

Causal Discovery with Mixed Latent Confounding via Precision Decomposition

Amir Asiaee

AMIR.ASIAEETAHERI@VUMC.ORG

Samhita Pal

SAMHITA.PAL@VUMC.ORG

*Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN 37232, USA*

James O’quinn

OQUINNJM@PROTON.ME

*Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA*

James P. Long

JPLONG@MDANDERSON.ORG

*Department of Biostatistics
MD Anderson Cancer Center
Houston, TX 77030, USA*

Abstract

We study causal discovery from observational data in linear Gaussian systems affected by *mixed latent confounding*, where some unobserved factors act broadly across many variables while others influence only small subsets. This setting is common in practice and poses a challenge for existing methods: differentiable and score-based DAG learners can misinterpret global latent effects as causal edges, while latent-variable graphical models recover only undirected structure.

We propose DCL-DECOR, a modular, precision-led pipeline that separates these roles. The method first isolates pervasive latent effects by decomposing the observed precision matrix into a structured component and a low-rank component. The structured component corresponds to the conditional distribution after accounting for pervasive confounders and retains only local dependence induced by the causal graph and localized confounding. A correlated-noise DAG learner is then applied to this deconfounded representation to recover directed edges while modeling remaining structured error correlations, followed by a simple reconciliation step to enforce bow-freeness.

We provide identifiability results that characterize the recoverable causal target under mixed confounding and show how the overall problem reduces to well-studied subproblems with modular guarantees. Synthetic experiments that vary the strength and dimensionality of pervasive confounding demonstrate consistent improvements in directed edge recovery over applying correlated-noise DAG learning directly to the confounded data.

Keywords: causal discovery, latent confounding, precision matrix, graphical models, identifiability, correlated noise, deconfounding

1. Introduction

Causal graphs are the stories we tell about systems we cannot intervene on. In the linear Gaussian setting, that story is often interrupted by latent confounders: unobserved variables that nudge many observables in concert (pervasive effects) or quietly perturb a few of them at a time (sparse, local effects). The result is familiar to practitioners: covariance that looks too global to be explained by

a small set of edges, and precision patterns that look too structured to be pure noise. If we ignore these latent forces, causal discovery tends to latch onto spurious correlations or hedge with large equivalence classes (Spirites et al., 2000). If we over-correct, we erase real signal.

Two types of confounding can generally arise. Pervasive confounding is driven by a small number of latent factors that load broadly across many measured variables (e.g., batch effects, global cell state, or shared environmental signals), and this regime has been extensively studied through sparse-plus-low-rank / latent-variable graphical models and approximate factor models (Chandrasekaran et al., 2012; Frot et al., 2019). In contrast, sparse (localized) confounding induces correlation only among a relatively small subset of variable pairs (e.g., pathway-specific hidden regulators or unmeasured co-regulators that affect only a few proteins/genes at a time), and while it is naturally represented using mixed-graph / correlated-error formalisms, it has received comparatively less systematic attention in scalable score-based causal discovery (Pal et al., 2025b). Our paper tackles *causal discovery with mixed confounding*: both *pervasive* low-rank effects and *sparse* low-rank effects co-exist. We work with linear Gaussian structural equation models (SEMs), $\mathbf{x} = \mathbf{B}^\top \mathbf{x} + \boldsymbol{\varepsilon}$, and model the exogenous noise as a homoskedastic idiosyncratic term plus two latent components: a dense low-rank factor $\mathbf{U}\mathbf{u}$ (few hidden causes that touch many nodes) and a column-sparse low-rank factor $\mathbf{V}\mathbf{v}$ (many hidden causes, each touching a small subset). This hybrid regime is common in practice: batch or device drift (\mathbf{U}) co-exists with pathway/module-specific influences (\mathbf{V}) in biology, finance, and recommendation.

Why mixed confounding is hard. In linear Gaussian models, latent variables generally render the DAG unidentifiable from a single observational environment (Spirites et al., 2000). Classical methods that assume causal sufficiency (no hidden nodes) return only a Markov-equivalence class (Chickering, 2002). Non-Gaussian or nonlinear assumptions can sometimes restore identifiability (Shimizu et al., 2006; Hoyer et al., 2009), but our focus is the strictly Gaussian, linear case with both pervasive and sparse confounders. In this regime, undirected analogs have long exploited sparse+low-rank decompositions of precision or covariance to separate conditional structure from latent factors (Chandrasekaran et al., 2012), yet turning those decompositions into *directed* graphs that are robust to mixed confounding is non-trivial.

Our view: deconfound in precision, learn in data. We develop a modular pipeline that begins where latent-variable graphical modeling is strongest *in the precision* and ends where modern continuous directed acyclic graph (DAG) learning thrives *on per-sample residuals*. First, we decompose the observed precision $\boldsymbol{\Theta}$ into a sparse SPD matrix \mathbf{S} and a positive semidefinite low-rank matrix \mathbf{L} (so $\boldsymbol{\Theta} \approx \mathbf{S} - \mathbf{L}$). At the population level, \mathbf{L} captures the pervasive component, while \mathbf{S} encodes the DAG convolved with the remaining (sparse) confounding. We then invert \mathbf{S} to obtain a conditioned (pervasive-adjusted) covariance estimate, and apply the DECOR-GL algorithm Pal et al. (2025b) to separate the remaining confounding from directed effects and learn a weighted DAG.

A structural condition that pays off. The key structural assumption underpinning our decomposition and conditioning is that the three components of our noise model are mutually independent. Not only does this assumption translate to additivity in the relevant matrix identities, but, more importantly, it enables clean removal of pervasive effects when we condition on the pervasive component.

Contributions.

- **D–C–L mixed-confounding formulation and precision decomposition.** We introduce a three-component Gaussian noise model composed of mutually independent (i) diagonal noise, (ii) localized low-rank confounders, and (iii) pervasive low-rank confounders. From that, we derive a D–C–L decomposition of the observed precision into a structured sparse component and a low-rank component.
- **Structural characterization of the components.** We prove that the pervasive component is PSD and low-rank and that the non-pervasive component is exactly local under disjoint confounding support and approximately local under controlled overlap. We further characterize how this locality translates to the population-level precision induced by the model.
- **Precision-led deconfounding pipeline (DCL–DECOR).** We propose a three-stage pipeline:
 1. estimate the structured–low-rank precision split via latent-variable graphical lasso with a configurable locality regularizer;
 2. invert the structured component to obtain a pervasive-adjusted (conditional) covariance estimate;
 3. run a correlated-noise continuous DAG learner (DECOR-GL) that jointly learns the DAG and the residual error precision, followed by a simple post-hoc bow reconciliation rule to enforce bow-freeness in the output.
- **Identifiability guarantees.** We show that, under standard transversality assumptions, the conditioned precision is identifiable from the population precision. Under additional mild conditions, we further identify the minimal bow-free target determined by the conditioned model; bow-freeness is a standard structural condition under which linear Gaussian SEM parameters become identifiable from observational covariances and has also been leveraged in recent bow-free covariance search procedures (Drton et al., 2011; Grassi and Tarantino, 2024).

Positioning with prior art. Constraint-based and score-based methods provide the classical backdrop (Spirtes et al., 2000; Chickering, 2002). Continuous formulations such as NOTEARS and its descendants bring differentiability and scalability to DAG learning (Zheng et al., 2018). Latent-variable graphical modeling separates sparse conditional structure from low-rank latent effects in undirected models (Chandrasekaran et al., 2012). Our contribution is to *bridge* these: use the precision domain to isolate pervasive directions and then invoke a DAG learner that explicitly models the remaining sparse confounding (Pal et al., 2025b). The result is a practical, provably grounded route to causal discovery in the linear Gaussian mixed-confounding regime—without interventions and without stepping outside the Gaussian world.

2. Related Work

Classical approaches without latent confounders. The modern literature on causal discovery for observational data begins with constraint-based procedures that leverage conditional independence (CI) tests to recover a Markov-equivalence class of DAGs. The PC algorithm and its variants are consistent under appropriate Markov and faithfulness assumptions when all relevant variables are observed (Spirtes et al., 2000). Score-based search supplies a complementary view: Greedy Equivalence Search (GES) optimizes a penalized likelihood (e.g., BIC) over equivalence classes of DAGs and is consistent in large samples (Chickering, 2002). These families are powerful under *causal sufficiency* but do not resolve directions within an equivalence class and typically degrade in the presence of unobserved confounding.

Latent variables and mixed graphs. When some causes are unobserved, the induced conditional independences on the observed margin can be represented by acyclic directed mixed graphs (ADMGs). Fast Causal Inference (FCI) and its relatives return a partial ancestral graph (PAG) encoding an equivalence class of ADMGs that is sound in the presence of hidden variables and selection bias (Spirtes et al., 2000; Spirtes, 2001; Richardson and Spirtes, 2002). For high-dimensional regimes and partial observability, RFCI improves sample efficiency by reducing the size of conditioning sets (Colombo et al., 2012). Hybrid methods like GFCI (Ogarrio et al., 2016) and PFCI (Pal et al., 2025a) combine a score step with CI testing and provide consistency guarantees while often improving empirical accuracy. Although these methods accommodate latent confounding, they generally leave many edges unoriented from a single observational environment in linear Gaussian settings.

Continuous optimization for DAG learning. A major development is to replace combinatorial search with continuous objectives that encode acyclicity via smooth constraints. NOTEARS uses a trace-exponential characterization to enforce DAGness and optimizes a penalized Gaussian least-squares score by gradient methods (Zheng et al., 2018). Subsequent work broadened the modeling scope and improved scaling: nonparametric mechanisms and nonlinear SEMs (Zheng et al., 2020), time-series structure via DYNOTEARS (Pamfil et al., 2020), likelihood-based scores in GOLEM (Ng et al., 2020), and log-determinant-based acyclicity in DAGMA (Bello et al., 2022). While these methods scale gracefully and often achieve strong accuracy on curated benchmarks, they typically assume independent errors and can be sensitive to latent confounding.

Differentiable discovery under latent confounding. A line of work incorporates latent confounding directly into continuous formulations. Bhattacharya et al. (2021) derive differentiable algebraic constraints for ADMGs and optimize a likelihood subject to mixed-graph structure, enabling discovery of bidirected edges. This treats latent dependence largely as *dense* nuisance covariance (flexible but potentially under-identified) and does not exploit structured decompositions of the noise that enable identifiability in linear Gaussian models from a single environment. Other recent works in this direction include (Prashant et al., 2024; Ma et al., 2024).

