THE STATISTICAL COST OF CAUSAL DISCOVERY WITH FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

What determines the unavoidable sample cost of learning cyclic causal structure? For cyclic linear non-Gaussian models, we study exact condensation recovery from observational data: identifying the strongly connected component (SCC) partition and all edges between components. We establish the first information-theoretic lower bounds on sample complexity for this target. For p variables, maximum SCC size $s_{\max}$, and maximum external-parent count d_B, any estimator requires order $s_{\max}\log(ep/s_{\max}) + d_B\log(ep/d_B)$ samples in the worst case over a regular model class. These bounds distinguish the costs of SCC membership and external-parent selection. Under principal invertibility and without correlation faithfulness, we establish a population block-exogeneity principle that identifies unknown root SCCs through residual independence and inclusion minimality. A sparse-adjustment characterization shows that small adjustment sets suffice to identify SCCs and their direct external parents, without regressing on all previously recovered variables. These characterizations yield BlockExo, which attains a structurally matching sample bound without knowing $s_{\max}$ or d_B under suitable conditions. Simulations support the structural dependence of our sample bound and demonstrate BlockExo's sample-efficient recovery in comparisons with other methods for cyclic causal discovery.

1 INTRODUCTION

Feedback is common in gene regulation, economic systems, and dynamical systems (Dai et al., 2024; Richardson & Robins, 2014; Bongers et al., 2021). For cyclic linear non-Gaussian (LiNG) models, a natural observational target is the condensation: the strongly connected component (SCC) partition and all edges between components. This target is invariant across the relevant observational equivalence class (Sharifian et al., 2025; Madaleno et al., 2026). We ask which structural sample costs are unavoidable and which population principles enable recovery with a structurally matching sample bound.

We establish the first sample-complexity lower bounds for exact condensation recovery in cyclic LiNG models. For p variables, let $s_{\max}$ be the maximum SCC size and d_B the maximum number of external parent variables of an SCC. Over regular model classes, these bounds imply the necessary structural sample cost $s_{\max}\log\left(ep/s_{\max}\right) + d_B\log\left(ep/d_B\right)$, where $e = \exp(1)$ and $\log$ denotes the natural logarithm. The two terms capture unavoidable costs of identifying SCC membership and selecting external parents, respectively. The first term remains even when there are no edges between SCCs: identifying which variables belong to the same SCC alone incurs a sample cost.

We complement these lower bounds with population principles for identifying unknown SCCs and their external parents. Under principal invertibility, our *block-exogeneity* shows that, after adjusting for previously recovered variables, residual independence characterizes blocks with no incoming edges from the remaining variables, without correlation faithfulness. The inclusion-minimal nonempty blocks satisfying this condition are exactly the remaining root SCCs, allowing their identification without a known SCC partition. Small adjustment sets further suffice to identify both SCCs and their direct external parents. This sparse-adjustment characterization avoids regressing on all previously recovered variables and controls regression dimensions through SCC size and external-parent count. These principles yield BlockExo, an algorithm that attains a structurally matching sample bound under regularity and separation conditions, even when $s_{\max}$ and d_B are unknown.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Related work. GroupLiNGAM characterizes exogenous sets through residual independence under correlation faithfulness (Kawahara et al., 2010). For exact condensation recovery, Coarsening provides an ICA-based method with a finite-sample recovery guarantee and polynomial-time computation (Madaleno et al., 2026). OptLiNGAM establishes necessary and achievable structural sample costs for LiNGAMs (Oh et al., 2025). Distributional equivalence has also been characterized for LiNG models with latent variables and cycles (Dai et al., 2026). Appendix A gives detailed comparisons.

Our contributions are summarized as follows.

- **Information-theoretic lower bounds.** We establish the first sample-complexity lower bounds for exact condensation recovery in cyclic LiNG models, separating the costs of SCC membership and external-parent selection over regular model classes (Theorem 3.1).
- **Block exogeneity for root SCC identification.** We characterize ancestral blocks by residual independence and identify root SCCs through inclusion minimality, under principal invertibility and without correlation faithfulness (Theorem 4.1 and Corollary 4.2).
- **Sample-efficient recovery through sparse adjustment.** We show that small adjustment sets suffice to identify ancestral blocks and their external parents (Proposition 4.3). BlockExo attains a structurally matching sample bound without knowing $s_{\max}$ or d_B under suitable conditions (Theorem 5.2), and is minimax optimal in sample complexity on a separated model class for fixed structural bounds (Corollary 5.3).

Proofs of all theoretical results are provided in the appendix.

2 Model and Condensation Recovery Target

Cyclic LiNG model. Let $G = (V, E)$ be a directed graph on $V = [p] := \{1, \dots, p\}$. Its weighted adjacency matrix $B \in \mathbb{R}^{p \times p}$ satisfies $\mathrm{diag}(B) = 0$, with $B_{ij} \neq 0$ if and only if $j \to i \in E$. We consider cyclic LiNG models (Lacerda et al., 2008) of the form

$$X = BX + \varepsilon, \tag{1}$$

where $X, \varepsilon \in \mathbb{R}^p$. The coordinates of ε are mutually independent and centered, with positive finite variances, and at most one is Gaussian. With I denoting the identity matrix, we assume the following.

Assumption 2.1 (Principal invertibility)**.** *Every principal submatrix of $I - B$ is invertible.*

Such well-posedness conditions are standard in the literature on linear cyclic causal models (Hyttinen et al., 2012; Bongers et al., 2021; Dai et al., 2024). This condition is automatic for DAGs and is implied by absolute stability, $\rho(|B|) < 1$ (Misiakos & Püschel, 2026), where ρ denotes the spectral radius and $|B|$ is the entrywise absolute value of B. We observe n i.i.d. copies of $X = (I - B)^{-1}\varepsilon$ and write $\Sigma = \mathrm{Cov}(X) = A\,\mathrm{Cov}(\varepsilon)A^\top \succ 0$, where $A := (I - B)^{-1}$.

For index sets $S, T \subseteq V$, write X_S for the corresponding coordinate subvector, $G[S]$ for the induced subgraph, and $B_{S,T}$ for the submatrix with rows in S and columns in T. We use the same indexing convention for other matrices and omit braces for singleton indices.

Recovery target. An SCC is an inclusion-maximal set of mutually reachable nodes. Write $\Pi = \{S_1, \dots, S_m\}$ for the SCC partition of G. The target is the condensation $G^{\mathrm{sc}} = (\Pi, E^{\mathrm{sc}})$, where

$$E^{\mathrm{sc}} = \{S \to T : S, T \in \Pi,\ S \neq T,\ B_{T,S} \neq 0\}.$$

Recovery requires the SCC partition and every edge of the condensation, including direct edges not determined by reachability alone. This target is invariant across the observational equivalence class of a cyclic LiNG model (Sharifian et al., 2025; Madaleno et al., 2026). For a directed acyclic graph (DAG), every SCC is a singleton, so exact condensation recovery is equivalent to graph recovery.

Structural parameters. For a nonempty set $C \subseteq V$, define its external parent set by

$$P_{\mathrm{ext}}(C) = \{k \in V \setminus C : B_{C,k} \neq 0\}.$$

The maximum SCC size and maximum number of external parent variables per SCC are, respectively,

$$s_{\max} = \max_{S \in \Pi} |S|, \qquad d_B = \max_{S \in \Pi} |P_{\mathrm{ext}}(S)|.$$

We count distinct external parent variables, regardless of how many edges they send into the SCC or whether they belong to the same parent SCC. For DAGs, $s_{\max} = 1$ and d_B is the maximum indegree.

Our sample complexity results use the following regularity condition.

Assumption 2.2 (Sub-Gaussianity and covariance eigenvalue bounds)**.** *There exist constants $K > 0$ and $\lambda \geq 1$, independent of $p, s_{\max}, d_B$, such that, for every $a \in \mathbb{R}^p$,*

$$\|a^\top X\|_{\psi_2} \leq K\sqrt{a^\top \Sigma a}, \qquad \lambda^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \lambda.$$

Here $\|U\|_{\psi_2} := \inf\{a > 0 : \mathbb{E}\exp(U^2/a^2) \leq 2\}$ denotes the sub-Gaussian norm of a random variable U. $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues, respectively.

3 Necessary Structural Sample Costs

Exact condensation recovery requires recovering the SCC partition and identifying external parents. We establish separate information-theoretic lower bounds for these two learning tasks. Their combined structural dependence is attained by the recovery guarantee in Section 5 under suitable conditions.

For integers $s \geq 1$ and $d \geq 0$, let $\mathcal{M}_{p,s,d}(K, \lambda)$ be the class of cyclic LiNG models on p variables satisfying Assumptions 2.1 and 2.2 with constants (K, λ), and $s_{\max} \leq s$, $d_B \leq d$. For each model M, let P_M denote the law of one observation, write $P_M^n := P_M^{\otimes n}$ for the law of the n i.i.d. observations, and let $G^{\mathrm{sc}}(M)$ be its true condensation.

Theorem 3.1 (Sample-complexity lower bound)**.** *There exist universal constants $K, c_{\mathrm{lb}} > 0$ and $\lambda \geq 1$ such that the following holds. For any integers $p \geq 4$, $1 \leq s, d \leq p/2$, and $n \geq 1$, and any $0 < \delta < 1$, if*

$$n \leq \frac{1-\delta}{c_{\mathrm{lb}}}\left[s \log \frac{ep}{s} + d \log \frac{ep}{d}\right], \tag{2}$$

then every estimator $\widehat{G}^{\mathrm{sc}}$ satisfies

$$\sup_{M \in \mathcal{M}_{p,s,d}(K,\lambda)} P_M^n\left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M)\right) \geq \delta.$$

When the sample size satisfies (2), no estimator can achieve worst-case error probability below δ. For $s \geq 2$, identifying SCC membership requires the first cost even when the condensation has no edges. The second cost arises from external-parent selection even on DAGs, where all SCCs are singletons. In this case, $s = 1$ and $\log(ep) \leq d \log(ep/d)$, so the bound has the same order as the parent-selection cost.

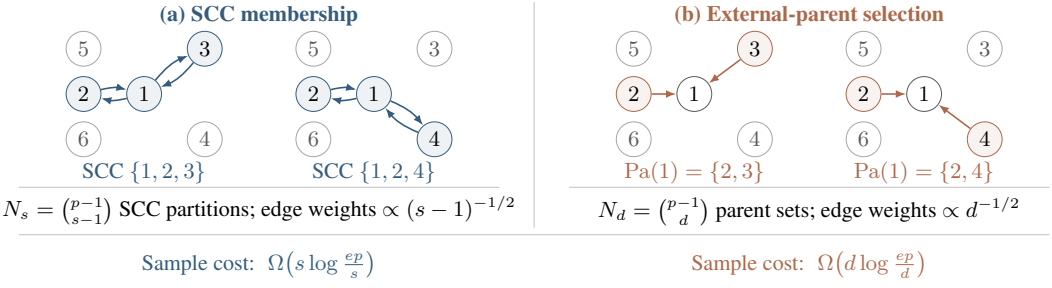

Figure 1: Two lower-bound constructions at $p = 6$: (a) SCC size $s = 3$; (b) parent count $d = 2$.

Establishing the SCC term requires varying the partition itself, since changing internal edges can leave the condensation unchanged. Figure 1(a) varies SCC membership through reciprocal connections to a fixed center, with no edges between SCCs (Lemma B.3). Figure 1(b) varies the parents of one node in a DAG, keeping the SCC partition fixed (Lemma B.4). In both families, scaling the connections limits the information provided by each observation, yielding the costs through Fano's inequality.

4 POPULATION PRINCIPLES FOR EXACT CONDENSATION RECOVERY

4.1 BLOCK EXOGENEITY AND UNKNOWN ROOT SCCS

Discovering a root SCC requires identifying its unknown boundary. Its nodes need not be individually exogenous, and ancestry alone does not distinguish the SCC from larger ancestral sets.

A set is ancestral if it contains every parent, equivalently every ancestor, of each of its nodes. Let $F \subseteq V$ be the already recovered ancestral set and $R = V \setminus F$. An SCC is called a root SCC if its node in the condensation has no incoming edge. For candidate blocks in R, ancestry and root status refer to $G[R]$; a current root SCC may therefore have external parents in F.

For centered random vectors Y and W with $\mathrm{Cov}(W) \succ 0$, define the population least-squares projection by

$$\mathrm{Proj}(Y \mid W) := \mathrm{Cov}(Y, W)\,\mathrm{Cov}(W)^{-1}W.$$

For disjoint index sets $Q, C, D \subseteq V$, with $C \neq \varnothing$, define

$$e_C(Q) := X_C - \mathrm{Proj}(X_C \mid X_Q), \qquad r_{Q,C}(D) := e_D(Q) - \mathrm{Proj}\big(e_D(Q) \mid e_C(Q)\big).$$

We denote their coordinates by $e_c(Q)$ for $c \in C$ and $r_{Q,C}(\ell)$ for $\ell \in D$. Projection onto an empty predictor set is zero. The predictor covariance matrices are positive definite because $\Sigma \succ 0$.

We first take $Q = F$. Since F is ancestral, $B_{F,R} = 0$, and the structural equations give

$$X_R = H_R X_F + (I - B_{R,R})^{-1}\varepsilon_R, \qquad H_R = (I - B_{R,R})^{-1}B_{R,F}. \tag{3}$$

The vector X_F depends only on ε_F, so $\varepsilon_R \perp\!\!\!\perp X_F$. The population projection onto X_F is therefore $H_R X_F$, so

$$e_R(F) = (I - B_{R,R})^{-1}\varepsilon_R. \tag{4}$$

Thus $e_R(F)$ follows the induced cyclic LiNG model on R, so it suffices to characterize its root SCCs.

Theorem 4.1 (Block exogeneity)**.** *Suppose the cyclic LiNG model (1) satisfies Assumption 2.1. Let $F \subseteq V$ be an ancestral set and $R = V \setminus F$. For every nonempty $C \subseteq R$, the following are equivalent:*

(i) C is ancestral in $G[R]$;

(ii) $e_C(F) \perp\!\!\!\perp r_{F,C}(R \setminus C)$;

(iii) $e_c(F) \perp\!\!\!\perp r_{F,C}(\ell)$ for all $c \in C$ and $\ell \in R \setminus C$.

Corollary 4.2 (Root SCC characterization)**.** *Under the conditions of Theorem 4.1, with $R \neq \varnothing$, the inclusion-minimal nonempty sets $C \subseteq R$ satisfying condition (ii) are exactly the root SCCs of $G[R]$.*

After adjustment for F, the residual of an ancestral block C involves only its own noises, and regressing the adjusted complement on $e_C(F)$ removes the effects transmitted from C. The resulting residual vectors depend on disjoint sets of independent noises, giving $(i) \Rightarrow (ii)$. Marginalization gives $(ii) \Rightarrow (iii)$. Conversely, the Darmois–Skitovitch theorem (Darmois, 1953; Skitovitch, 1953), combined with a rank argument based on principal invertibility, shows that any incoming edge to C in $G[R]$ forces $e_c(F)$ and $r_{F,C}(\ell)$ to be dependent for some $c \in C$ and $\ell \in R \setminus C$, giving $(iii) \Rightarrow (i)$. The equivalence of (ii) and (iii) allows block independence to be assessed through scalar independence tests.

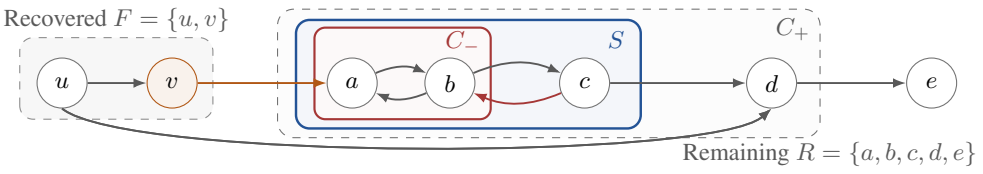

Figure 2: Illustration of block exogeneity and sparse adjustment with overlapping feedback cycles.

In Figure 2, adjustment for $F = \{u, v\}$ removes their upstream effects, so ancestry is evaluated in $G[R]$. Although $C_- = \{a, b\}$ contains the cycle $a \leftrightarrow b$, the red edge $c \to b$ enters from outside C_-. Because of this incoming edge, $e_b(F)$ and $r_{F,C_-}(c)$ share contributions from ε_b and ε_c and

are dependent, so C_- fails condition (ii). For $S = \{a, b, c\}$, $e_S(F)$ depends only on $\varepsilon_a, \varepsilon_b, \varepsilon_c$. Regressing the adjusted variables d, e on $e_S(F)$ removes the effects transmitted from S, leaving residuals driven only by $\varepsilon_d, \varepsilon_e$. These residuals are independent of $e_S(F)$, so S satisfies condition (ii).

The larger candidate $C_+ = S \cup \{d\}$ also has no incoming edge in $G[R]$ and satisfies condition (ii), but is not inclusion-minimal because its proper subset S satisfies the same condition. In contrast, every nonempty proper subset of S has an incoming edge from the rest of S and fails the condition, so S is inclusion-minimal. Corollary 4.2 therefore identifies S as a root SCC.

GroupLiNGAM characterizes exogenous sets through block residual independence under correlation faithfulness (Kawahara et al., 2010). Our block-exogeneity principle shows that independence of every pair of coordinates, one from each of $e_C(F)$ and $r_{F,C}(R \setminus C)$, already suffices to establish ancestry without correlation faithfulness, under principal invertibility and independent structural errors. This characterization identifies the root SCCs of $G[R]$ as exactly the inclusion-minimal nonempty blocks satisfying these scalar independence conditions.

4.2 Sparse adjustment and external-parent identification

Theorem 4.1 uses the full recovered set F, whose size grows over successive recovery rounds. To keep regression dimensions small, we seek a small subset $P \subseteq F$ that still permits root-SCC detection and external-parent identification.

For nonempty $C \subseteq R$ and $P \subseteq F$, we use two conditions. UP checks whether adjustment for P yields the same residual for C as adjustment for all of F. Once this holds, OUT checks whether C is ancestral in $G[R]$.

$$e_c(P) \perp\!\!\!\perp X_k \qquad \forall c \in C,\ k \in F, \tag{UP}$$

$$e_c(P) \perp\!\!\!\perp r_{P,C}(\ell) \qquad \forall c \in C,\ \ell \in R \setminus C. \tag{OUT}$$

We call (C, P) passing if it satisfies (UP) and (OUT), and write $\mathsf{PASS}(C, P) = 1$ for passing pairs and zero otherwise. We call C passing when such a P exists.

Proposition 4.3 (Sparse adjustment and external-parent identification)**.** *Under the conditions of Theorem 4.1, for every $\varnothing \neq C \subseteq R$ and every $P \subseteq F$, the pair (C, P) satisfies (UP)–(OUT) if and only if C is ancestral in $G[R]$ and $P_{\mathrm{ext}}(C) \subseteq P$. Consequently, for an ancestral block C, $P_{\mathrm{ext}}(C)$ is the unique minimum-cardinality passing adjustment.*

Covariance alone suffices to verify UP, even without Gaussianity. Linearity and independent components of ε ensure that any remaining upstream effect is detectable through covariance with X_F:

$$\mathsf{UP} \iff \mathrm{Cov}\big(e_C(P), X_F\big) = 0 \iff e_C(P) = e_C(F).$$

Once this holds, OUT is equivalent to ancestry in $G[R]$, even if upstream effects persist in $R \setminus C$. Both conditions use regressions only on P or $P \cup C$. Variables in $F \setminus P$ enter only the UP covariance checks and are not included as regression predictors.

To see why the minimum adjustment identifies external parents, suppose C is ancestral in $G[R]$. Then

$$X_C = (I - B_{C,C})^{-1} B_{C,F} X_F + e_C(F).$$

The invertible factor $(I - B_{C,C})^{-1}$ preserves which columns of $B_{C,F}$ are nonzero. Feedback may change individual coefficients, but each external parent still appears in at least one of the block's regression equations. Since $\mathrm{Cov}(X_F) \succ 0$, these regression coefficients are unique, so an omitted external parent cannot be replaced by other variables in F. Including all external parents in P, on the other hand, reproduces the regression of X_C on X_F, giving $e_C(P) = e_C(F)$. Therefore the unique minimum-cardinality passing adjustment is exactly $P_{\mathrm{ext}}(C)$.

In Figure 2, adjusting for u leaves the noise ε_v in S, whereas $e_S(\{v\}) = e_S(F)$. The complement residuals retain contributions from u but remain independent of $e_S(\{v\})$:

$$r_{\{v\},S}(d) = B_{d,u} e_u(\{v\}) + \varepsilon_d \quad \perp\!\!\!\perp \quad e_S(\{v\}).$$

The residual $r_{\{v\},S}(e)$ likewise contains a contribution from u transmitted through d and is independent of $e_S(\{v\})$. Thus $\{v\}$ satisfies UP and OUT, while $\varnothing$ and $\{u\}$ fail UP. Proposition 4.3 identifies the minimum adjustment $\{v\}$ as $P_{\mathrm{ext}}(S)$, recovering the incoming condensation edge $\{v\} \to S$ with adjustment for v alone rather than all of F.

4.3 EXACT CONDENSATION RECOVERY

Algorithm 1 applies these principles sequentially: adding a root SCC of $G[R]$ to F yields another ancestral set, so the characterizations continue to apply. A passing pair of minimum total size consists of a root SCC and its exact external parent set. Algorithm 1 recovers the condensation by repeatedly selecting the first passer in increasing order of $|C| + |P|$.

