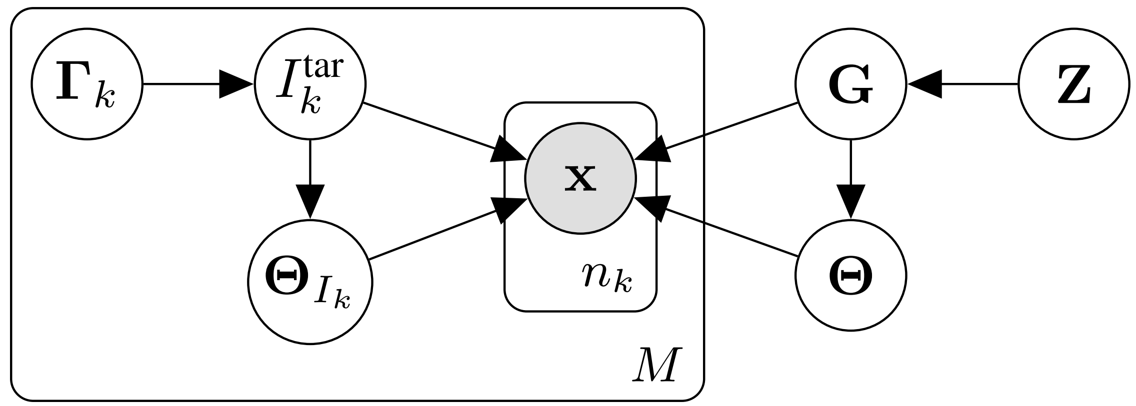
Full factorization



$$p(\mathbf{Z}, \mathbf{G}, \boldsymbol{\Theta}, \boldsymbol{\Gamma}, \mathcal{I}, \mathbf{D}) = \underbrace{p(\mathbf{Z})p(\mathbf{G}|\mathbf{Z})p(\boldsymbol{\Theta}|\mathbf{G})}_{\text{generative process CBN}} \tag{5}$$

$$\cdot \prod_{k=1}^{M} \underbrace{p(\boldsymbol{\Gamma}_k)p(I_k^{\text{tar}}|\boldsymbol{\Gamma}_k)p(\boldsymbol{\Theta}_{I_k}|I_k^{\text{tar}})}_{\text{generative process intervention}} \underbrace{p(\mathcal{D}_k|\mathbf{G}, \boldsymbol{\Theta}, I_k^{\text{tar}}, \boldsymbol{\Theta}_{I_k})}_{\text{interventional likelihood}}$$

