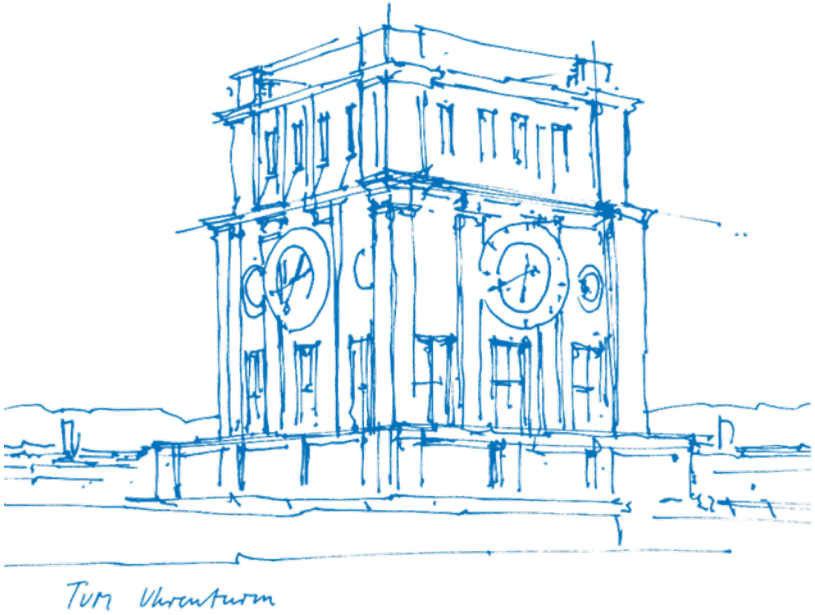
Unpaired Multi-Domain Causal Representation Learning

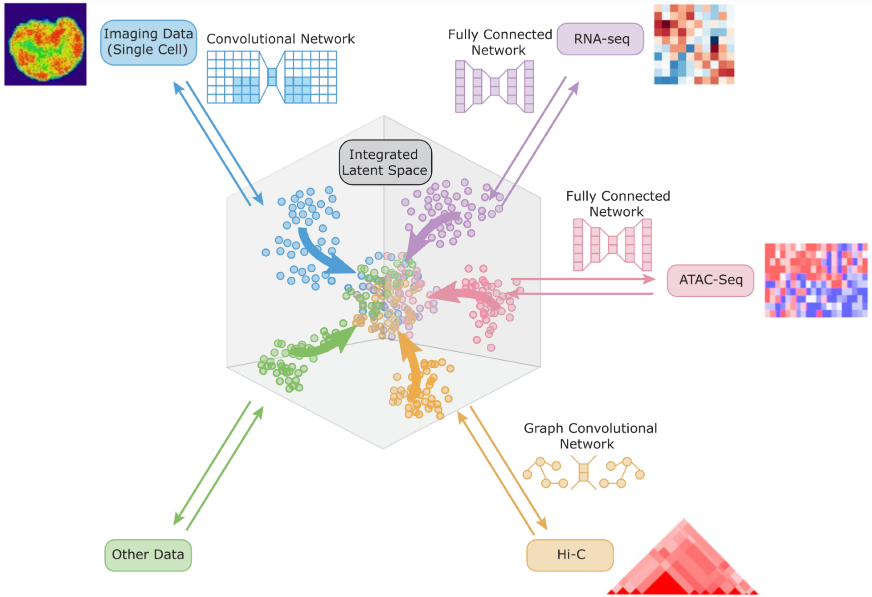
Nils Sturma

Research group Mathematical Statistics
TUM School of Computation, Information and Technology
Technical University of Munich

(joint work with Chandler Squires, Mathias Drton and Caroline Uhler)