Algorithm 1 BlockExo: exact condensation recovery

Require: Search bounds $s \geq s_{\max}$ and $d \geq d_B$
$F \leftarrow \varnothing$; $\widehat{\Pi} \leftarrow \varnothing$; $\widehat{E}^{\mathrm{sc}} \leftarrow \varnothing$
while $F \neq V$ **do**
  $R \leftarrow V \setminus F$; $S \leftarrow \varnothing$
  for $t = 1, \ldots, \min\{s, |R|\} + \min\{d, |F|\}$ **do**
    for all $C \subseteq R$, $P \subseteq F$ with $1 \leq |C| \leq s$, $|P| \leq d$, $|C| + |P| = t$, in increasing $|C|$ **do**
      if $\mathsf{PASS}(C, P) = 1$ **then**
        $(S, \widehat{P}(S)) \leftarrow (C, P)$; stop all search loops
      end if
    end for
  end for
  if $S = \varnothing$ **then return** failure
  $\widehat{E}^{\mathrm{sc}} \leftarrow \widehat{E}^{\mathrm{sc}} \cup \left\{ T \to S : T \in \widehat{\Pi},\ T \cap \widehat{P}(S) \neq \varnothing \right\}$
  $\widehat{\Pi} \leftarrow \widehat{\Pi} \cup \{S\}$; $F \leftarrow F \cup S$
end while
return $\widehat{G}^{\mathrm{sc}} = (\widehat{\Pi}, \widehat{E}^{\mathrm{sc}})$

Theorem 4.4 (Population recovery)**.** *Suppose the cyclic LiNG model* (1) *satisfies Assumption 2.1, and $s \geq s_{\max}$ and $d \geq d_B$. Algorithm 1 exactly recovers the condensation G^{sc}.*

Search bounds. The procedure does not require the exact values of $s_{\max}$ and d_B. At the population level, the procedure completes if and only if $s \geq s_{\max}$ and $d \geq d_B$ (Proposition C.4). With valid search bounds and correct decisions, every regression uses at most $s_{\max} + d_B$ predictors, and the procedure examines $p^{O(s_{\max}+d_B)}$ candidate pairs in total; neither bound depends on $|F|$, s, or d. In finite samples, estimation error can cause failure even with valid search bounds, but these bounds still limit the total number of candidate pairs examined to $p^{O(s+d)}$. Appendix F.6 compares sample costs under exact and larger valid search bounds, confirming exact recovery under both.

5 FINITE-SAMPLE RECOVERY AND STRUCTURAL OPTIMALITY

We establish a sample bound for exact condensation recovery by Algorithm 1 with any valid search bounds. The sample bound depends on the actual SCC size and external-parent count and matches the structural dependence of the lower bound in Theorem 3.1.

5.1 SCORES AND SEPARATION

The population characterization allows us to assess UP through covariance and OUT through residual dependence. For random variables Y, Z with positive finite variances, let $\mathcal{D}(Y, Z)$ be a nonnegative dependence functional evaluated after centering and scaling to unit variance, satisfying $\mathcal{D}(Y, Z) = 0$ if and only if $Y \perp\!\!\!\perp Z$. Taking empty maxima as zero, define for an ancestral set F, with $R = V \setminus F$, and a candidate pair (C, P),

$$D_F^{\mathrm{up}}(C, P) := \max_{c \in C,\ k \in F} \left| \mathrm{Cov}\big(e_c(P), X_k\big) \right|, \quad D_F^{\mathrm{out}}(C, P) := \max_{c \in C,\ \ell \in R \setminus C} \mathcal{D}\big(e_c(P), r_{P,C}(\ell)\big).$$

Then

$$D_F^{\mathrm{up}}(C, P) = D_F^{\mathrm{out}}(C, P) = 0 \quad \Longleftrightarrow \quad \mathsf{PASS}(C, P) = 1.$$

By Proposition 4.3, this is equivalent to C being ancestral in $G[R]$ and $P_{\mathrm{ext}}(C) \subseteq P$.

We define the candidate class over all ancestral recovery states and the population separation margins:

$$\mathcal{A} := \{(F, C, P) : F \subseteq V \text{ is ancestral in } G,\ \varnothing \neq C \subseteq V \setminus F,\ P \subseteq F,\ |C| + |P| \leq s_{\max} + d_B\}.$$

$$\Delta_{\text{par}} = \min_{(F,C,P)\in\mathcal{A}} \{D_F^{\text{up}}(C, P) : e_C(P) \neq e_C(F)\},$$

$$\Delta_{\mathcal{D}} = \min_{(F,C,P)\in\mathcal{A}} \{D_F^{\text{out}}(C, P) : e_C(P) = e_C(F),\ C \text{ is non-ancestral in } G[V \setminus F]\}. \tag{5}$$

Here, empty minima are $+\infty$, with reciprocal zero. The margin Δ_{par} quantifies the upstream effect remaining after an insufficient adjustment, and $\Delta_{\mathcal{D}}$ quantifies residual dependence when adjustment is sufficient but the block is non-ancestral.

Each insufficient adjustment has a positive UP score. For a sufficiently adjusted non-ancestral block, Proposition 4.3 guarantees at least one dependent scalar pair, so the maximum defining its OUT score is positive. Since $\mathcal{A}$ is finite for a fixed model, both margins are positive. This pointwise positivity does not imply uniform separation over a model class. Proposition E.1 makes this distinction concrete: as the coefficients of a two-variable cycle vanish, its distribution approaches that of the edgeless model while its SCC partition remains different.

5.2 Main recovery guarantee

We use the tail and covariance conditions in Assumption 2.2, as is common in high-dimensional linear SEM analysis (Gao et al., 2022; Ghoshal & Honorio, 2018). Sub-Gaussianity provides the concentration rate used below; consistency can also hold under weaker tail conditions. The covariance eigenvalue bounds control regression estimation and residual standardization. Every regression needed for candidates in $\mathcal{A}$ involves at most $s_{\max} + d_B$ predictors and one response, so the analysis only requires uniform eigenvalue bounds for principal submatrices of size at most $s_{\max} + d_B + 1$. The global condition is a concise sufficient assumption.

We fit the regressions and compute the empirical statistics using all n observations, following the construction in Appendix D. For each scalar pair in OUT, $\mathcal{D}$ is evaluated on the population residuals, whereas $\widehat{\mathcal{D}}$ is computed from the corresponding fitted residuals. The uniform estimation bound below uses $\ell(t)$ defined in Appendix D for confidence level $1 - \delta$, where $t = |C| + |P|$.

Assumption 5.1 (Uniform estimation error for dependence scores)**.** *There exists $c_{\mathcal{D}} > 0$ such that, uniformly over models satisfying Assumption 2.2 with fixed (K, λ), for every $1 \leq t \leq p$, $0 < \delta < 1$, and $n \geq c_{\mathcal{D}}\ell(t)$, with probability at least $1 - \delta/3$,*

$$\sup_{\substack{(F,C,P):\,|C|+|P|=t \\ c\in C,\ \ell\in V\setminus(F\cup C)}} \left|\widehat{\mathcal{D}}\big(e_c(P), r_{P,C}(\ell)\big) - \mathcal{D}\big(e_c(P), r_{P,C}(\ell)\big)\right| \leq c_{\mathcal{D}}\sqrt{\frac{\ell(t)}{n}},$$

where the supremum ranges over ancestral $F \subseteq V$, $\varnothing \neq C \subseteq V \setminus F$, and $P \subseteq F$.

This bound accounts jointly for errors in estimating the residuals and their dependence. The constant $c_{\mathcal{D}}$ may depend on (K, λ) and the chosen estimator, including its fixed tuning parameters, but not on $p, s_{\max}, d_B, n, \delta$. Lemmas D.5 and D.6 show that the distance-covariance and Gaussian-HSIC estimators, respectively, satisfy this assumption.

Using this estimation rate, we set $\tau_{\text{up}}(t) = \kappa_{\text{up}}\sqrt{\ell(t)/n}$ and $\tau_{\text{out}}(t) = \kappa_{\text{out}}\sqrt{\ell(t)/n}$. Let $\widehat{D}_F^{\text{up}}$ and $\widehat{D}_F^{\text{out}}$ denote the empirical counterparts of the population scores. Define $\widehat{\mathsf{PASS}}(C, P) = 1$ if

$$\widehat{D}_F^{\text{up}}(C, P) \leq \tau_{\text{up}}(t) \quad \text{and} \quad \widehat{D}_F^{\text{out}}(C, P) \leq \tau_{\text{out}}(t),$$

and zero otherwise.

For $0 < \delta < 1$, define the complexity term used in the sample bound below:

$$L := s_{\max} \log \frac{ep}{s_{\max}} + d_B \log \frac{ep}{d_B} + \log p + \log s_{\max} + \log \frac{64}{\delta},$$

with $d_B \log(ep/d_B) = 0$ when $d_B = 0$. The first two terms describe the block and adjustment searches. Although $\mathcal{A}$ ranges over ancestral recovery states, each scalar statistic depends only on C, P, and the tested variables, without further dependence on F. Thus the number of ancestral states contributes no additional counting term. Lemma D.1 gives $\ell(\min\{p, s_{\max} + d_B\}) \asymp L$, connecting the candidate-dependent thresholds to the structural sample bound below.

Theorem 5.2 (Achievability of structural sample complexity)**.** *Consider the cyclic LiNG model (1). Suppose Assumptions 2.1, 2.2, and 5.1 hold, with $\mathcal{D}$ as in Section 5.1. Let $\widehat{G}^{\mathrm{sc}}$ be the output of Algorithm 1 with $s \geq s_{\max}$, $d \geq d_B$, and $\widehat{\mathsf{PASS}}$ in place of PASS. There exist fixed threshold multipliers $\kappa_{\mathrm{up}}, \kappa_{\mathrm{out}} > 0$ and a constant $c_0 > 0$, such that, for every $0 < \delta < 1$,*

$$n \geq c_0 L \max\{1, \Delta_{\mathrm{par}}^{-2}, \Delta_{\mathcal{D}}^{-2}\} \quad \Longrightarrow \quad \Pr(\widehat{G}^{\mathrm{sc}} = G^{\mathrm{sc}}) \geq 1 - \delta. \tag{6}$$

The constants $\kappa_{\mathrm{up}}, \kappa_{\mathrm{out}}, c_0$ depend only on $K, \lambda, c_{\mathcal{D}}$.

The choice of $\mathcal{D}$ affects $\Delta_{\mathcal{D}}$ and $c_{\mathcal{D}}$, but not the structural dependence of the sample bound.

5.3 Structural optimality

We establish a uniform recovery guarantee over model classes whose separation margins are bounded away from zero. Fix $\mathcal{D}$ as in Section 5.1. For $\gamma_{\mathrm{par}}, \gamma_{\mathcal{D}} > 0$, define

$$\mathcal{M}^{\mathrm{sep}}_{p,s,d}(K, \lambda; \gamma_{\mathrm{par}}, \gamma_{\mathcal{D}}) := \big\{M \in \mathcal{M}_{p,s,d}(K, \lambda) : \Delta_{\mathrm{par}}(M) \geq \gamma_{\mathrm{par}},\ \Delta_{\mathcal{D}}(M) \geq \gamma_{\mathcal{D}}\big\}.$$

Here (s, d) denotes the structural bounds defining the model class. The margins are evaluated over $\mathcal{A}$ with $s_{\max} + d_B$ replaced by $s + d$. In the corollary below, L is taken at (s, d) instead of $(s_{\max}, d_B)$.

Corollary 5.3 (Uniform recovery and optimality)**.** *Fix $1 \leq s \leq p/2$ and $0 \leq d \leq p/2$. Suppose Assumption 5.1 holds over $\mathcal{M}^{\mathrm{sep}}_{p,s,d}(K, \lambda; \gamma_{\mathrm{par}}, \gamma_{\mathcal{D}})$. Let $\widehat{G}^{\mathrm{sc}}$ be the output of Algorithm 1 with search bounds at least s and d, respectively, using $\widehat{\mathsf{PASS}}$ and the threshold multipliers of Theorem 5.2. With c_0 as in that theorem, for every $0 < \delta < 1$,*

$$n \geq c_0 L \max\{1, \gamma_{\mathrm{par}}^{-2}, \gamma_{\mathcal{D}}^{-2}\} \quad \Longrightarrow \quad \sup_{M \in \mathcal{M}^{\mathrm{sep}}_{p,s,d}(K, \lambda; \gamma_{\mathrm{par}}, \gamma_{\mathcal{D}})} P_M^n\Big(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M)\Big) \leq \delta. \tag{7}$$

Under Assumption 5.1, the separated class of Proposition E.3 has minimax sample complexity $\Theta(\log p)$ for fixed structural bounds and fixed target error probability $0 < \delta < 1/2$.

For fixed $\delta, K, \lambda, \gamma_{\mathrm{par}}, \gamma_{\mathcal{D}}$, the sample bound in (7) matches the structural dependence $s \log(ep/s) + d \log(ep/d)$ of the lower bound in Theorem 3.1. For fixed structural bounds, we further establish minimax sample complexity $\Theta(\log p)$ on the separated model class of Proposition E.3. Larger valid search bounds also preserve this optimal rate: uniform accuracy over the candidate class ensures that each round selects a passing pair with $|C| + |P| \leq s_{\max} + d_B$.

For DAGs, $s = 1$ and d bounds the maximum indegree, so the guarantee applies to exact DAG recovery with structural dependence $\log(ep) + d \log(ep/d) \asymp d \log(ep/d)$ for $d \geq 1$. This agrees with the optimal dimension and sparsity dependence for LiNGAMs in Oh et al. (2025).

6 Simulations

We empirically examine sample cost and compare sample efficiency across methods using sub-Gaussian linear SEMs. We consider disjoint 2-cycles for $s_{\max} = 2$ and overlapping 2-cycles sharing a node for $s_{\max} = 3, 4$. Each point reports the exact-recovery rate across 20 replicates. Implementation details and additional experiments are provided in Appendix F.

6.1 Structural sample costs

Figure 3 provides empirical support for the structural and separation dependence of our sample bound. Recovery curves that differ on the raw sample-size axis align more closely after normalization by the corresponding structural and separation factors. The alignment is stronger in panel (b), which varies dimension at fixed SCC size, than in panel (a), which varies SCC size at fixed dimension. Appendix F.2 details this normalization.

6.2 Sample efficiency comparison across methods

BlockExo achieves sample-efficient exact condensation recovery across disjoint and overlapping cycle structures (Figure 4). At $p = 50$, we compare it with Coarsening (Madaleno et al., 2026) at

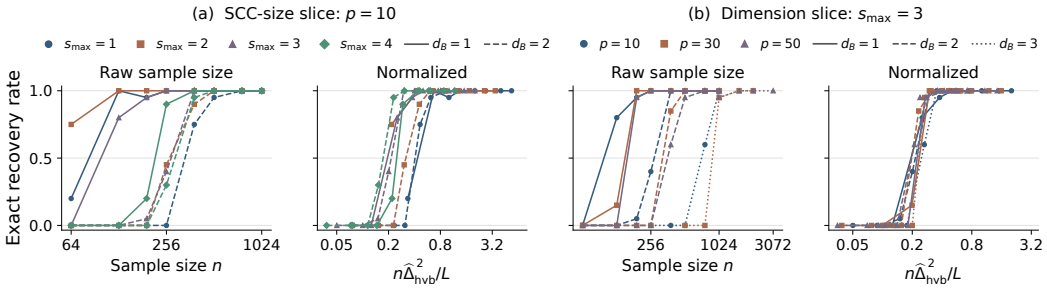

Figure 3: Structural sample costs for exact condensation recovery with exact search bounds.

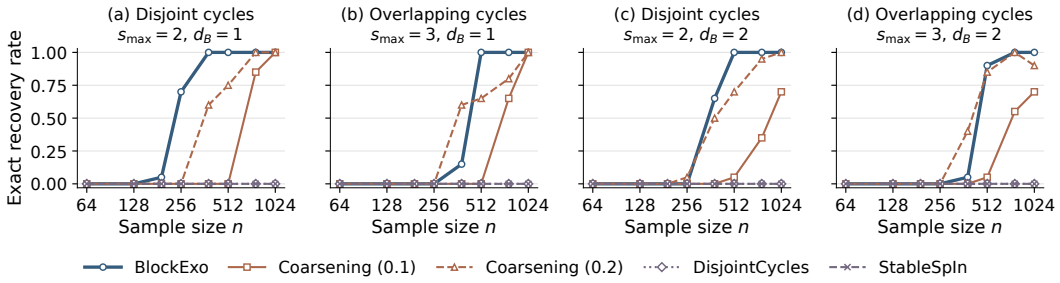

Figure 4: Exact condensation recovery comparison at $p = 50$.

fixed thresholds 0.1 and 0.2, DisjointCycles (Drton et al., 2025), and StableSpIn (Misiakos & Püschel, 2026). BlockExo uses common search bounds $(s, d) = (3, 2)$ across all four settings. DisjointCycles and StableSpIn achieve no exact recoveries over the evaluated grid. GroupLiNGAM (Kawahara et al., 2010) is infeasible under our memory budget (Appendix F.1).

BlockExo achieves sample-efficient exact recovery even without the exact values of $s_{\max}$ and d_B. With $d_B = 1$, it achieves exact recovery in all 20 replicates at $n = 384$ for disjoint cycles and $n = 512$ for overlapping cycles, compared with 12 and 13 of 20 for Coarsening 0.2, respectively. With $d_B = 2$, BlockExo reaches 20 of 20 at $n = 512$ for disjoint cycles, compared with 14 of 20 for Coarsening 0.2, and at $n = 768$ for overlapping cycles, matching Coarsening 0.2. BlockExo's median runtimes range from 8 to 34 seconds across the four settings: longer than Coarsening, shorter than or comparable to StableSpIn, and comparable to DisjointCycles on disjoint cycles (Appendix F.7).

7 Discussion

We establish information-theoretic lower bounds for exact condensation recovery and a procedure whose sample bound matches their structural dependence. The lower bounds distinguish the unavoidable sample costs of SCC membership and external-parent selection. BlockExo achieves this matching under suitable conditions, even when $s_{\max}$ and d_B are unknown. These necessary and achievable structural sample costs provide a concrete target for the design and analysis of methods for exact condensation recovery.

Block exogeneity provides a foundation for developing new algorithms for cyclic causal discovery. Scalar residual independence and inclusion minimality identify root SCCs without a known partition or correlation faithfulness. Sparse adjustment further permits identification of SCCs and their direct external parents without regressing on all previously recovered variables. These characterizations open a path to new algorithms that are both sample-efficient and computationally efficient.

Characterizing the separation margins Δ_{par} and $\Delta_{\mathcal{D}}$ in terms of graph structure, coefficients, and noise distributions is an important next step in explaining sample cost. Our sample bound distinguishes the structural costs from the additional difficulty captured by the separation margins. Relating these margins to model properties would help explain differences in condensation recovery difficulty and identify conditions that guarantee uniform separation.

AI USE STATEMENT

We used generative AI tools to assist with drafting proofs, designing simulation experiments, developing code for synthetic-data generation and method implementation, interpreting empirical results, preparing figures, and revising the manuscript. The authors supplied key prior work, including references needed for the proofs, and guidance on the intended mathematical results and assumptions. The authors reviewed and revised the AI-assisted proofs, refining their mathematical arguments, exposition, and logical organization. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work develops theoretical results for causal discovery and evaluates them using synthetically generated data. It does not involve human participants or personal data. Applications of the proposed recovery principles should assess the stated modeling assumptions, including causal sufficiency, before using the inferred structure to inform substantive decisions.

REPRODUCIBILITY STATEMENT

Sections 2–5 specify the model assumptions and recovery procedures. Appendices B–E provide proofs of the theoretical results, including the separation and optimality analysis. Appendix F describes the synthetic graph constructions, noise distributions, estimator and baseline settings, evaluation protocol, and reference calculations used for normalization. We will publicly release the implementation, experiment configurations, and scripts and data for reproducing the figures.

REFERENCES

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021. doi: 10.1214/21-AOS2064.

Abhishek Chakrabortty, Preetam Nandy, and Hongzhe Li. Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*, 2018. URL https://arxiv.org/abs/1809.10652.

Haoyue Dai, Ignavier Ng, Yujia Zheng, Zhengqing Gao, and Kun Zhang. Local causal discovery with linear non-Gaussian cyclic models. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 154–162, 2024. URL https://proceedings.mlr.press/v238/dai24a.html.

Haoyue Dai, Immanuel Albrecht, Peter Spirtes, and Kun Zhang. Distributional equivalence in linear non-Gaussian latent-variable cyclic causal models: Characterization and learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=b8TlYh6PN6.

Georges Darmois. Analyse générale des liaisons stochastiques: étude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique*, 21(1/2):2–8, 1953.

Mathias Drton, Marina Garrote-López, Niko Nikov, Elina Robeva, and Y. Samuel Wang. Causal discovery for linear non-Gaussian models with disjoint cycles. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1064–1083, 2025. URL https://proceedings.mlr.press/v286/drton25a.html.

Doris Entner and Patrik O. Hoyer. Estimating a causal order among groups of variables in linear models. In *Artificial Neural Networks and Machine Learning – ICANN 2012*, volume 7553 of *Lecture Notes in Computer Science*, pp. 84–91. Springer, 2012. doi: 10.1007/978-3-642-33266-1_11.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933.

Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of Gaussian DAG models. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8738–8757. PMLR, 2022. URL https://proceedings.mlr.press/v151/gao22a.html.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1466–1475. PMLR, 2018. URL https://proceedings.mlr.press/v84/ghoshal18a.html.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pp. 63–77. Springer, 2005. doi: 10.1007/11564089_7.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

Xiaoming Huo and Gábor J. Székely. Fast computing for distance covariance. *Technometrics*, 58(4): 435–447, 2016.

Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(109):3387–3439, 2012. URL https://jmlr.org/papers/v13/hyttinen12a.html.

Abram M. Kagan, Yuri V. Linnik, and C. Radhakrishna Rao. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

Yoshinobu Kawahara, Kenneth A. Bollen, Shohei Shimizu, and Takashi Washio. Grouplingam: Linear non-gaussian acyclic models for sets of variables. *arXiv preprint arXiv:1006.5041*, 2010.

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Edward I. George, and Linda Zhao. Uniform-in-submodel bounds for linear regression in a model-free framework. *Econometric Theory*, 39(6):1202–1248, 2023. doi: 10.1017/S0266466621000219.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 366–374, 2008.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

Francisco Madaleno, Francisco C. Pereira, and Alex Markham. Coarsening linear non-Gaussian causal models with cycles. *arXiv preprint arXiv:2605.10163*, 2026. URL https://arxiv.org/abs/2605.10163.

Panagiotis Misiakos and Markus Püschel. Learning stable digraphs from sparse-input linear structural causal models. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4561–4594, 2026. URL https://proceedings.mlr.press/v337/misiakos26a.html.

Sunmin Oh, Seungsu Han, and Gunwoong Park. Optimal estimation of linear non-Gaussian structure equation models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 748–756, 2025. URL https://proceedings.mlr.press/v258/oh25a.html.

Yannik Pitcan. Concentration inequalities for U-statistics: A survey with explicit constants, 2026. URL https://arxiv.org/abs/1712.06160v3. Version 3.

Carlos Ramos-Carreño and José L. Torrecilla. dcor: Distance correlation and energy statistics in Python. *SoftwareX*, 22:101326, 2023. doi: 10.1016/j.softx.2023.101326. URL https://doi.org/10.1016/j.softx.2023.101326.

Thomas S Richardson and James M Robins. Ace bounds; sems with equilibrium conditions. *Statistical Science*, 29(3):363–366, 2014.

Ehsan Sharifian, Saber Salehkaleybar, and Negar Kiyavash. Near-optimal experiment design in linear non-gaussian cyclic models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70881–70899, 2025. doi: 10.52202/085713-2133.

Viktor P. Skitovitch. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89: 217–219, 1953.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012. URL https://www.jmlr.org/papers/v13/song12a.html.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018. URL https://jmlr.org/papers/v18/17-492.html.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2nd edition, 2026.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 423–435. Springer, New York, 1997.

Ruixuan Zhao, Xin He, and Junhui Wang. Learning linear non-Gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research*, 23(269):1–34, 2022. URL https://www.jmlr.org/papers/v23/21-1173.html.

A Related work: recovery targets and statistical guarantees

Table 1 summarizes related LiNG methods.

Table 1: Recovery targets, population principles, and statistical guarantees of related LiNG methods.

Method	Recovery target	Population basis	Statistical guarantees
OptLiNGAM	sparse DAG	residual independence; parent selection	Structurally matching bounds: $d_{\text{in}} \log(ep/d_{\text{in}})$
GroupLiNGAM	block partition and order	exogenous-set residual independence	Population characterization
DisjointCycles	cycle-disjoint structure	moment relations for root cycles	Structure recovery consistency
StableSpIn	stable graph representative	stability; sparse-input conditions	MLE consistency
Coarsening	condensation	ICA; SCC invariance	Sample bound: $p/(\beta_{\min}^2\sqrt{\delta})$
BlockExo	condensation	block exogeneity; sparse adjustment	Structurally matching bounds: $s\log(ep/s) + d\log(ep/d)$

The displayed rates omit confidence and distributional factors for OptLiNGAM and BlockExo, and other model-dependent constants for Coarsening. The Coarsening bound uses threshold $\beta_{\min}/2$, where $\beta_{\min} := \min_{B_{ij}\neq 0} |B_{ij}|$ (Madaleno et al., 2026).

Acyclic sample complexity. Our structural sample bound matches OptLiNGAM's dimension and sparsity dependence for DAGs and adds the cost of identifying SCC membership in cyclic graphs. OptLiNGAM establishes optimal sample complexity for sparse DAG recovery under sub-Gaussianity, without faithfulness or a known indegree (Oh et al., 2025). For DAGs, $s_{\max} = 1$ and $d_B = d_{\text{in}}$, the maximum indegree, so our structural factor reduces to order $d_{\text{in}} \log(ep/d_{\text{in}})$ for $d_{\text{in}} \geq 1$.

Population characterizations of exogenous blocks. GroupLiNGAM characterizes exogenous sets through block residual independence, whereas our block-exogeneity principle establishes ancestry through scalar residual independence. GroupLiNGAM tests independence between a candidate variable subset and the residual obtained by regressing the remaining variables on it; recursive splitting identifies an ordered partition (Kawahara et al., 2010). Our characterization requires only independence of every scalar pair between $e_C(F)$ and $r_{F,C}(R \setminus C)$ after ancestral adjustment. Inclusion minimality then identifies root SCCs without a known partition. Sparse adjustment further permits ancestry testing and direct-parent identification using small adjustment sets (Proposition 4.3).

GroupLiNGAM assumes correlation faithfulness and permits dependent external influences within a group, allowing within-group latent confounding. Our characterization uses principal invertibility and independent structural errors, without correlation faithfulness. Related work estimates a causal order through residual independence when the multivariate groups are specified in advance (Entner & Hoyer, 2012).

Recovery with disjoint cycles. DisjointCycles recovers structure when directed cycles share no nodes, whereas our block-exogeneity characterization also applies to SCCs with overlapping cycles. For the disjoint-cycle setting, Drton et al. (2025) characterize distributional equivalence and use quadratic and cubic moment relations to locate root cycles. Decorrelation and multivariate regression then yield a block-topological order and a consistent structure-learning procedure. Our approach identifies root SCCs through residual independence and inclusion minimality.

Exact condensation recovery. Coarsening provides ICA-based condensation recovery, while our results establish information-theoretic lower bounds for this target and attain a structurally matching sample bound. Early ICA-based work enumerates cyclic LiNG models compatible with an observed distribution (Lacerda et al., 2008). Sharifian et al. (2025, Theorem 4.4 and Corollary 4.5) establish that the SCC partition and condensation graph are invariant across distribution-equivalent graphs, establishing identifiability of the condensation from observational data. Coarsening recovers the condensation without enumerating the variable-level equivalence class; its finite-sample guarantee follows from control of unmixing-matrix and edge-support estimation (Madaleno et al., 2026). Our lower bounds distinguish the unavoidable costs of SCC membership and external-parent selection (Theorem 3.1), and our recovery guarantee matches their structural dependence under regularity and separation conditions (Corollary 5.3).

StableSpIn targets a stable variable-level graph representative under sparse-input conditions and establishes MLE consistency (Misiakos & Püschel, 2026); our target is the SCC partition and all edges between components. Distributional equivalence has also been characterized for LiNG models with latent variables and cycles (Dai et al., 2026), whereas we study condensation sample complexity under causal sufficiency.

Computational complexity. The exact GroupLiNGAM procedure examines exponentially many subsets in p (Kawahara et al., 2010). For OptLiNGAM, the distance-covariance implementation has cost $O(n \log n\, p^{d_{\text{in}}+1} d_{\text{in}}^4 (p - d_{\text{in}}))$ when all independence decisions are correct (Oh et al., 2025, Lemma 18). With valid search bounds and correct decisions, BlockExo examines $p^{O(s_{\max}+d_B)}$ candidate pairs and uses at most $s_{\max} + d_B$ predictors in each regression, independently of the input search bounds. Its empirical worst-case candidate count is $p^{O(s+d)}$, with regressions and scalar tests for each pair. Coarsening has cost $O(knp^2 + p^3)$, where k is the number of FastICA iterations (Madaleno et al., 2026, Theorem 2). StableSpIn's optimization costs $O(np^2 + p^3)$ per iteration for a fixed power-iteration count (Misiakos & Püschel, 2026).

B STRUCTURAL LOWER-BOUND PROOFS

For a minimax lower bound, it suffices to establish the claim on a suitable subclass of the model class. Throughout this section, we therefore restrict attention to models whose structural noises are i.i.d. from the following fixed distribution. Let ϕ_σ denote the $N(0,\sigma^2)$ density and set

$$f(x) = \frac{1}{2}\phi_{1/\sqrt{2}}\left(x - \frac{1}{\sqrt{2}}\right) + \frac{1}{2}\phi_{1/\sqrt{2}}\left(x + \frac{1}{\sqrt{2}}\right).$$

If $\eta \sim f$, then $\eta = Z + R/\sqrt{2}$ in distribution, where $Z \sim N(0,1/2)$, R is independent of Z, and $\mathbb{P}(R=1) = \mathbb{P}(R=-1) = 1/2$. Hence $\mathbb{E}\eta = 0$, $\mathrm{Var}(\eta) = 1$, $\mathbb{E}e^{t\eta} = e^{t^2/4}\cosh(t/\sqrt{2}) \le e^{t^2/2}$, and $\mathbb{E}\eta^4 = 5/2 \neq 3$. Thus f is centered, variance one, non-Gaussian, and sub-Gaussian. Writing $g = \log f$, we have $g(x) = \mathrm{const} - x^2 + \log\cosh(\sqrt{2}x)$ and $g''(x) = -2\tanh^2(\sqrt{2}x)$, so $\|g''\|_\infty = 2$.

For each coefficient matrix B, let P_B denote the law of

$$X = (I - B)^{-1}\varepsilon,$$

and write $G^{\mathrm{sc}}(B)$ for the condensation of the graph determined by B. We first establish common KL and regularity bounds, then construct separate packings for SCC membership and external-parent selection. The generalized-Fano argument follows the sparse-DAG lower-bound framework of Gao et al. (2022) and Oh et al. (2025).

B.1 COMMON KL AND REGULARITY BOUNDS

Lemma B.1 (Local KL control)**.** *For every constant $\rho_0 \in (0,1)$ independent of dimension p and every zero-diagonal matrix B satisfying $\|B\|_{\mathrm{op}} \le \rho_0$,*

$$\mathrm{KL}(P_B \| P_0) \le \left\{\frac{1}{1-\rho_0} + \frac{1}{(1-\rho_0)^2}\right\}\|B\|_F^2.$$

Proof. The change of variables $\varepsilon = (I-B)x$ gives $p_B(x) = |\det(I-B)|\prod_{i=1}^p f(((I-B)x)_i)$ and $p_0(x) = \prod_{i=1}^p f(x_i)$. For every $t \in [0,1]$, the bound $\|tB\|_{\mathrm{op}} < 1$ implies $\det(I - tB) \neq 0$. Continuity of $t \mapsto \det(I-tB)$ and $\det(I) = 1$ therefore give $\det(I-B) > 0$.

Let $\Gamma = (I-B)^{-1} - I$. Under P_B, $X = \varepsilon + \Gamma\varepsilon$, and hence

$$\log\frac{p_B(X)}{p_0(X)} = \log\det(I-B) + \sum_{i=1}^p \{g(\varepsilon_i) - g(\varepsilon_i + (\Gamma\varepsilon)_i)\}.$$

Taylor's theorem and $-g''(x) = 2\tanh^2(\sqrt{2}x) \le 2$ give

$$g(\varepsilon_i) - g(\varepsilon_i + (\Gamma\varepsilon)_i) \le -g'(\varepsilon_i)(\Gamma\varepsilon)_i + (\Gamma\varepsilon)_i^2.$$

Since $f(x), xf(x) \to 0$ as $|x| \to \infty$, integration by parts gives $\mathbb{E}[g'(\varepsilon_i)] = 0$ and $\mathbb{E}[g'(\varepsilon_i)\varepsilon_i] = -1$. Together with independence, centering, and $\mathrm{Cov}(\varepsilon) = I$, this yields

$$\mathbb{E}[g'(\varepsilon_i)\varepsilon_j] = -\mathbf{1}\{i=j\}, \quad -\sum_{i=1}^p \mathbb{E}[g'(\varepsilon_i)(\Gamma\varepsilon)_i] = \mathrm{tr}(\Gamma), \quad \mathbb{E}\|\Gamma\varepsilon\|_2^2 = \mathrm{tr}(\Gamma\Gamma^\top) = \|\Gamma\|_F^2.$$

Taking expectations in the Taylor bound and applying these identities gives

$$\mathrm{KL}(P_B\|P_0) \le \log\det(I-B) + \mathrm{tr}(\Gamma) + \|\Gamma\|_F^2. \tag{8}$$

Since $\|B\|_{\mathrm{op}} < 1$, the log-determinant and Neumann series give

$$\log\det(I-B) = -\sum_{k\ge1}\frac{\mathrm{tr}(B^k)}{k}, \qquad \Gamma = \sum_{k\ge1} B^k.$$

For $k \ge 1$, we have

$$\|B^k\|_F \le \|B\|_F\|B\|_{\mathrm{op}}^{k-1} \le \|B\|_F\rho_0^{k-1},$$
$$|\mathrm{tr}(B^k)| = |\langle B^\top, B^{k-1}\rangle_F| \le \|B\|_F\|B^{k-1}\|_F.$$

These bounds give

$$|\log\det(I-B)+\operatorname{tr}(\Gamma)| = \left|\sum_{k\geq 2}\left(1-\frac{1}{k}\right)\operatorname{tr}(B^k)\right| \leq \|B\|_F^2 \sum_{k\geq 2}\rho_0^{k-2} = \frac{\|B\|_F^2}{1-\rho_0},$$

$$\|\Gamma\|_F \leq \sum_{k\geq 1}\|B^k\|_F \leq \|B\|_F \sum_{k\geq 1}\rho_0^{k-1} = \frac{\|B\|_F}{1-\rho_0}.$$

Substituting these bounds into (8) yields

$$\mathrm{KL}(P_B\|P_0) \leq \frac{\|B\|_F^2}{1-\rho_0} + \frac{\|B\|_F^2}{(1-\rho_0)^2} = \left\{\frac{1}{1-\rho_0}+\frac{1}{(1-\rho_0)^2}\right\}\|B\|_F^2.$$

□

Lemma B.2 (Uniform packing regularity)**.** *Under the fixed noise distribution above, suppose* $\|B\|_{\mathrm{op}} \leq \rho_0 < 1$*. Then*

$$(1+\rho_0)^{-2} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq (1-\rho_0)^{-2},$$

where

$$\Sigma = (I-B)^{-1}(I-B)^{-\top}.$$

Moreover, there exists a universal constant $K > 0$ *such that, for every deterministic* $a \in \mathbb{R}^p$*,*

$$\|a^\top X\|_{\psi_2} \leq K\sqrt{a^\top \Sigma a}.$$

Proof. For every $x \in \mathbb{R}^p$,

$$(1-\rho_0)\|x\|_2 \leq \|(I-B)x\|_2 \leq (1+\rho_0)\|x\|_2.$$

Thus the singular values of $(I-B)^{-1}$ lie in $[(1+\rho_0)^{-1},(1-\rho_0)^{-1}]$. Hence the eigenvalues of Σ lie in the squared interval $[(1+\rho_0)^{-2},(1-\rho_0)^{-2}]$.

For any deterministic $a \in \mathbb{R}^p$, let $c = (I-B)^{-\top}a$. The moment-generating-function bound above implies $\|\varepsilon_i\|_{\psi_2} \leq C_0$ for a universal constant C_0. The standard inequality for sums of independent centered sub-Gaussian variables therefore gives

$$\|a^\top X\|_{\psi_2} = \|c^\top \varepsilon\|_{\psi_2} \leq C\Big(\sum_{i=1}^p c_i^2\|\varepsilon_i\|_{\psi_2}^2\Big)^{1/2} \leq CC_0\|c\|_2 = CC_0\sqrt{a^\top\Sigma a}.$$

Taking $K := CC_0$ proves the claim. □

For the remainder of this section, fix universal constants $c_{\mathrm{KL}}, K > 0$ and $\lambda \geq 1$ such that every B with $\|B\|_{\mathrm{op}} \leq 1/2$ satisfies

$$\mathrm{KL}(P_B\|P_0) \leq c_{\mathrm{KL}}\|B\|_F^2, \qquad \lambda^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \lambda,$$

as well as the sub-Gaussian bound in Lemma B.2. Such constants exist by the two lemmas above.

B.2 SCC AND EXTERNAL-PARENT PACKINGS

Lemma B.3 (Reciprocal-star SCC packing)**.** *For every* $2 \leq s \leq p/2$ *and every* $S \subseteq \{2,\dots,p\}$ *with* $|S| = s-1$*, define* B_S *by*

$$(B_S)_{1j} = (B_S)_{j1} = \frac{1}{2\sqrt{s-1}}, \qquad j \in S,$$

with all other entries equal to zero. Then:

(i) The model defined by B_S *and* f *belongs to* $\mathcal{M}_{p,s,d}(K,\lambda)$ *for every* $d \geq 0$*. Its only nontrivial SCC is* $\{1\} \cup S$*, and* $d_B = 0$*.*

(ii) Writing $N_s := \binom{p-1}{s-1}$, the packing contains $N_s \geq 3$ distinct condensations and

$$\log N_s \geq \frac{\log 2}{2(1+\log 2)} s \log \frac{ep}{s}.$$

(iii) The observation laws satisfy

$$\max_S \mathrm{KL}(P^n_{B_S} \| P^n_0) \leq \frac{c_{\mathrm{KL}} n}{2}.$$

Proof. We first verify statement (i). The matrix B_S represents a weighted reciprocal star with center 1 and leaves S, with all other nodes isolated. Let e_j denote the standard coordinate vectors and set $u_S = (s-1)^{-1/2} \sum_{j \in S} e_j$. Then

$$B_S e_1 = \tfrac{1}{2} u_S, \qquad B_S u_S = \tfrac{1}{2} e_1, \qquad B_S v = 0 \quad \text{for } v \perp \mathrm{span}\{e_1, u_S\}.$$

Thus the symmetric matrix B_S has nonzero eigenvalues $1/2$ and $-1/2$, giving $\|B_S\|_{\mathrm{op}} = 1/2$. Every principal submatrix of B_S also has operator norm at most $1/2$, so every principal submatrix of $I - B_S$ is invertible. Thus Assumption 2.1 holds.

The nodes in $\{1\} \cup S$ are mutually reachable through node 1, whereas all other nodes are isolated. Hence $\{1\} \cup S$ is the only nontrivial SCC and no SCC has an external parent. Since $\|B_S\|_{\mathrm{op}} = 1/2$, Lemma B.2 with $\rho_0 = 1/2$ and the constants fixed above verifies Assumption 2.2. Therefore the model belongs to $\mathcal{M}_{p,s,d}(K, \lambda)$ for every $d \geq 0$.

For statement (ii), distinct sets S produce distinct nontrivial SCCs and therefore distinct condensations. For $1 \leq k \leq N$,

$$\binom{N}{k} = \prod_{j=0}^{k-1} \frac{N-j}{k-j} \geq \left(\frac{N}{k}\right)^k, \tag{9}$$

because $(N-j)/(k-j) \geq N/k$ is equivalent to $j(N-k) \geq 0$. Applying (9) with $(N, k) = (p-1, s-1)$ gives

$$\log N_s \geq (s-1) \log \frac{p-1}{s-1} \geq \frac{s}{2} \log \frac{p}{s} \geq \frac{\log 2}{2(1+\log 2)} s \log \frac{ep}{s}.$$

The second inequality uses $s - 1 \geq s/2$ and $(p-1)/(s-1) \geq p/s$. The third follows from $p/s \geq 2$ and $\log(p/s)/\log(ep/s) \geq \log 2/(1+\log 2)$, proving statement (ii).

For statement (iii), $\|B_S\|_F^2 = 1/2$, so Lemma B.1 and tensorization give

$$\mathrm{KL}(P^n_{B_S} \| P^n_0) \leq \frac{c_{\mathrm{KL}} n}{2}.$$

This bound holds for every S, completing the proof. □

Lemma B.4 (Parent-star external-parent packing)**.** *For every $p \geq 4$, $1 \leq d \leq p/2$, and $Q \subseteq \{2, \ldots, p\}$ with $|Q| = d$, define B_Q by*

$$(B_Q)_{1q} = \frac{1}{2\sqrt{d}}, \qquad q \in Q,$$

with all other entries equal to zero. Then:

(i) The model defined by B_Q and f belongs to $\mathcal{M}_{p,s,d}(K, \lambda)$ for every $s \geq 1$. It is acyclic, satisfies $s_{\max} = 1$, and has $d_B = d$.