Low-rank + sparse decompositions for latent structure (undirected). In Gaussian graphical modeling, a now-standard approach separates conditional structure from shared latent effects by decomposing the *precision* matrix as a sparse part minus a low-rank positive semidefinite part. The convex latent-variable graphical lasso of Chandrasekaran et al. (2012) provides identifiability and consistency under incoherence and transversality conditions. Related matrix decompositions—most prominently robust PCA (Candès et al., 2011)—formalize when a low-rank component can be separated from a sparse component. On the covariance side, factor-plus-sparse estimators such as POET (Fan et al., 2013) and the graphical lasso (Friedman et al., 2008) support scalable estimation in high dimensions. These ideas, however, target undirected structure and do not, by themselves, orient edges.

Non-Gaussian and nonlinear identifiability. Stronger distributional assumptions can break the observational symmetry. In linear non-Gaussian SEMs, LiNGAM identifies a unique ordering and orientation from a single environment (Shimizu et al., 2006). Nonlinear additive-noise models can also be identifiable by exploiting asymmetries in functional mechanisms (Hoyer et al., 2009). With latent confounding, recent work shows that in linear *non-Gaussian* models satisfying bow-free restrictions, the exact causal graph is identifiable (Wang and Drton, 2023). These advances are compelling but move outside the linear Gaussian setting that underlies many pipelines in practice.

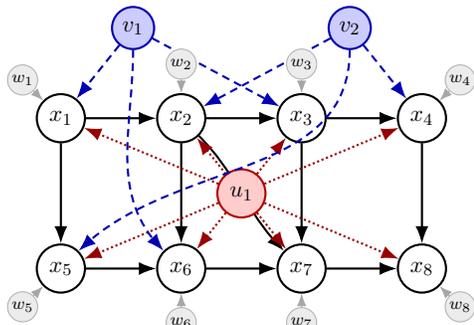


Figure 1: The D-C-L noise model. Observed variables x_1, \dots, x_8 follow a DAG (solid arrows). Each receives idiosyncratic noise w_i (gray). Localized confounders v_j (blue, dashed) each affect three variables with no direct DAG edges among them (bow-free). The pervasive confounder u_1 (red, dotted) in the center affects all variables, inducing low-rank correction \mathbf{L}_ε .

Position of the present work. Our setting is linear Gaussian with both *pervasive* (dense low-rank) and *sparse* low-rank confounding. We build on latent-variable precision decompositions from the undirected literature (Chandrasekaran et al., 2012; Fan et al., 2013; Friedman et al., 2008; Candès et al., 2011) and marry them with differentiable DAG learning (Zheng et al., 2018; Ng et al., 2020; Bello et al., 2022; Pamfil et al., 2020). The key ingredient is to use the learned structured component of the precision to form a pervasive-adjusted covariance estimate, after which a DAG learner that explicitly models *sparse* confounding can be applied (Pal et al., 2025b). This yields identifiability and a practical algorithm under homoskedastic idiosyncratic noise, incoherence of pervasive factors, bounded-degree sparsity, and a mild separated-touch (or dominance) condition—while remaining within the linear Gaussian world and a single observational environment.

3. Problem Setup

We consider a linear Gaussian structural equation model (SEM) on p observed variables $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbf{x} = \mathbf{B}^\top \mathbf{x} + \boldsymbol{\varepsilon}, \quad \mathbf{T} := \mathbf{I} - \mathbf{B}, \quad (1)$$

where \mathbf{B} encodes a directed acyclic graph (DAG) and \mathbf{T} is invertible under some causal ordering (unit-diagonal and triangular in that order), with $\det(\mathbf{T}) = 1$. Writing $\mathbf{x} = \mathbf{T}^{-\top} \boldsymbol{\varepsilon}$, all randomness is in the exogenous noise $\boldsymbol{\varepsilon}$. The population covariance and precision of \mathbf{x} are $\boldsymbol{\Sigma} = \mathbf{T}^{-\top} \boldsymbol{\Omega} \mathbf{T}^{-1}$, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = \mathbf{T} \boldsymbol{\Omega}^{-1} \mathbf{T}^\top$.

Subscript convention (noise vs. observed level). We use a subscript ε for quantities defined at the *noise level* (i.e., functions of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$) and a subscript x for their *observed-level* counterparts (i.e., functions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$). The two levels are related by congruence through \mathbf{T} : $\mathbf{M}_x := \mathbf{T} \mathbf{M}_\varepsilon \mathbf{T}^\top$ for any noise-level matrix \mathbf{M}_ε .

3.1. Noise Model and Mixed Confounding

In many real-world systems, unobserved confounders exhibit heterogeneous structure: some affect only small subsets of variables while others influence nearly all measurements. For instance, in gene expression, microRNAs can induce localized correlations across targeted gene sets, whereas chromatin state can act pervasively across the genome.

To capture this heterogeneity, we model the noise as a sum of three independent components,

$$\boldsymbol{\varepsilon} = \mathbf{W}\mathbf{w} + \mathbf{V}\mathbf{v} + \mathbf{U}\mathbf{u}, \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_S}), \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_L}), \quad (2)$$

where $\mathbf{W}, \mathbf{V}, \mathbf{U}$ encode how each type of noise acts on the observed variables. The diagonal matrix $\mathbf{W} \in \mathbb{R}^{p \times p}$ represents heteroskedastic idiosyncratic noise at each node. The matrix $\mathbf{V} \in \mathbb{R}^{p \times r_S}$

captures *localized confounding*: each column $\mathbf{V}_{\cdot j}$ has small support (e.g., $|\text{supp}(\mathbf{V}_{\cdot j})| \leq s \ll p$). Finally, $\mathbf{U} \in \mathbb{R}^{p \times r_L}$ captures *pervasive confounding* with dense loadings and $r_L \ll p$.

The noise covariance is $\boldsymbol{\Omega} = \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top + \mathbf{U}\mathbf{U}^\top$.

3.2. Covariance, Precision, and a D–C–L Decomposition

Our first step is to analyze the structure of $\boldsymbol{\Omega}^{-1}$. Define the diagonal idiosyncratic precision $\mathbf{D}_\varepsilon := (\mathbf{W}\mathbf{W}^\top)^{-1} = \text{diag}(d_1, \dots, d_p)$, $d_i > 0$, and the overlap matrix $\mathbf{A} := \mathbf{I} + \mathbf{V}^\top \mathbf{D}_\varepsilon \mathbf{V} \in \mathbb{R}^{r_S \times r_S}$. Applying the Sherman–Morrison–Woodbury (SMW) identity to $\mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top$ yields the *structured non-pervasive precision*

$$\mathbf{S}_\varepsilon := (\mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top)^{-1} = \mathbf{D}_\varepsilon - \underbrace{\mathbf{D}_\varepsilon \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top \mathbf{D}_\varepsilon}_{\mathbf{C}_\varepsilon}.$$

Here \mathbf{C}_ε (for Coupling) collects the pairwise dependencies induced by localized confounders. Importantly, the *structured* object we aim to exploit is $\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon - \mathbf{C}_\varepsilon$, which need not be strictly sparse: depending on the geometry of confounder supports, \mathbf{S}_ε can be row-sparse, banded, block-diagonal, etc. (see Remark 1 and Appendix A.4).

Now write $\boldsymbol{\Omega} = \mathbf{S}_\varepsilon^{-1} + \mathbf{U}\mathbf{U}^\top$ and apply SMW again. Defining $\mathbf{L}_\varepsilon := \mathbf{S}_\varepsilon \mathbf{U} (\mathbf{I} + \mathbf{U}^\top \mathbf{S}_\varepsilon \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{S}_\varepsilon$, the noise precision admits the D–C–L decomposition

$$\boldsymbol{\Omega}^{-1} = \underbrace{\mathbf{D}_\varepsilon}_{\text{diagonal}} - \underbrace{\mathbf{C}_\varepsilon}_{\text{localized coupling}} - \underbrace{\mathbf{L}_\varepsilon}_{\text{pervasive (low-rank)}} = \mathbf{S}_\varepsilon - \mathbf{L}_\varepsilon. \quad (3)$$

This decomposition is not directly observable because we do not observe $\boldsymbol{\Omega}$ itself. Crucially, however, the same split appears at the observed level.

Observed-level decomposition. Congruencing (3) by \mathbf{T} yields

$$\boldsymbol{\Theta} = \mathbf{T} \boldsymbol{\Omega}^{-1} \mathbf{T}^\top = \underbrace{\mathbf{T} \mathbf{D}_\varepsilon \mathbf{T}^\top}_{\mathbf{D}_x} - \underbrace{\mathbf{T} \mathbf{C}_\varepsilon \mathbf{T}^\top}_{\mathbf{C}_x} - \underbrace{\mathbf{T} \mathbf{L}_\varepsilon \mathbf{T}^\top}_{\mathbf{L}_x} = \underbrace{\mathbf{T} \mathbf{S}_\varepsilon \mathbf{T}^\top}_{\mathbf{S}_x} - \mathbf{L}_x. \quad (4)$$

The compact form $\boldsymbol{\Theta} = \mathbf{S}_x - \mathbf{L}_x$ matches the standard *structured-minus-low-rank* precision model used in latent graphical lasso approaches (Chandrasekaran et al., 2012). This alignment motivates our estimation strategy: we estimate $(\mathbf{S}_x, \mathbf{L}_x)$ from data, then use \mathbf{S}_x (after deconfounding) to recover the DAG structure.

Remark 1 (Beyond sparsity: other local structure classes) *While much of our algorithmic development focuses on row-sparse structure (amenable to ℓ_1 -regularization), the same decomposition supports other “local” structures for \mathbf{S}_ε , such as banded or block-diagonal patterns induced by contiguous or group-confined confounder supports. The key requirement is that the structure class be (approximately) preserved under \mathbf{T} -congruence; see Proposition 10 in Appendix A.4.*

3.3. Relation to existing formulations

The components in (3) recover several settings studied previously:

- **Independent errors (D only).** If $\mathbf{V} = \mathbf{U} = \mathbf{0}$, then $\mathbf{C}_\varepsilon = \mathbf{0}$, $\mathbf{L}_\varepsilon = \mathbf{0}$, and $\boldsymbol{\Theta} = \mathbf{D}_x = \mathbf{T} \mathbf{D}_\varepsilon \mathbf{T}^\top$. This reduces to a standard SEM with heteroskedastic but uncorrelated errors; Loh and Bühlmann (2014) relate the support of \mathbf{D}_x to the moralized graph.

