Discrete Causal Representations from Heterogeneous Domains

A Bayesian Approach with Social Survey Applications

Ankur Garg

Department of Statistics, University of Chicago

2026 ISBA World Meeting

Michael Stettler
University of Tübingen

Aaron Schein
University of Chicago

Julius von Kügelgen
ETH Zürich

- Randomized controlled trials are the gold standard for recovering causal structure.

[1] Schölkopf et al. *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021.

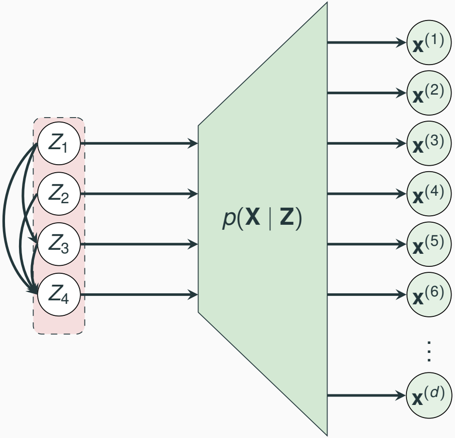
WHAT IS CAUSAL REPRESENTATION LEARNING?

- Randomized controlled trials are the gold standard for recovering causal structure.
- Interventions may be impossible, unethical, or impractical.
- Variables of interest may not be directly measurable.

[1] Schölkopf et al. *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021.

WHAT IS CAUSAL REPRESENTATION LEARNING?

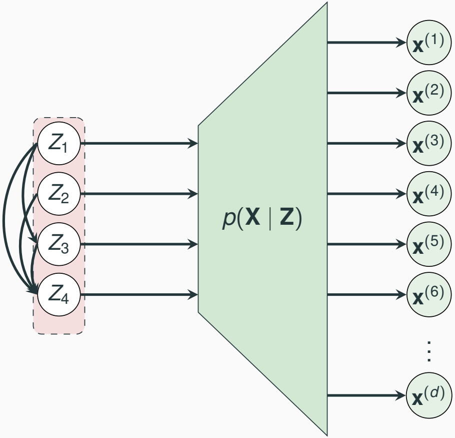
- Randomized controlled trials are the gold standard for recovering causal structure.
- Interventions may be impossible, unethical, or impractical.
- Variables of interest may not be directly measurable.



We observe only $\mathbf{x}$

[1] Schölkopf et al. *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021.

- Randomized controlled trials are the gold standard for recovering causal structure.
- Interventions may be impossible, unethical, or impractical.
- Variables of interest may not be directly measurable.
- CRL[1] aims to infer:
  - Latent Causal Variables **Z**
  - Latent Causal Graph

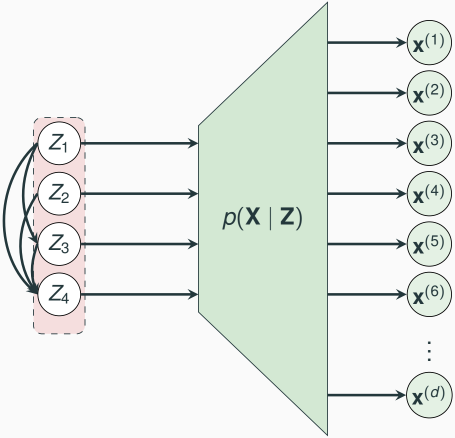


We observe only **x**

[1] Schölkopf et al. *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021.

Independent Causal Mechanisms[2]

$$p(\mathbf{z}) = p(z_1, \ldots, z_L) = \prod_{\ell=1}^{L} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$



We observe only $\mathbf{x}$

[2] Peters et al. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

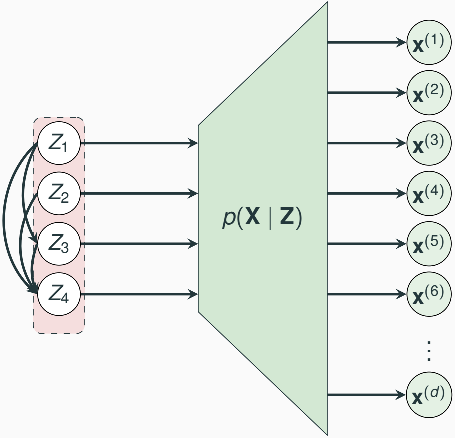
Independent Causal Mechanisms[2]

$$p(\mathbf{z}) = p(z_1, \ldots, z_L) = \prod_{\ell=1}^{L} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

Interventions replace mechanisms

$$\widetilde{p}_{\mathcal{I}}(\mathbf{z}) = \prod_{\ell \in \mathcal{I}} \widetilde{p}_\ell(z_\ell \mid \mathbf{pa}_\ell) \prod_{\ell \notin \mathcal{I}} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

What makes the model causal is that it also describes a family of *interventional* distributions



We observe only $\mathbf{x}$

[2] Peters et al. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, 2017.

Independent Causal Mechanisms[2]

$$p(\mathbf{z}) = p(z_1, \ldots, z_L) = \prod_{\ell=1}^{L} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

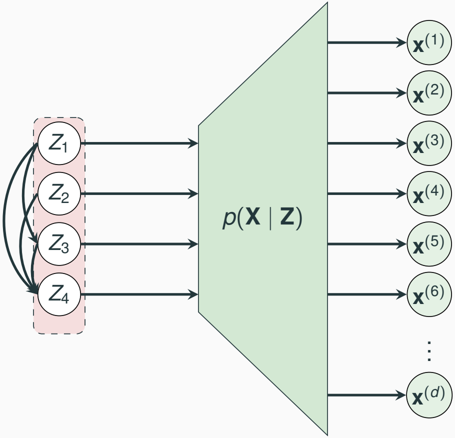
Interventions replace mechanisms

$$\widetilde{p}_{\mathcal{I}}(\mathbf{z}) = \prod_{\ell \in \mathcal{I}} \widetilde{p}_\ell(z_\ell \mid \mathbf{pa}_\ell) \prod_{\ell \notin \mathcal{I}} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

What makes the model causal is that it also describes a family of *interventional* distributions

Marginal observational distribution

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})$$

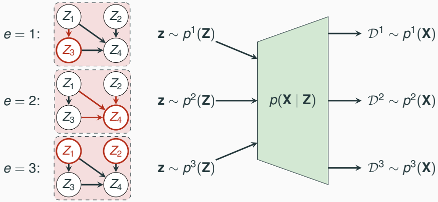


We observe only $\mathbf{x}$

[2] Peters et al. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

MULTI-DOMAIN DATA ASSUMPTIONS

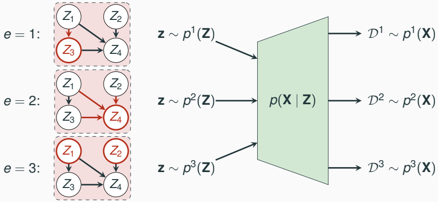
- We observe datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}}$.
- $\mathcal{D}^e$ is a set of i.i.d. samples from a domain-specific distribution $p^e(\mathbf{X})$



$$\mathcal{D}^e = \{\mathbf{x}^{e,1}, \ldots, \mathbf{x}^{e,N_e}\} \overset{\text{i.i.d.}}{\sim} p^e(\mathbf{X})$$

MULTI-DOMAIN DATA ASSUMPTIONS

- We observe datasets $\{\mathcal{D}^e\}_{e \in \mathcal{E}}$.
- $\mathcal{D}^e$ is a set of i.i.d. samples from a domain-specific distribution $p^e(\mathbf{X})$



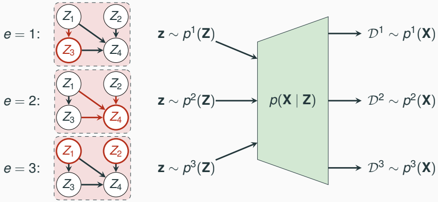
$$\mathcal{D}^e = \{\mathbf{x}^{e,1}, \ldots, \mathbf{x}^{e,N_e}\} \overset{\text{i.i.d.}}{\sim} p^e(\mathbf{X})$$

$\mathbf{X}$: survey responses $\quad$ $\mathbf{Z}$: values/opinions

e: geographic region

MULTI-DOMAIN DATA ASSUMPTIONS

- We observe datasets $\{\mathcal{D}^e\}_{e \in \mathcal{E}}$.
- $\mathcal{D}^e$ is a set of i.i.d. samples from a domain-specific distribution $p^e(\mathbf{X})$



$$\mathcal{D}^e = \{\mathbf{x}^{e,1}, \ldots, \mathbf{x}^{e,N_e}\} \overset{\text{i.i.d.}}{\sim} p^e(\mathbf{X})$$

X: survey responses **Z**: values/opinions

e: geographic region

A1: Invariant measurement model

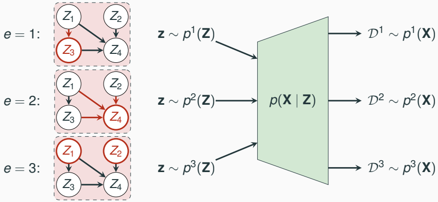
Only the latent distribution changes across domains:

$$\mathbf{z}^{e,j} \sim p^e(\mathbf{Z}), \qquad \mathbf{x}^{e,j} \sim p(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}^{e,j})$$

$$p^e(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z})\, p^e(\mathbf{z})$$

MULTI-DOMAIN DATA ASSUMPTIONS

- We observe datasets $\{\mathcal{D}^e\}_{e \in \mathcal{E}}$.
- $\mathcal{D}^e$ is a set of i.i.d. samples from a domain-specific distribution $p^e(\mathbf{X})$



$$\mathcal{D}^e = \{\mathbf{x}^{e,1}, \dots, \mathbf{x}^{e,N_e}\} \overset{\text{i.i.d.}}{\sim} p^e(\mathbf{X})$$

