

Research Summary: Statistical Algorithms for Complex Data

Junhyoung Chung

Department of Statistics, Seoul National University



Research 001: Operator Fused Optimal Transport

Research question: Is it possible to design a fused transport framework that is simultaneously (i) **convex** and computationally tractable, (ii) sensitive to **feature information**, and (iii) capable of preserving the intrinsic **geometric** structure of the domains?

Contributions

- **Convex objective design.** Develop a new loss function formulated as a **convex objective**, guaranteeing efficient and globally optimal solutions.
- **Graph-to-Metric space extension.** Extend the convex relaxation techniques of graph matching problems to the operator level, thereby generalizing the problem from aligning two graphs to aligning two **metric spaces**.
- **Scalable solver with guarantees.** Use a projection-free Frank–Wolfe algorithm for the empirical convex quadratic program, and derive an optimization-statistical error bound.

Convex Structural Penalty

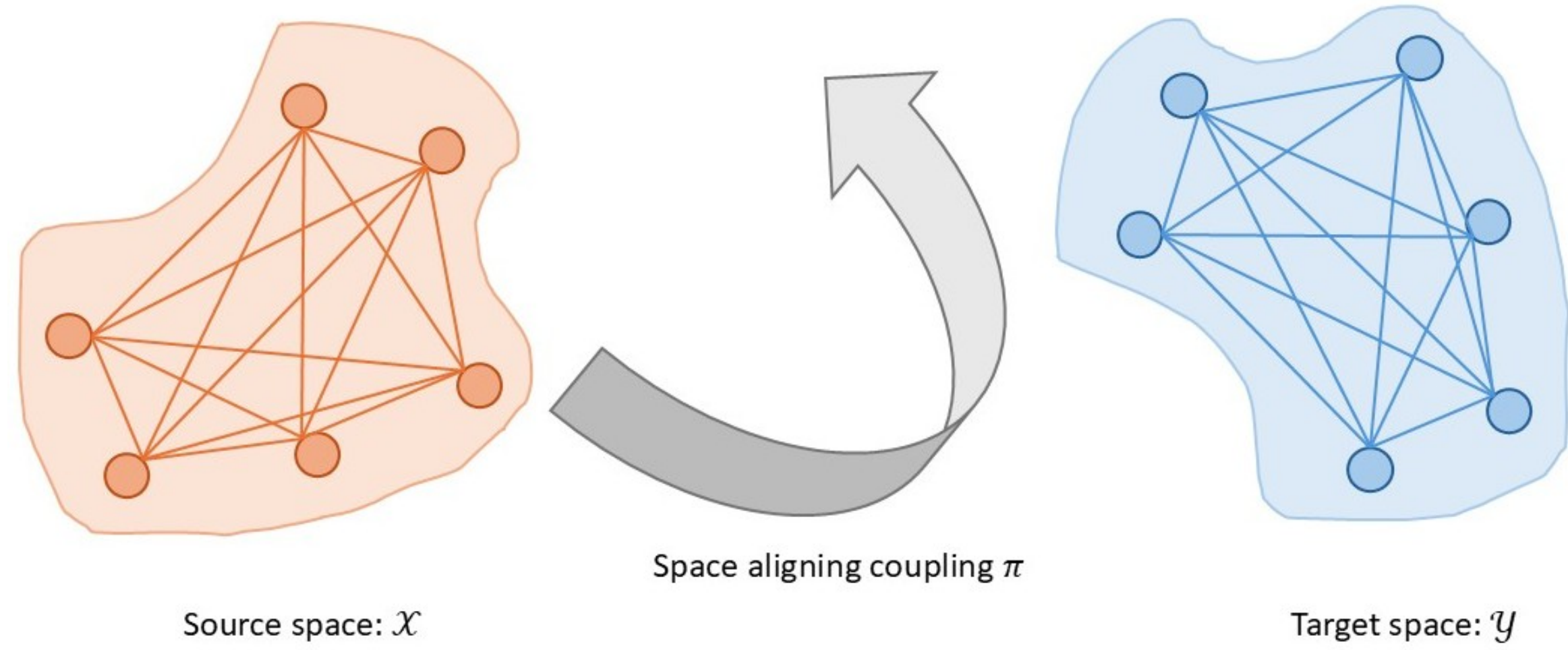


Figure: A coupling π that aligns two spaces \mathcal{X} and \mathcal{Y}

- Let $(\mathcal{X}, d_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \mathbb{P}_{\mathcal{Y}})$ be connected and compact metric measure spaces.
- The Gromov–Wasserstein (GW) discrepancy is powerful for matching problems between heterogeneous spaces:

$$\pi^* = \arg \min_{\pi \in \Pi(\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\mathcal{Y}})} \mathbb{E}_{\pi \otimes \pi} \left[|d_{\mathcal{X}}(X, X') - d_{\mathcal{Y}}(Y, Y')|^2 \right].$$

- However, the GW loss is highly **non-convex** with respect to π .