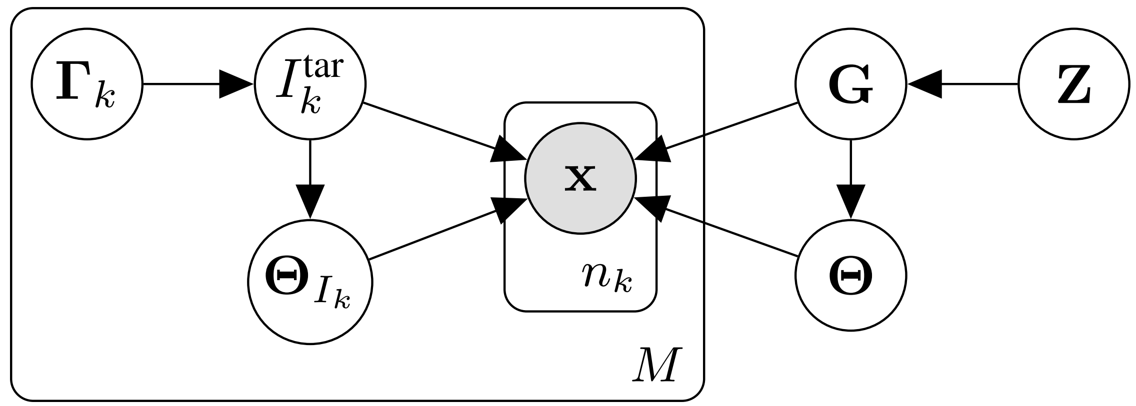
Generative model of graph



$$p_\alpha(\mathbf{G}|\mathbf{Z}) = \prod_{i=1}^{d}\prod_{j\neq i}^{d} p_\alpha(g_{ij}|\mathbf{u}_i, \mathbf{v}_j) \quad \text{with} \quad g_{ij}|\mathbf{u}_i, \mathbf{v}_j \sim \text{Bern}(\sigma_\alpha(\mathbf{u}_i^\top \mathbf{v}_j)) \tag{7}$$

$$p_\beta(\mathbf{Z}) = p(\mathbf{U}, \mathbf{V}) \propto \underbrace{\exp\left(-\beta \mathbb{E}_{p(\mathbf{G}|\mathbf{Z})}[h(\mathbf{G})]\right)}_{\text{acyclicity prior}} \cdot \prod_{i=1}^{d} \underbrace{\mathcal{N}(\mathbf{u}_i|\mathbf{0}, \eta_Z^2\mathbf{I})\mathcal{N}(\mathbf{v}_i|\mathbf{0}, \eta_Z^2\mathbf{I})}_{\text{inference stability}} \tag{8}$$

Generative model of interventions



$$p(\mathcal{I}^{\text{tar}}|\mathbf{\Gamma}) = \prod_{k=1}^{M}\prod_{i=1}^{d} p_\alpha(I_{k,i}^{\text{tar}}|\gamma_{k,i}) \quad \text{with} \quad I_{k,i}^{\text{tar}}|\gamma_{k,i} \sim \text{Bern}(\sigma_\alpha(\gamma_{k,i})) \tag{9}$$

$$p(\mathbf{\Gamma}) \propto \prod_{k=1}^{M} \underbrace{\exp\left(-\lambda\|\sigma_\alpha(\mathbf{\Gamma}_k)\|_1\right)}_{\text{sparse masks}} \cdot \prod_{i=1}^{d} \underbrace{\text{Beta}(\sigma_\alpha(\gamma_{k,i});\zeta_1,\zeta_2)}_{\text{sharp masks}} \underbrace{\mathcal{N}(\gamma_k|\mathbf{0},\eta_\gamma^2\mathbf{I})}_{\text{inference stability}} \tag{10}$$

Log posterior score

$$
\begin{aligned}
&\nabla_{\boldsymbol{\Gamma}} \log p(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Theta}_{\mathcal{I}} | \mathbf{D}) \\
&= \nabla_{\boldsymbol{\Gamma}} \log p(\boldsymbol{\Gamma}) + \frac{\nabla_{\boldsymbol{\Gamma}} \mathbb{E}_{p(\mathbf{G}|\mathbf{Z})} \mathbb{E}_{p(\mathcal{I}^{\text{tar}}|\boldsymbol{\Gamma})} \left[\boldsymbol{\Psi} \right]}{\mathbb{E}_{p(\mathbf{G}|\mathbf{Z})} \mathbb{E}_{p(\mathcal{I}^{\text{tar}}|\boldsymbol{\Gamma})} \left[\boldsymbol{\Psi} \right]}
\end{aligned}
$$

with weighting $\boldsymbol{\Psi} = p(\boldsymbol{\Theta}|\mathbf{G}) p(\boldsymbol{\Theta}_{\mathcal{I}}|\mathcal{I}^{\text{tar}}) p(\mathbf{D}|\mathbf{G}, \mathcal{I}, \boldsymbol{\Theta})$ *and* $p(\mathbf{D}|\mathbf{G}, \mathcal{I}, \boldsymbol{\Theta}) = \prod_{k=1}^{M} p(\mathcal{D}_k|\mathbf{G}, \boldsymbol{\Theta}, I_k^{\text{tar}}, \boldsymbol{\Theta}_{I_k})$.