$\mathbf{X}$: survey responses $\quad$ $\mathbf{Z}$: values/opinions

e: geographic region

A1: Invariant measurement model

Only the latent distribution changes across domains:

$$\mathbf{z}^{e,j} \sim p^e(\mathbf{Z}), \qquad \mathbf{x}^{e,j} \sim p(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}^{e,j})$$

$$p^e(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z})\, p^e(\mathbf{z})$$

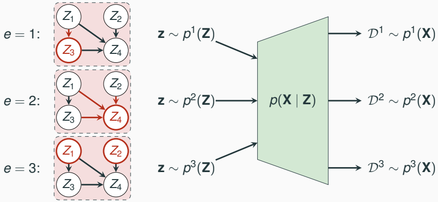
A2: Shared mechanisms[3]

Unintervened mechanisms are shared across domains.

$$p^e(\mathbf{z}) = \prod_{\ell \in \mathcal{I}^e} \widetilde{p}^e_\ell(z_\ell \mid \mathbf{pa}_\ell) \prod_{\ell \notin \mathcal{I}^e} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

MULTI-DOMAIN DATA ASSUMPTIONS

- We observe datasets $\{\mathcal{D}^e\}_{e\in\mathcal{E}}$.
- $\mathcal{D}^e$ is a set of i.i.d. samples from a domain-specific distribution $p^e(\mathbf{X})$



$$\mathcal{D}^e = \{\mathbf{x}^{e,1}, \dots, \mathbf{x}^{e,N_e}\} \overset{\text{i.i.d.}}{\sim} p^e(\mathbf{X})$$

X: survey responses **Z**: values/opinions

e: geographic region

A1: Invariant measurement model

Only the latent distribution changes across domains:

$$\mathbf{z}^{e,j} \sim p^e(\mathbf{Z}), \qquad \mathbf{x}^{e,j} \sim p(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}^{e,j})$$

$$p^e(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z})\, p^e(\mathbf{z})$$

A2: Shared mechanisms[3]

Unintervened mechanisms are shared across domains.

$$p^e(\mathbf{z}) = \prod_{\ell\in\mathcal{I}^e} \tilde{p}^e_\ell(z_\ell \mid \mathbf{pa}_\ell) \prod_{\ell\notin\mathcal{I}^e} p_\ell(z_\ell \mid \mathbf{pa}_\ell)$$

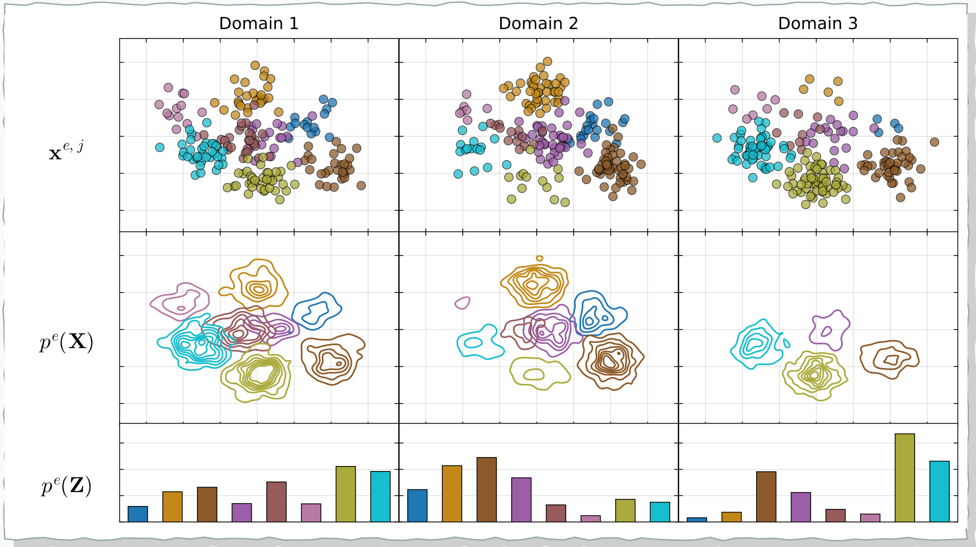
A3: Sparse Mechanism Shift[1,3]

Each domain changes only a small subset of mechanisms:

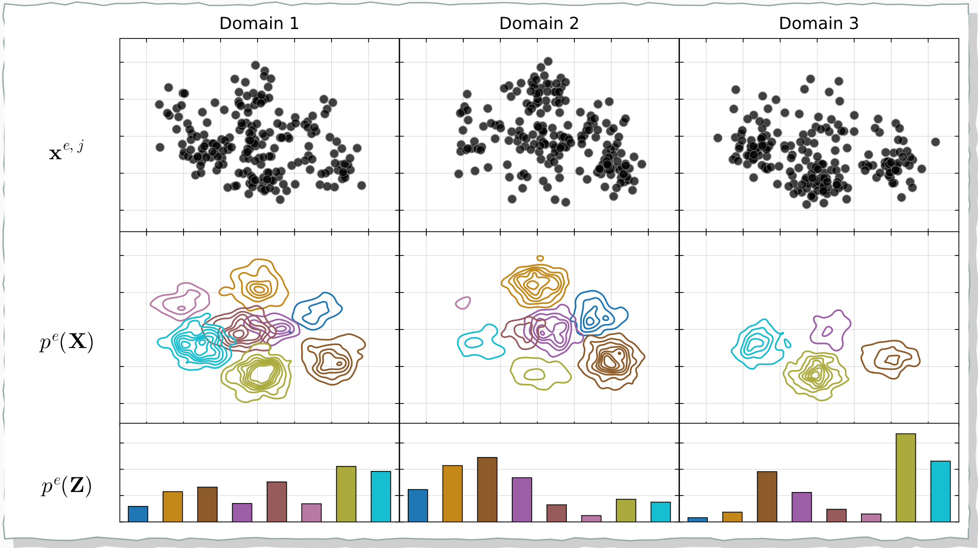
$$|\mathcal{I}^e| \ll L$$

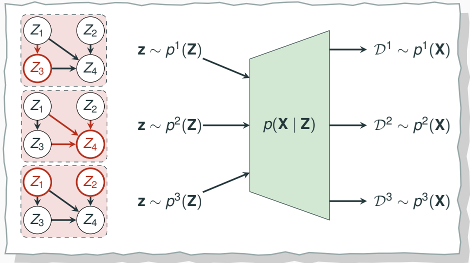
[3] Perry et al. *Causal Discovery in Heterogeneous Environments under the Sparse Mechanism Shift Hypothesis*. NeurIPS, 2022.
[1] Schölkopf et al. *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021.

MULTI-DOMAIN DATA



MULTI-DOMAIN DATA





Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_{\ell \in [L]} \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv}{-}\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms

$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Theta\right)$$

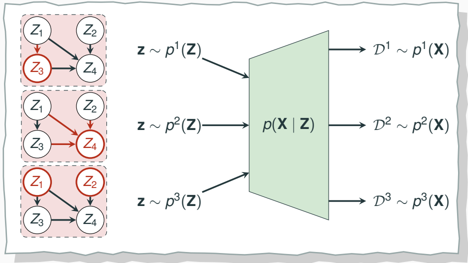
$$\boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell) \sim p\left(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Delta\right) \cdot \mathbb{1}\left\{\frac{1}{2}\boldsymbol{\delta}^\top \mathbf{H} \boldsymbol{\delta} \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\mathrm{softmax}\left(\boldsymbol{\theta}_\ell\left(\mathbf{pa}_\ell^{e,j}\right) + \mathcal{I}_\ell^e\, \boldsymbol{\delta}_\ell^e\left(\mathbf{pa}_\ell^{e,j}\right)\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\boldsymbol{\sigma}^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \text{Inv}-\Gamma(\alpha_\sigma, \beta_\sigma)$$



Latent Mechanisms

$$c_\ell \sim \text{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \text{Bernoulli}(c_\ell)$$

$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \Sigma_\Theta)$$

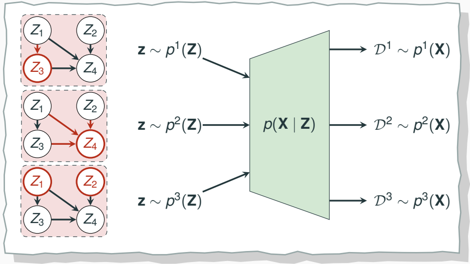
$$\delta_\ell^e(\mathbf{pa}_\ell) \sim p(\delta \mid \Sigma_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \Sigma_\Delta) \cdot \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \text{Categorical}\left(\operatorname{softmax}\left(\theta_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv}\!-\!\Gamma(\alpha_\sigma, \beta_\sigma)$$



Latent Mechanisms

$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \Sigma_\Theta)$$

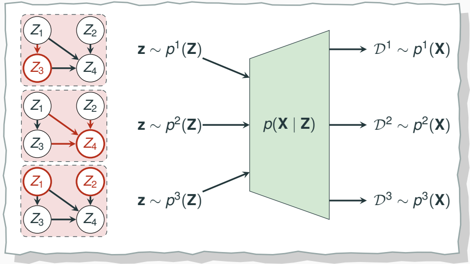
$$\delta_\ell^e(\mathbf{pa}_\ell) \sim p(\delta \mid \Sigma_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \Sigma_\Delta) \cdot \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\mathrm{softmax}\left(\theta_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\boldsymbol{\sigma}^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv\!-\!\Gamma}(\alpha_\sigma, \beta_\sigma)$$