Motivation: Convex Relaxation in Graph Matching

- Let A_X and A_Y be the adjacency matrices of G_X and G_Y , respectively. The standard graph matching problem of finding a permutation matrix P such that $A_X \approx PA_Y P^T$ can be written as minimizing $\|A_X - PA_Y P^T\|_F^2$, which is equivalent to $\|A_X P - PA_Y\|_F^2$.
- Relaxing P to a soft assignment matrix Π in the Birkhoff polytope then yields the convex quadratic program $\min_{\Pi} \|A_X \Pi - \Pi A_Y\|_F^2$.
- We lift this idea from the graph domain to the **operator level alignment** for general metric spaces.

Our Penalty: $\|D_{\mathbb{P}_X} T_{\pi} - T_{\pi} D_{\mathbb{P}_Y}\|_{\text{HS}}^2$.

- $D_{\mathbb{P}_X}$ (**Distance operator**):
 - ▶ This operator encodes the distance information within the metric space $(\mathcal{X}, d_{\mathcal{X}})$, analogous to the adjacency matrix (A_X) in graph matching.
 - ▶ Definition: $(D_{\mathbb{P}_X} f)(x) = \mathbb{E}_{\mathbb{P}_X}[d_{\mathcal{X}}(x, X)f(X)]$.
- T_{π} (**Alignment operator**):
 - ▶ This operator represents the **soft assignment** or alignment between the two spaces, generalizing the permutation matrix (P) or soft assignment matrix (Π).
 - ▶ Definition: $(T_{\pi} g)(x) = \mathbb{E}_{\pi}[g(Y) \mid X = x]$.

Main Results

Theorem 1 (Convexity). For $0 \leq \alpha \leq 1$, the following is a convex optimization problem:

$$\inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \underbrace{(1 - \alpha) \mathbb{E}_{\pi} [\|f_X(X) - f_Y(Y)\|_2^2] + \frac{\alpha}{2} \|D_{\mathbb{P}_X}^{\kappa} T_{\pi} - T_{\pi} D_{\mathbb{P}_Y}^{\kappa}\|_{\text{HS}}^2}_{=\mathcal{L}(\pi)}. \quad (1)$$

- We additionally introduce an feature space $M \subset \mathbb{R}^k$, into which the source $X \sim \mathbb{P}_X$ and target $Y \sim \mathbb{P}_Y$ are mapped via continuous feature functions $f_X: \mathcal{X} \rightarrow M$ and $f_Y: \mathcal{Y} \rightarrow M$.

Proposition 1 (Isometry consistency). Let $T: \mathcal{X} \rightarrow \mathcal{Y}$ be a bijective measurable map, and consider $\pi = (\text{Id}, T)_{\#} \mathbb{P}_X$. Then,

$$\|D_{\mathbb{P}_X} T_{\pi} - T_{\pi} D_{\mathbb{P}_Y}\|_{\text{HS}}^2 = 0 \iff d_Y(T(x), T(x')) = d_X(x, x') \text{ for } \mathbb{P}_X \otimes \mathbb{P}_X\text{-a.e. } (x, x').$$

- The above proposition shows that the proposed structural penalty favors isometry transport plans, while ensuring convexity.
- More generally, the penalty vanishes iff $D_{\mathbb{P}_X} T_{\pi} = T_{\pi} D_{\mathbb{P}_Y}$, that is, if φ is an eigenfunction of $D_{\mathbb{P}_Y}$ with eigenvalue λ , then

$$D_{\mathbb{P}_X}(T_{\pi} \varphi) = T_{\pi}(D_{\mathbb{P}_Y} \varphi) = \lambda T_{\pi} \varphi,$$

forcing an alignment of their geometric eigenstructures.

Theorem 2 (Consistency). Under regularity conditions, the error of the solution $\hat{\pi}$ from the empirical loss $\mathcal{L}_n(\pi)$ relative to the true optimal loss $\mathcal{L}(\pi)$ is bounded by:

$$\left| \mathcal{L}_n(\hat{\pi}) - \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_Y)} \mathcal{L}(\pi) \right| \leq \underbrace{\frac{8\alpha n}{(T+1)}}_{\text{Optimization error}} + C \underbrace{\left(W_2^{d_X}(\mathbb{P}_X, \hat{\mathbb{P}}_X) + W_2^{d_Y}(\mathbb{P}_Y, \hat{\mathbb{P}}_Y) \right)}_{\text{Statistical error}},$$

Research 002: Graphical Models under Data Contamination

Research question: Can we design a robust statistical algorithm to estimate causal structures using graphical models, given that the data often suffers from **measurement errors** and other forms of **contamination** in fields like biology, social science, and environmental science?