- **Pervasive confounding (D–L).** If $\mathbf{V} = \mathbf{0}$, then $\mathbf{C}_\varepsilon = \mathbf{0}$, $\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon$, and $\Theta = \mathbf{D}_x - \mathbf{L}_x$, the sparse-minus-low-rank regime central to latent-variable Gaussian graphical models (Chandrasekaran et al., 2012) and confounded DAG learning methods such as Frot et al. (2019); Agrawal et al. (2023).
- **Localized confounding (D–C).** If $\mathbf{U} = \mathbf{0}$, then $\mathbf{L}_\varepsilon = \mathbf{0}$ and $\Theta = \mathbf{S}_x = \mathbf{D}_x - \mathbf{C}_x$, corresponding to a DAG with localized (structured) error coupling. This setting is closely related to recent work such as DAG-DECOR (Pal et al., 2025b).
- **Mixed confounding (D–C–L, this work).** In the full model, both localized coupling and pervasive latent factors co-occur, yielding $\Theta = \mathbf{S}_x - \mathbf{L}_x$ with $\mathbf{S}_x = \mathbf{D}_x - \mathbf{C}_x$. Our goal is to develop theory and algorithms that handle this combined regime.

3.4. Structural properties of precision matrix components

We summarize the key properties that underpin our estimation procedure; full statements and proofs are given in Appendix A.

Proposition 2 (Low-rankness of \mathbf{L}_ε and locality of \mathbf{S}_ε) *Assume the model in §3 with $\mathbf{T} = \mathbf{I} - \mathbf{B}$ unit-diagonal and triangular under some causal order, and define $\mathbf{D}_\varepsilon, \mathbf{C}_\varepsilon, \mathbf{S}_\varepsilon, \mathbf{L}_\varepsilon$ as above.*

- Low-rank, PSD pervasive correction.** $\mathbf{L}_\varepsilon \succeq \mathbf{0}$ and $\text{rank}(\mathbf{L}_\varepsilon) \leq r_L$. Consequently, $\mathbf{L}_x = \mathbf{T}\mathbf{L}_\varepsilon\mathbf{T}^\top \succeq \mathbf{0}$ with $\text{rank}(\mathbf{L}_x) \leq r_L$.
- Exact locality under disjoint supports.** If the columns of \mathbf{V} have disjoint supports of size at most s , and each variable participates in at most c such supports, then each row of \mathbf{S}_ε has at most $c(s-1)$ off-diagonal nonzeros (equivalently, \mathbf{C}_ε is supported on a union of small cliques).
- Controlled leakage under overlap.** If the confounder supports overlap mildly, then \mathbf{S}_ε remains approximately local: each row has only $\mathcal{O}(cs)$ entries above a fixed threshold, provided the overlap-induced leakage is controlled. A sufficient condition is given by Proposition 9 (Appendix A.3).
- Propagation through the DAG.** If the DAG \mathbf{B} has bounded degree, then $\mathbf{S}_x = \mathbf{T}\mathbf{S}_\varepsilon\mathbf{T}^\top$ preserves locality up to a constant inflation factor (row-sparsity in the sparse case; bandedness/block-structure in ordered or grouped settings as in Proposition 10).

4. From Precision Decomposition to a Deconfounding Algorithm

At the population level, §3 implies an observed-level precision decomposition

$$\Theta = \Sigma^{-1} = \underbrace{\mathbf{T}\mathbf{S}_\varepsilon\mathbf{T}^\top}_{\mathbf{S}_x} - \underbrace{\mathbf{T}\mathbf{L}_\varepsilon\mathbf{T}^\top}_{\mathbf{L}_x}, \quad \mathbf{S}_\varepsilon = (\mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top)^{-1} = \Gamma_\varepsilon^{-1}, \quad (5)$$

where $\mathbf{L}_x \succeq \mathbf{0}$ has rank at most r_L , while \mathbf{S}_x inherits a *local* structure from \mathbf{S}_ε under bounded-degree \mathbf{T} -congruence (Proposition 2). The decomposition (5) has a direct probabilistic interpretation: \mathbf{S}_x is the precision matrix of \mathbf{x} after conditioning on the pervasive factors \mathbf{u} .

Proposition 3 (Conditional precision) *Under (1)–(2), let $\Gamma_\varepsilon := \mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top$ and $\mathbf{S}_\varepsilon = \Gamma_\varepsilon^{-1}$. Then $\text{Cov}(\mathbf{x} \mid \mathbf{u}) = \mathbf{T}^{-\top}\Gamma_\varepsilon\mathbf{T}^{-1}$, $\text{Cov}(\mathbf{x} \mid \mathbf{u})^{-1} = \mathbf{T}\mathbf{S}_\varepsilon\mathbf{T}^\top = \mathbf{S}_x$.*

Proposition 3 reduces mixed confounding to a correlated-noise SEM: after conditioning on \mathbf{u} , the DAG \mathbf{B} is unchanged and the remaining noise covariance $\mathbf{\Gamma}_\varepsilon$ encodes only idiosyncratic and localized confounding.

4.1. A three-stage pipeline

We operationalize the reduction in three stages: (I) estimate the structured–low-rank precision split, (II) invert the structured component to obtain the deconfounded covariance, and (III) learn the DAG via a correlated-noise estimator.

Stage I: Structured–low-rank precision split. Given the sample covariance $\widehat{\Sigma} = n^{-1}\mathbf{X}^\top\mathbf{X}$, we estimate $(\mathbf{S}_x, \mathbf{L}_x)$ via the latent-variable graphical lasso (Chandrasekaran et al., 2012):

$$(\widehat{\mathbf{S}}_x, \widehat{\mathbf{L}}_x) \in \arg \min_{\substack{\mathbf{S} > 0, \mathbf{L} \succeq 0 \\ \mathbf{S} - \mathbf{L} \succ 0}} \left\{ -\log \det(\mathbf{S} - \mathbf{L}) + \text{tr}(\widehat{\Sigma}(\mathbf{S} - \mathbf{L})) + \lambda_s \mathcal{R}_{\text{loc}}(\mathbf{S}) + \lambda_* \text{tr}(\mathbf{L}) \right\}, \quad (6)$$

where \mathcal{R}_{loc} is a convex regularizer encoding the chosen locality class (e.g., $\|\mathbf{S}\|_{1,\text{off}}$ for sparsity; see Remark 1 for banded and block-sparse alternatives).

Stage II: Deconfounded covariance. By Proposition 3, the conditional covariance $\Sigma_{\text{cond}} = \text{Cov}(\mathbf{x} \mid \mathbf{u})$ equals \mathbf{S}_x^{-1} . We estimate it by inverting the structured component: $\widehat{\Sigma}_{\text{cond}} := \widehat{\mathbf{S}}_x^{-1}$. This inversion is computed via sparse Cholesky factorization and triangular solves, exploiting the local structure of $\widehat{\mathbf{S}}_x$ without forming an explicit dense inverse.

Stage III: Correlated-noise DAG learning. By Proposition 3, conditioning on \mathbf{u} removes pervasive confounding at the second-moment level. The *conditional residual* $\mathbf{x}^\perp := \mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{u}]$ obeys the same linear SEM with *correlated* errors driven only by idiosyncratic and localized confounding:

$$\mathbf{x}^\perp = \mathbf{B}^\top \mathbf{x}^\perp + \boldsymbol{\varepsilon}^\perp, \quad \boldsymbol{\varepsilon}^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_\varepsilon), \quad \Sigma_{\text{cond}} = \mathbf{T}^{-\top} \mathbf{\Gamma}_\varepsilon \mathbf{T}^{-1}, \quad (7)$$

where $\mathbf{T} = \mathbf{I} - \mathbf{B}$ and $\mathbf{S}_\varepsilon = \mathbf{\Gamma}_\varepsilon^{-1}$.

We instantiate Stage III with DECOR-GL (Pal et al., 2025b), a precision-penalized variant that directly enforces sparsity in \mathbf{S}_ε :

$$\min_{\mathbf{B}, \mathbf{S}_\varepsilon > 0} \underbrace{\text{tr}(\widehat{\Sigma}_{\text{cond}} \mathbf{T} \mathbf{S}_\varepsilon \mathbf{T}^\top)}_{\text{negative log-likelihood}} - \log \det \mathbf{S}_\varepsilon + \lambda_B \|\mathbf{B}\|_1 + \lambda_S \|\mathbf{S}_\varepsilon\|_{1,\text{off}} + \rho h(\mathbf{B}), \quad (8)$$

where $h(\mathbf{B}) = \text{tr}(e^{\mathbf{B} \odot \mathbf{B}}) - p$ is the smooth acyclicity surrogate (Zheng et al., 2018). For any acyclic \mathbf{B} , \mathbf{T} is unit-triangular with $\det(\mathbf{T}) = 1$, so no log-determinant term in \mathbf{T} appears.

Remark 4 (Connection to NOTEARS) Interpreting $\widehat{\Sigma}_{\text{cond}}$ as the empirical covariance of deconfounded samples $\mathbf{X}^\perp \in \mathbb{R}^{n \times p}$, the trace term satisfies $\text{tr}(\widehat{\Sigma}_{\text{cond}} (\mathbf{I} - \mathbf{B}) \mathbf{S}_\varepsilon (\mathbf{I} - \mathbf{B})^\top) \approx \frac{1}{n} \|(\mathbf{X}^\perp - \mathbf{X}^\perp \mathbf{B}) \mathbf{S}_\varepsilon^{1/2}\|_F^2$. When $\mathbf{S}_\varepsilon = \mathbf{I}$ (independent errors), the \mathbf{B} -update reduces to NOTEARS. Stage III thus generalizes NOTEARS to correlated-error SEMs, where the residual whitening matrix is learned jointly.

Algorithm 1: DCL-DECOR: Deconfounding via D–C–L Decomposition

Input: Data $\mathbf{X} \in \mathbb{R}^{n \times p}$; penalties $\lambda_S, \lambda_*, \lambda_B, \lambda_S$; thresholds τ_B, τ_Γ ; reconciliation constant $c > 0$.