Latent Mechanisms

$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \Sigma_\Theta)$$

$$\delta_\ell^e(\mathbf{pa}_\ell) \sim p(\delta \mid \Sigma_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \Sigma_\Delta) \cdot \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\operatorname{softmax}\left(\theta_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\sigma^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv}\!-\!\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms

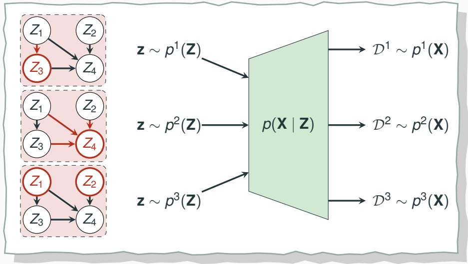
$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Theta\right)$$

$$\boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell) \sim p\left(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Delta\right) \cdot \mathbb{1}\left\{\tfrac{1}{2}\boldsymbol{\delta}^\top \mathbf{H}\boldsymbol{\delta} \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\mathrm{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

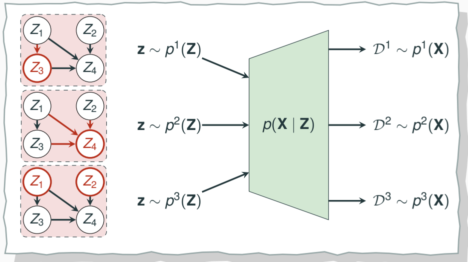


Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_{\ell \in [L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\sigma^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv\!-\!}\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms



$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e\, \delta^e(\mathbf{pa}_\ell)_k}$$

$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Theta\right)$$

$$\boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell) \sim p\left(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Delta\right) \cdot \mathbb{1}\left\{\tfrac{1}{2}\boldsymbol{\delta}^\top \mathbf{H} \boldsymbol{\delta} \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\operatorname{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell \in [L]} \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\sigma^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv}\!-\!\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms

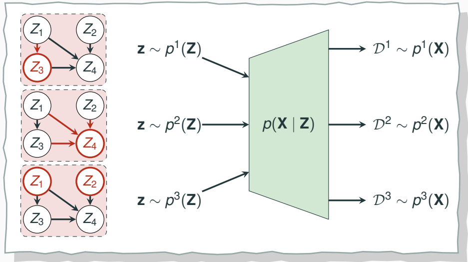
$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \Sigma_\Theta\right)$$

$$\delta_\ell^e(\mathbf{pa}_\ell) \sim p\left(\delta \mid \Sigma_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \Sigma_\Delta\right) \cdot \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\mathrm{softmax}\left(\theta_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

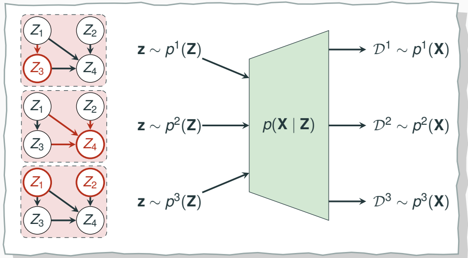


$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e\, \delta^e(\mathbf{pa}_\ell)_k}$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell \in [L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\sigma^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \operatorname{Inv-\Gamma}(\alpha_\sigma, \beta_\sigma)$$



Latent Mechanisms

$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e \, \delta^e(\mathbf{pa}_\ell)_k}$$

$$c_\ell \sim \operatorname{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \operatorname{Bernoulli}(c_\ell)$$

$$\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\Theta)$$

$$\boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell) \sim p\left(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon\right) \propto \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\Delta) \cdot \mathbb{1}\left\{\tfrac{1}{2}\boldsymbol{\delta}^\top \mathbf{H} \boldsymbol{\delta} \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \operatorname{Categorical}\left(\operatorname{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e \, \boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \textstyle\sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\sigma^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \mathrm{Inv-}\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms

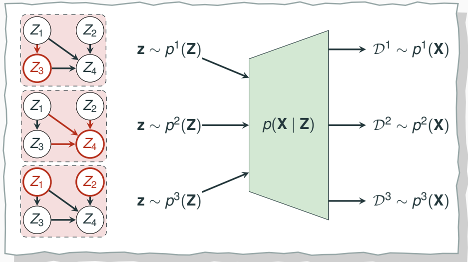
$$c_\ell \sim \mathrm{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \mathrm{Bernoulli}(c_\ell)$$

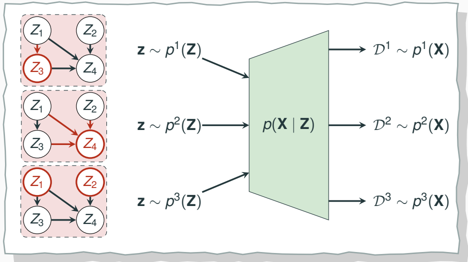
$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \Sigma_\Theta\right)$$

$$\delta_\ell^e(\mathbf{pa}_\ell) \sim p\left(\delta \mid \Sigma_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \Sigma_\Delta\right) \cdot \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \mathrm{Categorical}\left(\mathrm{softmax}\left(\theta_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$



$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e\, \delta^e(\mathbf{pa}_\ell)_k}$$



Measurement Model

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_{\ell\in[L]} \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\boldsymbol{\sigma}^2)\right)$$

$$\mathbf{m}_\ell \sim \mathcal{N}_D(\mathbf{0}, \mathbf{I}), \quad \sigma_d^2 \sim \text{Inv}-\Gamma(\alpha_\sigma, \beta_\sigma)$$

Latent Mechanisms

$$c_\ell \sim \text{Beta}(\alpha_c, \beta_c)$$

$$\mathcal{I}_\ell^e \sim \text{Bernoulli}(c_\ell)$$

$$\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Theta\right)$$

$$\boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell) \sim p\left(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon\right) \propto \mathcal{N}_K\left(\mathbf{0}, \boldsymbol{\Sigma}_\Delta\right) \cdot \mathbb{1}\left\{\tfrac{1}{2}\boldsymbol{\delta}^\top \mathbf{H}\boldsymbol{\delta} \geq \varepsilon\right\}$$

$$z_\ell^{e,j} \sim \text{Categorical}\left(\operatorname{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

SHIFT PRIOR: DETECTABILITY AND KL APPROXIMATION

$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e \delta^e(\mathbf{pa}_\ell)_k}$$

ε-intervention-detectability

For each candidate shift $(\ell, e, \mathbf{pa}_\ell)$, either the mechanism is unchanged,

$$p^e(Z_\ell \mid \mathbf{pa}_\ell) = p(Z_\ell \mid \mathbf{pa}_\ell),$$

or the change is bounded away from zero:

$$\mathrm{KL}\big(p(Z_\ell \mid \mathbf{pa}_\ell) \,\big\|\, p^e(Z_\ell \mid \mathbf{pa}_\ell)\big) \geq \varepsilon.$$

Shift prior: detectability and KL approximation

$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e \, \delta^e(\mathbf{pa}_\ell)_k}$$

ε-intervention-detectability

For each candidate shift $(\ell, e, \mathbf{pa}_\ell)$, either the mechanism is unchanged,

$$p^e(Z_\ell \mid \mathbf{pa}_\ell) = p(Z_\ell \mid \mathbf{pa}_\ell),$$

or the change is bounded away from zero:

$$\mathrm{KL}\big(p(Z_\ell \mid \mathbf{pa}_\ell) \,\big\|\, p^e(Z_\ell \mid \mathbf{pa}_\ell)\big) \geq \varepsilon.$$

Second-order approximation

$$\begin{aligned} D(\boldsymbol{\delta}) &= \mathrm{KL}\big(\mathrm{softmax}(\boldsymbol{\theta}) \,\big\|\, \mathrm{softmax}(\boldsymbol{\theta} + \boldsymbol{\delta})\big) \\ &\approx \tfrac{1}{2}\boldsymbol{\delta}^\top H_{\boldsymbol{\theta}} \boldsymbol{\delta} \\ &\approx \tfrac{1}{2}\boldsymbol{\delta}^\top H_0 \boldsymbol{\delta}, \qquad H_{\mathbf{0}} = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right). \end{aligned}$$

SHIFT PRIOR: DETECTABILITY AND KL APPROXIMATION

$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e\, \delta^e(\mathbf{pa}_\ell)_k}$$

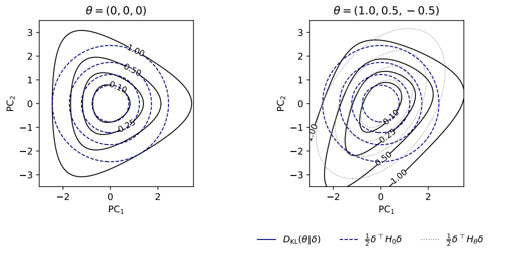
ε-intervention-detectability

For each candidate shift $(\ell, e, \mathbf{pa}_\ell)$, either the mechanism is unchanged,

$$p^e(Z_\ell \mid \mathbf{pa}_\ell) = p(Z_\ell \mid \mathbf{pa}_\ell),$$

or the change is bounded away from zero:

$$\mathrm{KL}\big(p(Z_\ell \mid \mathbf{pa}_\ell) \,\big\|\, p^e(Z_\ell \mid \mathbf{pa}_\ell)\big) \geq \varepsilon.$$