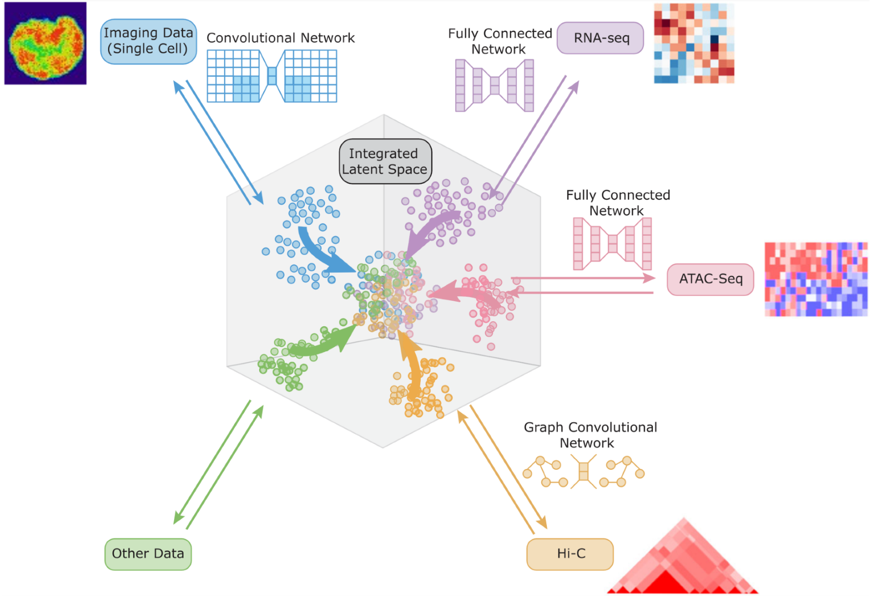


Motivation: Single-Cell Biology



[Yang et al., Nat. Commun. 2021]

Motivation: Single-Cell Biology



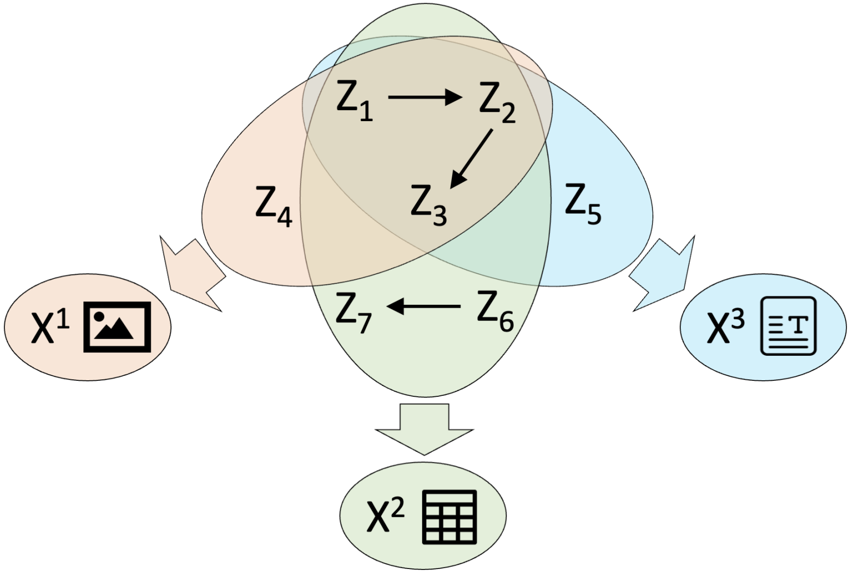
[Yang et al., Nat. Commun. 2021]

- Unpaired observations.
- Observations are of “different nature”.
- “High-level”, latent causal features that determine cell states.

  ⟶ Invariant to modality.

Different data modalities provide multiple “views” on shared latent space.

Multi-Domain Causal Representation Learning



Causal Representation

- Latent variables Z.
- Structural Causal Model.
- Shared variables $Z_{\mathcal{L}}$ capture key causal relations.

Observed Data

- $X^e = g_e(Z_{S_e})$ such that $\mathcal{L} \subseteq S_e$.
- Joint distribution of X^e, X^f unknown.

Integrate data from different modalities to identify causal representation.

Identifiability

Suppose, we are in the "infinite data limit", that is, we know the true observational distribution in each domain.

Questions:

- How large is the shared latent space?
- Can we identify the joint distribution?
- Can we identify the graph of the shared latent space?

Topic of this talk: Identifiability in the **linear** case.