Stage I: Structured–low-rank precision split

Compute $\widehat{\Sigma} \leftarrow n^{-1} \mathbf{X}^\top \mathbf{X}$
 Solve LVGLASSO (6) for $(\widehat{\mathbf{S}}_x, \widehat{\mathbf{L}}_x)$

Stage II: Deconfounded covariance

Compute $\widehat{\Sigma}_{\text{cond}} \leftarrow \widehat{\mathbf{S}}_x^{-1}$ via Cholesky factorization

Stage III: DECOR-GL

Initialize $\mathbf{S}_\varepsilon^{(0)} \leftarrow \text{diag}(\widehat{\mathbf{S}}_x)$, $\mathbf{B}^{(0)} \leftarrow \mathbf{0}$

repeat

Graph step: Update $\mathbf{B}^{(k+1)}$ via (9) (proximal augmented Lagrangian)

Noise step: Form $\widehat{\Sigma}_B \leftarrow (\mathbf{T}^{(k+1)})^\top \widehat{\Sigma}_{\text{cond}} \mathbf{T}^{(k+1)}$; solve (10) for $\mathbf{S}_\varepsilon^{(k+1)}$

until convergence at iteration t

Post-processing:

Hard-threshold: $\widetilde{\mathbf{B}} \leftarrow \text{HT}(\mathbf{B}^{(t)}; \tau_B)$, $\widetilde{\Gamma}_\varepsilon \leftarrow \text{HT}_{\text{off}}((\mathbf{S}_\varepsilon^{(t)})^{-1}; \tau_\Gamma)$

Bow reconciliation: $(\widehat{\mathbf{B}}, \widehat{\Gamma}_\varepsilon) \leftarrow \text{apply (11) to } (\widetilde{\mathbf{B}}, \widetilde{\Gamma}_\varepsilon)$ for all conflicting pairs

Output: DAG estimate $\widehat{\mathbf{B}}$; structured noise covariance $\widehat{\Gamma}_\varepsilon$.

DECOR-GL alternation. Problem (8) is solved by alternating updates:

- **Graph step (given \mathbf{S}_ε).** With \mathbf{S}_ε fixed, update \mathbf{B} by proximal-gradient descent on

$$\min_{\mathbf{B}} \text{tr}(\widehat{\Sigma}_{\text{cond}} (\mathbf{I} - \mathbf{B}) \mathbf{S}_\varepsilon (\mathbf{I} - \mathbf{B})^\top) + \lambda_B \|\mathbf{B}\|_1 + \rho h(\mathbf{B}), \quad (9)$$

using an augmented-Lagrangian schedule for $h(\mathbf{B}) \approx 0$ as in NOTEARS.

- **Noise step (given \mathbf{B}).** With \mathbf{B} fixed, form the residual covariance $\widehat{\Sigma}_B := \mathbf{T}^\top \widehat{\Sigma}_{\text{cond}} \mathbf{T}$ and solve the graphical lasso

$$\widehat{\mathbf{S}}_\varepsilon \in \arg \min_{\mathbf{S} \succ 0} \left\{ -\log \det \mathbf{S} + \text{tr}(\widehat{\Sigma}_B \mathbf{S}) + \lambda_S \|\mathbf{S}\|_{1, \text{off}} \right\}. \quad (10)$$

Post-hoc bow reconciliation. A *bow* occurs when a pair (i, j) has both a directed edge ($B_{ij} \neq 0$ or $B_{ji} \neq 0$) and a bidirected edge ($[\Gamma_\varepsilon]_{ij} \neq 0$). Bows create fundamental non-identifiability even in two-node models (see §5.3). Following Pal et al. (2025b), we apply a post-hoc reconciliation step after convergence at iteration t . For each pair (i, j) with both edge types present after thresholding, we compare the directed signal strength against the normalized error correlation:

$$\text{keep directed edge} \iff \max\{|B_{ij}^{(t)}|, |B_{ji}^{(t)}|\} \geq c \cdot |\Gamma_{\varepsilon, ij}^{(t)}| / \sqrt{\Gamma_{\varepsilon, ii}^{(t)} \Gamma_{\varepsilon, jj}^{(t)}}, \quad (11)$$

where $c > 0$ is a tuning constant (we use $c = 1$). The weaker channel is zeroed, ensuring bow-freeness in the final output without complicating the optimization.

5. Identifiability and Modular Guarantees

This section establishes what is identifiable from the observed covariance Σ in the D–C–L model and characterizes the recovery target for Stage III once pervasive confounding has been removed.

5.1. Identifiability of the structured–low-rank split

The observed precision satisfies $\Theta = \mathbf{S}_x - \mathbf{L}_x$ with $\mathbf{L}_x \succeq 0$ low-rank and \mathbf{S}_x in a local structure class (sparse, banded, block-sparse, etc.). Under standard transversality conditions between the structured and low-rank tangent cones (Chandrasekaran et al., 2012), the pair $(\mathbf{S}_x, \mathbf{L}_x)$ is identifiable from Θ .

Assumption 1 (Transversality) *The tangent cones of the structured class \mathcal{S} and the low-rank manifold \mathcal{L}_{r_L} at the true parameters intersect trivially: $T_{\mathcal{S}}(\mathbf{S}_x) \cap T_{\mathcal{L}_{r_L}}(\mathbf{L}_x) = \{\mathbf{0}\}$.*

Under Assumption 1 LVGLASSO consistently recovers $(\mathbf{S}_x, \mathbf{L}_x)$, yielding estimation error $\delta_{S,n} := \|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2 \rightarrow 0$ as $n \rightarrow \infty$ (Chandrasekaran et al., 2012).

5.2. Stability of the conditional covariance

The inversion $\Sigma_{\text{cond}} = \mathbf{S}_x^{-1}$ is Lipschitz-stable on well-conditioned SPD matrices:

Lemma 5 (Inversion perturbation bound) *If $\|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2 < \lambda_{\min}(\mathbf{S}_x)$, then $\|\widehat{\Sigma}_{\text{cond}} - \Sigma_{\text{cond}}\|_2 \leq \frac{\|\mathbf{S}_x^{-1}\|_2^2 \cdot \|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2}{1 - \|\mathbf{S}_x^{-1}\|_2 \cdot \|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2}$.*

The proof follows from standard matrix perturbation theory (Appendix B.2). In words: if Stage I achieves small error $\delta_{S,n}$ and \mathbf{S}_x is well-conditioned (bounded condition number $\kappa(\mathbf{S}_x) = \lambda_{\max}(\mathbf{S}_x)/\lambda_{\min}(\mathbf{S}_x)$), then $\widehat{\Sigma}_{\text{cond}}$ is a stable estimate of Σ_{cond} .

5.3. Identifiability from the conditional covariance

After conditioning on \mathbf{u} , the model (7) is a linear Gaussian SEM with correlated errors. Such models are represented as *acyclic directed mixed graphs* (ADMGs) $G = (V, E^{\rightarrow}, E^{\leftrightarrow})$, where $E^{\rightarrow} = \text{supp}(\mathbf{B})$ encodes directed edges and $E^{\leftrightarrow} = \text{supp}_{\text{off}}(\Gamma_{\varepsilon})$ encodes bidirected edges from error correlation.

Bow obstruction. A *bow* on pair (i, j) occurs when both a directed edge ($B_{ij} \neq 0$ or $B_{ji} \neq 0$) and a bidirected edge ($[\Gamma_{\varepsilon}]_{ij} \neq 0$) are present. Bows create fundamental non-identifiability that motivates the bow-free condition at the conditional level: $\text{supp}(\mathbf{B}) \cap \text{supp}_{\text{off}}(\Gamma_{\varepsilon}) = \emptyset$.

Connection to the D–C–L model. In terms of the loading matrix \mathbf{V} , bow-freeness requires that no localized confounder simultaneously affects a parent-child pair in the DAG. Formally, for each column \mathbf{v}_k of \mathbf{V} : if $v_{ik} \neq 0$ and $v_{jk} \neq 0$, then $B_{ij} = B_{ji} = 0$ (Figure 1.)

Identifiability within a fixed bow-free ADMG. For a fixed bow-free acyclic mixed graph G , Drton et al. (2011) established that the covariance map $(\mathbf{B}, \Gamma_{\varepsilon}) \mapsto \Sigma_{\text{cond}}$ is injective on the parameter space restricted to G 's zero pattern, provided a uniform eigenvalue margin $\Gamma_{\varepsilon} \succeq m\mathbf{I}$ for some $m > 0$. This supplies the identifiability foundation for Stage III.

Identifiability across graphs: bow-free equivalence classes. Even when parameters are identifiable *given* a graph, the same Σ_{cond} can arise from multiple distinct bow-free ADMGs (distributional equivalence). Let $\mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}}) := \{(\mathbf{B}, \Gamma_{\varepsilon}) : \mathbf{B} \text{ acyclic, } \Gamma_{\varepsilon} \succ 0, (\mathbf{B}, \Gamma_{\varepsilon}) \text{ bow-free, } \Sigma(\mathbf{B}, \Gamma_{\varepsilon}) = \Sigma_{\text{cond}}\}$. Following Pal et al. (2025b), the natural estimation target is a *minimal* (sparsest) representative: $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}}) := \arg \min_{(\mathbf{B}, \Gamma_{\varepsilon}) \in \mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}})} (\|\mathbf{B}\|_0 + \|\Gamma_{\varepsilon}\|_{\text{off}})$. When Γ_{ε} is diagonal, \mathcal{E}_{bow} reduces to the standard Markov equivalence class of DAGs.

What DECOR-GL estimates. Stage III (DECOR-GL) is a sparsity-regularized likelihood method over $(\mathbf{B}, \mathbf{S}_\varepsilon)$. Because $\mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}})$ may contain multiple elements, the statistically meaningful target is $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}})$ rather than a unique ground-truth DAG. Under suitable initialization and regularization, DECOR-GL with post-hoc bow reconciliation consistently finds an element of $\mathcal{E}_{\text{bow}}^{\min}$ (Pal et al., 2025b).

5.4. End-to-end modular guarantee

The D–C–L model admits a clean modular target: first recover the pervasive-adjusted (conditional) covariance $\Sigma_{\text{cond}} = \mathbf{S}_x^{-1}$ from the structured–low-rank precision split, then recover the appropriate bow-free correlated-noise SEM object from Σ_{cond} .