(ii) Writing $N_d := \binom{p-1}{d}$, the packing contains $N_d \geq 3$ distinct condensations and

$$\log N_d \geq \frac{\log 2}{2(1+\log 2)} d \log \frac{ep}{d}.$$

(iii) The observation laws satisfy

$$\max_Q \mathrm{KL}(P^n_{B_Q} \| P^n_0) \leq \frac{c_{\mathrm{KL}} n}{4}.$$

Proof. We first verify statement (i). The only directed edges are $q \to 1$ for $q \in Q$, so the graph is acyclic with $s_{\max} = 1$ and $d_B = d$. Acyclicity also implies Assumption 2.1: after a topological ordering, every principal submatrix of $I - B_Q$ is triangular with unit diagonal. Since B_Q has only one nonzero row,

$$\|B_Q\|_{\mathrm{op}}^2 = \|B_Q\|_F^2 = d\left(\frac{1}{2\sqrt{d}}\right)^2 = \frac{1}{4}.$$

Lemma B.2 at $\rho_0 = 1/2$ and the constants fixed above therefore imply that the model defined by B_Q and f belongs to $\mathcal{M}_{p,s,d}(K,\lambda)$ for every $s \geq 1$.

For statement (ii), distinct sets Q give different parent sets of node 1, so the packing contains exactly $N_d = \binom{p-1}{d}$ distinct condensations. Since $1 \leq d \leq p-2$, $N_d \geq p-1 \geq 3$. Applying (9) with $(N,k) = (p-1,d)$ yields

$$\begin{aligned}\log N_d \geq d \log \frac{p-1}{d} &\geq d\left\{\log(3/2) + \log\frac{p}{2d}\right\} \\ &\geq \frac{\log 2}{2(1+\log 2)} d\left\{1 + \log 2 + \log\frac{p}{2d}\right\} = \frac{\log 2}{2(1+\log 2)} d \log\frac{ep}{d}.\end{aligned}$$

The second inequality uses $p - 1 \geq 3p/4$ for $p \geq 4$. The third uses $\frac{1}{2}\log 2 < \log(3/2)$, $\log 2/[2(1+\log 2)] < 1$, and $\log(p/(2d)) \geq 0$.

For statement (iii), Lemma B.1 at $\rho_0 = 1/2$ and tensorization give

$$\mathrm{KL}(P_{B_Q}^n \| P_0^n) = n\,\mathrm{KL}(P_{B_Q}\|P_0) \leq c_{\mathrm{KL}} n \|B_Q\|_F^2 = \frac{c_{\mathrm{KL}} n}{4}.$$

This bound holds for every Q, completing the proof. □

B.3 Proof of Theorem 3.1

Proof of Theorem 3.1. The parent family lies in $\mathcal{M}_{p,s,d}(K,\lambda)$ for every $s \geq 1$, and the SCC family does so for $s \geq 2$, with the universal constants fixed above. Write $h_s = s\log(ep/s)$ and $h_d = d\log(ep/d)$. Lemmas B.3 and B.4 give the following bounds, with the SCC bound applying only for $s \geq 2$:

$$\log N_s \geq \frac{\log 2}{2(1+\log 2)} h_s, \qquad \max_S \mathrm{KL}(P_{B_S}^n \| P_0^n) \leq \frac{c_{\mathrm{KL}} n}{2}, \tag{10}$$

$$\log N_d \geq \frac{\log 2}{2(1+\log 2)} h_d, \qquad \max_Q \mathrm{KL}(P_{B_Q}^n \| P_0^n) \leq \frac{c_{\mathrm{KL}} n}{4}. \tag{11}$$

Set

$$c_{\mathrm{lb}} := \frac{2(1+\log 2)c_{\mathrm{KL}}}{\log 2} + 4(1+\log 2).$$

Since the models within each packing have distinct condensations, generalized Fano's inequality (Yu, 1997; Tsybakov, 2009), applied separately to (10) and (11), gives

$$\begin{aligned}\max_S P_{B_S}^n\left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(B_S)\right) &\geq 1 - \frac{(1+\log 2)c_{\mathrm{KL}} n}{\log 2\, h_s} - \frac{2(1+\log 2)}{h_s} \geq 1 - \frac{c_{\mathrm{lb}} n}{2h_s}, \qquad s \geq 2, \\ \max_Q P_{B_Q}^n\left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(B_Q)\right) &\geq 1 - \frac{(1+\log 2)c_{\mathrm{KL}} n}{2\log 2\, h_d} - \frac{2(1+\log 2)}{h_d} \geq 1 - \frac{c_{\mathrm{lb}} n}{2h_d}.\end{aligned}$$

The last inequalities use $n \geq 1$. The first bound is obtained on models with $d_B = 0$, and the second is obtained on acyclic models; hence they also establish the two separate necessity claims in Section 3.

It remains to combine these bounds over the full model class. If $s \geq 2$, use the packing corresponding to $\max\{h_s, h_d\}$. Since both packings are subsets of $\mathcal{M}_{p,s,d}(K,\lambda)$,

$$\sup_{M \in \mathcal{M}_{p,s,d}(K,\lambda)} P_M^n\left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M)\right) \geq 1 - \frac{c_{\mathrm{lb}} n}{2\max\{h_s, h_d\}} \geq 1 - \frac{c_{\mathrm{lb}} n}{h_s + h_d}.$$

If $s = 1$, the SCC packing is unavailable, but the function $x \mapsto x \log(ep/x)$ is increasing on $[1, p/2]$, and therefore $h_s = \log(ep) \le h_d$. The parent-packing bound then gives

$$\sup_{M \in \mathcal{M}_{p,1,d}(K,\lambda)} P_M^n \left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M) \right) \ge 1 - \frac{c_{\mathrm{lb}} n}{2h_d} \ge 1 - \frac{c_{\mathrm{lb}} n}{h_s + h_d}.$$

Thus the same bound holds for every $1 \le s \le p/2$. Under the sample-size condition in (2), its right-hand side is at least δ. □

C Population proof details

Throughout this section, we consider the cyclic LiNG model (1) on $V = [p]$ under Assumption 2.1, using the notation of Section 2. Thus B is the weighted adjacency matrix, $A = (I - B)^{-1}$, and the coordinates of ε are mutually independent, with at most one Gaussian coordinate.

C.1 Block exogeneity and root SCCs

The following lemma isolates the argument used to rule out incoming edges.

Lemma C.1 (Residual independence forces ancestry)**.** *Let $R \subseteq V$ be partitioned as $R = C \cup D$, with $C \neq \varnothing$, and write $A^{(R)} = (I - B_{R,R})^{-1}$. For matrices Q and L of compatible dimensions, let*

$$Y = A^{(R)}_{C,R} \varepsilon_R, \qquad Z = (A^{(R)}_{D,R} - Q A^{(R)}_{C,R}) \varepsilon_R + L \varepsilon_{V \setminus R}.$$

If $\mathrm{Cov}(Y, Z) = 0$ and $Y_c \perp\!\!\!\perp Z_\ell$ for every $c \in C$ and $\ell \in D$, then C is ancestral in $G[R]$.

Proof. The claim is immediate when $D = \varnothing$, so assume $D \neq \varnothing$. Assumption 2.1 and Jacobi's complementary-minor identity (Horn & Johnson, 2013, Section 0.8.5) imply that the principal submatrices $A^{(R)}_{C,C}$ and $A^{(R)}_{D,D}$ are nonsingular: indeed, $\det(A^{(R)}_{C,C}) = \det(I - B_{D,D}) / \det(I - B_{R,R})$ and $\det(A^{(R)}_{D,D}) = \det(I - B_{C,C}) / \det(I - B_{R,R})$ are nonzero.

Since Y depends only on ε_R, independence of the sources and $\mathrm{Cov}(Y, Z) = 0$ give

$$\sum_{k \in R} \mathrm{Var}(\varepsilon_k) A^{(R)}_{C,k} \big(A^{(R)}_{D,k} - Q A^{(R)}_{C,k} \big)^\top = 0.$$

The Darmois–Skitovitch theorem (Darmois, 1953; Skitovitch, 1953; Kagan et al., 1973), applied to each pair (Y_c, Z_ℓ), gives

$$A^{(R)}_{C,k} \big(A^{(R)}_{D,k} - Q A^{(R)}_{C,k} \big)^\top = 0$$

for every non-Gaussian source ε_k, $k \in R$. There is at most one Gaussian source. Hence either every summand already vanishes, or the covariance identity forces the sole remaining summand to vanish. Therefore, for every $k \in R$,

$$A^{(R)}_{C,k} \neq 0 \quad \Longrightarrow \quad A^{(R)}_{D,k} - Q A^{(R)}_{C,k} = 0. \tag{12}$$

Nonsingularity of $A^{(R)}_{C,C}$ and (12) imply

$$A^{(R)}_{D,C} - Q A^{(R)}_{C,C} = 0.$$

If C were non-ancestral, then $B_{C,D} \neq 0$. The (C, D) block of $I = (I - B_{R,R}) A^{(R)}$ is zero, which gives

$$0 = (I - B_{C,C}) A^{(R)}_{C,D} - B_{C,D} A^{(R)}_{D,D}.$$

Since $A^{(R)}_{D,D}$ is nonsingular, $A^{(R)}_{C,D} \neq 0$. Choose $k \in D$ with $A^{(R)}_{C,k} \neq 0$. Equation (12) then yields

$$A^{(R)}_{R, C \cup \{k\}} = \begin{bmatrix} I_{|C|} \\ Q \end{bmatrix} A^{(R)}_{C, C \cup \{k\}}.$$

The matrix $A^{(R)}_{C, C \cup \{k\}}$ has rank $|C|$ because it contains the nonsingular submatrix $A^{(R)}_{C,C}$, and $\left[\begin{smallmatrix} I_{|C|} \\ Q \end{smallmatrix} \right]$ has full column rank. Hence the right-hand side has rank $|C|$. In contrast, the left-hand side consists of $|C| + 1$ columns of the invertible matrix $A^{(R)}$ and therefore has rank $|C| + 1$, a contradiction. Thus $B_{C,D} = 0$, so C is ancestral in $G[R]$. □

Proof of Theorem 4.1. Statement (ii) implies statement (iii) by marginalization. We prove the other two implications. Write $D = R \setminus C$. Since F is ancestral, (4) gives

$$e_C(F) = A_{C,R}\varepsilon_R, \qquad e_D(F) = A_{D,R}\varepsilon_R, \qquad A_{R,R} = (I - B_{R,R})^{-1}. \tag{13}$$

Let $\Omega_R = \mathrm{Cov}(\varepsilon_R)$, which is diagonal and positive definite, and define

$$Q_C := \mathrm{Cov}(e_D(F), e_C(F))\,\mathrm{Cov}(e_C(F))^{-1} = A_{D,R}\Omega_R A_{C,R}^\top \big(A_{C,R}\Omega_R A_{C,R}^\top\big)^{-1}. \tag{14}$$

The inverse exists because $A_{C,C}$ is nonsingular and $\Omega_R \succ 0$. By definition,

$$r_{F,C}(D) = (A_{D,R} - Q_C A_{C,R})\varepsilon_R. \tag{15}$$

For $(i) \Rightarrow (ii)$, suppose that C is ancestral in $G[R]$. Then $B_{C,D} = 0$ and

$$A_{R,R} = \begin{bmatrix} A_{C,C} & 0 \\ A_{D,C} & A_{D,D} \end{bmatrix}.$$

Consequently,

$$e_C(F) = A_{C,C}\varepsilon_C, \qquad Q_C = A_{D,C}A_{C,C}^{-1}, \qquad r_{F,C}(D) = A_{D,D}\varepsilon_D.$$

$e_C(F)$ and $r_{F,C}(D)$ depend on disjoint independent source vectors, proving statement (ii).

For $(iii) \Rightarrow (i)$, suppose that statement (iii) holds. Since $r_{F,C}(D)$ is the projection residual, $\mathrm{Cov}(e_C(F), r_{F,C}(D)) = 0$. The representations in (13) and (15) satisfy Lemma C.1 with $Q = Q_C$ in (14) and no additional source term. The lemma therefore shows that C is ancestral in $G[R]$, proving statement (i). □

The graph-theoretic part of the characterization will also be used below.

Lemma C.2 (Ancestral sets and root SCCs)**.** *In any finite directed graph, the root SCCs are exactly the inclusion-minimal nonempty ancestral sets. Moreover, every nonempty ancestral set contains a root SCC.*

Proof. A root SCC is ancestral because it has no incoming edge from another SCC. For every nonempty proper subset of a root SCC, strong connectivity provides a path from outside the subset to a node inside it, so the subset is not ancestral.

Now let U be a nonempty ancestral set. Then U is a union of SCCs, and the condensation subgraph induced by these SCCs has a source S. No SCC inside U enters S by the source property, and no SCC outside U enters S by ancestry. Thus S is a root SCC of the full graph. Every nonempty ancestral set therefore contains a root SCC, and inclusion minimality forces equality. □

Proof of Corollary 4.2. By Theorem 4.1, the nonempty sets satisfying condition (ii) are exactly the nonempty ancestral sets of $G[R]$. Lemma C.2 identifies their inclusion-minimal elements as the root SCCs of $G[R]$. □

C.2 Sparse adjustment and external parents

Proof of Proposition 4.3. If $F = \varnothing$, then $P = \varnothing$ and UP holds automatically. Theorem 4.1 gives the equivalence between OUT and ancestry, and an ancestral C has $P_{\mathrm{ext}}(C) = \varnothing$. Thus all claims hold in this case. Assume henceforth that $F \neq \varnothing$.

Write $D = R \setminus C$. For $T = C, D$, let H_T denote the rows of H_R indexed by T in (3). Then

$$X_T = H_T X_F + e_T(F), \qquad T = C, D. \tag{16}$$

Here $e_T(F) \perp\!\!\!\perp X_F$ by the ancestry of F. Define $G_T = G_T(P)$ by

$$G_T X_F := H_T X_F - \mathrm{Proj}(H_T X_F \mid X_P).$$

Since X_P is a subvector of X_F, independence gives $\mathrm{Proj}(e_T(F) \mid X_P) = 0$. Therefore

$$\begin{aligned} e_T(P) &= X_T - \mathrm{Proj}(X_T \mid X_P) \\ &= e_T(F) + H_T X_F - \mathrm{Proj}(H_T X_F \mid X_P) \\ &= e_T(F) + G_T X_F. \end{aligned}$$

If UP holds, the independence in that condition gives

$$0 = \mathrm{Cov}(e_C(P), X_F) = G_C \,\mathrm{Cov}(X_F).$$

Positive definiteness of $\mathrm{Cov}(X_F)$ gives $G_C = 0$. Conversely, $e_C(P) = e_C(F)$ implies $e_C(P) \perp\!\!\!\perp X_F$. Hence

$$\mathsf{UP} \iff e_C(P) = e_C(F). \tag{17}$$

Under UP, we have $e_C(P) = e_C(F)$. Since $X_F \perp\!\!\!\perp e_C(F)$, the projection coefficients of $e_D(P)$ and $e_D(F)$ onto the common candidate residual are both Q_C from (14). Consequently,

$$r_{P,C}(D) = e_D(F) + G_D X_F - Q_C e_C(F) = r_{F,C}(D) + G_D X_F. \tag{18}$$

For the forward implication, suppose that UP and OUT hold. Since F is ancestral, $X_F = A_{F,F}\varepsilon_F$. The representations (13), (15), and (18) have the form in Lemma C.1. Condition OUT supplies the required scalar independences, while the projection normal equation gives $\mathrm{Cov}(e_C(P), r_{P,C}(D)) = 0$. Thus Lemma C.1 applies with $Q = Q_C$ and $L = G_D A_{F,F}$. Hence C is ancestral in $G[R]$.

For this ancestral C, block triangularity gives $H_C = (I - B_{C,C})^{-1} B_{C,F}$. Because $(I - B_{C,C})^{-1}$ is invertible and ancestry rules out parents in D,

$$\{j \in F : H_{C,j} \neq 0\} = \{j \in F : B_{C,j} \neq 0\} = P_{\mathrm{ext}}(C).$$

Moreover, UP and (17) give

$$\mathrm{Proj}(X_C \mid X_P) = H_C X_F.$$

Positive definiteness of $\mathrm{Cov}(X_F)$ makes the linear coefficients in X_F unique. Since the coefficient matrix of the left-hand side has zero columns outside P, every nonzero column of H_C must be indexed by P. Using the support identity above, we obtain $P_{\mathrm{ext}}(C) \subseteq P$.

For the reverse implication, suppose that C is ancestral in $G[R]$ and $P_{\mathrm{ext}}(C) \subseteq P$. The support identity above uses only ancestry, so $H_C X_F$ is a linear function of X_P. By (16) and $e_C(F) \perp\!\!\!\perp X_F$, $\mathrm{Proj}(X_C \mid X_P) = H_C X_F$, and hence $e_C(P) = e_C(F)$. Equation (17) gives UP.

The proof of Theorem 4.1 gives

$$e_C(F) = A_{C,C}\varepsilon_C, \qquad r_{F,C}(D) = A_{D,D}\varepsilon_D.$$

Since $X_F = A_{F,F}\varepsilon_F$, equation (18) gives

$$r_{P,C}(D) = A_{D,D}\varepsilon_D + G_D A_{F,F}\varepsilon_F, \qquad e_C(P) = A_{C,C}\varepsilon_C.$$

Independence of the noise coordinates therefore gives $e_C(P) \perp\!\!\!\perp r_{P,C}(D)$, proving OUT.

The minimum-cardinality claim follows because the passing adjustments for an ancestral C are exactly the supersets of $P_{\mathrm{ext}}(C)$. □

C.3 Minimum passers and population recovery

Lemma C.3 (Minimum passer)**.** *Fix an ancestral F with $R = V \setminus F \neq \varnothing$. Let (C, P) minimize $|C| + |P|$ among passing pairs. Then C is a root SCC of $G[R]$ and $P = P_{\mathrm{ext}}(C)$.*

Proof. Let (C, P) be any passing pair. By Proposition 4.3, C is ancestral in $G[R]$ and $P_{\mathrm{ext}}(C) \subseteq P$. By Lemma C.2, C contains a root SCC S of $G[R]$. Every external parent of S lies in F and is also an external parent of C, so

$$P_{\mathrm{ext}}(S) \subseteq P_{\mathrm{ext}}(C) \subseteq P, \qquad |S| + |P_{\mathrm{ext}}(S)| \leq |C| + |P|.$$

The pair $(S, P_{\mathrm{ext}}(S))$ passes by Proposition 4.3. If (C, P) is a minimizer, equality must hold, forcing $C = S$ and $P = P_{\mathrm{ext}}(S)$. This proves the claim.

Since every root SCC S yields the passing pair $(S, P_{\mathrm{ext}}(S))$, the same argument also gives

$$\min_{(C,P):\,\mathsf{PASS}(C,P)=1} (|C| + |P|) = \min_{S:\,\text{root SCC of } G[R]} (|S| + |P_{\mathrm{ext}}(S)|).$$

The equality case above identifies all minimizers. □

Proof of Theorem 4.4. We prove by induction that the following invariants hold before every round:

$$\widehat{\Pi} \subseteq \Pi, \qquad F = \bigcup_{T \in \widehat{\Pi}} T, \qquad F \text{ is ancestral in } G, \tag{I1}$$

$$\widehat{E}^{\mathrm{sc}} = \{T \to U \in E^{\mathrm{sc}} : T, U \in \widehat{\Pi}\}. \tag{I2}$$

They hold initially because F, $\widehat{\Pi}$, and $\widehat{E}^{\mathrm{sc}}$ are empty.

Suppose that (I1) and (I2) hold before a round and that $F \neq V$. Write $R = V \setminus F$. Since F is an ancestral union of true SCCs, the SCCs of $G[R]$ are exactly the unrecovered SCCs

$$\Pi_R := \{T \in \Pi : T \subseteq R\} = \Pi \setminus \widehat{\Pi}. \tag{19}$$

Indeed, removing F preserves each remaining SCC and cannot create new directed paths between distinct SCCs.

Every root SCC S of $G[R]$ satisfies

$$|S| \le s_{\max} \le s, \qquad P_{\mathrm{ext}}(S) \subseteq F, \qquad |P_{\mathrm{ext}}(S)| \le d_B \le d.$$

Thus the minimizing pair in Lemma C.3 is eligible. Write the first passing pair selected by the algorithm as $(S, \widehat{P}(S))$. By the search order and Lemma C.3, S is a root SCC of $G[R]$, $\widehat{P}(S) = P_{\mathrm{ext}}(S)$, and

$$|S| + |\widehat{P}(S)| = \min_{U\text{: root SCC of } G[R]} (|U| + |P_{\mathrm{ext}}(U)|) \le s_{\max} + d_B. \tag{20}$$

Thus the failure branch is not reached. For every $T \in \widehat{\Pi}$,

$$T \to S \in E^{\mathrm{sc}} \quad \Longleftrightarrow \quad T \cap P_{\mathrm{ext}}(S) \neq \varnothing \quad \Longleftrightarrow \quad T \cap \widehat{P}(S) \neq \varnothing.$$

The update therefore adds exactly the condensation edges involving S and previously recovered SCCs: the displayed equivalence identifies all incoming edges, while ancestry of F rules out edges from S into F. Thus (I2) is preserved. Since S is a root SCC of $G[R]$, $F \cup S$ remains ancestral. Together with adding the unrecovered true SCC S to $\widehat{\Pi}$, this preserves (I1).