Setup: Linear Model

Causal Model in Latent Space

Latent variables:

$$Z = (Z_i)_{i \in \mathcal{H}}$$

Structural equation model:

$$Z = AZ + \varepsilon$$

- (sparse) parameter matrix A
- error variables ε_i are independent

Setup: Linear Model

Causal Model in Latent Space

Latent variables:

$$Z = (Z_i)_{i \in \mathcal{H}}$$

Structural equation model:

$$Z = AZ + \varepsilon$$

- (sparse) parameter matrix A
- error variables ε_i are independent

Observed Domains

Observed random vectors:

$$X^e \in \mathbb{R}^{d_e} \text{ for each domain } e = 1, \dots, m$$

Linear mixing:

$$X^e = G^e \cdot Z_{S_e},$$

such that $S_e = \mathcal{L} \cup I_e$, where

- $\mathcal{L} \subseteq \mathcal{H}$ indexes the *shared* latent variables and
- $I_e \subseteq \mathcal{H} \setminus \mathcal{L}$ indexes the *domain-specific* latent variables.

Graphical Perspective

m-Domain Graph

- Nodes $\mathcal{H} \cup V_1 \cup \cdots \cup V_m$, where $|V_e| = d_e$.
- Edges in $\mathcal{H}$ encode sparsity in A (acyclic).

  (Recall: $Z = AZ + \varepsilon$.)
- Edges from $\mathcal{H}$ to V_e encode sparsity in G^e.

  (Recall: $X^e = G^e \cdot Z_{S_e}$.)
- The set $\mathcal{L} \subseteq \mathcal{H}$ consists of the shared latent nodes.
- Assumption: No edges from domain-specific to shared latent nodes.

Graphical Perspective

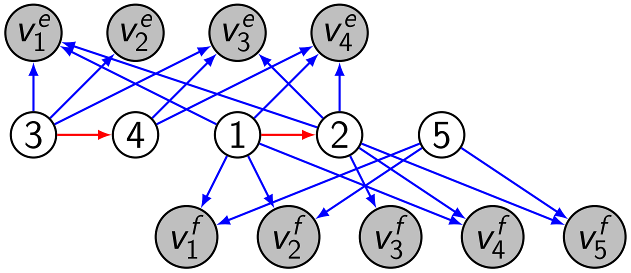
m-Domain Graph

- Nodes $\mathcal{H} \cup V_1 \cup \cdots \cup V_m$, where $|V_e| = d_e$.
- Edges in $\mathcal{H}$ encode sparsity in A (acyclic).

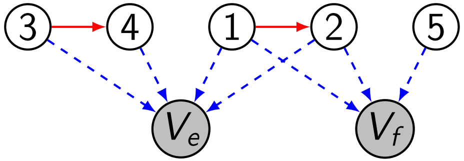
  (Recall: $Z = AZ + \varepsilon$.)
- Edges from $\mathcal{H}$ to V_e encode sparsity in G^e.

  (Recall: $X^e = G^e \cdot Z_{S_e}$.)
- The set $\mathcal{L} \subseteq \mathcal{H}$ consists of the shared latent nodes.
- Assumption: No edges from domain-specific to shared latent nodes.

Example



Compact version:



Latent variables: $\mathcal{L} = \{1, 2\}$ are shared and $I_e = \{3, 4\}$, $I_f = \{5\}$ are domain-specific.

Graphical Perspective

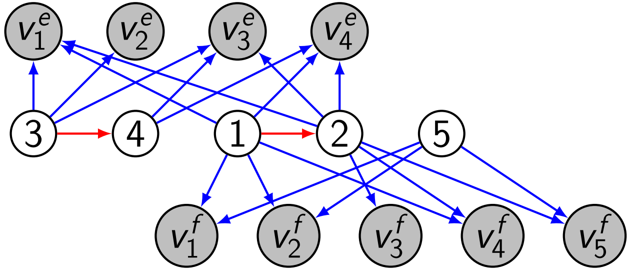
m-Domain Graph

- Nodes $\mathcal{H} \cup V_1 \cup \cdots \cup V_m$, where $|V_e| = d_e$.
- Edges in $\mathcal{H}$ encode sparsity in A (acyclic).