Contributions

- **Identifiability.** Propose two complementary sets of conditions that identify true causal graph up to its Markov equivalence class (MEC), even in the presence of data contamination.
 - ▶ **Condition 1 (Anchored-frugality):** Requires the Gaussian assumption on the true data distribution, but does not require prior knowledge of the contamination process.
 - ▶ **Condition 2 (Geometry-faithfulness):** Is distribution-free, but requires prior knowledge of the contamination process (e.g., the structure or type of noise).
- **Consistency.** Design consistent MEC learning algorithms.

Anchored Directed Acyclic Graphical (DAG) Models

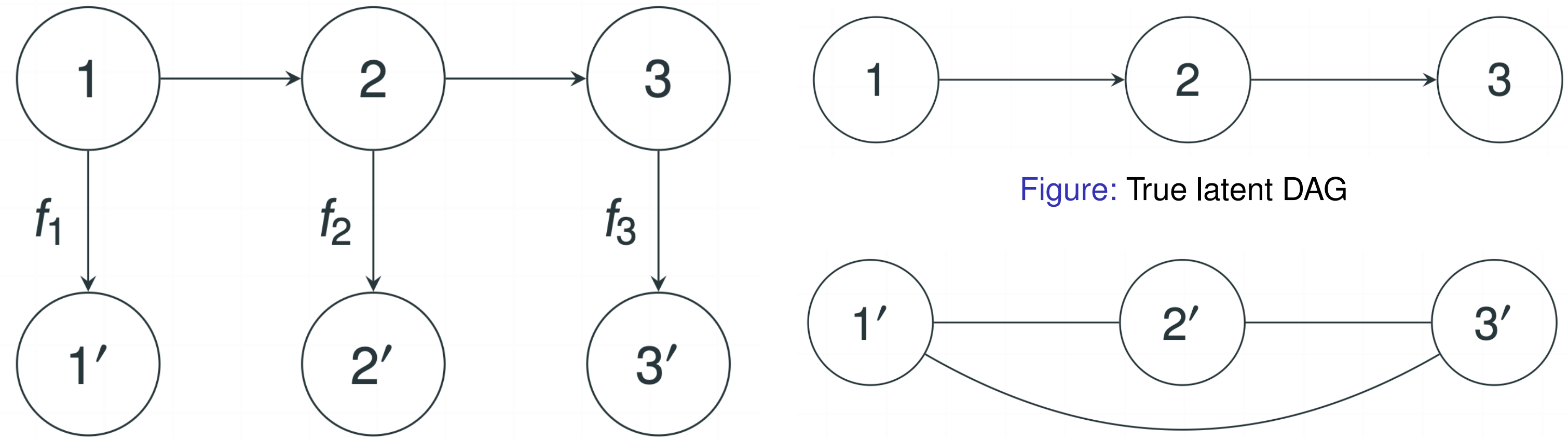


Figure: 3-node anchored DAG

Figure: True latent DAG

Figure: Contaminated graph

- Let $Z \in \mathbb{R}^d$ be a latent random vector generated by a linear structural equation model (SEM):

$$Z = BZ + E,$$

where B is the edge weight matrix, and E is a mean zero random vector with finite variance.

- We assume that B is strictly lower-triangular, excluding cyclic relationships within Z .
- In anchored graphical models, we do not observe Z directly, but rather its imperfect realizations, denoted by the observed random vector $X \in \mathbb{R}^d$.
- The relationship is defined element-wise:

$$X_j = f_j(Z_j), \quad \forall j \in \{1, \dots, d\},$$

where each f_j can be either deterministic or a stochastic mapping.

- Anchored DAG models encompass a wide range of contamination models:
 - ▶ **Additive measurement error models.** $X_j = Z_j + \Psi_j$ with $\mathbb{E}(\Psi_j) = 0$ and $\mathbb{E}(\Psi_j^2) < \infty$.
 - ▶ **Dropout models.** $X_j = \Psi_j Z_j$ with $\Psi_j \sim \text{Bernoulli}(p_j)$.
 - ▶ **Discretized models.** $X_j = \sum_{k=1}^K a_k I(Z_j \in S_k)$, where S_1, \dots, S_K form a partition of \mathbb{R} .