Second-order approximation

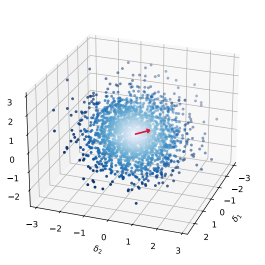
$$D(\delta) = \mathrm{KL}\big(\mathrm{softmax}(\boldsymbol{\theta}) \,\big\|\, \mathrm{softmax}(\boldsymbol{\theta} + \boldsymbol{\delta})\big)$$
$$\approx \tfrac{1}{2}\boldsymbol{\delta}^\top H_{\boldsymbol{\theta}} \boldsymbol{\delta}$$
$$\approx \tfrac{1}{2}\boldsymbol{\delta}^\top H_0 \boldsymbol{\delta}, \qquad H_0 = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right).$$

$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e \, \delta^e(\mathbf{pa}_\ell)_k}$$

Log-space Gaussian priors

$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\Theta),$$

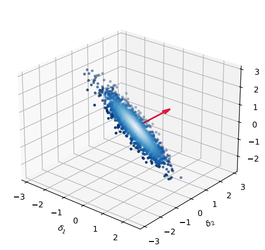
$$\boldsymbol{\Sigma}_\Theta = \operatorname{diag}(\boldsymbol{\sigma}_\Theta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Theta \boldsymbol{\sigma}_\Theta^\top$$



$$p(\boldsymbol{\delta} \mid \boldsymbol{\Sigma}_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\Delta)\, \mathbb{1}\left\{\tfrac{1}{2}\boldsymbol{\delta}^\top \mathbf{H} \boldsymbol{\delta} \geq \varepsilon\right\}$$

$$\boldsymbol{\Sigma}_\Delta = \operatorname{diag}(\boldsymbol{\sigma}_\Delta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Delta \boldsymbol{\sigma}_\Delta^\top$$

$$\mathbf{H} = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)$$



$$p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e\, \delta^e(\mathbf{pa}_\ell)_k}$$

Log-space Gaussian priors

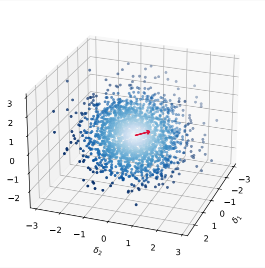
$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \Sigma_\Theta),$$

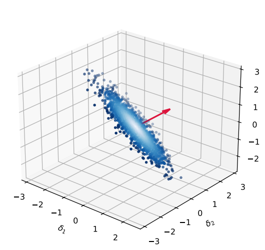
$$\Sigma_\Theta = \operatorname{diag}(\boldsymbol{\sigma}_\Theta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Theta \boldsymbol{\sigma}_\Theta^\top$$

$$p(\delta \mid \Sigma_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \Sigma_\Delta)\, \mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}$$

$$\Sigma_\Delta = \operatorname{diag}(\boldsymbol{\sigma}_\Delta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Delta \boldsymbol{\sigma}_\Delta^\top$$

$$\mathbf{H} = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)$$





SHIFT PRIOR: COVARIANCE AND TRUNCATION

$$\boxed{p(Z_\ell = k) \propto e^{\theta(\mathbf{pa}_\ell)_k + \mathcal{I}_\ell^e \, \delta^e(\mathbf{pa}_\ell)_k}}$$

Log-space Gaussian priors

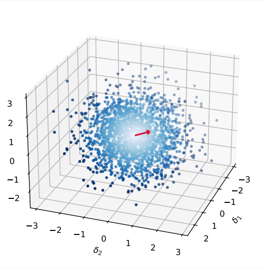
$$\theta_\ell(\mathbf{pa}_\ell) \sim \mathcal{N}_K(\mathbf{0}, \Sigma_\Theta),$$

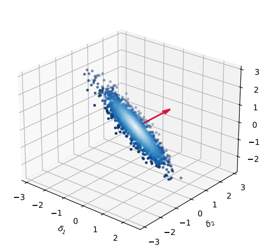
$$\boldsymbol{\Sigma}_\Theta = \operatorname{diag}(\boldsymbol{\sigma}_\Theta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Theta \boldsymbol{\sigma}_\Theta^\top$$

$$p(\delta \mid \boldsymbol{\Sigma}_\Delta, \varepsilon) \propto \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma}_\Delta)\, \boxed{\mathbb{1}\left\{\tfrac{1}{2}\delta^\top \mathbf{H}\delta \geq \varepsilon\right\}}$$

$$\boldsymbol{\Sigma}_\Delta = \operatorname{diag}(\boldsymbol{\sigma}_\Delta^2) - \frac{1}{K}\boldsymbol{\sigma}_\Delta \boldsymbol{\sigma}_\Delta^\top$$

$$\boxed{\mathbf{H} = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)}$$





Measurement model symmetry

For any permutation $\pi \in S_L$ of latent coordinates,

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{M}, \sigma) = p(\mathbf{x} \mid \mathbf{z}_\pi, \mathbf{M}_\pi, \sigma)$$

$$\mathbf{z}_\pi = (z_{\pi(1)}, \ldots, z_{\pi(L)})$$

$$\mathbf{M}_\pi = \begin{bmatrix} \mathbf{m}_0 & \mathbf{m}_{\pi(1)} & \cdots & \mathbf{m}_{\pi(L)} \end{bmatrix}$$

In the additive Gaussian model, column $\mathbf{m}_\ell$ determines how Z_ℓ affects the mean of $\mathbf{x}$.

Measurement model symmetry

For any permutation $\pi \in S_L$ of latent coordinates,

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{M}, \sigma) = p(\mathbf{x} \mid \mathbf{z}_\pi, \mathbf{M}_\pi, \sigma)$$

$$\mathbf{z}_\pi = (z_{\pi(1)}, \ldots, z_{\pi(L)})$$

$$\mathbf{M}_\pi = \begin{bmatrix} \mathbf{m}_0 & \mathbf{m}_{\pi(1)} & \cdots & \mathbf{m}_{\pi(L)} \end{bmatrix}$$

In the additive Gaussian model, column $\mathbf{m}_\ell$ determines how Z_ℓ affects the mean of $\mathbf{x}$.

Why this matters

- The causal graph is fixed to a complete DAG ordered

$$Z_1 \preceq \cdots \preceq Z_L.$$

- Inference over causal order happens through which concept is encoded in each latent coordinate.
- Permuting columns of $\mathbf{M}$ permutes the causal order of latent concepts.

Measurement model symmetry

For any permutation $\pi \in S_L$ of latent coordinates,

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{M}, \sigma) = p(\mathbf{x} \mid \mathbf{z}_\pi, \mathbf{M}_\pi, \sigma)$$

$$\mathbf{z}_\pi = (z_{\pi(1)}, \ldots, z_{\pi(L)})$$

$$\mathbf{M}_\pi = \begin{bmatrix} \mathbf{m}_0 & \mathbf{m}_{\pi(1)} & \cdots & \mathbf{m}_{\pi(L)} \end{bmatrix}$$

In the additive Gaussian model, column $\mathbf{m}_\ell$ determines how Z_ℓ affects the mean of $\mathbf{x}$.

Why this matters

- The causal graph is fixed to a complete DAG ordered

$$Z_1 \preceq \cdots \preceq Z_L.$$

- Inference over causal order happens through which concept is encoded in each latent coordinate.
- Permuting columns of **M** permutes the causal order of latent concepts.

The measurement model is symmetric; the full model is not, because the priors over **Z** break this symmetry.

World Values Survey

- Large-scale cross-national survey measuring values, beliefs, and norms[4].
- Covers topics such as democracy, religion, gender equality, societal trust, work, and economic security.
- We focus on wave 7, conducted from 2017–2022.

[4] Haerpfer et al. *World Values Survey: Round Seven Country-Pooled Datafile Version 5.0.* 2022.

World Values Survey

- Large-scale cross-national survey measuring values, beliefs, and norms[4].
- Covers topics such as democracy, religion, gender equality, societal trust, work, and economic security.
- We focus on wave 7, conducted from 2017–2022.

CRL setup

- $\mathbf{x}$: responses to survey questions.
- $\mathbf{z}$: latent cultural values and demographic concepts.
- e: country / survey environment.
- Direct interventions on values are infeasible, so country-level heterogeneity provides weak supervision.

[4] Haerpfer et al. *World Values Survey: Round Seven Country-Pooled Datafile Version 5.0.* 2022.

World Values Survey

- Large-scale cross-national survey measuring values, beliefs, and norms[4].
- Covers topics such as democracy, religion, gender equality, societal trust, work, and economic security.
- We focus on wave 7, conducted from 2017–2022.

CRL setup

- $\mathbf{x}$: responses to survey questions.
- $\mathbf{z}$: latent cultural values and demographic concepts.
- e: country / survey environment.
- Direct interventions on values are infeasible, so country-level heterogeneity provides weak supervision.

$D = 37$
survey questions

$85{,}851$
total responses

$L = 3,\ K = 3$
discrete latent variables

[4] Haerpfer et al. *World Values Survey: Round Seven Country-Pooled Datafile Version 5.0.* 2022.