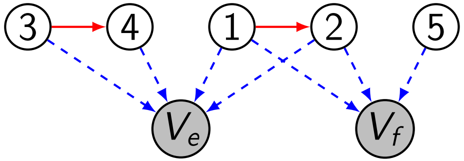
  (Recall: $Z = AZ + \varepsilon$.)
- Edges from $\mathcal{H}$ to V_e encode sparsity in G^e.

  (Recall: $X^e = G^e \cdot Z_{S_e}$.)
- The set $\mathcal{L} \subseteq \mathcal{H}$ consists of the shared latent nodes.
- Assumption: No edges from domain-specific to shared latent nodes.

Example



Compact version:



Latent variables: $\mathcal{L} = \{1, 2\}$ are shared and $I_e = \{3, 4\}$, $I_f = \{5\}$ are domain-specific.

Important: The graph, the set $\mathcal{L} \subseteq \mathcal{H}$ and the joint distribution (X^e, X^f) for $e \neq f$ are *unknown*.

Identifiability of the Joint Distribution

Joint Observations: Denote G the "large" mixing matrix, that is, $G_{V_e,S_e} = G^e$. Then

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^m \end{pmatrix} = G \cdot Z = \underbrace{G \cdot (I - A)^{-1}}_{=B} \cdot \varepsilon$$

Identifiability of the Joint Distribution

Joint Observations: Denote G the "large" mixing matrix, that is, $G_{V_e,S_e} = G^e$. Then

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^m \end{pmatrix} = G \cdot Z = \underbrace{G \cdot (I - A)^{-1}}_{=B} \cdot \varepsilon$$

One Domain:

$$X^e = G_{V_e,S_e} \cdot Z_{S_e} = G_{V_e,S_e} \cdot (I - A)^{-1}_{S_e} \cdot \varepsilon_{S_e} = B_{V_e,S_e} \cdot \varepsilon_{S_e} = \left(\, B_{V_e,\mathcal{L}} \, \middle| \, B_{V_e,I_e} \, \right) \cdot \begin{pmatrix} \varepsilon_{\mathcal{L}} \\ \varepsilon_{I_e} \end{pmatrix}.$$

Identifiability of the Joint Distribution

Joint Observations: Denote G the "large" mixing matrix, that is, $G_{V_e,S_e} = G^e$. Then

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^m \end{pmatrix} = G \cdot Z = \underbrace{G \cdot (I - A)^{-1}}_{=B} \cdot \varepsilon = \left(\begin{array}{c|ccc} B_{V_1,\mathcal{L}} & B_{V_1,I_1} & & \\ \vdots & & \ddots & \\ B_{V_m,\mathcal{L}} & & & B_{V_m,I_m} \end{array} \right) \cdot \begin{pmatrix} \varepsilon_{\mathcal{L}} \\ \varepsilon_{I_1} \\ \vdots \\ \varepsilon_{I_m} \end{pmatrix}.$$

One Domain:

$$X^e = G_{V_e,S_e} \cdot Z_{S_e} = G_{V_e,S_e} \cdot (I - A)^{-1}_{S_e} \cdot \varepsilon_{S_e} = B_{V_e,S_e} \cdot \varepsilon_{S_e} = \left(\begin{array}{c|c} B_{V_e,\mathcal{L}} & B_{V_e,I_e} \end{array} \right) \cdot \begin{pmatrix} \varepsilon_{\mathcal{L}} \\ \varepsilon_{I_e} \end{pmatrix}.$$

Approach/ Algorithm:

1. Apply linear ICA in each domain.
2. Identify shared columns and shared ε_i by matching distributions.
3. Reconstruct B up to unknown (block)-permutation of the columns.

Identifiability Result for the Joint Distribution

Assumptions

(C1) (Different distributions P_i of errors ε_i.)

- Non-degenerate, mean zero, unit variance and independent.
- Non-symmetric ($\Longrightarrow$ non-Gaussian), $P_i \neq P_j$ and $P_i \neq -P_j$ for all $i, j \in \mathcal{H}$ with $i \neq j$.

(C2) (Full rank of mixing.)

The matrix G_{V_e,S_e} is of full column rank for each $e = 1, \ldots, m$.