Identifiability

Condition 1 (Anchored-frugality). Let Z be Gaussian, and suppose that X is contaminated by additive measurement errors, such that its covariance matrix is $\Sigma^X = \Sigma^Z + \Sigma^{\Psi}$, where Σ^{Ψ} is diagonal. Among all possible corrections $\Sigma^X - \text{diag}(\eta^2) \in \mathcal{S}_{++}^d$, the graph induced by the resulting covariance matrix Σ^Z exhibits the sparsest structure. Here, \mathcal{S}_{++}^d is the set of $d \times d$ positive definite matrices.

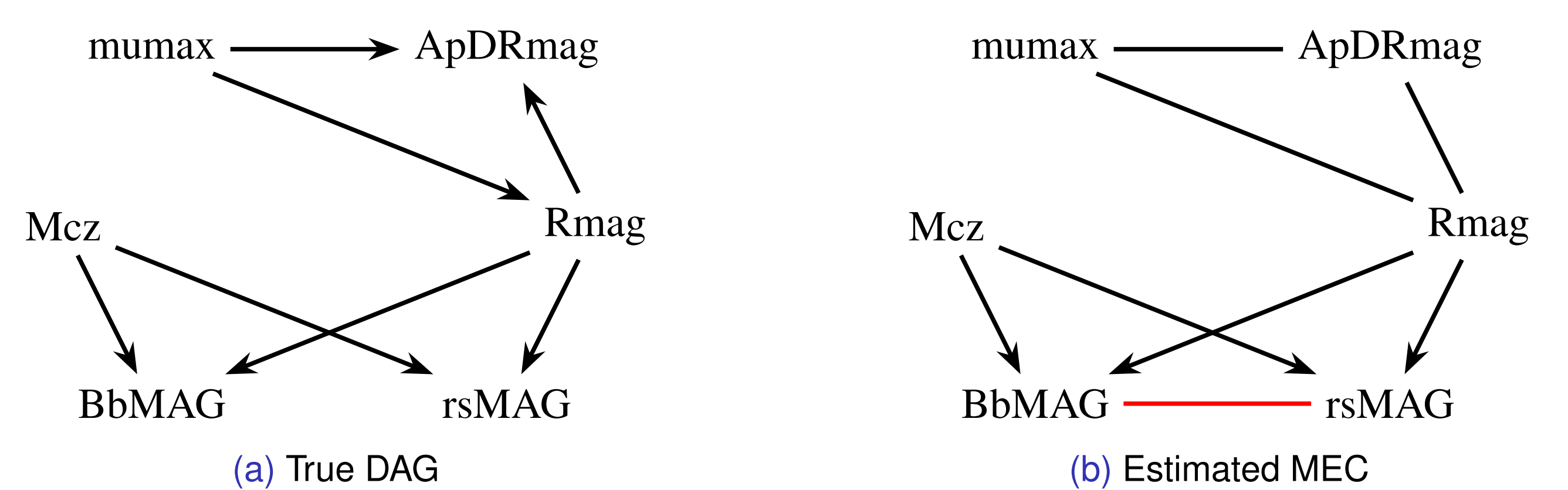
Condition 2 (Geometry-faithfulness). Assume that the latent covariance matrix Σ^Z can be recovered from the known moment relationships between X and Z . The geometry-faithfulness requires that the d-separation relationships between nodes perfectly encode the orthogonal relationships among the latent random vector Z , that is,

$$i \text{ and } j \text{ are d-separated by a set } S \iff Z_i - \Sigma_{iS}^Z (\Sigma_{SS}^Z)^{-1} Z_S \text{ and } Z_j - \Sigma_{jS}^Z (\Sigma_{SS}^Z)^{-1} Z_S \text{ are uncorrelated.}$$

- Anchored-frugality is deeply aligned with Occam's razor: among all candidate structures obtained after correcting for variability, the simplest one reveals the true relationships.
- Geometry-faithfulness replaces the conditional independence relationships in the standard faithfulness by linear orthogonality.
- Under linear SEMs, both conditions are valid except for a set of Lebesgue measure zero.

Theorem 1 (Identifiability). Under Condition 1 or 2, the latent graph is identifiable up to its MEC.

Real-World Application: Galaxy Brightness Measurements



(a) True DAG

(b) Estimated MEC

References

- Chung, J., Ahn, Y., Shin, D., & Park, G. (2025). Learning distribution-free anchored linear structural equation models in the presence of measurement error. *Journal of the Korean Statistical Society*.
- Shin, J., Chung, J., Hwang, S., & Park, G. (2025). Discovering causal structures in corrupted data: frugality in anchored Gaussian DAG models. *Computational Statistics & Data Analysis*.