Theorem 6 (Modular reduction) *Assume (split identifiability) $(\mathbf{S}_x, \mathbf{L}_x)$ is identifiable from Σ under Assumption 1; (conditioning) \mathbf{S}_x is well-conditioned ($\kappa(\mathbf{S}_x) < \infty$); and (bow-freeness) the conditional SEM (7) satisfies bow-freeness. Then the identifiable causal target of the D–C–L model from Σ is the minimal bow-free equivalence class $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}})$. The proof is given in Appendix B.3.*

Theorem 7 (End-to-end consistency, informal) *Under the assumptions of Theorem 6, suppose Stage I is consistent in operator norm, $\delta_{S,n} = \|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2 \rightarrow 0$ as $n \rightarrow \infty$, and the Stage III procedure is stable (continuous) to small perturbations of its covariance input around Σ_{cond} . Then $\widehat{\Sigma}_{\text{cond}} = \widehat{\mathbf{S}}_x^{-1} \rightarrow \Sigma_{\text{cond}}$, and the Stage III output converges to an element of $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}})$. The proof is provided in Appendix B.4.*

6. Synthetic Experiments: Mixed Confounding (Pervasive Rank/Strength)

We evaluate causal discovery under *mixed confounding*, where a few *pervasive* latent factors affect many variables while additional *localized* latent factors induce structured, sparse dependence.

Baselines. We compare our three-stage DCL–DECOR-GL pipeline against: (i) DECOR-GL, which learns a DAG with correlated noise under bow-free identifiability (Pal et al., 2025b); (ii) DECAF-LIN, a factor/residualization-style baseline for *pervasive* confounding (provided the true latent rank in our sweep) (Agrawal et al., 2023); (iii) continuous, causal-sufficiency DAG learners NOTEARS and GOLEM (Zheng et al., 2018; Ng et al., 2020); and (iv) classical score/ICA baselines GES and LINGAM (Chickering, 2002; Shimizu et al., 2006). We report directed-edge recovery under a shared thresholding rule and emphasize robustness as pervasive and localized confounding vary.

Simulator. We simulate linear Gaussian SEMs on $p = 40$ observed variables, $\mathbf{x} = \mathbf{B}^\top \mathbf{x} + \varepsilon$, where \mathbf{B} is acyclic with approximately 8% directed edge density and edge weights sampled with random sign and magnitude $\text{Unif}(0.5, 2)$. The exogenous noise follows the D–C–L mixed-confounding model $\varepsilon = \mathbf{V}\mathbf{v} + \mathbf{U}\mathbf{u} + \mathbf{w}$, where $\mathbf{w} \sim \mathcal{N}(0, 0.36 \mathbf{I})$, localized confounding is represented by $\mathbf{V} \in \mathbb{R}^{p \times q}$ with $q = 15$ column-sparse loadings (each column has 6 active entries drawn from $\mathcal{N}(0, 0.3^2)$), and pervasive confounding is represented by $\mathbf{U} \in \mathbb{R}^{p \times q_P}$ with dense loadings sampled as $U_{ik} \sim \mathcal{N}(0, (U_d/\sqrt{p})^2)$. This scaling makes each pervasive factor have ℓ_2 -norm on the order of U_d , so the overall magnitude of the pervasive covariance contribution grows roughly with $q_P U_d^2$. To align with our identifiability target, we enforce bow-freeness *only with respect to localized confounders*: the simulator removes bows induced by \mathbf{V} , but does not remove bows induced by the pervasive component \mathbf{U} .

DCL1: sweeping pervasive rank and strength. We vary pervasive rank and strength over the grid $q_P \in \{1, 3, 5\}$, $U_d \in \{0.5, 1.0, 2.0\}$, using $n = 600$ samples and 10 replicates per grid cell, running all methods listed above. All methods use standardized inputs and a shared reporting threshold $\tau = 0.30$ to form the estimated directed support. For fairness, all non-DCL methods use $\lambda_B = 0.10$; DCL–DECOR–GL uses a DCL-specific $\lambda_B = 0.01$ (as in our synthetic setting), while sharing $\lambda_\Theta = 0.01$ with DECOR–GL. DCL–DECOR–GL first estimates a structured–low-rank precision split (LVGLASSO) with $(\lambda_{S_x}, \lambda_{L_x}) = (0.001, 0.005)$, then inverts the structured component to obtain a pervasive-adjusted covariance estimate and runs DECOR–GL on that input. Both DECOR–GL and DCL–DECOR–GL use identical post-hoc bow reconciliation (strict mode with the same thresholds and constant).

Figure 2a reports the mean $\Delta F1$ relative to DECOR–GL across the grid. DCL–DECOR–GL improves over DECOR–GL consistently as pervasive confounding strengthens (larger q_P and/or U_d), matching the intended regime where separating the pervasive component before modeling localized dependence is beneficial. Aggregated over the full DCL1 grid, DCL–DECOR–GL achieves higher directed-edge recovery and substantially lower SHD than DECOR–GL (mean F1 0.417 vs. 0.280; mean SHD 55.3 vs. 74.9). Methods that ignore latent structure (NOTEARS, GOLEM, GES, LINGAM) degrade as pervasive confounding increases, typically producing denser graphs and larger SHD. DECAF–LIN shows mixed performance in this mixed regime, suggesting that explicitly modeling *both* confounding types (pervasive plus localized) is important.

DCL2: fixing pervasive confounding and varying localized strength. To complement DCL1, we fix pervasive confounding at $(q_P, U_d) = (3, 1.0)$ and vary the localized confounding density $L_d \in \{0, 0.05, 0.10, 0.15, 0.20\}$ (10 replicates per setting). Figure 2b–c shows that DCL–DECOR–GL remains robust as localized confounding increases, while DECOR–GL and factor-only baselines degrade. Averaged across DCL2, DCL–DECOR–GL improves both F1 and SHD relative to DECOR–GL (mean F1 0.431 vs. 0.266; mean SHD 55.0 vs. 76.2), supporting our central claim that removing pervasive effects in the precision domain and then learning under correlated local noise is a strong strategy for mixed confounding.

References

- Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. The decamfounder: nonlinear causal discovery in the presence of hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1639–1658, 2023.
- Kevin Bello, Bryon Aragam, and Pradeep K. Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.
- Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proc. of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 2314–2322, 2021.
- Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–37, 2011. doi: 10.1145/1970392.1970395.

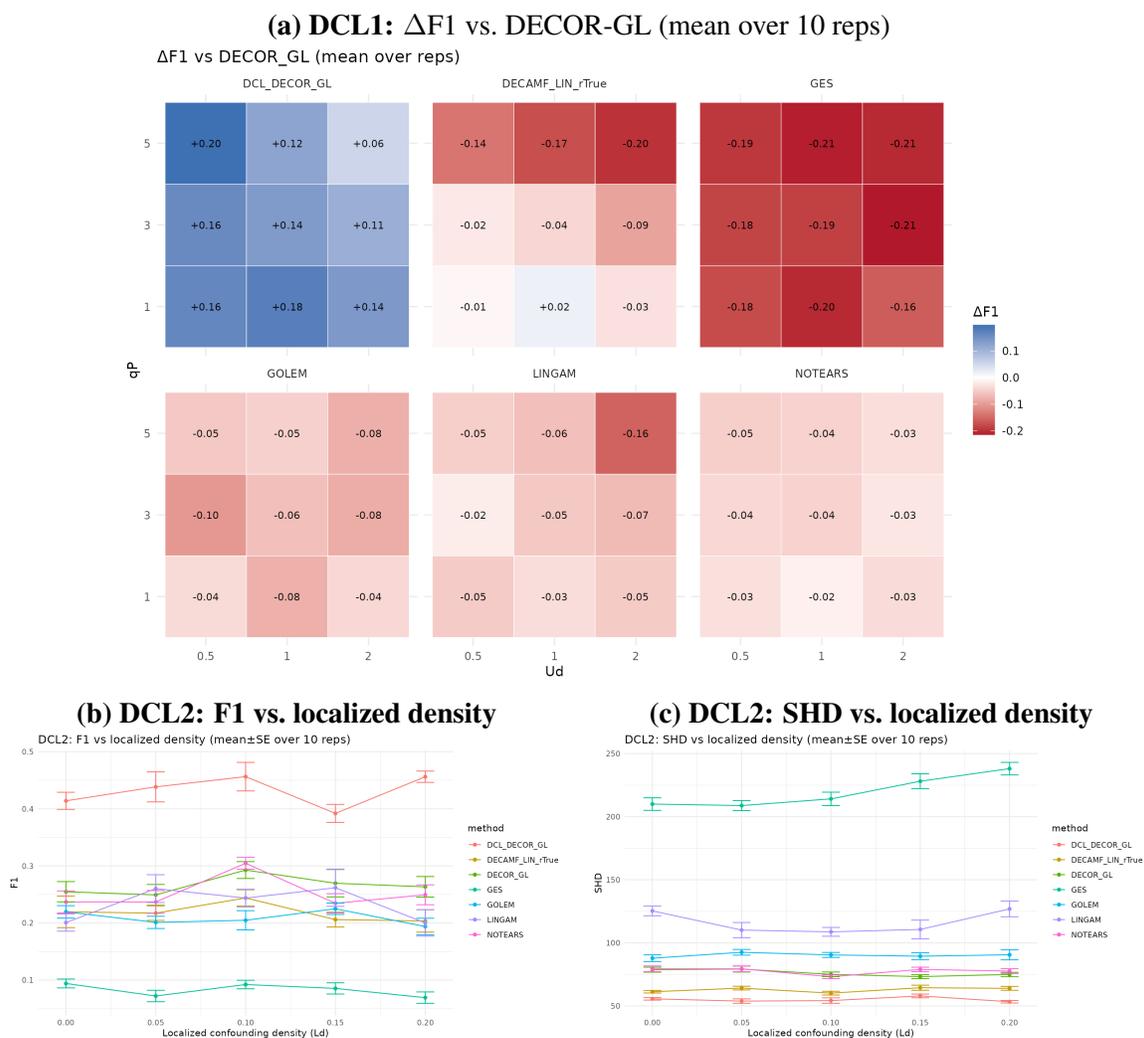


Figure 2: Mixed-confounding synthetic experiments. (a) Mean $\Delta F1$ relative to DECOR-GL across pervasive rank (q_p) and strength (U_d). (b–c) With pervasive confounding fixed, performance as localized confounding density (L_d) increases (mean over 10 replicates).