Identifiability Result for the Joint Distribution

Assumptions

(C1) (Different distributions P_i of errors ε_i.)

- Non-degenerate, mean zero, unit variance and independent.
- Non-symmetric ($\Longrightarrow$ non-Gaussian), $P_i \neq P_j$ and $P_i \neq -P_j$ for all $i, j \in \mathcal{H}$ with $i \neq j$.

(C2) (Full rank of mixing.)

The matrix G_{V_e, S_e} is of full column rank for each $e = 1, \dots, m$.

Theorem

Let $\mathcal{G}_m$ be an m-domain graph with shared latent nodes $\mathcal{L} = [\ell]$, and let $P_X \in \mathcal{M}(\mathcal{G}_m)$ with representation (B, P). Suppose that $m \geq 2$ and that Conditions (C1) and (C2) are satisfied. Let $(\widehat{\ell}, \widehat{B}, \widehat{P})$ be the output of our algorithm. Then $\widehat{\ell} = \ell$ and

$$\widehat{B} = B \cdot \Psi \quad \textit{and} \quad \widehat{P} = \Psi^\top \# P,$$

for a signed permutation block matrix $\Psi \in \Pi$.

$$\Pi = \left\{ \begin{pmatrix} \Psi_{\mathcal{L}} & & & \\ & \Psi_{I_1} & & \\ & & \ddots & \\ & & & \Psi_{I_m} \end{pmatrix} : \begin{array}{l} \Psi_{\mathcal{L}} \in SP(|\mathcal{L}|), \\ \Psi_{I_e} \in SP(|I_e|) \end{array} \right\}.$$

Identifiability Result for the Joint Distribution

Assumptions

(C1) (Different distributions P_i of errors ε_i.)

- Non-degenerate, mean zero, unit variance and independent.
- Non-symmetric ($\implies$ non-Gaussian), $P_i \neq P_j$ and $P_i \neq -P_j$ for all $i, j \in \mathcal{H}$ with $i \neq j$.

(C2) (Full rank of mixing.)

The matrix G_{V_e, S_e} is of full column rank for each $e = 1, \ldots, m$.

Theorem

Let $\mathcal{G}_m$ be an m-domain graph with shared latent nodes $\mathcal{L} = [\ell]$, and let $P_X \in \mathcal{M}(\mathcal{G}_m)$ with representation (B, P). Suppose that $m \geq 2$ and that Conditions (C1) and (C2) are satisfied. Let $(\widehat{\ell}, \widehat{B}, \widehat{P})$ be the output of our algorithm. Then $\widehat{\ell} = \ell$ and

$$\widehat{B} = B \cdot \Psi \quad \textit{and} \quad \widehat{P} = \Psi^\top \# P,$$

for a signed permutation block matrix $\Psi \in \Pi$.

$$\Pi = \left\{ \begin{pmatrix} \Psi_{\mathcal{L}} & & & \\ & \Psi_{I_1} & & \\ & & \ddots & \\ & & & \Psi_{I_m} \end{pmatrix} : \begin{array}{l} \Psi_{\mathcal{L}} \in SP(|\mathcal{L}|), \\ \Psi_{I_e} \in SP(|I_e|) \end{array} \right\}.$$

✔ Number of shared latent vars $\ell = |\mathcal{L}|$.

✔ Joint distribution.

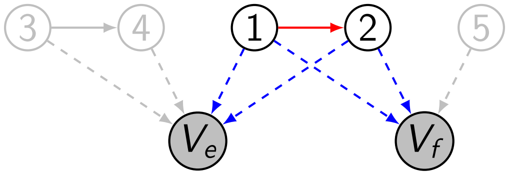
Identifiability of the Shared Latent Graph

Goal: Identify the DAG of the shared latent space $\mathcal{G}_{\mathcal{L}}$.