$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_\ell \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right)$$

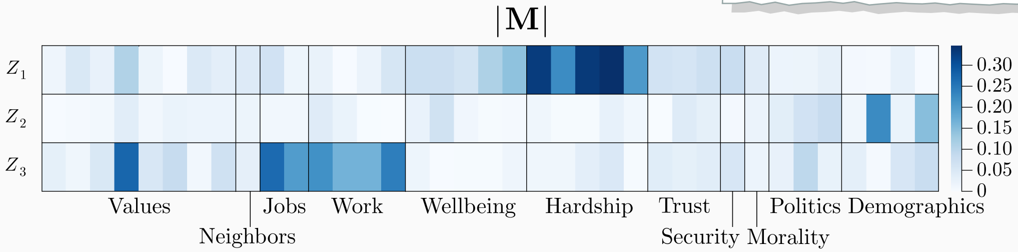
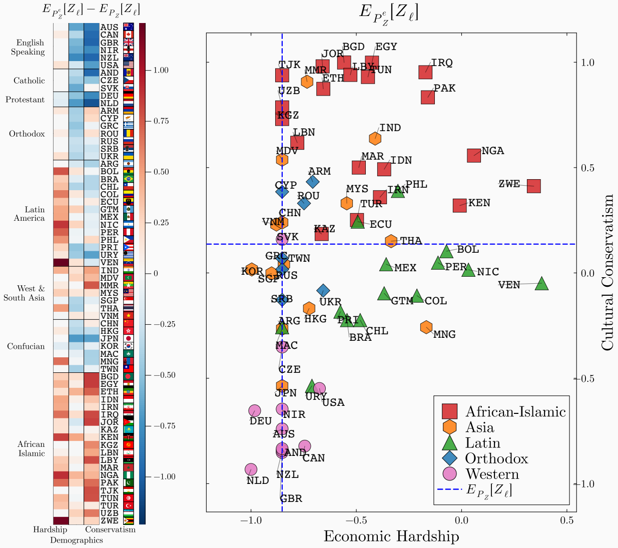


Figure 1: Measurement matrix from latent variables to survey responses.

$Z_1 \mapsto$ economic hardship, $Z_2 \mapsto$ demographics, $Z_3 \mapsto$ cultural conservatism

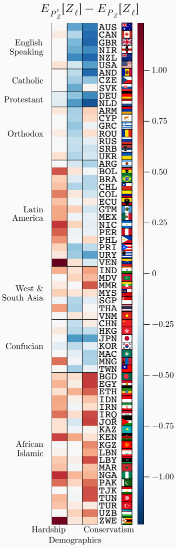
LATENT MAP

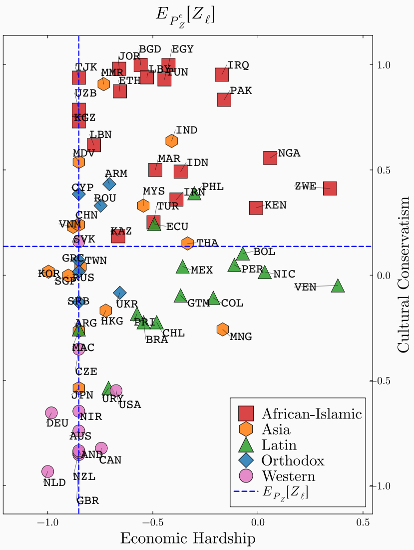


$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_\ell \mathbf{m}_\ell z_\ell^{e,j}, \operatorname{diag}(\boldsymbol{\sigma}^2)\right)$$

$$z_\ell^{e,j} \sim \mathrm{Cat}\left(\mathrm{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \boldsymbol{\delta}_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$

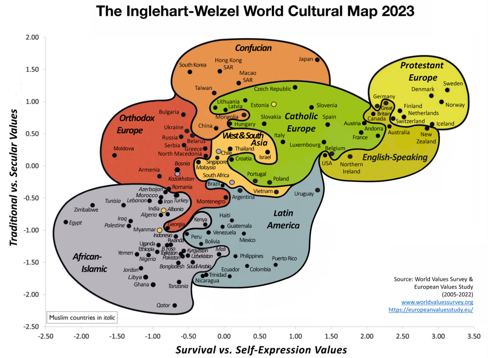
LATENT MAP





$$\mathbf{x}^{e,j} \sim \mathcal{N}_D\left(\mathbf{m}_0 + \sum_\ell \mathbf{m}_\ell z_\ell^{e,j}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right)$$

$$z_\ell^{e,j} \sim \mathrm{Cat}\left(\mathrm{softmax}\left(\boldsymbol{\theta}_\ell(\mathbf{pa}_\ell^{e,j}) + \mathcal{I}_\ell^e\, \delta_\ell^e(\mathbf{pa}_\ell^{e,j})\right)\right)$$



Ordering of latent concepts

hardship $\preceq$ demographics $\preceq$ cultural conservatism

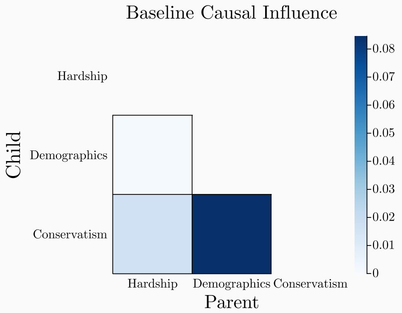
Post-hoc edge strength

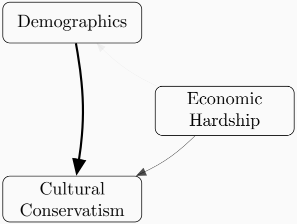
Causal influence[5] measures the change in the joint distribution when the direct influence of Z_j on Z_ℓ is removed.

$$\mathrm{CI}_p^{j\to\ell} := D_{\mathrm{KL}}\left(p(\mathbf{Z}) \,\|\, p^{j\to\ell}(\mathbf{Z})\right),$$

$$p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j) = \int p(z_\ell \mid \mathbf{pa}_\ell)\, p(z_j)\, dz_j$$

and $p^{j\to\ell}(\mathbf{Z})$ is the interventional distribution arising from replacing $p(z_\ell \mid \mathbf{pa}_\ell)$ by $p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j)$.





[5] Janzing et al. *Quantifying Causal Influences.* The Annals of Statistics, 2013.

Ordering of latent concepts

hardship $\preceq$ demographics $\preceq$ cultural conservatism

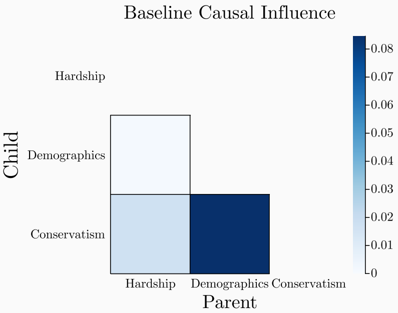
Post-hoc edge strength

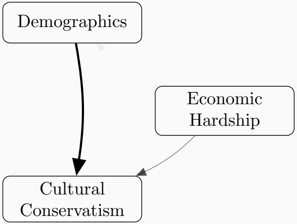
Causal influence[5] measures the change in the joint distribution when the direct influence of Z_j on Z_ℓ is removed.

$$\mathrm{CI}_p^{j\to\ell} := D_{\mathrm{KL}}\left(p(\mathbf{Z}) \,\|\, p^{j\to\ell}(\mathbf{Z})\right),$$

$$p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j) = \int p(z_\ell \mid \mathbf{pa}_\ell) p(z_j) dz_j$$

and $p^{j\to\ell}(\mathbf{Z})$ is the interventional distribution arising from replacing $p(z_\ell \mid \mathbf{pa}_\ell)$ by $p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j)$.





[5] Janzing et al. *Quantifying Causal Influences*. The Annals of Statistics, 2013.

Ordering of latent concepts

hardship $\preceq$ demographics $\preceq$ cultural conservatism

Post-hoc edge strength

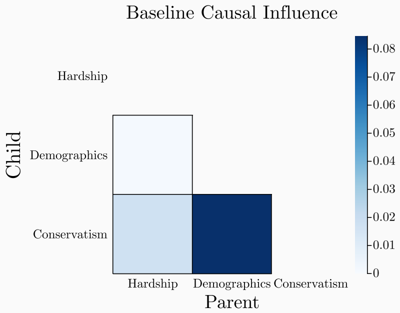
Causal influence[5] measures the change in the joint distribution when the direct influence of Z_j on Z_ℓ is removed.

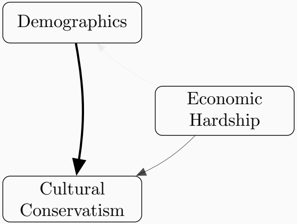
$$\mathrm{CI}_p^{j\to\ell} := D_{\mathrm{KL}}\left(p(\mathbf{Z}) \parallel p^{j\to\ell}(\mathbf{Z})\right),$$

$$p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j) = \int p(z_\ell \mid \mathbf{pa}_\ell) p(z_j) dz_j$$

and $p^{j\to\ell}(\mathbf{Z})$ is the interventional distribution arising from replacing $p(z_\ell \mid \mathbf{pa}_\ell)$ by $p^{j\to\ell}(z_\ell \mid \mathbf{pa}_\ell \setminus z_j)$.

We also can look at intervention indicators($\mathcal{I}$), direction($\mathbf{\Delta}$), and relations per environment.





[5] Janzing et al. *Quantifying Causal Influences*. The Annals of Statistics, 2013.

Evidence across causal orderings

$$\log \mathrm{BF}(\pi) = \log p(\mathbf{X}, \boldsymbol{\Psi}_{\pi}) - \log p(\mathbf{X}, \boldsymbol{\Psi}_{\pi^\star})$$

Red points are latent-order permutations; blue paths interpolate. The true ordering is the reference.