Each round adds one unrecovered SCC, so the algorithm reaches $F = V$ after at most $|\Pi|$ rounds. The two invariants then give $\widehat{\Pi} = \Pi$ and $\widehat{E}^{\mathrm{sc}} = E^{\mathrm{sc}}$. □

C.4 Population recovery with arbitrary search bounds

Proposition C.4 (Insufficient search bounds)**.** *Suppose the cyclic LiNG model (1) and Assumption 2.1 hold. Run Algorithm 1 with arbitrary integers $s \ge 1$, $d \ge 0$ and population* PASS *decisions. The algorithm returns failure if and only if $s < s_{\max}$ or $d < d_B$.*

Proof. If $s \ge s_{\max}$ and $d \ge d_B$, Theorem 4.4 rules out failure. For the converse, we show that termination without failure requires $s \ge s_{\max}$ and $d \ge d_B$.

We use an argument similar to the proof of Theorem 4.4. The invariants hold initially; suppose (I1) and (I2) hold before a round with $F \neq V$, and write $R = V \setminus F$. If the round completes, let (C, P) be its first passing pair. As in the proof of Lemma C.3, C contains a root SCC S of $G[R]$ such that

$$S \subseteq C, \qquad P_{\mathrm{ext}}(S) \subseteq P, \qquad \mathsf{PASS}(S, P_{\mathrm{ext}}(S)) = 1.$$

Since

$$|S| \le |C| \le s, \qquad |P_{\mathrm{ext}}(S)| \le |P| \le d,$$

this pair is also eligible, and the search order gives

$$|C| + |P| \le |S| + |P_{\mathrm{ext}}(S)| \le |C| + |P|.$$

Equality and the inclusions above force $C = S$ and $P = P_{\mathrm{ext}}(S)$. Under the invariants, S is an unrecovered true SCC. The update therefore preserves both invariants by the same argument as in Theorem 4.4.

If the algorithm does not return failure, it reaches $F = V$ because each completed round adds an unrecovered true SCC. Every true SCC and its exact external parent set have therefore been selected within the search bounds. Hence

$$s_{\max} = \max_{S \in \Pi} |S| \le s, \qquad d_B = \max_{S \in \Pi} |P_{\mathrm{ext}}(S)| \le d.$$

□

Computational count. With valid search bounds and correct decisions, every examined pair has $|C| + |P| \le s_{\max} + d_B$ by (20). At a round with sets F, R, the number of eligible pairs of total size t is at most

$$\sum_{\substack{1\le q\le |R| \\ 0\le t-q\le |F|}} \binom{|R|}{q}\binom{|F|}{t-q} \le \binom{p}{t}.$$

Over at most p rounds, the total number of examined pairs is at most

$$p \sum_{t=1}^{\min\{p, s_{\max}+d_B\}} \binom{p}{t} \le p^{O(s_{\max}+d_B)}.$$

Each candidate evaluation requires regressions and scalar statistics, with at most $s_{\max} + d_B$ predictors in any regression. In general, the same counting argument gives a worst-case bound of $p^{O(s+d)}$.

D FINITE-SAMPLE PROOF DETAILS

This section first counts the scalar tests at each candidate size and establishes uniform bounds for residuals fitted on the same observations. We then prove Theorem 5.2 and Corollary 5.3. Finally, we verify Assumption 5.1 for normalized distance covariance and Gaussian HSIC. The upper-bound analysis requires only the natural bounds $s_{\max} \le p$ and $d_B \le p - 1$.

We establish uniform bounds at each total candidate size and then combine them over the sizes needed for recovery. For $1 \le h \le p$, let

$$\mathcal{A}(h) := \{(F, C, P) : F \text{ ancestral},\ \varnothing \ne C \subseteq V \setminus F,\ P \subseteq F,\ |C| + |P| = h\},$$

and define

$$\mathcal{I}_{\text{up}}(h) := \{(F, C, P, c, k) : (F, C, P) \in \mathcal{A}(h),\ c \in C,\ k \in F\},$$
$$\mathcal{I}_{\text{out}}(h) := \{(F, C, P, c, \ell) : (F, C, P) \in \mathcal{A}(h),\ c \in C,\ \ell \in V \setminus (F \cup C)\}.$$

Thus $\mathcal{A} = \bigcup_{h=1}^{\min\{p, s_{\max}+d_B\}} \mathcal{A}(h)$. For $0 < \alpha < 1$, define

$$\ell(h; \alpha) := h \log \frac{ep}{h} + h \log 2 + \log p + \log h + \log \frac{64}{\alpha} + \log[h(h+1)].$$

For the target error probability δ, write $\ell(h) = \ell(h; \delta)$.

All sample regressions use empirically centered responses and predictors, and empirical covariances are computed from centered arrays. Fitted residual arrays are therefore already centered. Constants depend only on the stated tail, covariance, and fixed estimator parameters.

The recovery procedure returns failure if a required sample regression is singular or a residual scale is zero. For definiteness, the affected statistics are assigned zero; these cases are excluded on the regularity events used below.

D.1 CANDIDATE COUNT

Lemma D.1 (Candidate count and complexity)**.** *The following statements hold.*

(i) For every $1 \le h \le p$, the number of distinct scalar statistics indexed by $\mathcal{I}_{\text{up}}(h) \cup \mathcal{I}_{\text{out}}(h)$ satisfies

$$N_{\text{test}} \le 2ph \left(\frac{2ep}{h}\right)^h.$$

(ii) For every $1 \le h \le p$ and $0 < \alpha < 1$,

$$\log \frac{32(N_{\text{test}} \vee 1)}{\alpha} \le \ell(h; \alpha), \quad \text{and} \quad h + 1 \le \ell(h; \alpha).$$

(iii) For every fixed $0 < \alpha < 1$, $h \mapsto \ell(h; \alpha)$ is increasing on $[1, p]$.

(iv) For $1 \le s \le p$ *and* $0 \le d \le p-1$*, define*

$$L_{s,d} := s \log \frac{ep}{s} + d \log \frac{ep}{d} + \log p + \log s + \log \frac{64}{\delta},$$

with $d \log(ep/d) = 0$ *when* $d = 0$*. Then*

$$\frac{1}{2} L_{s,d} \le \ell(\min\{p, s+d\}; \delta) \le 3 L_{s,d}.$$

Proof. For statement (i), choose the union $C \cup P$ and then its nonempty subset C. Since C and P are disjoint, the number of candidate pairs is

$$\sum_{q=1}^{h} \binom{p}{q} \binom{p-q}{h-q} = \binom{p}{h} \sum_{q=1}^{h} \binom{h}{q} = (2^h - 1) \binom{p}{h}.$$

For each pair, there are at most h choices of c and p choices of the tested variable for each of the two scores. Since F only restricts eligibility,

$$N_{\text{test}} \le 2ph(2^h - 1) \binom{p}{h} \le 2ph \left(\frac{2ep}{h} \right)^h,$$

where the last inequality uses $2^h - 1 \le 2^h$ and $\binom{p}{h} \le (ep/h)^h$. This proves statement (i).

For statement (ii), taking logarithms in statement (i) gives the first inequality; the case $N_{\text{test}} = 0$ is immediate. The second follows from $h \log(ep/h) \ge h$ and $\log(64/\alpha) > 1$.

For statement (iii), the derivative of $h \log(ep/h)$ is $\log(p/h)$. All remaining h-dependent terms in $\ell(h; \alpha)$ are strictly increasing, so the claim follows.

For statement (iv), put $u = \min\{p, s+d\}$ and $J = s \log(ep/s) + d \log(ep/d)$. Then $L_{s,d} = J + \log p + \log s + \log(64/\delta)$. Since $s, d \le p$, we have $J \ge s + d \ge u$.

If $s + d \le p$, then $u = s + d$ and

$$J - u \log(ep/u) = u \left(-\frac{s}{u} \log \frac{s}{u} - \frac{d}{u} \log \frac{d}{u} \right) \in [0, u \log 2].$$

Then the binary-entropy bound gives $u \log(ep/u) \le J \le u \log(ep/u) + u \log 2$.

If $s + d > p$, then $u = p$ and $u \log(ep/u) = p \le J$. Thus, in both cases, $u \log(ep/u) \le J$.

Also, $u \ge 1$ and $2 \log u \le u$ imply

$$\log[u(u+1)] \le 2 \log u + \log 2 \le u + \log 2.$$

Using $u \le p$ and $u \le J$, we obtain

$$\begin{aligned} \ell(u; \delta) &\le J + (1 + \log 2)u + 2 \log p + \log(64/\delta) + \log 2 \\ &\le (2 + \log 2)J + 2 \log p + \log(64/\delta) + \log 2 \le 3 L_{s,d}, \end{aligned}$$

where the last inequality uses $\log s \ge 0$, $\log 2 < 1$, and $\log(64/\delta) \ge \log 2$.

For the reverse inequality, if $s + d \le p$, the entropy bound and $\log s \le \log u$ give

$$L_{s,d} \le u \log(ep/u) + u \log 2 + \log p + \log u + \log(64/\delta) \le \ell(u; \delta).$$

If $s + d > p$, the function $x \log(ep/x)$ is increasing on $(0, p]$, so $J \le 2p$. Then

$$L_{s,d} \le 2p + 2 \log p + \log(64/\delta) \le 2\ell(p; \delta).$$

□

D.2 Residual estimation on the same observations

The next lemma establishes uniform residual and scale bounds over our candidate collection.

The required response–predictor pairs are

$$\begin{aligned}\mathcal{R}(h) := &\big\{(c,P) : (F,C,P,c,k) \in \mathcal{I}_{\mathrm{up}}(h) \text{ for some } (F,C,k)\big\} \\ &\cup\big\{(c,P) : (F,C,P,c,\ell) \in \mathcal{I}_{\mathrm{out}}(h) \text{ for some } (F,C,\ell)\big\} \\ &\cup\big\{(\ell,P\cup C) : (F,C,P,c,\ell) \in \mathcal{I}_{\mathrm{out}}(h)\big\}.\end{aligned}$$

For $u \in \mathbb{R}^n$, write

$$\|u\|_n := \left(\frac{1}{n}\sum_{i=1}^{n} u_i^2\right)^{1/2}.$$

For any $(v,Q) \in \mathcal{R}(h)$, write

$$Y = e_v(Q), \qquad y = (Y^{(i)} - \overline{Y})_{i=1}^n,$$

let $\widehat{y}$ be the residual array from least-squares regression of X_v on X_Q with an intercept, and set

$$\sigma_Y^2 = \mathrm{Var}(Y), \qquad \widehat{\sigma}_y = \|y\|_n, \qquad \widehat{\sigma}_{\widehat{y}} = \|\widehat{y}\|_n.$$

Lemma D.2 (Uniform residual prediction)**.** *Under Assumption 2.2, for $0 < \alpha < 1$, there is a constant $c_{\mathrm{res}} > 0$, depending only on (K,λ), such that, if $n \geq c_{\mathrm{res}}\ell(h;\alpha)$, there is an event $\mathcal{E}_{\mathrm{res}}(\alpha)$ with $\Pr(\mathcal{E}_{\mathrm{res}}(\alpha)) \geq 1-\alpha$ on which the following hold simultaneously:*

(i) Every required sample regression is nonsingular, and for every $(v,Q) \in \mathcal{R}(h)$,

$$\|\widehat{y} - y\|_n \leq c_{\mathrm{res}}\sqrt{\ell(h;\alpha)/n}.$$

(ii) For every $(v,Q) \in \mathcal{R}(h)$,

$$\widehat{\sigma}_y^2, \widehat{\sigma}_{\widehat{y}}^2 \in [\lambda^{-1}/2, 2\lambda], \qquad |\widehat{\sigma}_y/\sigma_Y - 1| \leq c_{\mathrm{res}}\sqrt{\ell(h;\alpha)/n}.$$

(iii) For every observed variable X_k entering a required UP *statistic, writing σ_{X_k} and $\widehat{\sigma}_{X_k}$ for its population and empirical standard deviations,*

$$\widehat{\sigma}_{X_k}^2 \in [\lambda^{-1}/2, 2\lambda], \qquad |\widehat{\sigma}_{X_k}/\sigma_{X_k} - 1| \leq c_{\mathrm{res}}\sqrt{\ell(h;\alpha)/n}.$$

Proof. Assume $\mathcal{R}(h) \neq \varnothing$ so that $N_{\mathrm{test}} \geq 1$, since otherwise all three statements hold automatically. We first construct a covariance event common to the three statements.

Write $\widehat{\Sigma}$ for covariance centered by the empirical mean, with divisor n. Let $c_{\mathrm{cov}} > 0$ denote the universal numerical constant in the mean-centered sub-Gaussian covariance concentration inequality (Vershynin, 2026; Chakrabortty et al., 2018). Thus, for every index set H with $m_H := |H| \leq h+1$ and every $t > 0$,

$$\Pr\left(\|\widehat{\Sigma}_{H,H} - \Sigma_{H,H}\|_{\mathrm{op}} \leq c_{\mathrm{cov}}K^2\lambda\left(\sqrt{\frac{m_H+t}{n}} + \frac{m_H+t}{n}\right)\right) \geq 1 - 2e^{-t}. \tag{21}$$

Associate to each scalar UP statistic the covariance matrix on $P \cup \{c,k\}$ and to each scalar OUT statistic the covariance matrix on $P \cup C \cup \{\ell\}$. These matrices contain all submatrices needed for their regressions. After identifying duplicates, there are at most N_{test} required index sets; denote their collection by $\mathcal{H}$.

Set $t = \log(2N_{\mathrm{test}}/\alpha)$. By Lemma D.1(ii), $t \leq \ell(h;\alpha)$ and $m_H \leq h+1 \leq \ell(h;\alpha)$, so $m_H + t \leq 2\ell(h;\alpha)$. Hence, by a union bound,

$$\Pr\left(\max_{H\in\mathcal{H}}\|\widehat{\Sigma}_{H,H} - \Sigma_{H,H}\|_{\mathrm{op}} \leq c_{\mathrm{cov}}K^2\lambda\left(\sqrt{\frac{2\ell(h;\alpha)}{n}} + \frac{2\ell(h;\alpha)}{n}\right)\right) \geq 1 - 2N_{\mathrm{test}}e^{-t} \geq 1-\alpha.$$

Denote this event by $\mathcal{E}_{\mathrm{res}}(\alpha)$.

Define the covariance error bound

$$\eta_\Sigma := c_{\mathrm{cov}} K^2 \lambda \left(\sqrt{\frac{2\ell(h;\alpha)}{n}} + \frac{2\ell(h;\alpha)}{n} \right).$$

Assume $n \geq C_{\mathrm{samp}} \ell(h;\alpha)$, where $C_{\mathrm{samp}} := \max\{2, 32 c_{\mathrm{cov}}^2 K^4 \lambda^4\}$. Then

$$\eta_\Sigma \leq 2\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda \sqrt{\frac{\ell(h;\alpha)}{n}} \leq \frac{\lambda^{-1}}{2}. \tag{22}$$

Since $\lambda_{\min}(\Sigma_{H,H}) \geq \lambda^{-1}$, this operator-norm bound gives

$$\frac{1}{2}\Sigma_{H,H} \preceq \widehat{\Sigma}_{H,H} \preceq \frac{3}{2}\Sigma_{H,H}$$

on every required submatrix. In particular, every required sample regression is nonsingular.

For statement (i), fix $(v, Q) \in \mathcal{R}(h)$. Here, assume $Q \neq \varnothing$, since $\widehat{y} = y$ when $Q = \varnothing$. For response X_v and predictors X_Q, write $\beta = \Sigma_{Q,Q}^{-1}\Sigma_{Q,v}$ and $\widehat{\beta} = \widehat{\Sigma}_{Q,Q}^{-1}\widehat{\Sigma}_{Q,v}$. On the event $\mathcal{E}_{\mathrm{res}}(\alpha)$, for $H = Q \cup \{v\}$, $\max\left\{\|\widehat{\Sigma}_{Q,Q} - \Sigma_{Q,Q}\|_{\mathrm{op}}, \|\widehat{\Sigma}_{Q,v} - \Sigma_{Q,v}\|_2\right\} \leq \eta_\Sigma \leq \lambda^{-1}/2$. Using $\|\beta\|_2 \leq \|\Sigma_{Q,Q}^{-1}\|_{\mathrm{op}}\|\Sigma_{Q,v}\|_2 \leq \lambda^2$, Theorem 3.1 of Kuchibhotla et al. (2023) gives

$$\|\widehat{\beta} - \beta\|_2 \leq \frac{\eta_\Sigma(1 + \|\beta\|_2)}{\lambda^{-1} - \eta_\Sigma} \leq 2\lambda\eta_\Sigma\left(1 + \lambda^2\right).$$

By empirical centering, $\widehat{y} - y = -X_Q(\widehat{\beta} - \beta)$. Hence

$$\|\widehat{y} - y\|_n = \left\{(\widehat{\beta} - \beta)^\top \widehat{\Sigma}_{Q,Q}(\widehat{\beta} - \beta)\right\}^{1/2} \leq \sqrt{\frac{3\lambda}{2}}\, \|\widehat{\beta} - \beta\|_2 \leq \sqrt{\frac{3\lambda}{2}}\, 2\lambda\eta_\Sigma\left(1 + \lambda^2\right).$$

Then for $C_{\mathrm{pred}} := \sqrt{\frac{3\lambda}{2}}\, 4\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda^2 (1 + \lambda^2)$, the inequality (22) gives

$$\|\widehat{y} - y\|_n \leq C_{\mathrm{pred}} \sqrt{\ell(h;\alpha)/n}.$$

Since the covariance event holds simultaneously for every required submatrix, this prediction bound holds uniformly over $\mathcal{R}(h)$.

For statement (ii), if $Q = \varnothing$, then $\widehat{y} = y$, and the variance and scale bounds follow from the covariance event on $\{v\}$. Assume $Q \neq \varnothing$ and let $H = Q \cup \{v\}$. Let $a, \widehat{a} \in \mathbb{R}^H$ be the oracle and fitted residual coefficient vectors, respectively, with $a_v = \widehat{a}_v = 1$, $a_Q = -\beta$, and $\widehat{a}_Q = -\widehat{\beta}$. Consequently,

$$\widehat{\sigma}_y^2 = a^\top \widehat{\Sigma}_{H,H} a \geq (\lambda^{-1}/2)\|a\|_2^2 \geq \lambda^{-1}/2,$$
$$\widehat{\sigma}_{\widehat{y}}^2 = \widehat{a}^\top \widehat{\Sigma}_{H,H} \widehat{a} \geq (\lambda^{-1}/2)\|\widehat{a}\|_2^2 \geq \lambda^{-1}/2.$$

The population residual variance is the Schur complement $\sigma_Y^2 = \Sigma_{vv} - \Sigma_{v,Q}\Sigma_{Q,Q}^{-1}\Sigma_{Q,v} \leq \Sigma_{vv} \leq \lambda$. The Loewner upper bound gives $\widehat{\sigma}_y^2 \leq (3/2)\sigma_Y^2 \leq 3\lambda/2$, and empirical least squares minimizes squared error, so $\widehat{\sigma}_{\widehat{y}}^2 \leq \widehat{\sigma}_y^2$. This proves the stated two-sided empirical variance bounds.

Since $\sigma_Y^2 = a^\top \Sigma_{H,H} a \geq \lambda^{-1}\|a\|_2^2$,

$$\left|\frac{\widehat{\sigma}_y^2}{\sigma_Y^2} - 1\right| = \frac{|a^\top(\widehat{\Sigma}_{H,H} - \Sigma_{H,H})a|}{\sigma_Y^2} \leq \frac{\eta_\Sigma \|a\|_2^2}{\sigma_Y^2} \leq \lambda\eta_\Sigma \leq C_{\mathrm{scale}} \sqrt{\frac{\ell(h;\alpha)}{n}},$$

where $C_{\mathrm{scale}} := 2\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda^2$. Since $|\sqrt{x} - 1| \leq |x - 1|$ for $x \geq 0$, the same bound holds for $|\widehat{\sigma}_y/\sigma_Y - 1|$.