Starting point: We know the columns corresponding to the shared latent space:

$$\widehat{B}_{\mathcal{L}} = B_{\mathcal{L}} \cdot \Psi_{\mathcal{L}} = G_{\mathcal{L}} \cdot (I - A_{\mathcal{L},\mathcal{L}})^{-1} \cdot \Psi_{\mathcal{L}}, \qquad \text{where } G_{\mathcal{L}} = \begin{pmatrix} G_{V_1,\mathcal{L}} \\ \vdots \\ G_{V_m,\mathcal{L}} \end{pmatrix}.$$

Example



Given the matrix $\widehat{B}_{\mathcal{L}}$, when is it possible to identify the causal graph $\mathcal{G}_{\mathcal{L}}$? (Or the matrix $A_{\mathcal{L},\mathcal{L}}$)?

Partial Pure Children

Literature

Sufficient conditions in recent work are based on sparsity assumptions on the mixing matrix (“pure children”).

[Xie et al., ICML 2022; Dai et al. NeurIPS 2022].

Definitions

$v \in V$ is a *pure child* of $h \in \mathcal{H}$ if $\mathrm{pa}(v) = \{h\}$.

$v \in V$ is a *partial pure child* of $h \in \mathcal{H}$ if $\mathrm{pa}(v) \cap \mathcal{L} = \{h\}$.

Partial Pure Children

Literature

Sufficient conditions in recent work are based on sparsity assumptions on the mixing matrix (“pure children”).

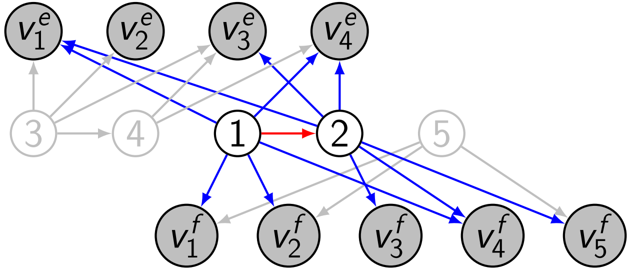
[Xie et al., ICML 2022; Dai et al. NeurIPS 2022].

Definitions

$v \in V$ is a *pure child* of $h \in \mathcal{H}$ if $\mathrm{pa}(v) = \{h\}$.

$v \in V$ is a *partial pure child* of $h \in \mathcal{H}$ if $\mathrm{pa}(v) \cap \mathcal{L} = \{h\}$.

Example



v_1^f is a partial pure child but not a pure child of 1.

Identifiability Result for the Shared Latent Graph

Observation

$\text{rank}(B_{\{v,w\},\mathcal{L}}) = 1$ if and only if there is a node $h \in \mathcal{L}$ such that both v and w are partial pure children of h.

(trek separation, vertex cuts)

Identifiability Result for the Shared Latent Graph

Observation

$\text{rank}(B_{\{v,w\},\mathcal{L}}) = 1$ if and only if there is a node $h \in \mathcal{L}$ such that both v and w are partial pure children of h.

(trek separation, vertex cuts)

Algorithm

1. For each $h \in \mathcal{L}$ find two corresponding partial pure children (rank constraints).
2. Consider $\widehat{B}_{I,\mathcal{L}}$, where $I = \{i_1, \dots, i_{|\mathcal{L}|}\}$ and i_h is a pure children of $h \in \mathcal{L}$.
3. Find permutation matrices R_1, R_2 such that $W = R_1 \widehat{B}_{I,\mathcal{L}} R_2$ lower triangular.
4. Ensure that all diagonal entries are equal to 1. This yields a new matrix $\widetilde{W}$.
5. $\widehat{A}_{\mathcal{L},\mathcal{L}} = I - \widetilde{W}^{-1}$.

Identifiability Result for the Shared Latent Graph

Observation

$\mathrm{rank}(B_{\{v,w\},\mathcal{L}}) = 1$ if and only if there is a node $h \in \mathcal{L}$ such that both v and w are partial pure children of h.

(trek separation, vertex cuts)

Algorithm

1. For each $h \in \mathcal{L}$ find two corresponding partial pure children (rank constraints).
2. Consider $\widehat{B}_{I,\mathcal{L}}$, where $I = \{i_1, \dots, i_{|\mathcal{L}|}\}$ and i_h is a pure children of $h \in \mathcal{L}$.
3. Find permutation matrices R_1, R_2 such that $W = R_1 \widehat{B}_{I,\mathcal{L}} R_2$ lower triangular.
4. Ensure that all diagonal entries are equal to 1. This yields a new matrix $\widetilde{W}$.
5. $\widehat{A}_{\mathcal{L},\mathcal{L}} = I - \widetilde{W}^{-1}$.