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4):1935–1967, 2012.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional DAGs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, 2012.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

- Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013. doi: 10.1111/rssb.12016.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.
- Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):459–487, 2019.
- Mario Grassi and Barbara Tarantino. Sembap: Bow-free covariance search and data de-correlation. *PLOS Computational Biology*, 20(9):e1012448, 2024.
- Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Non-linear causal discovery with additive noise models. *Proceedings of the National Academy of Sciences*, 106(51):21887–21892, 2009.
- Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(140):3065–3105, 2014.
- Pingchuan Ma, Rui Ding, Qiang Fu, Jiaru Zhang, Shuai Wang, Shi Han, and Dongmei Zhang. Scalable differentiable causal discovery in the presence of latent confounders with skeleton posterior. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2141–2152, 2024.
- Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems (NeurIPS) 33*, pages 22236–22247, 2020.
- Juan M. Ogarrio, Peter Spirtes, and Joseph Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models (PGM 2016)*, volume 52 of *Proceedings of Machine Learning Research*, pages 368–379. PMLR, 2016.
- Samhita Pal, Dhruvajyoti Ghosh, and Shu Yang. Penalized fci for causal structure learning in a sparse dag for biomarker discovery in parkinson’s disease. *arXiv preprint arXiv:2507.00173*, 2025a.
- Samhita Pal, James O’quinn, Kaveh Aryan, Heather Pua, James P. Long, and Amir Asiaee. Dag decoration: Continuous optimization for structure learning under hidden confounding. *arXiv preprint arXiv:2510.02117*, 2025b.
- Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Bryon Aragam, and Devavrat Shah. DYNOTEARS: Structure learning from time-series data. In *Proc. of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 1595–1605, 2020.
- Parjanya Prashant, Ignavier Ng, Kun Zhang, and Biwei Huang. Differentiable causal discovery for latent hierarchical causal models. *arXiv preprint arXiv:2411.19556*, 2024.

- Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30(4):962–1030, 2002. doi: 10.1214/aos/1031689015.
- Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
- Peter Spirtes. An anytime algorithm for causal inference. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics (AISTATS)*, volume R3 of *Proceedings of Machine Learning Research*, pages 278–285. PMLR, 2001.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search (2nd ed.)*. MIT Press, 2000.
- Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.
- Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, pages 9472–9483, 2018.
- Xun Zheng, Chen Dan, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *Proc. of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425, 2020.

Appendix A. Structural Properties of the D–C–L Precision Decomposition

In this appendix we provide a full version of Proposition 2 together with complete proofs. We use the notation from §3.

Recall the noise model

$$\boldsymbol{\varepsilon} = \mathbf{W}\mathbf{w} + \mathbf{V}\mathbf{v} + \mathbf{U}\mathbf{u},$$

with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_S})$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_L})$ independent, and

$$\boldsymbol{\Omega} = \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top + \mathbf{U}\mathbf{U}^\top.$$

A.1. Noise-level components

Let

$$\mathbf{D}_\varepsilon := (\mathbf{W}\mathbf{W}^\top)^{-1} = \text{diag}(d_1, \dots, d_p), \quad d_i > 0,$$

and define

$$\mathbf{A} := \mathbf{I} + \mathbf{V}^\top \mathbf{D}_\varepsilon \mathbf{V} \in \mathbb{R}^{r_S \times r_S}, \quad \mathbf{C}_\varepsilon := \mathbf{D}_\varepsilon \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top \mathbf{D}_\varepsilon.$$

The non-pervasive noise covariance and precision are

$$\boldsymbol{\Gamma}_\varepsilon := \mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top, \quad \mathbf{S}_\varepsilon := \boldsymbol{\Gamma}_\varepsilon^{-1} = \mathbf{D}_\varepsilon - \mathbf{C}_\varepsilon.$$

Writing $\boldsymbol{\Omega} = \boldsymbol{\Gamma}_\varepsilon + \mathbf{U}\mathbf{U}^\top = \mathbf{S}_\varepsilon^{-1} + \mathbf{U}\mathbf{U}^\top$, the SMW identity gives

$$\boldsymbol{\Omega}^{-1} = \mathbf{S}_\varepsilon - \mathbf{S}_\varepsilon \mathbf{U} (\mathbf{I} + \mathbf{U}^\top \mathbf{S}_\varepsilon \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{S}_\varepsilon. \quad (12)$$

We define the pervasive (low-rank) correction

$$\mathbf{L}_\varepsilon := \mathbf{S}_\varepsilon \mathbf{U} (\mathbf{I} + \mathbf{U}^\top \mathbf{S}_\varepsilon \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{S}_\varepsilon, \quad (13)$$

so that $\boldsymbol{\Omega}^{-1} = \mathbf{D}_\varepsilon - \mathbf{C}_\varepsilon - \mathbf{L}_\varepsilon = \mathbf{S}_\varepsilon - \mathbf{L}_\varepsilon$.

A.2. Full statement and proof

Proposition 8 (Low-rankness of \mathbf{L}_ε and locality of \mathbf{S}_ε) *Assume the model in §3 with $\mathbf{T} = \mathbf{I} - \mathbf{B}$ unit-diagonal and triangular under some causal order, and define $\mathbf{D}_\varepsilon, \mathbf{A}, \mathbf{C}_\varepsilon, \mathbf{S}_\varepsilon, \mathbf{L}_\varepsilon$ as above.*

(a) *Low-rankness and PSD of \mathbf{L}_ε . Let $\mathbf{M} := (\mathbf{I} + \mathbf{U}^\top \mathbf{S}_\varepsilon \mathbf{U})^{-1}$. Then*

$$\mathbf{L}_\varepsilon = (\mathbf{S}_\varepsilon^{1/2} \mathbf{U}) \mathbf{M} (\mathbf{S}_\varepsilon^{1/2} \mathbf{U})^\top \succeq 0, \quad \text{rank}(\mathbf{L}_\varepsilon) \leq r_L.$$

Consequently, $\mathbf{L}_x = \mathbf{T} \mathbf{L}_\varepsilon \mathbf{T}^\top \succeq 0$ with $\text{rank}(\mathbf{L}_x) \leq r_L$.

(b) *Exact locality of \mathbf{S}_ε under disjoint column supports. Suppose the columns of \mathbf{V} have disjoint supports $S_j := \text{supp}(\mathbf{V}_{\cdot j})$ with $|S_j| \leq s$, and each row i belongs to at most c such supports. Then*

$$\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon - \sum_{j=1}^{r_S} \frac{1}{A_{jj}} (\mathbf{D}_\varepsilon \mathbf{V}_{\cdot j}) (\mathbf{D}_\varepsilon \mathbf{V}_{\cdot j})^\top,$$

and each row of \mathbf{S}_ε has at most $c(s-1)$ off-diagonal nonzeros. Equivalently, \mathbf{C}_ε is supported on a union of cliques $\{S_j \times S_j\}$.

(c) **Approximate locality under controlled leakage.** Assume each row belongs to at most c supports and each column has $|S_j| \leq s$. Let

$$\|\mathbf{A}^{-1}\|_{\text{off},\infty} := \max_j \sum_{k \neq j} |(\mathbf{A}^{-1})_{jk}| \leq \nu$$

for some $\nu \geq 0$. Then for any $i \neq \ell$,

$$|(S_\varepsilon)_{i\ell}| \leq (D_\varepsilon)_{ii}(D_\varepsilon)_{\ell\ell} \left(\sum_{j: i \in S_j} \frac{|V_{ij}V_{\ell j}|}{A_{jj}} + \nu \sum_{j: i \in S_j} \sum_{k \neq j} |V_{ij}V_{\ell k}| \right).$$

In particular, for any threshold $\epsilon > 0$, the number of indices ℓ with $|(S_\varepsilon)_{i\ell}| \geq \epsilon$ is at most $\mathcal{O}(cs)$ provided ν is small enough relative to ϵ , $\max_{i,j} |V_{ij}|$, and $\max_i (D_\varepsilon)_{ii}$.

(d) **Propagation to $\mathbf{S}_x = \mathbf{T}\mathbf{S}_\varepsilon\mathbf{T}^\top$ under bounded degree.** Let

$$\text{deg}_T^{\text{row}} := \max_i \text{NNZ}(\mathbf{T}_i), \quad \text{deg}_T^{\text{col}} := \max_i \text{NNZ}(\mathbf{T}_i), \quad \text{deg}_{S_\varepsilon} := \max_i \text{NNZ}((\mathbf{S}_\varepsilon)_i),$$

where NNZ counts entries whose magnitude exceeds a fixed small threshold. Then each row of $\mathbf{S}_x = \mathbf{T}\mathbf{S}_\varepsilon\mathbf{T}^\top$ has at most

$$\mathcal{O}(\text{deg}_T^{\text{row}} \text{deg}_{S_\varepsilon} \text{deg}_T^{\text{col}})$$

entries above a (slightly larger) threshold. In particular, if \mathbf{B} has bounded (in+out) degree d , then $\text{deg}_T^{\text{row}}, \text{deg}_T^{\text{col}} \leq 1 + d$, and \mathbf{S}_x inherits row-locality from \mathbf{S}_ε .

Proof

(a) From (12) and (13),

$$\mathbf{L}_\varepsilon = \mathbf{S}_\varepsilon \mathbf{U} \mathbf{M} \mathbf{U}^\top \mathbf{S}_\varepsilon, \quad \mathbf{M} = (\mathbf{I} + \mathbf{U}^\top \mathbf{S}_\varepsilon \mathbf{U})^{-1} \succ 0.$$

Factor $\mathbf{S}_\varepsilon = \mathbf{S}_\varepsilon^{1/2} \mathbf{S}_\varepsilon^{1/2}$ to obtain

$$\mathbf{L}_\varepsilon = (\mathbf{S}_\varepsilon^{1/2} \mathbf{U}) \mathbf{M} (\mathbf{S}_\varepsilon^{1/2} \mathbf{U})^\top,$$

which is positive semidefinite. Its rank is at most $\text{rank}(\mathbf{U}) \leq r_L$. Congruence by invertible \mathbf{T} preserves PSD and rank, so the same holds for $\mathbf{L}_x = \mathbf{T} \mathbf{L}_\varepsilon \mathbf{T}^\top$.