For statement (iii), on $\mathcal{E}_{\mathrm{res}}(\alpha)$, the bounds for an observed variable X_k give

$$|\widehat{\Sigma}_{kk} - \Sigma_{kk}| \leq \eta_\Sigma, \qquad \widehat{\Sigma}_{kk} \in [\lambda^{-1}/2, 3\lambda/2], \qquad \left|\frac{\widehat{\sigma}_{X_k}}{\sigma_{X_k}} - 1\right| \leq C_{\mathrm{scale}} \sqrt{\frac{\ell(h;\alpha)}{n}}.$$

Taking

$$c_{\mathrm{res}} := \max\{C_{\mathrm{samp}}, C_{\mathrm{pred}}, C_{\mathrm{scale}}\}$$

ensures the required sample-size condition and all three bounds. □

D.3 UNIFORM UP APPROXIMATION

Lemma D.3 (UP covariance approximation)**.** *There exists a constant $c_{\mathrm{up}} > 0$, depending only on (K, λ), such that for every $0 < \delta < 1$, if $n \geq c_{\mathrm{up}} \ell(h; \delta)$, then*

$$\Pr\left(\max_{(F,C,P,c,k) \in \mathcal{I}_{\mathrm{up}}(h)} \left| \widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k) - \mathrm{Cov}(e_c(P), X_k) \right| \leq c_{\mathrm{up}} \sqrt{\frac{\ell(h;\delta)}{n}} \right) \geq 1 - \frac{\delta}{3}.$$

Proof. Assume $\mathcal{I}_{\mathrm{up}}(h) \neq \varnothing$, since otherwise the assertion holds automatically. Then $N_{\mathrm{test}} \geq 1$. For every indexed test, the triangle inequality gives

$$\begin{aligned} &|\widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k) - \mathrm{Cov}(e_c(P), X_k)| \\ &\quad \leq |\widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k) - \widehat{\mathrm{Cov}}(e_c(P), X_k)| + |\widehat{\mathrm{Cov}}(e_c(P), X_k) - \mathrm{Cov}(e_c(P), X_k)|. \end{aligned}$$

Since $\ell(h; \delta) \geq \log(32/\delta) \geq \log 32$,

$$\ell(h; \delta/6) = \ell(h; \delta) + \log 6 \leq \left(1 + \frac{\log 6}{\log 32}\right) \ell(h; \delta).$$

Apply Lemma D.2(i) and (iii) with $\alpha = \delta/6$. On that event, Cauchy–Schwarz gives

$$|\widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k) - \widehat{\mathrm{Cov}}(e_c(P), X_k)| \leq \|\widehat{e}_c(P) - e_c(P)\|_n \widehat{\sigma}_{X_k} \leq \sqrt{2\lambda}\, c_{\mathrm{res}} \sqrt{1 + \frac{\log 6}{\log 32}} \sqrt{\frac{\ell(h;\delta)}{n}}.$$

It remains to control the oracle covariance estimation error. For each indexed test, set $H = P \cup \{c, k\}$. If $P = \varnothing$, then $e_c(P) = X_c$, so take $a = e_c$, with $\|a\|_2 = 1 \leq \sqrt{1 + \lambda^2}$. If $P \neq \varnothing$, then

$$\beta = \Sigma_{P,P}^{-1} \Sigma_{P,c}, \qquad e_c(P) = X_c - \beta^\top X_P = a^\top X_H,$$

where $a_c = 1,\ a_P = -\beta,\ a_{H \setminus (P \cup \{c\})} = 0$. Moreover,

$$\lambda^{-1} \|\beta\|_2^2 \leq \beta^\top \Sigma_{P,P} \beta \leq \mathrm{Var}(X_c) \leq \lambda.$$

Here the second inequality follows from the orthogonal variance decomposition $\mathrm{Var}(X_c) = \mathrm{Var}(\beta^\top X_P) + \mathrm{Var}(e_c(P))$. Thus $\|a\|_2 \leq \sqrt{1 + \lambda^2}$.

Applying (21) with $t = \ell(h; \delta)$ and taking a union bound over at most N_{test} sets H, each of size at most $h + 1$, Lemma D.1(ii) gives, for $n \geq 2\ell(h; \delta)$,

$$\Pr\left(\max_H \|\widehat{\Sigma}_{H,H} - \Sigma_{H,H}\|_{\mathrm{op}} \leq 2\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda \sqrt{\frac{\ell(h;\delta)}{n}} \right) \geq 1 - 2N_{\mathrm{test}} e^{-\ell(h;\delta)} \geq 1 - \frac{\delta}{16}.$$

Consequently, on this event,

$$|\widehat{\mathrm{Cov}}(e_c(P), X_k) - \mathrm{Cov}(e_c(P), X_k)| = |a^\top (\widehat{\Sigma}_{H,H} - \Sigma_{H,H}) e_k| \leq 2\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda \sqrt{1 + \lambda^2} \sqrt{\frac{\ell(h;\delta)}{n}}$$

uniformly over $\mathcal{I}_{\mathrm{up}}(h)$.

Set

$$c_{\mathrm{up}} := \max\left\{ \left(1 + \frac{\log 6}{\log 32}\right) c_{\mathrm{res}},\ 2,\ \sqrt{2\lambda}\, c_{\mathrm{res}} \sqrt{1 + \frac{\log 6}{\log 32}} + 2\sqrt{2}\, c_{\mathrm{cov}} K^2 \lambda \sqrt{1 + \lambda^2} \right\}.$$

Combining the two bounds and applying a union bound gives

$$\begin{aligned} &\Pr\left(\max_{(F,C,P,c,k) \in \mathcal{I}_{\mathrm{up}}(h)} \left| \widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k) - \mathrm{Cov}(e_c(P), X_k) \right| \leq c_{\mathrm{up}} \sqrt{\frac{\ell(h;\delta)}{n}} \right) \\ &\qquad \geq 1 - \frac{\delta}{6} - \frac{\delta}{16} > 1 - \frac{\delta}{3}, \end{aligned}$$

which proves the claim. □

D.4 Proof of Theorem 5.2

Proof of Theorem 5.2. Let c_{res} and c_{up} be the constants in Lemmas D.2 and D.3. Choose fixed multipliers satisfying

$$\kappa_{\mathrm{up}} \geq 2\sqrt{2}c_{\mathrm{up}}, \qquad \kappa_{\mathrm{out}} \geq 2\sqrt{2}c_{\mathcal{D}}.$$

For each $1 \leq h \leq \min\{p, s_{\max} + d_B\}$, put $\delta_h = \delta/[h(h+1)]$. By the definition of ℓ,

$$\ell(h; \delta_h) = \ell(h; \delta) + \log[h(h+1)] \leq 2\ell(h; \delta),$$
$$\ell(h; \delta_h/3) \leq 2\ell(h; \delta) + \log 3 \leq 3\ell(h; \delta).$$

Lemma D.1(iii)–(iv) therefore gives

$$\ell(h; \delta_h) \leq 2\ell(\min\{p, s_{\max} + d_B\}; \delta) \leq 6L, \qquad \ell(h; \delta_h/3) \leq 3\ell(\min\{p, s_{\max} + d_B\}; \delta) \leq 9L.$$

Write $\widehat{D}_F^{\mathrm{up}}(C, P)$ and $\widehat{D}_F^{\mathrm{out}}(C, P)$ for the empirical counterparts of the two population scores

$$\widehat{D}_F^{\mathrm{up}}(C, P) := \max_{c \in C,\, k \in F} |\widehat{\mathrm{Cov}}(\widehat{e}_c(P), X_k)|, \qquad \widehat{D}_F^{\mathrm{out}}(C, P) := \max_{c \in C,\, \ell \in R \setminus C} \widehat{\mathcal{D}}\big(e_c(P), r_{P,C}(\ell)\big),$$

with empty maxima equal to zero.

Set

$$c_0 := \max\{9c_{\mathrm{res}}, 6c_{\mathrm{up}}, 6c_{\mathcal{D}}, 27\kappa_{\mathrm{up}}^2, 27\kappa_{\mathrm{out}}^2\}.$$

Since the sample-size condition gives $n \geq c_0 L$, for every h under consideration,

$$n \geq 9c_{\mathrm{res}}L \geq c_{\mathrm{res}}\ell(h; \delta_h/3), \qquad n \geq 6c_{\mathrm{up}}L \geq c_{\mathrm{up}}\ell(h; \delta_h), \qquad n \geq 6c_{\mathcal{D}}L \geq c_{\mathcal{D}}\ell(h; \delta_h).$$

Hence, for each such h, we may apply Lemma D.2(i) and (ii) with $\alpha = \delta_h/3$, and apply Lemma D.3 and Assumption 5.1 with δ replaced by δ_h. Their failure probabilities sum to at most δ_h. Since

$$\sum_{h=1}^{\min\{p, s_{\max}+d_B\}} \delta_h = \delta\left(1 - \frac{1}{\min\{p, s_{\max} + d_B\} + 1}\right) \leq \delta,$$

there is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which all required regressions for $\mathcal{A}$ are nonsingular, all required residual scales are positive, and both scalar score approximation bounds hold. The inequalities $|\max_i a_i - \max_i b_i| \leq \max_i |a_i - b_i|$ and $||a| - |b|| \leq |a - b|$ therefore give, for every candidate of size h,

$$|\widehat{D}_F^{\mathrm{up}}(C, P) - D_F^{\mathrm{up}}(C, P)| \leq \sqrt{2}c_{\mathrm{up}}\sqrt{\ell(h)/n} \leq \tau_{\mathrm{up}}(h)/2,$$
$$|\widehat{D}_F^{\mathrm{out}}(C, P) - D_F^{\mathrm{out}}(C, P)| \leq \sqrt{2}c_{\mathcal{D}}\sqrt{\ell(h)/n} \leq \tau_{\mathrm{out}}(h)/2. \tag{23}$$

If a margin is infinite, the corresponding violating family is empty. For each finite margin, the last two terms in the definition of c_0 give

$$\tau_{\mathrm{up}}(h) \leq \kappa_{\mathrm{up}}\sqrt{3L/n} \leq \Delta_{\mathrm{par}}/3, \qquad \tau_{\mathrm{out}}(h) \leq \kappa_{\mathrm{out}}\sqrt{3L/n} \leq \Delta_{\mathcal{D}}/3.$$

A population passer has both scores zero and therefore passes empirically. For a nonpasser, Proposition 4.3 gives either insufficient adjustment or non-ancestry after exact adjustment. Respectively,

$$\widehat{D}_F^{\mathrm{up}}(C, P) \geq \Delta_{\mathrm{par}} - \tau_{\mathrm{up}}(h)/2 > \tau_{\mathrm{up}}(h),$$
$$\widehat{D}_F^{\mathrm{out}}(C, P) \geq \Delta_{\mathcal{D}} - \tau_{\mathrm{out}}(h)/2 > \tau_{\mathrm{out}}(h).$$

Thus on $\mathcal{E}$,

$$\widehat{\mathsf{PASS}}(C, P) = \mathsf{PASS}(C, P) \qquad \text{for every } (F, C, P) \in \mathcal{A}. \tag{24}$$

On $\mathcal{E}$, we follow the induction in the proof of Theorem 4.4. Suppose the invariants (I1) and (I2) hold before a round with $F \neq V$; in particular, F is ancestral. By (20) and the search order, every candidate examined up to and including the first population passer belongs to $\mathcal{A}$. Equation (24) ensures that the empirical search rejects all preceding candidates and accepts this passer. The invariants hold initially, so the same induction and termination argument gives exact recovery on $\mathcal{E}$. □

Proof of Corollary 5.3. In this proof, write $L_{s,d}$ for L evaluated at the class bounds (s, d), and reserve L for its value at the actual sizes $(s_{\max}, d_B)$ of a model. Fix a model in the separated class. Its actual sizes satisfy $s_{\max} \le s$ and $d_B \le d$. Hence the model-specific candidate class $\mathcal{A}$ is contained in the candidate class defined by $|C| + |P| \le s + d$. Taking minima over the larger set gives

$$\Delta_{\mathrm{par}}(M) \ge \gamma_{\mathrm{par}}, \qquad \Delta_{\mathcal{D}}(M) \ge \gamma_{\mathcal{D}}$$

for the model-specific margins used in Theorem 5.2. Since $x \log(ep/x)$ is increasing on $0 < x \le p$, the model-specific complexity satisfies $L \le L_{s,d}$. The search bounds are valid for every model in the class, so Theorem 5.2 gives

$$n \ge c_0 L_{s,d} \max\{1, \gamma_{\mathrm{par}}^{-2}, \gamma_{\mathcal{D}}^{-2}\} \quad \Longrightarrow \quad P_M^n(\widehat{G}^{\mathrm{sc}} \ne G^{\mathrm{sc}}(M)) \le \delta.$$

The multipliers and constants are common throughout the class. Taking the supremum proves the uniform recovery bound with the same constant c_0.

For optimality, fix $s_0 \ge 2$, $d_0 \ge 1$, and $0 < \delta < 1/2$. Take $K_0, \lambda_0, \tau_0, c_{\mathrm{lb},0}$ from Proposition E.3. The uniform recovery bound with $(s, d, K, \lambda) = (s_0, d_0, K_0, \lambda_0)$ and $\gamma_{\mathrm{par}} = \gamma_{\mathcal{D}} = \tau_0$ gives sample complexity $O(\log p)$.

Applying that proposition with δ replaced by $(1 + \delta)/2$ shows that every estimator has worst-case error probability at least $(1 + \delta)/2 > \delta$ on the same class whenever

$$n \le \frac{1 - \delta}{2c_{\mathrm{lb},0}} \left\{ s_0 \log \frac{ep}{s_0} + d_0 \log \frac{ep}{d_0} \right\}.$$

Since the expression in braces is $\asymp \log p$ for fixed s_0, d_0, the minimax sample complexity on this class is $\Theta(\log p)$. □

D.5 Concrete dependence-score estimators

We verify Assumption 5.1 for normalized distance covariance and Gaussian HSIC using complete order-four U-statistic formulas evaluated on fitted residuals.

Shared residuals. For each $(F, C, P, c, \ell) \in \mathcal{I}_{\mathrm{out}}(h)$, use the local notation

$$Y := e_c(P), \qquad Z := r_{P,C}(\ell) = e_\ell(P \cup C),$$

where the equality follows from the Frisch–Waugh–Lovell identity (Frisch & Waugh, 1933; Lovell, 1963). Let y, z be the empirically centered arrays obtained from the population residuals Y, Z, and let $\widehat{y}, \widehat{z}$ be the fitted residual arrays from the regressions indexed by (c, P) and $(\ell, P \cup C)$, respectively.

For each estimator, we first bound the estimation error of the complete U-statistic on i.i.d. population residuals, then control the effects of residual estimation and empirical standardization. This decomposition follows the proof strategy of Zhao et al. (2022, Lemmas 4–6) and Oh et al. (2025, Lemmas 9–14). Both kernels depend only on pairwise differences, so empirical centering leaves their values unchanged.

D.5.1 Normalized distance covariance

For scalar random variables (Y, Z) with finite, positive variances, let

$$(Y_i, Z_i), \qquad i = 1, 2, \ldots,$$

be i.i.d. copies of (Y, Z). The squared distance covariance admits the representation

$$\mathcal{V}^2(Y, Z) := \mathbb{E}[|Y_1 - Y_2||Z_1 - Z_2|] + \mathbb{E}[|Y_1 - Y_2|]\mathbb{E}[|Z_3 - Z_4|] - 2\mathbb{E}[|Y_1 - Y_2||Z_1 - Z_3|]. \quad (25)$$

The standardized squared distance covariance is

$$\mathcal{D}_{\mathrm{dCov}}(Y, Z) := \frac{\mathcal{V}^2(Y, Z)}{\sqrt{\mathrm{Var}(Y)\,\mathrm{Var}(Z)}}.$$

It vanishes exactly under independence (Székely et al., 2007).

For arrays (u, v) and a quartet $q = (q_1, q_2, q_3, q_4)$, define

$$a^q_{ij}(u) := |u_{q_i} - u_{q_j}|, \qquad b^q_{ij}(v) := |v_{q_i} - v_{q_j}|, \qquad i, j \in [4].$$

The symmetrized quartet kernel is

$$\bar{h}_4(u, v; q) = \frac{1}{4!} \sum_{\pi \in S_4} \left\{ a^q_{\pi_1\pi_2}(u) b^q_{\pi_1\pi_2}(v) + a^q_{\pi_1\pi_2}(u) b^q_{\pi_3\pi_4}(v) - 2a^q_{\pi_1\pi_2}(u) b^q_{\pi_1\pi_3}(v) \right\}, \tag{26}$$

where S_4 is the set of permutations of $[4]$. For $n \geq 4$, define

$$T_{4,n}(u, v) := \binom{n}{4}^{-1} \sum_{1 \leq q_1 < q_2 < q_3 < q_4 \leq n} \bar{h}_4(u, v; q).$$

Lemma D.4 (Empirical distance-covariance perturbation)**.** *For $n \geq 8$, let u, v, u', v' be centered arrays with empirical norms at most $\overline{\sigma}$. Then*

$$|T_{4,n}(u, v) - T_{4,n}(u', v')| \leq 10\overline{\sigma}(\|u - u'\|_n + \|v - v'\|_n).$$

If their norms are also at least $\underline{\sigma} > 0$, then

$$\left| \frac{T_{4,n}(u, v)}{\|u\|_n \|v\|_n} - \frac{T_{4,n}(u', v')}{\|u'\|_n \|v'\|_n} \right| \leq 20 \frac{\overline{\sigma}}{\underline{\sigma}^2} (\|u - u'\|_n + \|v - v'\|_n).$$

Proof. Let $\langle f(q) \rangle$ denote the uniform average over all ordered four-tuples q of distinct indices in $[n]$. For distinct positions $r, s \in [4]$, each ordered pair (i, j) occurs $(n-2)(n-3)$ times. Thus, for any centered array w,

$$\left\langle |w_{q_r} - w_{q_s}|^2 \right\rangle = \frac{1}{n(n-1)} \sum_{i \neq j} (w_i - w_j)^2 = \frac{2n}{n-1} \|w\|_n^2.$$

The reverse triangle inequality, with $w = u - u'$, gives

$$\left\langle |a^q_{rs}(u) - a^q_{rs}(u')|^2 \right\rangle \leq \left\langle \left| (u_{q_r} - u_{q_s}) - (u'_{q_r} - u'_{q_s}) \right|^2 \right\rangle = \frac{2n}{n-1} \|u - u'\|_n^2.$$

For any distinct $t, j \in [4]$, Cauchy–Schwarz and $ab - a'b' = (a - a')b + a'(b - b')$ therefore yield

$$\left\langle |a^q_{rs}(u) b^q_{tj}(v) - a^q_{rs}(u') b^q_{tj}(v')| \right\rangle \leq \frac{2n}{n-1} \{ \|u - u'\|_n \|v\|_n + \|u'\|_n \|v - v'\|_n \}.$$

Averaging over ordered four-tuples equals symmetrizing the kernel and averaging over increasing four-tuples. The absolute coefficient sum in (26) is 4, and $n \geq 8$ gives $8n/(n-1) \leq 10$. Hence,

$$\begin{aligned} |T_{4,n}(u, v) - T_{4,n}(u', v')| &\leq \frac{8n}{n-1} \{ \|u - u'\|_n \|v\|_n + \|u'\|_n \|v - v'\|_n \} \\ &\leq 10\overline{\sigma}(\|u - u'\|_n + \|v - v'\|_n). \end{aligned}$$

The same calculation without perturbation gives

$$|T_{4,n}(u, v)| \leq \frac{8n}{n-1} \|u\|_n \|v\|_n \leq 10 \|u\|_n \|v\|_n.$$

For the normalized statistics, the preceding inequalities and the lower bound $\underline{\sigma}$ on the empirical norms give

$$\begin{aligned} \left| \frac{T_{4,n}(u, v)}{\|u\|_n \|v\|_n} - \frac{T_{4,n}(u', v')}{\|u'\|_n \|v'\|_n} \right| &\leq \frac{|T_{4,n}(u, v) - T_{4,n}(u', v')|}{\|u\|_n \|v\|_n} + \frac{|T_{4,n}(u', v')| \, | \|u'\|_n \|v'\|_n - \|u\|_n \|v\|_n |}{\|u\|_n \|v\|_n \|u'\|_n \|v'\|_n}, \\ &\leq 10 \frac{\overline{\sigma}}{\underline{\sigma}^2} (\|u - u'\|_n + \|v - v'\|_n) + \frac{10}{\underline{\sigma}^2} | \|u'\|_n \|v'\|_n - \|u\|_n \|v\|_n | \end{aligned}$$

Then $| \|u'\|_n \|v'\|_n - \|u\|_n \|v\|_n | \leq \overline{\sigma}(\|u - u'\|_n + \|v - v'\|_n)$ proves the second assertion. □

Define the empirical dependence score by

$$\widehat{\mathcal{D}}_4\big(e_c(P), r_{P,C}(\ell)\big) := \frac{T_{4,n}(\widehat{y},\widehat{z})}{\widehat{\sigma}_{\widehat{y}}\widehat{\sigma}_{\widehat{z}}}.$$

Lemma D.5 (Normalized distance-covariance approximation)**.** *There exists a constant $c_{\mathrm{dCov}} > 0$, depending only on (K, λ), such that $\widehat{\mathcal{D}}_4$ satisfies Assumption 5.1 for $\mathcal{D} = \mathcal{D}_{\mathrm{dCov}}$ with $c_{\mathcal{D}} = c_{\mathrm{dCov}}$.*

Proof. Assume $\mathcal{I}_{\mathrm{out}}(h) \neq \varnothing$, since otherwise the assertion holds automatically. Then $N_{\mathrm{test}} \geq 1$. Since $\ell(h;\delta) \geq \log 32$,

$$\ell(h;\delta/6) \leq \left(1 + \frac{\log 6}{\log 32}\right)\ell(h;\delta).$$

For each index, write $\widehat{\mathcal{D}}_4$ and $\mathcal{D}_{\mathrm{dCov}}$ for the corresponding empirical and population scores, respectively. Let $\widetilde{\mathcal{D}}_{4,n}(Y,Z) := T_{4,n}(y,z)/(\widehat{\sigma}_y\widehat{\sigma}_z)$ be the statistic computed from oracle residuals on the same observations. Then

$$\begin{aligned}|\widehat{\mathcal{D}}_4 - \mathcal{D}_{\mathrm{dCov}}| &\leq |\widehat{\mathcal{D}}_4 - \widetilde{\mathcal{D}}_{4,n}| + |\widetilde{\mathcal{D}}_{4,n} - \mathcal{D}_{\mathrm{dCov}}| \\ &= \left|\frac{T_{4,n}(\widehat{y},\widehat{z})}{\widehat{\sigma}_{\widehat{y}}\widehat{\sigma}_{\widehat{z}}} - \frac{T_{4,n}(y,z)}{\widehat{\sigma}_y\widehat{\sigma}_z}\right| + \left|\frac{T_{4,n}(y,z)}{\widehat{\sigma}_y\widehat{\sigma}_z} - \frac{\mathcal{V}^2(Y,Z)}{\sigma_Y\sigma_Z}\right|.\end{aligned}$$

We first bound the estimation error of the oracle numerator. Empirical centering leaves pairwise differences unchanged, so $T_{4,n}(y,z)$ equals the complete U-statistic on the i.i.d. population residual pairs. The order-four representation of Huo & Székely (2016) therefore gives $\mathbb{E}T_{4,n}(y,z) = \mathcal{V}^2(Y,Z)$.