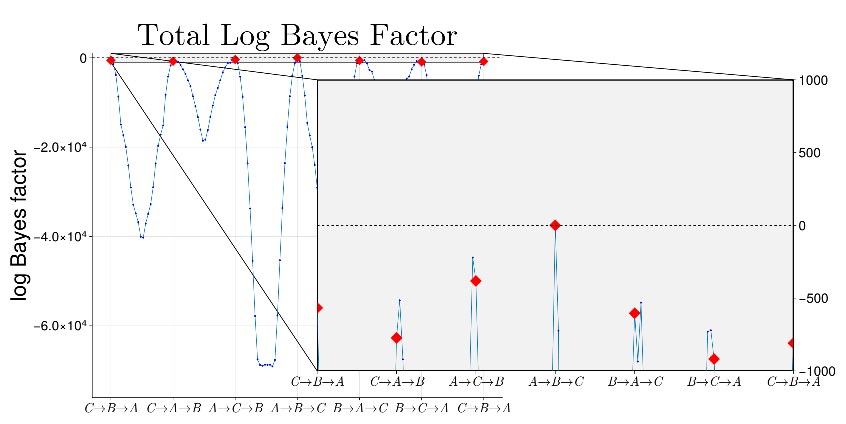


Figure 2: Multimodality across permutations of the causal ordering.

POSTERIOR LANDSCAPE

$$\log \mathrm{BF}(\pi) = \underbrace{\Delta \log p(\mathbf{Z}_\pi, \boldsymbol{\Phi}_\pi)}_{\text{latent mechanisms}} + \underbrace{\Delta \log p(\mathbf{X}, \boldsymbol{\Psi}_\pi \mid \mathbf{Z}_\pi)}_{\text{measurement model}}.$$

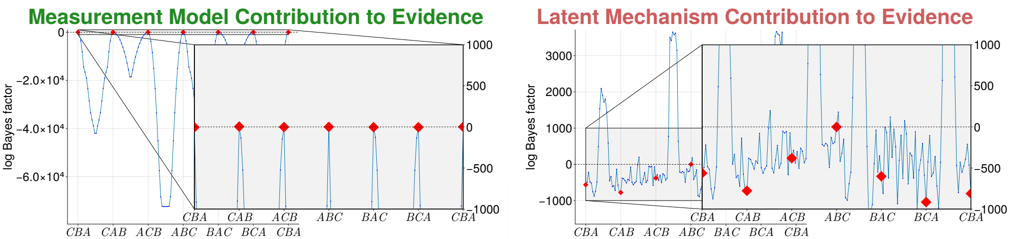


Figure 3: Measurement model creates permutation modes; latent mechanisms distinguish among them.

Inference strategy

- The model admits closed-form conjugate updates for most parameter blocks.
- The exception is $(\boldsymbol{\Theta}, \boldsymbol{\Delta})$; we exploit Pólya–Gamma augmentation[6].

[6] Polson et al. *Bayesian inference for logistic models using Pólya–Gamma latent variables*. JASA, 2013.
[7] Del Moral et al. *Sequential Monte Carlo Samplers*. JRSS Series B, 2006.

Inference strategy

- The model admits closed-form conjugate updates for most parameter blocks.
- The exception is $(\boldsymbol{\Theta}, \boldsymbol{\Delta})$; we exploit Pólya–Gamma augmentation[6].
- Standard MCMC chains can get trapped in one causal-ordering mode.
- Tempering lets latent mechanisms guide inference before the full likelihood dominates.

[6] Polson et al. *Bayesian inference for logistic models using Pólya–Gamma latent variables*. JASA, 2013.
[7] Del Moral et al. *Sequential Monte Carlo Samplers*. JRSS Series B, 2006.

Inference strategy

- The model admits closed-form conjugate updates for most parameter blocks.
- The exception is $(\boldsymbol{\Theta}, \boldsymbol{\Delta})$; we exploit Pólya–Gamma augmentation[6].
- Standard MCMC chains can get trapped in one causal-ordering mode.
- Tempering lets latent mechanisms guide inference before the full likelihood dominates.

Likelihood-tempered SMCS

Many weighted particles explore causal-ordering modes in parallel[7].

$$p_{T_k}(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z} \mid \mathbf{X}) \propto p(\boldsymbol{\Phi}) p(\boldsymbol{\Psi}) p(\mathbf{Z}, \mathbf{X} \mid \boldsymbol{\Phi}, \boldsymbol{\Psi})^{1/T_k}$$

$$\infty = T_0 > T_1 > \cdots > T_K = 1, \quad \boldsymbol{\Phi} = (\boldsymbol{\Theta}, \mathcal{I}, \boldsymbol{\Delta}), \quad \boldsymbol{\Psi} = (\mathbf{M}, \boldsymbol{\sigma}^2)$$

$(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_0^{(s)} \overset{\text{i.i.d.}}{\sim} p_{T_0}(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})$
$w_0^{(s)} \leftarrow 1/S$
for $k = 1$ **to** K **do**
  Set weights $w_k^{(s)} \propto w_{k-1}^{(s)}$
  $\left[p(\mathbf{Z}^{(s)}, \mathbf{X} \mid \boldsymbol{\Phi}^{(s)}, \boldsymbol{\Psi}^{(s)})\right]^{1/T_k - 1/T_{k-1}}$
  Resample particles using $\{w_k^{(s)}\}_{s=1}^S$
  $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)} \leftarrow \textsc{GibbsUpdate}((\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)}, T_k)$
end for
Return weighted particles at $T_K = 1$

[6] Polson et al. *Bayesian inference for logistic models using Pólya–Gamma latent variables*. JASA, 2013.
[7] Del Moral et al. *Sequential Monte Carlo Samplers*. JRSS Series B, 2006.

US SURVEY DATA: PREDICTWISE

PredictWise political opinion surveys

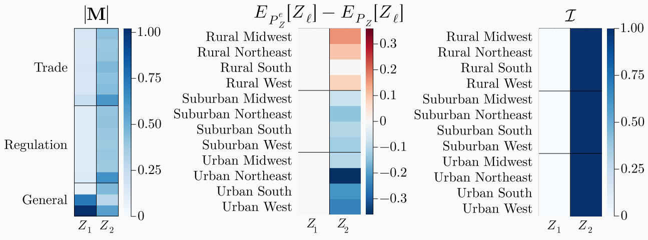
- Collected from U.S. respondents between 2017–2022.
- 7.8k respondents; 75.5k item-level responses.
- Each survey covers one political topic pair, with issue questions plus party/vote-choice questions.
- We analyze two surveys: *Trade & Regulation* and *Elites & Racism*.

CRL setup

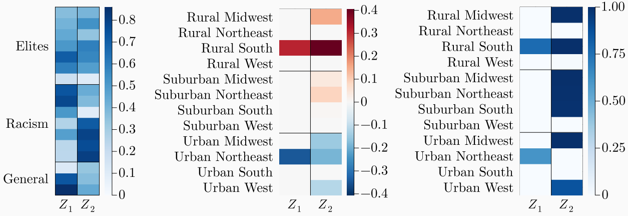
- $\mathbf{x}$: political issue, party, and vote-choice responses.
- $\mathbf{z}$: latent political attitudes within each survey.
- e: region $\times$ urbanicity survey environment.

7.8k	12	$L = 2, \; K = 5$
respondents	natural domains	discrete latent variables

US SURVEY DATA: MOSTLY ONE-DIMENSIONAL



(a) Trade and Regulation



(b) Elites and Racism

What the model finds

- Most cross-environment variation collapses onto a single left–right axis.
- *Trade & Regulation*: clear rural–urban gradient.
- Rural domains shift right, urban domains shift left, suburban domains lie in between.
- *Elites & Racism*: similar gradient, with larger deviations in the rural South and urban Northeast.

We use the PredictWise survey design, but replace natural domain shifts with controlled interventions on latent political topics.

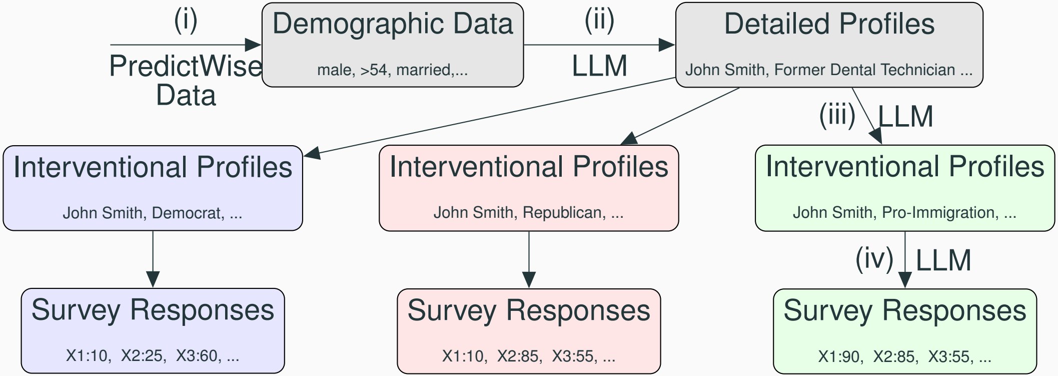
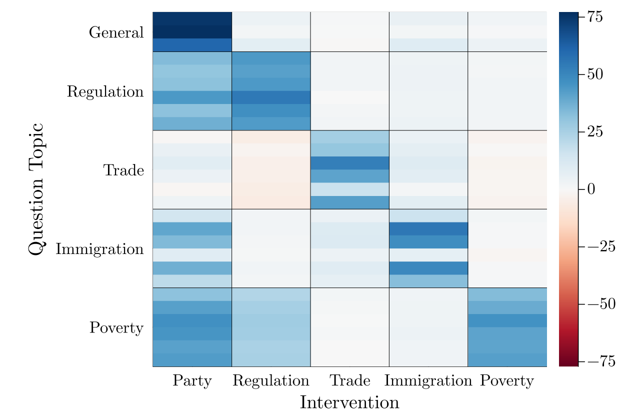
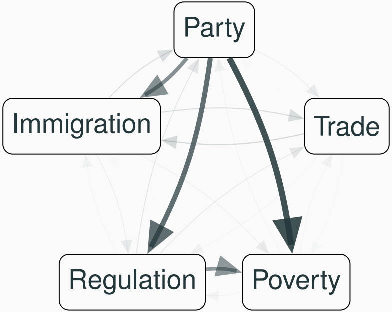


Figure 5: Four-stage pipeline: sample demographics, expand profiles, apply interventions, then generate survey responses.