Theorem

Suppose we are given $\widehat{B}_{\mathcal{L}}$. Assume rank faithfulness and that each shared latent node has at least two partial pure children (across domains). Then $A_{\mathcal{L},\mathcal{L}}$ is identifiable up to a signed permutation σ that "is consistent with the DAG $G_{\mathcal{L}}$", i.e., $\widehat{A}_{\mathcal{L},\mathcal{L}} = Q_\sigma^\top A_{\mathcal{L},\mathcal{L}} Q_\sigma$.

Finite Samples

1. Choose Linear ICA algorithm, “match” empirical distributions by non-parametric test.

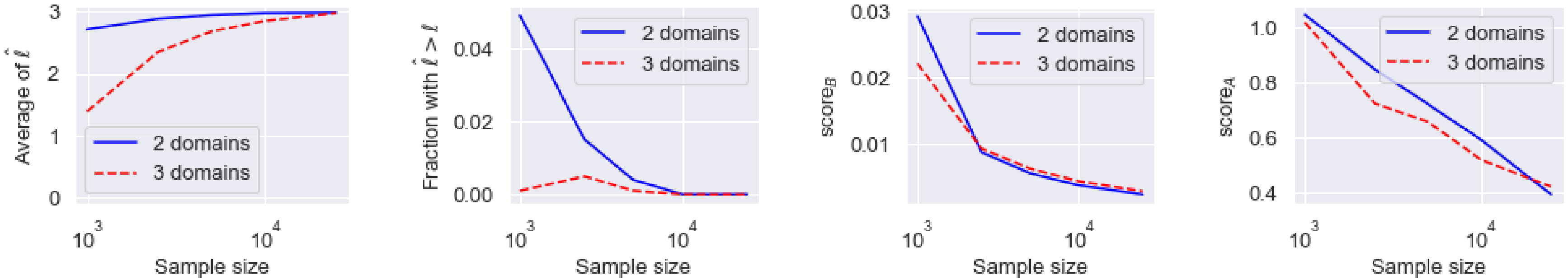
2. Determine the rank of a matrix as the number of singular values above a threshold.

Finite Samples

1. Choose Linear ICA algorithm, “match” empirical distributions by non-parametric test.

2. Determine the rank of a matrix as the number of singular values above a threshold.

Synthetic Data



- 1000 random models, $l = |\mathcal{L}| = 3$ shared and $|l_e| = 2$ domain-specific latent nodes, 10 observed nodes in each domain.
- m-domain graph is samplesd from Erdős-Rényi model with edge probability 0.75 (ensuring two pure children).
- Nonzero entries of A and G are samples from Unif($\pm[0.25, 1]$). Beta, Gumbel, Weibull, exponential, skew normal distributions for errors ε_i.

Conclusion

- First principled identifiability results for shared causal representations in an unpaired multi-domain setting.
- Two-step approach: (i) Joint distribution via linear ICA.
  (ii) Shared causal graph via rank deficiencies.
- Lots of things to explore...
  - Expand identifiability theory: Necessary conditions? Gaussian case? More direct approach?
  - Finite samples: Score based methods?
  - Address non-linear setup.
  - ...

Our paper:

Sturma, Squires, Drton, Uhler (2023).
Unpaired Multi-Domain Causal Representation Learning. arXiv:2302.00993.

European Research Council
Established by the European Commission

References

- Yang, Belyaeva, Venkatachalapathy, Damodaran, Katcoff, Radhakrishnan, Shivashankar, Uhler (2021).
  Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat. Commun. 12, no. 31.

- Xie, Huang, Chen, He, Geng, Zhang (2022).
  Identification of Linear Non-Gaussian Latent Hierarchical Structure. ICML.

- Dai, Spirtes, Zhang (2022).
  Independence Testing-Based Approach to Causal Discovery under Measurement Error and Linear Non-Gaussian Models. NeurIPS.