Write $\|U\|_{\psi_1} := \inf\{a > 0 : \mathbb{E}\exp(|U|/a) \leq 2\}$ for the sub-exponential norm. Let $c_\times > 0$ be a universal constant such that

$$\|UV\|_{\psi_1} \leq c_\times \|U\|_{\psi_2}\|V\|_{\psi_2}$$

for all sub-Gaussian random variables U, V. Choose a universal $c_{\mathrm{Ber}} \geq 1$ such that every centered ξ with $\|\xi\|_{\psi_1} \leq \nu$ satisfies

$$\mathbb{E}e^{t\xi} \leq \exp(c_{\mathrm{Ber}}^2\nu^2t^2/2), \qquad |t| \leq (c_{\mathrm{Ber}}\nu)^{-1}.$$

This is the sub-exponential moment-generating-function bound (Vershynin, 2026).

Population residuals are fixed sub-Gaussian linear forms with ψ_2 norm at most $K\sqrt{\lambda}$. Each product of two distances in (26) consequently has ψ_1 norm at most $4c_\times K^2\lambda$. Symmetrization, the total absolute coefficient sum of four, and centering give, for four i.i.d. population residual pairs,

$$\begin{aligned}\xi &:= \bar{h}_4\big((Y_i)_{i=1}^4, (Z_i)_{i=1}^4; (1,2,3,4)\big) - \mathcal{V}^2(Y,Z), \\ \|\xi\|_{\psi_1} &\leq 2\cdot 4\cdot(4c_\times K^2\lambda) = 32c_\times K^2\lambda =: \nu.\end{aligned}$$

By Pitcan (2026, Theorem 4.4), applied with sub-exponential parameters $(c_{\mathrm{Ber}}^2\nu^2, c_{\mathrm{Ber}}\nu)$, for every $x > \log 2$,

$$\Pr\left(|T_{4,n}(y,z) - \mathcal{V}^2(Y,Z)| > c_{\mathrm{Ber}}\nu\left\{\sqrt{\frac{2x}{\lfloor n/4\rfloor}} + \frac{2x}{\lfloor n/4\rfloor}\right\}\right) \leq 2e^{-x}.$$

With $x = 2\ell(h;\delta)$, $\lfloor n/4\rfloor \geq n/8$ and $n \geq 16\ell(h;\delta)$ give

$$\sqrt{\frac{4\ell(h;\delta)}{\lfloor n/4\rfloor}} + \frac{4\ell(h;\delta)}{\lfloor n/4\rfloor} \leq 4\sqrt{2}\sqrt{\frac{\ell(h;\delta)}{n}} + \frac{32\ell(h;\delta)}{n} \leq (4\sqrt{2}+8)\sqrt{\frac{\ell(h;\delta)}{n}} \leq 14\sqrt{\frac{\ell(h;\delta)}{n}}.$$

A union bound over at most N_{test} OUT statistics gives

$$\begin{aligned}&\Pr\left(\max_{\mathcal{I}_{\mathrm{out}}(h)} |T_{4,n}(y,z) - \mathcal{V}^2(Y,Z)| \leq 448c_{\mathrm{Ber}}c_\times K^2\lambda\sqrt{\frac{\ell(h;\delta)}{n}}\right) \\ &\qquad \geq 1 - 2N_{\mathrm{test}}e^{-2\ell(h;\delta)} \geq 1 - \frac{\delta}{128}.\end{aligned} \tag{27}$$

The last inequality uses Lemma D.1(ii) and $\ell(h;\delta) \geq \log 32$.

Now apply Lemma D.2(i) and (ii) with $\alpha = \delta/6$. On this event,

$$|\widehat{\sigma}_y/\sigma_Y - 1| + |\widehat{\sigma}_z/\sigma_Z - 1| \leq 2c_{\mathrm{res}}\sqrt{1 + \frac{\log 6}{\log 32}}\sqrt{\frac{\ell(h;\delta)}{n}}. \tag{28}$$

Moreover, $\mathbb{E}|Y_1 - Y_2|^2 = 2\sigma_Y^2$ and $\mathbb{E}|Z_1 - Z_2|^2 = 2\sigma_Z^2$, so Cauchy–Schwarz applied to (25) yields

$$|\mathcal{V}^2(Y,Z)| \leq 8\sigma_Y\sigma_Z. \tag{29}$$

If $n \geq 16(1 + \log 6/\log 32)c_{\mathrm{res}}^2\ell(h;\delta)$, the left-hand side of (28) is at most $1/2$.

Thus $\widehat{\sigma}_y/\sigma_Y, \widehat{\sigma}_z/\sigma_Z \in [1/2, 3/2]$, and

$$\left|\frac{\sigma_Y\sigma_Z}{\widehat{\sigma}_y\widehat{\sigma}_z} - 1\right| = \frac{|1 - (\widehat{\sigma}_y/\sigma_Y)(\widehat{\sigma}_z/\sigma_Z)|}{(\widehat{\sigma}_y/\sigma_Y)(\widehat{\sigma}_z/\sigma_Z)} \leq 4\left|1 - \frac{\widehat{\sigma}_y}{\sigma_Y}\frac{\widehat{\sigma}_z}{\sigma_Z}\right| \leq 6\left\{\left|\frac{\widehat{\sigma}_y}{\sigma_Y} - 1\right| + \left|\frac{\widehat{\sigma}_z}{\sigma_Z} - 1\right|\right\}.$$

On the intersection of the residual-prediction event and the event in (27), the preceding inequality, (28), (29), and $\widehat{\sigma}_y\widehat{\sigma}_z \geq \lambda^{-1}/2$ give

$$\begin{aligned}|\widetilde{\mathcal{D}}_{4,n} - \mathcal{D}_{\mathrm{dCov}}| &\leq \frac{|T_{4,n}(y,z) - \mathcal{V}^2(Y,Z)|}{\widehat{\sigma}_y\widehat{\sigma}_z} + \frac{|\mathcal{V}^2(Y,Z)|}{\sigma_Y\sigma_Z}\left|\frac{\sigma_Y\sigma_Z}{\widehat{\sigma}_y\widehat{\sigma}_z} - 1\right| \\ &\leq \left\{\frac{448c_{\mathrm{Ber}}c_\times K^2\lambda}{\lambda^{-1}/2} + 96c_{\mathrm{res}}\sqrt{1 + \frac{\log 6}{\log 32}}\right\}\sqrt{\frac{\ell(h;\delta)}{n}}. \end{aligned} \tag{30}$$

It remains to bound the difference between the empirical and oracle scores. On the same event, Lemma D.2(ii) bounds the empirical norms of all oracle and fitted residual arrays between $\sqrt{\lambda^{-1}/2}$ and $\sqrt{2\lambda}$. Lemma D.4 and Lemma D.2(i) therefore give, uniformly over $\mathcal{I}_{\mathrm{out}}(h)$,

$$|\widehat{\mathcal{D}}_4 - \widetilde{\mathcal{D}}_{4,n}| \leq \frac{20\sqrt{2\lambda}}{\lambda^{-1}/2}\left(\|\widehat{y} - y\|_n + \|\widehat{z} - z\|_n\right) \leq \frac{40\sqrt{2\lambda}}{\lambda^{-1}/2}c_{\mathrm{res}}\sqrt{1 + \frac{\log 6}{\log 32}}\sqrt{\frac{\ell(h;\delta)}{n}}. \tag{31}$$

Combining (31) and (30), choose

$$\begin{aligned}c_{\mathrm{dCov}} := \max\Bigg\{&\left(1 + \frac{\log 6}{\log 32}\right)c_{\mathrm{res}}, 16, 16\left(1 + \frac{\log 6}{\log 32}\right)c_{\mathrm{res}}^2, \\ &\frac{40\sqrt{2\lambda}}{\lambda^{-1}/2}c_{\mathrm{res}}\sqrt{1 + \frac{\log 6}{\log 32}} + \frac{448c_{\mathrm{Ber}}c_\times K^2\lambda}{\lambda^{-1}/2} + 96c_{\mathrm{res}}\sqrt{1 + \frac{\log 6}{\log 32}}\Bigg\}.\end{aligned}$$

This choice ensures all required sample-size conditions and yields the uniform error bound in Assumption 5.1 with failure probability at most $\delta/6 + \delta/128 < \delta/3$, proving the lemma. □

D.5.2 Gaussian HSIC

Fix, independently of the data, a Gaussian-kernel bandwidth $b \geq b_- > 0$, and write

$$k_b(x,x') := \exp\left\{-\frac{(x-x')^2}{2b^2}\right\}.$$

For scalar random variables (U,V), let (U_i,V_i), $i = 1,2,\ldots$, be i.i.d. copies of (U,V) and define

$$\mathrm{HSIC}_b(U,V) := \mathbb{E}[k_b(U_1,U_2)k_b(V_1,V_2)] + \mathbb{E}[k_b(U_1,U_2)k_b(V_3,V_4)] - 2\mathbb{E}[k_b(U_1,U_2)k_b(V_1,V_3)].$$

The population score is

$$\mathcal{D}_{\mathrm{HSIC}}(Y,Z) := \mathrm{HSIC}_b(Y/\sigma_Y, Z/\sigma_Z).$$

It vanishes exactly under independence (Gretton et al., 2005; Szabó & Sriperumbudur, 2018).

For arrays (u, v) and a quartet $q = (q_1, q_2, q_3, q_4)$, define the symmetrized HSIC kernel

$$\bar{h}^{\mathrm{HSIC}}_{4,b}(u, v; q) := \frac{1}{4!} \sum_{\pi \in S_4} \Big\{ k_b(u_{q_{\pi_1}}, u_{q_{\pi_2}}) k_b(v_{q_{\pi_1}}, v_{q_{\pi_2}}) + k_b(u_{q_{\pi_1}}, u_{q_{\pi_2}}) k_b(v_{q_{\pi_3}}, v_{q_{\pi_4}}) - 2k_b(u_{q_{\pi_1}}, u_{q_{\pi_2}}) k_b(v_{q_{\pi_1}}, v_{q_{\pi_3}}) \Big\}. \tag{32}$$

This is the symmetrized order-four kernel for the unbiased HSIC estimator of Song et al. (2012, Theorem 3). For $n \geq 4$, define the complete statistic

$$T^{\mathrm{HSIC}}_{4,n}(u, v) := \binom{n}{4}^{-1} \sum_{1 \leq q_1 < q_2 < q_3 < q_4 \leq n} \bar{h}^{\mathrm{HSIC}}_{4,b}(u, v; q).$$

Define the empirical dependence score by

$$\widehat{\mathcal{D}}_{\mathrm{HSIC},4}\big(e_c(P), r_{P,C}(\ell)\big) := T^{\mathrm{HSIC}}_{4,n}\left(\frac{\widehat{y}}{\widehat{\sigma}_{\widehat{y}}}, \frac{\widehat{z}}{\widehat{\sigma}_{\widehat{z}}}\right).$$

Lemma D.6 (Gaussian HSIC approximation)**.** *There exists a constant $c_{\mathrm{HSIC}} > 0$, depending only on (K, λ, b_-), such that $\widehat{\mathcal{D}}_{\mathrm{HSIC},4}$ satisfies Assumption 5.1 for $\mathcal{D} = \mathcal{D}_{\mathrm{HSIC}}$ with $c_{\mathcal{D}} = c_{\mathrm{HSIC}}$.*

Proof. Assume $\mathcal{I}_{\mathrm{out}}(h) \neq \varnothing$, since otherwise the assertion holds automatically. Then $N_{\mathrm{test}} \geq 1$. Since $\ell(h; \delta) \geq \log 32$,

$$\ell(h; \delta/6) \leq \left(1 + \frac{\log 6}{\log 32}\right) \ell(h; \delta).$$

For each index, write $\widehat{\mathcal{D}}_{\mathrm{HSIC},4}$ and $\mathcal{D}_{\mathrm{HSIC}}$ for the corresponding empirical and population scores, respectively. Define the oracle statistic on the same observations by

$$\widetilde{\mathcal{D}}_{\mathrm{HSIC},4} := T^{\mathrm{HSIC}}_{4,n}\left(\frac{y}{\sigma_Y}, \frac{z}{\sigma_Z}\right).$$

Then $|\widehat{\mathcal{D}}_{\mathrm{HSIC},4} - \mathcal{D}_{\mathrm{HSIC}}| \leq |\widehat{\mathcal{D}}_{\mathrm{HSIC},4} - \widetilde{\mathcal{D}}_{\mathrm{HSIC},4}| + |\widetilde{\mathcal{D}}_{\mathrm{HSIC},4} - \mathcal{D}_{\mathrm{HSIC}}|$.

We first bound the estimation error of the oracle score. Empirical centering leaves pairwise differences unchanged, so the oracle statistic equals the complete unbiased HSIC statistic on the i.i.d. pairs $(Y_i/\sigma_Y, Z_i/\sigma_Z)$. Since $0 \leq k_b \leq 1$, Song et al. (2012, Theorem 4) gives, for every $0 < \eta < 1$,

$$\Pr\left(|\widetilde{\mathcal{D}}_{\mathrm{HSIC},4} - \mathcal{D}_{\mathrm{HSIC}}| > 8\sqrt{\frac{\log(2/\eta)}{n}}\right) \leq \eta.$$

Take $\eta = 2e^{-2\ell(h;\delta)}$. By Lemma D.1(ii) and $\ell(h; \delta) \geq \log 32$, we have

$$N_{\mathrm{test}}\eta = 2N_{\mathrm{test}}e^{-2\ell(h;\delta)} \leq \delta/128.$$

A union bound over at most N_{test} OUT statistics therefore gives, with probability at least $1 - \delta/128$,

$$\max_{\mathcal{I}_{\mathrm{out}}(h)} |\widetilde{\mathcal{D}}_{\mathrm{HSIC},4} - \mathcal{D}_{\mathrm{HSIC}}| \leq 8\sqrt{2}\sqrt{\frac{\ell(h; \delta)}{n}}. \tag{33}$$

It remains to bound the difference between the empirical and oracle scores. Apply Lemma D.2(i) and (ii) with $\alpha = \delta/6$. On the residual-prediction event,

$$\left\| \frac{\widehat{y}}{\widehat{\sigma}_{\widehat{y}}} - \frac{y}{\sigma_Y} \right\|_n \leq \frac{\|\widehat{y} - y\|_n}{\widehat{\sigma}_{\widehat{y}}} + \frac{\|y\|_n |\widehat{\sigma}_{\widehat{y}} - \sigma_Y|}{\widehat{\sigma}_{\widehat{y}} \sigma_Y}.$$

The reverse triangle inequality gives $|\widehat{\sigma}_{\widehat{y}} - \sigma_Y| \leq \|\widehat{y} - y\|_n + |\widehat{\sigma}_y - \sigma_Y|$. Substituting the bounds from Lemma D.2(i) and (ii) into the preceding inequalities gives, uniformly over $\mathcal{I}_{\mathrm{out}}(h)$,

$$\left\| \frac{\widehat{y}}{\widehat{\sigma}_{\widehat{y}}} - \frac{y}{\sigma_Y} \right\|_n \leq c_{\mathrm{std}} \sqrt{\frac{\ell(h; \delta)}{n}}. \tag{34}$$

Here

$$c_{\mathrm{std}} := \sqrt{1+\frac{\log 6}{\log 32}}\left[\frac{c_{\mathrm{res}}}{\sqrt{\lambda^{-1}/2}}+\frac{\sqrt{2\lambda}}{\sqrt{\lambda^{-1}/2}\sqrt{\lambda^{-1}}}\{c_{\mathrm{res}}+c_{\mathrm{res}}\sqrt{\lambda}\}\right].$$

The same bound holds for z.

For each observation, define

$$\delta_i^Y := \left|\frac{\widehat{y}_i}{\widehat{\sigma}_{\widehat{y}}}-\frac{y_i}{\sigma_Y}\right|, \qquad \delta_i^Z := \left|\frac{\widehat{z}_i}{\widehat{\sigma}_{\widehat{z}}}-\frac{z_i}{\sigma_Z}\right|.$$

The mean-value theorem, together with $\sup_{t\in\mathbb{R}}\left|\frac{d}{dt}e^{-t^2/(2b^2)}\right| = e^{-1/2}/b$, gives

$$|k_b(u,v)-k_b(u',v')| \le \frac{e^{-1/2}}{b_-}(|u-u'|+|v-v'|).$$

Since Gaussian kernel values lie in $[0,1]$, $|ab-a'b'| \le |a-a'|+|b-b'|$ applies to each kernel product. For a quartet $q=(q_1,q_2,q_3,q_4)$, these inequalities bound the absolute difference between the fitted and oracle evaluations of the order-four kernel by

$$\frac{4e^{-1/2}}{b_-}\sum_{a=1}^{4}(\delta_{q_a}^Y+\delta_{q_a}^Z).$$

In the complete average, each index belongs to exactly $\binom{n-1}{3}$ quartets, so Cauchy–Schwarz gives

$$\begin{aligned}&\binom{n}{4}^{-1}\sum_{1\le q_1<q_2<q_3<q_4\le n}\sum_{a=1}^{4}(\delta_{q_a}^Y+\delta_{q_a}^Z)\\ &\quad=\frac{\binom{n-1}{3}}{\binom{n}{4}}\sum_{i=1}^{n}(\delta_i^Y+\delta_i^Z)=\frac{4}{n}\sum_{i=1}^{n}(\delta_i^Y+\delta_i^Z)\le 4(\|\delta^Y\|_n+\|\delta^Z\|_n).\end{aligned}$$

Applying (34) to both arrays then gives

$$|\widehat{\mathcal{D}}_{\mathrm{HSIC},4}-\widetilde{\mathcal{D}}_{\mathrm{HSIC},4}| \le \frac{16e^{-1/2}}{b_-}(\|\delta^Y\|_n+\|\delta^Z\|_n) \le \frac{32e^{-1/2}}{b_-}c_{\mathrm{std}}\sqrt{\frac{\ell(h;\delta)}{n}}. \tag{35}$$

On the intersection of the event from Lemma D.2(i) and (ii) with $\alpha=\delta/6$ and the event in (33), both error bounds hold. Combining (33) and (35), choose

$$c_{\mathrm{HSIC}} := \max\left\{\left(1+\frac{\log 6}{\log 32}\right)c_{\mathrm{res}}, \frac{4}{\log 32}, 8\sqrt{2}+\frac{32e^{-1/2}}{b_-}c_{\mathrm{std}}\right\}.$$

This choice ensures all required sample-size conditions and yields the uniform error bound in Assumption 5.1 with failure probability at most $\delta/6+\delta/128<\delta/3$, proving the lemma. □

E Separation and optimality

Vanishing cycles explain the need for quantitative separation. The lower bound with fixed separation supplies the necessity argument for the optimality conclusion in Corollary 5.3. We retain the noise density f, $g=\log f$, the laws P_B, and the constants $c_{\mathrm{KL}}, K, \lambda$ fixed in Appendix B. Class margins use candidates with $|C|+|P|\le s+d$, as in Section 5.3.

E.1 Vanishing cyclic separation

Proposition E.1 (No uniform recovery without cyclic separation)**.** *Let*

$$B_0 = 0_{2\times 2}, \qquad B_\gamma = \begin{bmatrix}0 & \gamma\\ \gamma & 0\end{bmatrix},$$

and write $\Pi(B)$ for the SCC partition generated by B. Then, for every $0<\gamma\le 1/2$ and every estimator $\widehat{\Pi}$ based on n observations,

$$n \le \frac{1}{16c_{\mathrm{KL}}\gamma^2} \quad\Longrightarrow\quad \max\left\{P_{B_0}^n\big(\widehat{\Pi}\neq\Pi(B_0)\big), P_{B_\gamma}^n\big(\widehat{\Pi}\neq\Pi(B_\gamma)\big)\right\} \ge \frac{3}{8}.$$

Proof. The two models have different SCC partitions: B_0 has two singleton SCCs, whereas B_γ has one two-node SCC. Moreover, $\|B_\gamma\|_{\mathrm{op}} = \gamma \le 1/2$ and $\det(I - B_\gamma) = 1 - \gamma^2 > 0$. Thus both models satisfy Assumption 2.1, and Lemma B.2 gives common constants for Assumption 2.2.

Since $\|B_\gamma\|_F^2 = 2\gamma^2$, Lemma B.1 and tensorization give

$$\mathrm{KL}(P_{B_\gamma}^n \| P_0^n) \le 2c_{\mathrm{KL}} n\gamma^2.$$

Any estimator of the SCC partition induces a test between the two models. Pinsker's inequality and Le Cam's two-point bound (Tsybakov, 2009, Ch. 2) therefore imply

$$\begin{aligned}\max\left\{P_{B_0}^n\big(\widehat{\Pi} \neq \Pi(B_0)\big), P_{B_\gamma}^n\big(\widehat{\Pi} \neq \Pi(B_\gamma)\big)\right\} &\ge \frac{1 - \mathrm{TV}(P_{B_\gamma}^n, P_0^n)}{2} \\ &\ge \frac{1 - \sqrt{\mathrm{KL}(P_{B_\gamma}^n \| P_0^n)/2}}{2} \\ &\ge \frac{1 - \sqrt{c_{\mathrm{KL}} n\gamma^2}}{2}.\end{aligned}$$

If $n \le (16c_{\mathrm{KL}}\gamma^2)^{-1}$, the last expression is at least $3/8$, which proves the proposition. □

E.2 Lower bound with fixed separation

We construct uniformly separated model families by adding isolated variables to fixed finite models.