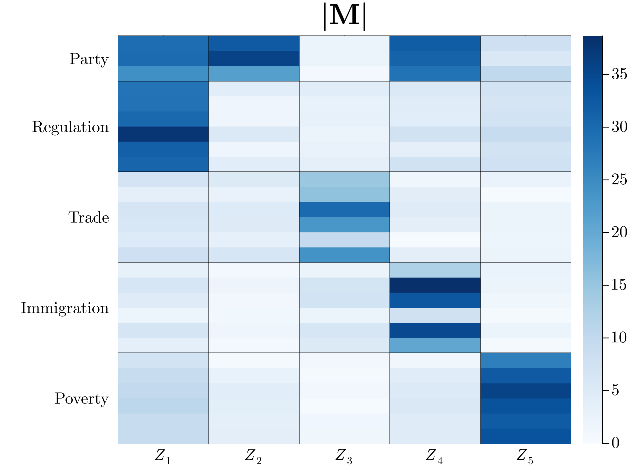
LLM REFERENCE EFFECTS

$$\mathrm{ATE}(d,\ell) = \mathbb{E}_{do(Z_\ell \to +0.5)}[X_d] - \mathbb{E}_{do(Z_\ell \to -0.5)}[X_d], \qquad w_{T,\ell} = \frac{1}{|T|}\sum_{d\in T} \mathrm{ATE}(d,\ell).$$





Reference effects computed directly from known LLM intervention pairs; the structure is nearly block-lower-triangular, with strong party/regulation effects and weak trade effects.



Interpreting latents

The posterior mean $|\mathbf{M}|$ recovers the topic blocks used to generate the synthetic survey.

$Z_1 \mapsto$ Regulation(& Party)

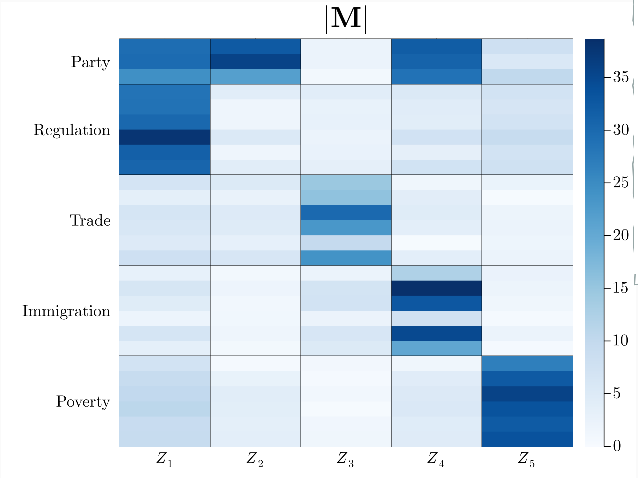
$Z_2 \mapsto$ Party

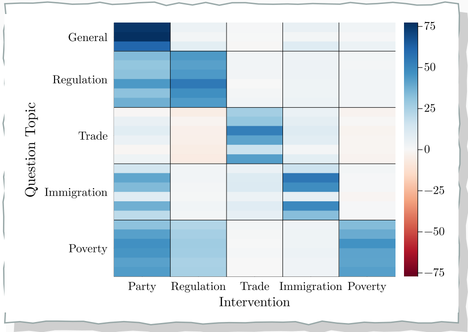
$Z_3 \mapsto$ Trade

$Z_4 \mapsto$ Immigration(& Party)

$Z_5 \mapsto$ Poverty

LLM MEASUREMENT RECOVERY





$Z_1 \mapsto$ Regulation(& Party)

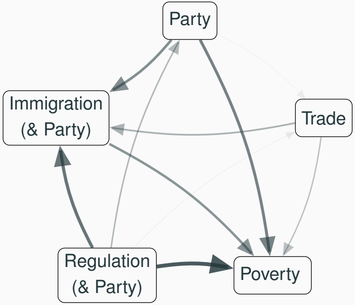
$Z_2 \mapsto$ Party

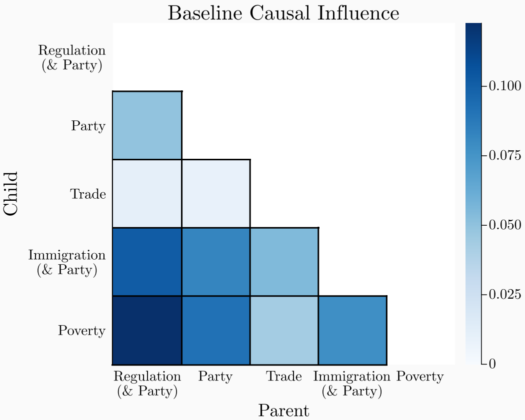
$Z_3 \mapsto$ Trade

$Z_4 \mapsto$ Immigration(& Party)

$Z_5 \mapsto$ Poverty

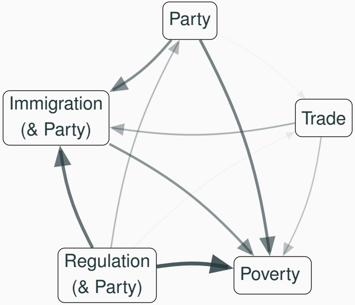
LLM CAUSAL INFLUENCE RECOVERY

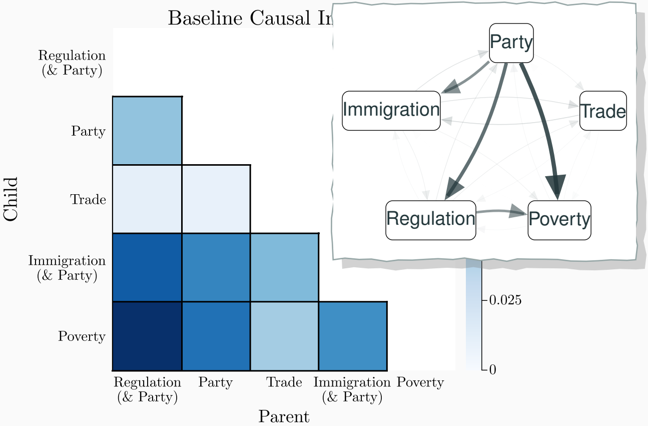




- Largest recovered edge: Regulation → Poverty.
- Trade is nearly independent of the other issue dimensions.
- Party signal is split across recovered latents, so party edges are less isolated.

LLM CAUSAL INFLUENCE RECOVERY

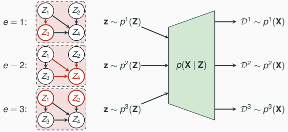


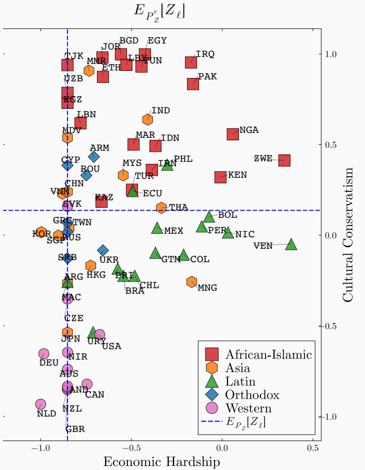


- Largest recovered edge: Regulation → Poverty.
- Trade is nearly independent of the other issue dimensions.
- Party signal is split across recovered latents, so party edges are less isolated.

Thank you

Questions?





APPENDIX: KL SHIFT PRIOR

Finite-sample lower bound

For testing $H_0 : p = p_0$ versus $H_1 : p = p_\delta$ with N_e samples,

$$p_0(T) + p_\delta(1 - T) \le \eta$$

$$\Rightarrow \quad \mathrm{KL}(p_0 \| p_\delta) \ge \frac{2(1-\eta)^2}{N_e}.$$

A detectable intervention therefore needs KL signal bounded away from zero.

Local KL expansion

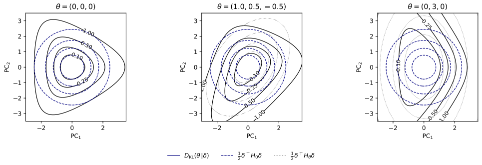
$$\mathrm{KL}\big(\mathrm{softmax}(\boldsymbol{\theta}) \big\| \mathrm{softmax}(\boldsymbol{\theta} + \boldsymbol{\delta})\big) \approx \tfrac{1}{2}\boldsymbol{\delta}^\top H_{\boldsymbol{\theta}} \boldsymbol{\delta},$$

$$H_{\boldsymbol{\theta}} = \mathrm{diag}(p) - pp^\top, \qquad p = \mathrm{softmax}(\boldsymbol{\theta}).$$

Fixed Hessian for tractability

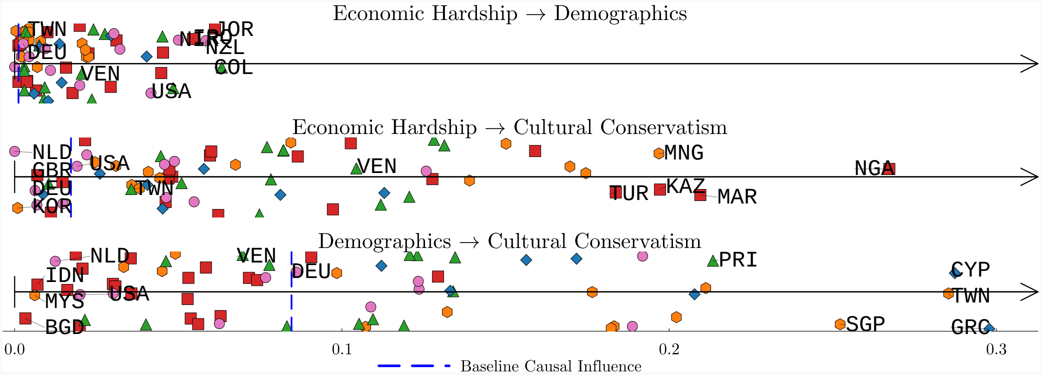
If the truncation depends on $\boldsymbol{\theta}$, its normalizing constant also depends on $\boldsymbol{\theta}$, breaking conjugate updates.

$$\tfrac{1}{2}\boldsymbol{\delta}^\top H_{\mathbf{0}} \boldsymbol{\delta} \ge \varepsilon, \qquad H_{\mathbf{0}} = \frac{1}{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right).$$



The approximation is best near $\boldsymbol{\theta} = \mathbf{0}$ and degrades for highly concentrated categorical distributions.