(b) Under disjoint supports, $\mathbf{V}^\top \mathbf{D}_\varepsilon \mathbf{V}$ is diagonal: for $j \neq k$,

$$(\mathbf{V}^\top \mathbf{D}_\varepsilon \mathbf{V})_{jk} = \sum_{i=1}^p V_{ij} (D_\varepsilon)_{ii} V_{ik} = 0,$$

since no row i belongs to both S_j and S_k . Hence \mathbf{A} is diagonal and $\mathbf{A}^{-1} = \text{diag}(A_{11}^{-1}, \dots, A_{rsrs}^{-1})$. Using $\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon - \mathbf{D}_\varepsilon \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top \mathbf{D}_\varepsilon$,

$$\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon - \sum_{j=1}^{rs} \mathbf{D}_\varepsilon \mathbf{V}_{\cdot j} A_{jj}^{-1} \mathbf{V}_{\cdot j}^\top \mathbf{D}_\varepsilon = \mathbf{D}_\varepsilon - \sum_{j=1}^{rs} \frac{1}{A_{jj}} (\mathbf{D}_\varepsilon \mathbf{V}_{\cdot j}) (\mathbf{D}_\varepsilon \mathbf{V}_{\cdot j})^\top.$$

Each rank-one term has support contained in $S_j \times S_j$. Fixing a row i , there are at most c indices j with $i \in S_j$, and within each such support S_j row i connects to at most $|S_j| - 1 \leq s - 1$ other indices. Therefore row i has at most $c(s - 1)$ off-diagonal nonzeros.

(c) From $\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon - \mathbf{D}_\varepsilon \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top \mathbf{D}_\varepsilon$, for $i \neq \ell$,

$$(S_\varepsilon)_{i\ell} = - \sum_{j,k} (D_\varepsilon)_{ii} V_{ij} (\mathbf{A}^{-1})_{jk} V_{\ell k} (D_\varepsilon)_{\ell\ell}.$$

Split into diagonal and off-diagonal parts in (j, k) :

$$\begin{aligned} |(S_\varepsilon)_{i\ell}| &\leq (D_\varepsilon)_{ii} (D_\varepsilon)_{\ell\ell} \left(\sum_j \frac{|V_{ij} V_{\ell j}|}{A_{jj}} + \sum_j \sum_{k \neq j} |V_{ij}| \cdot |(\mathbf{A}^{-1})_{jk}| \cdot |V_{\ell k}| \right) \\ &\leq (D_\varepsilon)_{ii} (D_\varepsilon)_{\ell\ell} \left(\sum_{j: i \in S_j} \frac{|V_{ij} V_{\ell j}|}{A_{jj}} + \sum_{j: i \in S_j} \sum_{k \neq j} |V_{ij}| \cdot |(\mathbf{A}^{-1})_{jk}| \cdot |V_{\ell k}| \right). \end{aligned}$$

Use $\sum_{k \neq j} |(\mathbf{A}^{-1})_{jk}| \leq \nu$ and pull out $|V_{ij}|$ to obtain the displayed bound. The row-locality conclusion follows by the same counting argument as in part (b), combined with the requirement that ν be small enough so that leakage terms fall below the chosen threshold except on $\mathcal{O}(cs)$ indices.

(d) Let $N_T(i) := \{k : T_{ik} \neq 0\}$ (row support of \mathbf{T}). Then

$$(\mathbf{S}_x)_{ij} = (\mathbf{T} \mathbf{S}_\varepsilon \mathbf{T}^\top)_{ij} = \sum_{k \in N_T(i)} \sum_{\ell \in N_T(j)} T_{ik} (S_\varepsilon)_{k\ell} T_{j\ell}.$$

Thus $(\mathbf{S}_x)_{ij}$ can be non-negligible only if there exist $k \in N_T(i)$ and $\ell \in N_T(j)$ with $(S_\varepsilon)_{k\ell}$ non-negligible. Fix row i . Each $k \in N_T(i)$ has at most \deg_{S_ε} such ℓ , and for each such ℓ , there are at most \deg_T^{col} indices j with $T_{j\ell} \neq 0$. Since $|N_T(i)| \leq \deg_T^{\text{row}}$, the stated bound follows by a union argument. \blacksquare

A.3. A sufficient condition for controlled leakage

The next result gives an interpretable sufficient condition for small $\|\mathbf{A}^{-1}\|_{\text{off},\infty}$, replacing hard ‘‘orthogonality’’ assumptions with overlap and dominance parameters.

Proposition 9 (Controlled leakage under relaxed orthogonality) *Let $\mathbf{A} = \mathbf{I} + \mathbf{V}^\top \mathbf{D}_\varepsilon \mathbf{V}$. Define*

$$m := \max_j |\{k \neq j : \text{supp}(\mathbf{V}_{\cdot j}) \cap \text{supp}(\mathbf{V}_{\cdot k}) \neq \emptyset\}|, \quad \eta := \max_{j \neq k} |\mathbf{V}_{\cdot j}^\top \mathbf{D}_\varepsilon \mathbf{V}_{\cdot k}|, \quad \tau_{\min} := \min_j A_{jj}.$$

Let $\rho := m\eta/\tau_{\min}$. If $\rho < 1$, then

$$\|\mathbf{A}^{-1}\|_{\text{off},\infty} \leq \frac{\rho}{\tau_{\min}(1-\rho)} = \frac{m\eta}{\tau_{\min}^2(1-m\eta/\tau_{\min})}.$$

Proof Write $\mathbf{A} = \mathbf{D} + \mathbf{E}$ where $\mathbf{D} := \text{diag}(\mathbf{A})$ and $\mathbf{E} := \mathbf{A} - \mathbf{D}$ (off-diagonal part). Then

$$\mathbf{A}^{-1} = (\mathbf{D}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{E}))^{-1} = (\mathbf{I} + \mathbf{D}^{-1}\mathbf{E})^{-1} \mathbf{D}^{-1}.$$

By definition, $\|\mathbf{D}^{-1}\|_\infty = 1/\tau_{\min}$. Moreover, for each row j ,

$$\sum_{k \neq j} |(\mathbf{D}^{-1}\mathbf{E})_{jk}| = \frac{1}{A_{jj}} \sum_{k \neq j} |A_{jk}| \leq \frac{1}{\tau_{\min}} \cdot m\eta = \rho,$$

so $\|\mathbf{D}^{-1}\mathbf{E}\|_\infty \leq \rho < 1$. Hence the Neumann series converges:

$$(\mathbf{I} + \mathbf{D}^{-1}\mathbf{E})^{-1} = \sum_{t=0}^{\infty} (-\mathbf{D}^{-1}\mathbf{E})^t.$$

The off-diagonal mass comes from $t \geq 1$, giving

$$\|\mathbf{A}^{-1}\|_{\text{off},\infty} \leq \|\mathbf{D}^{-1}\|_\infty \sum_{t=1}^{\infty} \|\mathbf{D}^{-1}\mathbf{E}\|_\infty^t \leq \frac{1}{\tau_{\min}} \cdot \frac{\rho}{1-\rho},$$

which is the desired bound. ■

A.4. Structure preservation under \mathbf{T} -congruence

The next proposition formalizes how local structure classes are propagated under \mathbf{TMT}^\top . This supports Remark 1.

Proposition 10 (Structure preservation under \mathbf{T} -congruence) *Let $\mathbf{T} = \mathbf{I} - \mathbf{B}$ where \mathbf{B} encodes a DAG. Consider $\mathbf{M} \in \mathbb{R}^{p \times p}$.*

- (a) **Row-sparse case.** *If \mathbf{M} is k -row-sparse and \mathbf{B} has maximum (in+out)-degree at most d , then \mathbf{TMT}^\top is $k(1+d)^2$ -row-sparse.*
- (b) **Banded case.** *Suppose the variables are ordered and \mathbf{M} is b -banded. If the DAG respects the ordering in the sense that $B_{ij} \neq 0 \Rightarrow |i-j| \leq d_{\text{DAG}}$, then \mathbf{TMT}^\top is $(b+2d_{\text{DAG}})$ -banded.*
- (c) **Block-diagonal case (approximate).** *Let $\mathcal{P} = \{B_1, \dots, B_K\}$ be a partition of $\{1, \dots, p\}$ and assume \mathbf{M} is block-diagonal w.r.t. \mathcal{P} . If each block has at most c cross-block edges incident to it (in either direction) in the DAG, then each off-diagonal block of \mathbf{TMT}^\top has at most $\mathcal{O}(c^2)$ nonzero entries.*

Proof

(a) Let $N_T^{\text{row}}(i) := \{k : T_{ik} \neq 0\}$ and $N_T^{\text{col}}(\ell) := \{j : T_{j\ell} \neq 0\}$. Under degree bound d , we have $|N_T^{\text{row}}(i)| \leq 1+d$ and $|N_T^{\text{col}}(\ell)| \leq 1+d$. For fixed i, j ,

$$(\mathbf{TMT}^\top)_{ij} = \sum_{k \in N_T^{\text{row}}(i)} \sum_{\ell \in N_T^{\text{row}}(j)} T_{ik} M_{k\ell} T_{j\ell}.$$

Fix row i . There are at most $1+d$ choices of k . For each k , row-sparsity of \mathbf{M} gives at most k indices ℓ with $M_{k\ell} \neq 0$. For each such ℓ , there are at most $1+d$ indices j with $T_{j\ell} \neq 0$. Thus row i has at most $(1+d) \cdot k \cdot (1+d) = k(1+d)^2$ nonzeros.

(b) If \mathbf{M} is b -banded, then $M_{k\ell} = 0$ whenever $|k-\ell| > b$. In addition, $T_{ik} \neq 0$ implies $|i-k| \leq d_{\text{DAG}}$, and similarly $T_{j\ell} \neq 0$ implies $|j-\ell| \leq d_{\text{DAG}}$. Therefore, a nonzero contribution to $(\mathbf{TMT}^\top)_{ij}$ requires $|i-k| \leq d_{\text{DAG}}$, $|k-\ell| \leq b$, $|j-\ell| \leq d_{\text{DAG}}$, which implies $|i-j| \leq b+2d_{\text{DAG}}$. Hence \mathbf{TMT}^\top is $(b+2d_{\text{DAG}})$ -banded.