APPENDIX: ENVIRONMENT-SPECIFIC CAUSAL INFLUENCE



Baseline CI is only one view.

Environment-specific interventions may strengthen or weaken the dependencies among latent concepts.

Country-level variation.

The overall pattern is similar, but $Z_{\text{hard}} \rightarrow Z_{\text{cons}}$ and $Z_{\text{demo}} \rightarrow Z_{\text{cons}}$ vary substantially across countries, while $Z_{\text{hard}} \rightarrow Z_{\text{demo}}$ stays small.

APPENDIX: LABEL SWITCHING

Sign-flip symmetry

For any $\mathbf{s} \in \{-1, 1\}^L$,

$$\mathbf{z}' = \mathbf{s} \odot \mathbf{z}, \qquad \mathbf{M}' = \begin{bmatrix} \mathbf{m}_0 & s_1\mathbf{m}_1 & \cdots & s_L\mathbf{m}_L \end{bmatrix}.$$

This leaves the full posterior representation unchanged.

Not causal-order ambiguity

Permutation modes correspond to different causal orderings; sign flips are nuisance label switching.

Per-sample sign alignment

$$d_\ell^\star = \arg\max_{d \in [D]} |M_{d,\ell}|, \qquad s_\ell = \mathrm{sign}(M_{d_\ell^\star,\ell})$$

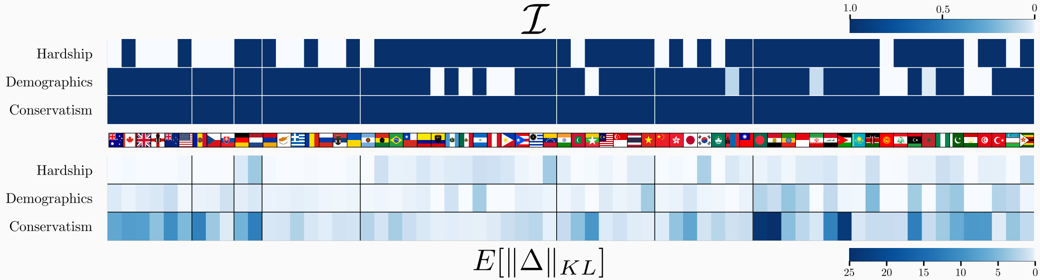
$$z_\ell' = s_\ell z_\ell.$$

With ground truth

$$s_\ell = \mathrm{sign}(M^*_{d_\ell^\star,\ell})\,\mathrm{sign}(M_{d_\ell^\star,\ell}), \qquad z_\ell' = s_\ell z_\ell.$$

All metrics and graphs use the aligned $\mathbf{z}'$ rather than raw $\mathbf{z}$.

APPENDIX: WVS INTERVENTION INDICATORS



Intervention presence

The posterior mean intervention matrix $\mathcal{I}$ gives

$$\Pr\left(\mathcal{I}_\ell^e = 1 \mid \mathcal{D}\right),$$

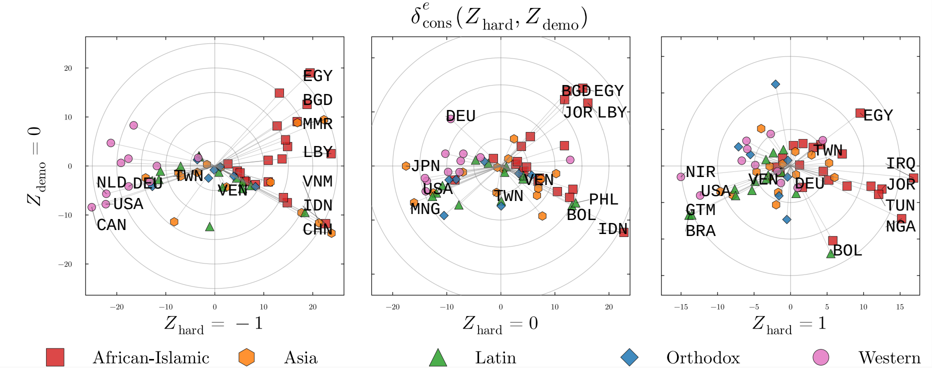
the probability that latent mechanism Z_ℓ is perturbed in country e.

Intervention magnitude

$$\mathbb{E}\left[\|\boldsymbol{\Delta}_\ell^e\|_{\mathrm{KL}}\right] = \mathbb{E}_{\mathbf{pa}_\ell \sim p^e}\left[D_{\mathrm{KL}}\left(p(Z_\ell \mid \mathbf{pa}_\ell) \| p^e(Z_\ell \mid \mathbf{pa}_\ell)\right)\right].$$

Many active interventions are small, so effective sparsity is stronger than $\mathcal{I}$ alone suggests.

APPENDIX: WVS INTERVENTION DIRECTIONS



Mechanism-level perturbations

For each country and parent configuration, $\delta^e_{\text{cons}}(Z_{\text{hard}}, Z_{\text{demo}})$ describes how $p^e(Z_{\text{cons}} \mid Z_{\text{hard}}, Z_{\text{demo}})$ shifts relative to baseline.

Interpretation

Many Western countries show a consistent downward shift in conservatism across parent states; countries such as Taiwan show more heterogeneous direction and magnitude.

APPENDIX: SMCS PSEUDOCODE

$$p_{T_k}(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z} \mid \mathbf{X}) \propto p(\boldsymbol{\Phi}) p(\boldsymbol{\Psi}) p(\mathbf{Z}, \mathbf{X} \mid \boldsymbol{\Phi}, \boldsymbol{\Psi})^{1/T_k}, \qquad \infty = T_0 > \cdots > T_K = 1.$$

Sequential Monte Carlo Sampler with Likelihood Tempering

Require: Temperatures $\{T_0, \ldots, T_K\}$, particles S, and tempered invariant kernels $\textsc{GibbsUpdate}$

$(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_0^{(s)} \overset{\text{i.i.d.}}{\sim} p_{T_0}(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})$ for all $s \in [S]$

$w_0^{(s)} \leftarrow 1/S$ for all $s \in [S]$

for $k = 1$ **to** K **do**

    Set weights $w_k^{(s)} \propto w_{k-1}^{(s)} \left[p(\mathbf{Z}^{(s)}, \mathbf{X} \mid \boldsymbol{\Phi}^{(s)}, \boldsymbol{\Psi}^{(s)}) \right]^{1/T_k - 1/T_{k-1}}$ ▷ Reweight

    $\{(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)}, w_k^{(s)}\}_{s=1}^S \leftarrow \textsc{Resample}(\{(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)}, w_k^{(s)}\}_{s=1}^S)$ ▷ Resample

    $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)} \leftarrow \textsc{GibbsUpdate}((\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_k^{(s)}, T_k)$ ▷ Mutate

end for

return $\{((\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathbf{Z})_K^{(s)}, w_K^{(s)})\}_{s=1}^S$

APPENDIX: SMCS DETAILS

Reweighting

$$\tilde{w}_k^{(s)} = w_{k-1}^{(s)} \left[p(\mathbf{Z}^{(s)} \mid \mathbf{\Phi}^{(s)}) p(\mathbf{X} \mid \mathbf{Z}^{(s)}, \mathbf{\Psi}^{(s)}) \right]^{\frac{1}{T_k} - \frac{1}{T_{k-1}}}$$

$$w_k^{(s)} = \frac{\tilde{w}_k^{(s)}}{\sum_{r=1}^{S} \tilde{w}_k^{(r)}}.$$

Resampling

$$\mathrm{ESS}_k = \frac{1}{\sum_{s=1}^{S} (w_k^{(s)})^2}, \qquad \text{resample if} \quad \mathrm{ESS}_k < \tau S.$$

We use $\tau = 0.5$ and stratified resampling.

Mutation

After resampling, each particle is moved with Gibbs updates that leave $p_{T_k}(\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{Z} \mid \mathbf{X})$ invariant.

Posterior approximation

$$p(\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{Z}^{\mathcal{E}} \mid \mathbf{X}^{\mathcal{E}}) \approx \sum_{s=1}^{S} w_K^{(s)} \delta_{(\mathbf{\Phi}, \mathbf{\Psi}, \mathbf{Z}^{\mathcal{E}})_K^{(s)}}.$$

The final particle population represents uncertainty over latent assignments, mechanisms, measurement parameters, and causal ordering modes.