(c) Write \mathbf{M} in $K \times K$ block form. Since \mathbf{M} is block-diagonal, any off-block entry of \mathbf{TMT}^\top must arise from multiplying \mathbf{M} by cross-block nonzeros of \mathbf{T} . If each block participates in at most c cross-block edges, then each block-row/column of \mathbf{T} has at most $\mathcal{O}(c)$ nonzeros outside the diagonal block. Expanding \mathbf{TMT}^\top and counting the ways an off-diagonal block can be hit by left- and right-multiplication yields at most $\mathcal{O}(c^2)$ induced nonzeros per off-diagonal block. \blacksquare

Moralized-graph special case (independent errors)

In the special case $\mathbf{V} = \mathbf{U} = \mathbf{0}$, we have $\mathbf{S}_\varepsilon = \mathbf{D}_\varepsilon$ diagonal and

$$\Theta = \mathbf{T}\mathbf{D}_\varepsilon\mathbf{T}^\top.$$

Under a causal order where $\mathbf{T} = \mathbf{I} - \mathbf{B}$ is unit-diagonal and upper-triangular, $T_{ik} \neq 0$ iff $k = i$ or there is a directed edge $i \rightarrow k$. Hence $\text{supp}(\Theta)$ coincides with the moralized graph of \mathbf{B} (undirected edges plus co-parent “marriages”) (Loh and Bühlmann, 2014). When $\mathbf{V} \neq \mathbf{0}$ but \mathbf{S}_ε is (approximately) local as in Proposition 8(b)–(c), Proposition 8(d) shows that $\mathbf{S}_x = \mathbf{TS}_\varepsilon\mathbf{T}^\top$ augments the moralized pattern only locally, while \mathbf{L}_x contributes a low-rank pervasive component.

Appendix B. Proofs for §4–§5

B.1. Proof of Proposition 3

Recall $\varepsilon = \mathbf{W}\mathbf{w} + \mathbf{V}\mathbf{v} + \mathbf{U}\mathbf{u}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_S})$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r_L})$ mutually independent. Let

$$\gamma := \mathbf{W}\mathbf{w} + \mathbf{V}\mathbf{v}, \quad \gamma \sim \mathcal{N}(\mathbf{0}, \Gamma_\varepsilon), \quad \Gamma_\varepsilon = \mathbf{W}\mathbf{W}^\top + \mathbf{V}\mathbf{V}^\top,$$

so that $\varepsilon = \gamma + \mathbf{U}\mathbf{u}$ with $\gamma \perp\!\!\!\perp \mathbf{u}$.

From (1), $\mathbf{x} = \mathbf{T}^{-\top}\varepsilon$, hence

$$\mathbf{x} = \mathbf{T}^{-\top}\gamma + \mathbf{T}^{-\top}\mathbf{U}\mathbf{u}.$$

Conditioning on \mathbf{u} , the second term is deterministic and the first term remains Gaussian with covariance

$$\text{Cov}(\mathbf{x} \mid \mathbf{u}) = \text{Cov}(\mathbf{T}^{-\top}\gamma) = \mathbf{T}^{-\top} \text{Cov}(\gamma) \mathbf{T}^{-1} = \mathbf{T}^{-\top}\Gamma_\varepsilon\mathbf{T}^{-1}.$$

Since $\Gamma_\varepsilon \succ 0$ and \mathbf{T} is invertible, the conditional covariance is SPD and

$$\text{Cov}(\mathbf{x} \mid \mathbf{u})^{-1} = \mathbf{T}\Gamma_\varepsilon^{-1}\mathbf{T}^\top = \mathbf{TS}_\varepsilon\mathbf{T}^\top = \mathbf{S}_x,$$

which completes the proof.

B.2. Proof of Lemma 5

Let $\widehat{\mathbf{S}}_x = \mathbf{S}_x + \Delta$ with $\|\Delta\|_2 = \delta_{S,n}$. Whenever $\|\mathbf{S}_x^{-1}\|_2 \|\Delta\|_2 < 1$, the matrix $\widehat{\mathbf{S}}_x$ is invertible and

$$\widehat{\mathbf{S}}_x^{-1} - \mathbf{S}_x^{-1} = (\mathbf{S}_x + \Delta)^{-1} - \mathbf{S}_x^{-1} = -\mathbf{S}_x^{-1}\Delta(\mathbf{S}_x + \Delta)^{-1}.$$

Taking operator norms and using $\|(\mathbf{S}_x + \mathbf{\Delta})^{-1}\|_2 \leq \|\mathbf{S}_x^{-1}\|_2 / (1 - \|\mathbf{S}_x^{-1}\|_2 \|\mathbf{\Delta}\|_2)$ (which follows from the Neumann series bound), we obtain

$$\|\widehat{\mathbf{S}}_x^{-1} - \mathbf{S}_x^{-1}\|_2 \leq \|\mathbf{S}_x^{-1}\|_2 \|\mathbf{\Delta}\|_2 \|(\mathbf{S}_x + \mathbf{\Delta})^{-1}\|_2 \leq \frac{\|\mathbf{S}_x^{-1}\|_2^2 \delta_{S,n}}{1 - \|\mathbf{S}_x^{-1}\|_2 \delta_{S,n}},$$

which is desired bound after substituting $\widehat{\Sigma}_{\text{cond}} = \widehat{\mathbf{S}}_x^{-1}$ and $\Sigma_{\text{cond}} = \mathbf{S}_x^{-1}$.

B.3. Proof of Theorem 6

Write the population precision as $\Theta := \Sigma^{-1}$. In the D–C–L model we have the population decomposition

$$\Theta = \mathbf{S}_x - \mathbf{L}_x, \quad \mathbf{S}_x \succ 0, \mathbf{L}_x \succeq 0 \text{ (low-rank)}. \quad (14)$$

Under Assumption 1, the structured–low-rank split is identifiable at the population level, i.e., the pair $(\mathbf{S}_x, \mathbf{L}_x)$ is uniquely determined by Θ (equivalently by Σ). In particular, \mathbf{S}_x is identifiable from Σ .

Step 1: Identifying Σ_{cond} . By Proposition 3, the conditional covariance after removing pervasive confounding satisfies

$$\Sigma_{\text{cond}} := \text{Cov}(\mathbf{x} \mid \mathbf{u}) = \mathbf{S}_x^{-1}. \quad (15)$$

Because \mathbf{S}_x is identifiable from Σ , the matrix $\Sigma_{\text{cond}} = \mathbf{S}_x^{-1}$ is also identifiable as a (deterministic) function of Σ . Assumption $\kappa(\mathbf{S}_x) < \infty$ is not needed for this population identification but will be used to control stability under estimation in Theorem 7.

Step 2: The identifiable object from Σ_{cond} . Conditioning on \mathbf{u} yields the correlated-noise SEM (7),

$$\mathbf{x}^\perp = \mathbf{B}^\top \mathbf{x}^\perp + \boldsymbol{\varepsilon}^\perp, \quad \boldsymbol{\varepsilon}^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_\varepsilon), \quad \Sigma_{\text{cond}} = \mathbf{T}^{-\top} \mathbf{\Gamma}_\varepsilon \mathbf{T}^{-1},$$

with $\mathbf{T} = \mathbf{I} - \mathbf{B}$. By the bow-free assumption, the true conditional parameters $(\mathbf{B}, \mathbf{\Gamma}_\varepsilon)$ belong to the bow-free model family.

Let $\mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}})$ be the set of *all* bow-free parameter pairs $(\mathbf{B}', \mathbf{\Gamma}'_\varepsilon)$ that reproduce the same conditional covariance Σ_{cond} , and let $\mathcal{E}_{\text{bow}}^{\text{min}}(\Sigma_{\text{cond}})$ be the sparsity-minimal subset. By construction, both sets depend on the data-generating process only through Σ_{cond} . Since Σ_{cond} is identifiable from Σ (Step 1), the sets $\mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}})$ and $\mathcal{E}_{\text{bow}}^{\text{min}}(\Sigma_{\text{cond}})$ are also identifiable from Σ .

Finally, because the true conditional SEM is bow-free, its (unknown) parameters lie in $\mathcal{E}_{\text{bow}}(\Sigma_{\text{cond}})$, and hence the statistically meaningful identifiable causal target—when the graph is not assumed known a priori—is the corresponding minimal bow-free equivalence class $\mathcal{E}_{\text{bow}}^{\text{min}}(\Sigma_{\text{cond}})$. This proves the claim.

B.4. Proof of Theorem 7

We prove the two conclusions in order.

Step 1: $\widehat{\Sigma}_{\text{cond}} \rightarrow \Sigma_{\text{cond}}$. By assumption, $\delta_{S,n} = \|\widehat{\mathbf{S}}_x - \mathbf{S}_x\|_2 \rightarrow 0$. Since $\mathbf{S}_x \succ 0$ and $\kappa(\mathbf{S}_x) < \infty$, we have $\lambda_{\min}(\mathbf{S}_x) > 0$. Therefore, for all sufficiently large n , $\delta_{S,n} < \lambda_{\min}(\mathbf{S}_x)$, so Lemma 5 applies and yields

$$\|\widehat{\Sigma}_{\text{cond}} - \Sigma_{\text{cond}}\|_2 = \|\widehat{\mathbf{S}}_x^{-1} - \mathbf{S}_x^{-1}\|_2 \leq \frac{\|\mathbf{S}_x^{-1}\|_2^2 \delta_{S,n}}{1 - \|\mathbf{S}_x^{-1}\|_2 \delta_{S,n}} \xrightarrow{n \rightarrow \infty} 0,$$

which proves $\widehat{\Sigma}_{\text{cond}} \rightarrow \Sigma_{\text{cond}}$ in operator norm.

Step 2: Stability of Stage III transfers this to the causal target. Let \mathcal{A} denote the (possibly set-valued) Stage III mapping that takes a covariance input Σ to a bow-free output (e.g., a particular representative $(\widehat{\mathbf{B}}, \widehat{\Gamma}_\varepsilon)$ after optimization and bow reconciliation). The stability assumption in the theorem is exactly that \mathcal{A} is continuous at Σ_{cond} with respect to the metric used to compare outputs (e.g., Frobenius distance between matrix representatives, or an appropriate distance between equivalence-class representatives). Hence, as $\widehat{\Sigma}_{\text{cond}} \rightarrow \Sigma_{\text{cond}}$, we obtain

$$\mathcal{A}(\widehat{\Sigma}_{\text{cond}}) \rightarrow \mathcal{A}(\Sigma_{\text{cond}}).$$

Under the bow-free assumption and the definition of the identifiability target, the population-level output $\mathcal{A}(\Sigma_{\text{cond}})$ is an element of $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}})$ (this is the object that Stage III is designed to return, up to distributional equivalence within the bow-free family). Therefore the Stage III output computed from $\widehat{\Sigma}_{\text{cond}}$ converges to an element of $\mathcal{E}_{\text{bow}}^{\min}(\Sigma_{\text{cond}})$, completing the proof.