Lemma E.2 (Isolated padding and fixed-motif separation)**.** *Let B° define a fixed LiNG model on a node set W of size m_0 satisfying Assumption 2.1. For $p \ge m_0$, embed this model in $[p]$ and let every coordinate in W^c be an isolated noise variable with density f. Fix integers $s \ge 1$ and $d \ge 0$, and a dependence functional $\mathcal{D}$ satisfying the independence zero-set condition in Section 5.1. Define Δ_{par} and $\Delta_{\mathcal{D}}$ as in (5), with the candidate-size bound $|C| + |P| \le s + d$. Then both margins belong to $(0, \infty]$ and depend only on $(B^\circ, s, d, \mathcal{D})$, not on p or the embedding of W.*

Proof. The coordinates in W^c are mutually independent and independent of X_W. For a candidate (F, C, P), set $F_W = F \cap W$, $C_W = C \cap W$, and $P_W = P \cap W$, and use a superscript $\circ$ for quantities in the motif model. If $C_W = \varnothing$, then $e_C(P) = e_C(F)$ and C is ancestral, so neither margin includes this candidate. Hence assume $C_W \neq \varnothing$.

Since

$$\mathrm{Cov}(X_P) = \begin{bmatrix}\mathrm{Cov}(X_{P_W}) & 0 \\ 0 & \mathrm{Cov}(X_{P \setminus W})\end{bmatrix},$$

the normal equations give, for $v \notin P$,

$$e_v(P) = \begin{cases} e_v^\circ(P_W), & v \in W, \\ X_v, & v \notin W. \end{cases}$$

Consequently,

$$e_C(P) = \big(e_{C_W}^\circ(P_W), X_{C \setminus W}\big),$$

and the two blocks are independent. The normal equations for the projection of $e_\ell(P)$ onto $e_C(P)$ therefore give

$$r_{P,C}(\ell) = \begin{cases} r_{P_W, C_W}^\circ(\ell), & \ell \in W, \\ X_\ell, & \ell \notin W. \end{cases}$$

Consequently, every UP or OUT score is zero if either tested variable lies outside W; otherwise, it equals the corresponding motif score. Taking empty maxima as zero,

$$D_F^{\mathrm{up}}(C, P) = D_{F_W}^{\mathrm{up},\circ}(C_W, P_W), \qquad D_F^{\mathrm{out}}(C, P) = D_{F_W}^{\mathrm{out},\circ}(C_W, P_W).$$

The first residual identity, applied with P and with F, also gives

$$e_C(P) = e_C(F) \iff e_{C_W}^\circ(P_W) = e_{C_W}^\circ(F_W).$$

Because every node outside W is isolated, ancestry is preserved under the same reduction:

$$\begin{aligned} F \text{ is ancestral in } G &\iff F_W \text{ is ancestral in } G^\circ, \\ C \text{ is ancestral in } G[V \setminus F] &\iff C_W \text{ is ancestral in } G^\circ[W \setminus F_W]. \end{aligned}$$

Thus restricting (F, C, P) to W preserves both the conditions defining each margin and the corresponding scores. The restricted candidates (F_W, C_W, P_W) satisfy

$$F_W \subseteq W \text{ ancestral}, \qquad \varnothing \neq C_W \subseteq W \setminus F_W, \qquad P_W \subseteq F_W, \qquad |C_W| + |P_W| \leq s + d.$$

Conversely, every such motif tuple is realized by taking $(F, C, P) = (F_W, C_W, P_W)$.

For every reduced tuple in the family defining Δ_{par},

$$e^\circ_{C_W}(P_W) \neq e^\circ_{C_W}(F_W) \quad \Longrightarrow \quad D^{\mathrm{up},\circ}_{F_W}(C_W, P_W) > 0,$$

because a zero UP score is equivalent to equality of the two residuals.

Likewise, consider a restricted candidate satisfying $e^\circ_{C_W}(P_W) = e^\circ_{C_W}(F_W)$ with C_W non-ancestral in $G^\circ[W \setminus F_W]$. The residual equality gives UP. If its OUT score were zero, the pair would pass, so Proposition 4.3 would imply that C_W is ancestral, a contradiction. Thus $D^{\mathrm{out},\circ}_{F_W}(C_W, P_W) > 0$.

The two margins are therefore minima over finite collections of positive motif scores depending only on $(B^\circ, s, d, \mathcal{D})$. Hence they belong to $(0, \infty]$, with an empty minimum equal to $+\infty$. □

We now apply this construction to the separated model class in Section 5.3, with a common separation level for the two margins.

Proposition E.3 (Lower bound with fixed separation)**.** *Fix $s_0 \geq 2$, $d_0 \geq 1$, and a normalized dependence functional $\mathcal{D}$ satisfying the independence zero-set condition. There exist constants*

$$K_0, \tau_0, c_{\mathrm{lb},0} > 0, \qquad \lambda_0 \geq 1,$$

depending only on $(s_0, d_0, f, \mathcal{D})$ and not on (p, n, δ), such that the following holds. For every $p \geq 2\max\{s_0, d_0\}$, sample size $n \geq 1$, and $0 < \delta < 1$, every estimator $\widehat{G}^{\mathrm{sc}}$ satisfies

$$n \leq \frac{1-\delta}{c_{\mathrm{lb},0}} \left\{ s_0 \log \frac{ep}{s_0} + d_0 \log \frac{ep}{d_0} \right\} \quad \Longrightarrow \quad \sup_{M \in \mathcal{M}^{\mathrm{sep}}_{p,s_0,d_0}(K_0,\lambda_0;\tau_0,\tau_0)} P^n_M \left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M) \right) \geq \delta.$$

Proof. We first construct a cyclic family. Set $\gamma = 1/8$. For every $S = \{j_1 < \cdots < j_{s_0}\} \subseteq [p]$, let B^{cyc}_S represent a directed cycle on S with coefficient γ on each edge and all nodes in S^c isolated:

$$(B^{\mathrm{cyc}}_S)_{j_{r+1},j_r} = \gamma \quad \text{for } 1 \leq r < s_0, \qquad (B^{\mathrm{cyc}}_S)_{j_1,j_{s_0}} = \gamma,$$

with all other entries equal to zero.

The principal block $(B^{\mathrm{cyc}}_S)_{S,S}$ is the permutation matrix of a directed s_0-cycle, scaled by γ. All entries outside this block are zero. Hence

$$\|B^{\mathrm{cyc}}_S\|_{\mathrm{op}} = \gamma, \qquad \|B^{\mathrm{cyc}}_S\|_F^2 = s_0\gamma^2 = \frac{s_0}{64}.$$

Every principal submatrix of B^{cyc}_S has operator norm at most $\gamma < 1$, so every principal submatrix of $I - B^{\mathrm{cyc}}_S$ is invertible. Hence every model in this family satisfies Assumption 2.1. By construction, $s_{\max} = s_0$ and $d_B = 0$.

We next construct the parent family using the packing in Lemma B.4 with $d = d_0$. Every model in that family is acyclic, satisfies $s_{\max} = 1$, and has $d_B = d_0$. Since both families have operator norm at most $1/2$, Lemma B.2 gives the common class constants $K_0 := K$ and $\lambda_0 := \lambda$. Thus both families lie in $\mathcal{M}_{p,s_0,d_0}(K_0, \lambda_0)$.

We next verify uniform separation. Each cyclic model is a fixed s_0-node motif padded with isolated coordinates, and each parent model is a fixed $(d_0 + 1)$-node motif padded with isolated coordinates. Lemma E.2 therefore shows that all corresponding margins are positive or $+\infty$ and independent of p and of the embedding.

For the cyclic family, there are no edges between distinct SCCs. Hence every ancestral peeling state F satisfies $X_F \perp\!\!\!\perp X_{V\setminus F}$, and $e_C(P) = X_C = e_C(F)$ for every candidate in the definition of the margins. Thus $\Delta^{\mathrm{cyc}}_{\mathrm{par}} = +\infty$. Let $\Delta^{\mathrm{cyc}}_{\mathcal{D}}$, $\Delta^{\mathrm{par}}_{\mathrm{par}}$, and $\Delta^{\mathrm{par}}_{\mathcal{D}}$ denote the remaining motif margins, and set

$$\tau_0 := \min\left\{1,\, \Delta^{\mathrm{cyc}}_{\mathcal{D}},\, \Delta^{\mathrm{par}}_{\mathrm{par}},\, \Delta^{\mathrm{par}}_{\mathcal{D}}\right\}.$$

Then $\tau_0 > 0$, and every model in either packing belongs to $\mathcal{M}^{\mathrm{sep}}_{p,s_0,d_0}(K_0, \lambda_0; \tau_0, \tau_0)$.

It remains to apply Fano's inequality. Write $h_{s_0} = s_0 \log(ep/s_0)$ and $h_{d_0} = d_0 \log(ep/d_0)$. The cyclic and parent families contain $N_{s_0} = \binom{p}{s_0}$ and $N_{d_0} = \binom{p-1}{d_0}$ distinct condensations, respectively. Since $\binom{p}{s_0} \geq \binom{p-1}{s_0-1}$, Lemmas B.3(ii) and B.4(ii) give

$$\log N_{s_0} \geq \frac{\log 2}{2(1+\log 2)} h_{s_0}, \qquad \log N_{d_0} \geq \frac{\log 2}{2(1+\log 2)} h_{d_0}.$$

Lemma B.1, tensorization, and Lemma B.4(iii) give

$$\max_S \mathrm{KL}(P^n_{B^{\mathrm{cyc}}_S} \| P^n_0) \leq \frac{c_{\mathrm{KL}} s_0 n}{64}, \qquad \max_Q \mathrm{KL}(P^n_{B_Q} \| P^n_0) \leq \frac{c_{\mathrm{KL}} n}{4}.$$

Choose the cyclic family if $h_{s_0} \geq h_{d_0}$ and the parent family otherwise, and let N_0 be its cardinality. Its log-cardinality is at least $\log 2\,(h_{s_0} + h_{d_0})/[4(1+\log 2)]$, and its average KL divergence from P^n_0 is at most $c_{\mathrm{KL}} \max\{s_0/64, 1/4\} n$. Set

$$c_{\mathrm{lb},0} := \frac{4(1+\log 2) c_{\mathrm{KL}} \max\{s_0/64, 1/4\}}{\log 2} + 4(1+\log 2).$$

Generalized Fano's inequality and $n \geq 1$ give

$$\begin{aligned} &\sup_{M \in \mathcal{M}^{\mathrm{sep}}_{p,s_0,d_0}(K_0,\lambda_0;\tau_0,\tau_0)} P^n_M\left(\widehat{G}^{\mathrm{sc}} \neq G^{\mathrm{sc}}(M)\right) \\ &\quad \geq 1 - \frac{4(1+\log 2) c_{\mathrm{KL}} \max\{s_0/64, 1/4\} n}{\log 2\,(h_{s_0} + h_{d_0})} - \frac{4(1+\log 2)}{h_{s_0} + h_{d_0}} \geq 1 - \frac{c_{\mathrm{lb},0} n}{h_{s_0} + h_{d_0}}. \end{aligned}$$

Under the sample-size condition in the proposition, the right-hand side is at least δ. □

F Experimental details and additional results

F.1 Experimental protocol

Graph structures and coefficients. The main experiments use two graph constructions with one or four cyclic SCCs. Within each cyclic SCC of size $s_{\max}$, a center is connected reciprocally to the other $s_{\max} - 1$ nodes, with coefficient $0.5/\sqrt{s_{\max} - 1}$ on each edge, keeping the spectral radius at 0.5 across SCC sizes. For $s_{\max} = 2$, this forms a 2-cycle; for $s_{\max} = 3, 4$, it forms overlapping 2-cycles sharing the center. Each external edge has coefficient $0.8/\sqrt{d_B}$.

In the one-SCC construction, nodes $1, \ldots, s_{\max}$ form a root SCC with center 1, and there are $d_B - 1$ additional singleton roots. Each of the remaining $p - s_{\max} - d_B + 1$ singleton children has node 1 and all additional roots as parents. Increasing p adds children while preserving the cyclic core. Figure 3 uses this construction.

The four-SCC construction has one root and three child cyclic SCCs, each of size $s_{\max}$, together with $d_B - 1$ singleton roots. The root SCC's center and all singleton roots are parents of each child SCC's center and all $p - 4s_{\max} - d_B + 1$ singleton children. Figure 4 uses this construction.

Data generation. Given a coefficient matrix B, we generate independent observations of $X = (I - B)^{-1}\varepsilon$. Unless otherwise specified, the coordinates of ε are independent and follow the standardized mixture $(W + 1.6)/\sqrt{1.45}$, where $W \sim 0.9\, N(-2, 0.01) + 0.1\, N(2, 0.01)$. Samples at different sample sizes are generated independently.

Method settings. BlockExo uses centered full-sample OLS and the absolute empirical covariances for UP. For OUT it uses the complete unbiased estimator of squared distance covariance, normalized by the empirical residual scales of the centered fitted residuals y, z:

$$\widehat{\mathcal{D}}(y,z) = \frac{\widehat{\mathrm{dCov}}_U^2(y,z)}{\sqrt{(n^{-1}\sum_i y_i^2)(n^{-1}\sum_i z_i^2)}}.$$

The signed estimate is used without absolute values or clipping. We compute the statistic using the `dcor` package (Ramos-Carreño & Torrecilla, 2023). Previously computed statistics are cached across recovery rounds. Thresholds are $\tau_j(t) = \kappa_j\sqrt{\ell(t)/n}$ at $\delta = 0.05$, with empirical multipliers $\kappa_{\mathrm{up}} = 1.68$ and $\kappa_{\mathrm{out}} = 0.22$, selected using separate simulated datasets and held fixed except in the noise-distribution comparisons below.

Baseline thresholds were selected using separate simulated datasets and held fixed during evaluation. For Coarsening, we report both the default threshold 0.1 and the selected threshold 0.2 (Madaleno et al., 2026). Each threshold sets the initial unmixing cutoff and the edge cutoff; only the former is relaxed if no admissible matching exists. DisjointCycles uses $\alpha = 0.05$ (Drton et al., 2025); the overlapping construction lies outside its disjoint-cycle assumption. StableSpIn uses 3,000 epochs and edge cutoff $\omega = 0.15$ (Misiakos & Püschel, 2026).

For the noise-distribution comparisons in Appendix F.5, κ_{up} remains fixed, while $\kappa_{\mathrm{out}} = 0.04$ for the symmetric Gaussian mixture and 0.01 for uniform noise. These values were selected using separate simulated datasets from each noise law and fixed before evaluation.

Using 36 selection datasets per distribution shared across the baselines, we evaluate Coarsening thresholds $\{0.025, 0.05, 0.1, 0.2, 0.3\}$, DisjointCycles levels $\{0.0001, 0.001, 0.01, 0.05, 0.1\}$, and StableSpIn cutoffs $\{0.025, 0.05, 0.09, 0.15, 0.25\}$. Selection prioritizes exact recovery, then partition recovery and the number of correctly recovered cyclic SCCs, with remaining ties resolved by preference for the default and proximity to it. The selected values are 0.2, 0.01, and 0.09, respectively, for both noise laws, with DisjointCycles and StableSpIn retaining their defaults because all three recovery criteria are zero throughout their grids. Coarsening 0.1 is also retained.

The exact GroupLiNGAM search (Kawahara et al., 2010) is infeasible at $p = 50$ under our 48 GiB memory budget. Its large-graph approximation, L-GroupLiNGAM, targets approximate block ordering rather than exact condensation recovery and is therefore omitted.

Evaluation. All methods are evaluated on the same datasets across 20 replicates. For methods returning variable-level graphs, we extract the SCC partition and condensation edges. A replicate counts as an exact recovery when both the SCC partition and all condensation edges are correct, with components compared by their node sets. Fits that do not complete, including those returning no passing block, count as recovery failures.

F.2 Separation and sample cost

We examine the contributions of structural complexity and separation to sample cost in the one-SCC construction. In the dimension comparison, the hybrid margin is constant within each d_B, while the external-parent comparison changes both structural complexity and separation.

To compare recovery curves on the scale suggested by the sample bound, we normalize sample size as

$$\widehat{z} = \frac{n\widehat{\Delta}_{\mathrm{hyb}}^2}{L}, \qquad \widehat{\Delta}_{\mathrm{hyb}} = \min\{\Delta_{\mathrm{par}}/\kappa_{\mathrm{up}}, \widehat{\Delta}_{\mathcal{D}}/\kappa_{\mathrm{out}}\}.$$

Parent margins are evaluated from population covariance, and dependence margins are estimated from reference samples generated independently of the recovery datasets. Each setting has a single reference margin for normalizing its recovery curve.

Dimension dependence. At fixed $s_{\max} = 3$ and d_B, the parent margin determines the same hybrid margin across $p = 10, 30, 50$ in Figure 3(b). Thus n/L and $\widehat{z}$ differ by a constant factor within each group, and the closer alignment after normalization supports the dimension dependence captured by L. The hybrid margins are approximately 0.201, 0.132, and 0.091 for $d_B = 1, 2, 3$, respectively, and agree across dimensions at numerical precision.

External-parent count. In this construction, increasing d_B raises the structural cost and reduces the hybrid margin. Moving from $d_B = 1$ to $d_B = 2$ expands the adjustment search and reduces each external-edge coefficient from 0.8 to $0.8/\sqrt{2}$. For $s_{\max} = 3$ and $p = 50$, L increases from approximately 28.5 to 32.0, while $\widehat{\Delta}_{\mathrm{hyb}}$ decreases from 0.201 to 0.132. Together, these changes increase $L/\widehat{\Delta}^2_{\mathrm{hyb}}$ by a factor of approximately 2.58, consistent with BlockExo's recovery transition at larger sample sizes for $d_B = 2$ in Figure 3(b).

Reference margin calculation. The reference calculations evaluate (5) over the full candidate class $\mathcal{A}$: all ancestral sets F and candidates satisfying $|C| + |P| \le s_{\max} + d_B$, without separate restrictions on $|C|$ and $|P|$. Dependence margins are estimated using the complete dCov statistic on sixteen independent reference samples of 8,192 observations, with population regression coefficients and residual scales. Scalar scores are averaged across these replicates before forming candidate maxima and the minimum over invalid candidates. These calculations cover every setting in Figure 3.

F.3 Dimension comparison with two cyclic SCCs

Normalization by L brings recovery transitions across dimensions closer together within each d_B, also for graphs with a root and a child cyclic SCC (Figure 5). This extends the dimension comparison in Figure 3(b).

We extend the one-SCC construction in Appendix F.1 at $s_{\max} = 3$ by adding a three-node child SCC with the same internal coefficients. The center of the root SCC and the $d_B - 1$ additional singleton roots are parents of the child SCC's center and all singleton children, with coefficient $0.8/\sqrt{d_B}$ on each external edge. Adding singleton children varies $p \in \{10, 30, 50\}$ while $s_{\max} = 3$ and the two cyclic SCCs remain fixed. We consider $d_B \in \{1, 2\}$ and use search bounds $(s, d) = (s_{\max}, d_B)$.

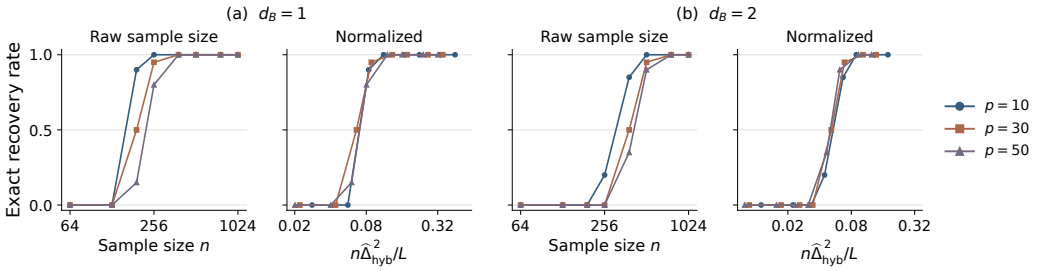

Figure 5: Exact condensation recovery across dimensions with two cyclic SCCs.

The parent margin determines the hybrid margin, which is approximately 0.0946 for $d_B = 1$ and 0.0624 for $d_B = 2$ across $p = 10, 30, 50$. These values remain unchanged despite the decrease in the estimated dependence margin with dimension at $d_B = 2$. Thus L accounts for the observed differences in sample cost across dimensions within each d_B.

Reference margins are computed over $\mathcal{A}$ following Appendix F.2, with dependence scores estimated from four independent reference samples of 16,384 observations.

F.4 External-edge strength and dependence separation

Dependence separation cannot be explained by external-edge strength or covariance alone. We examine these distinctions by varying external-edge coefficients and noise distributions in two separate experiments.

Both experiments use the candidate class $\mathcal{A}$ and scalar-score calculations described in Appendix F.2. Parent margins are computed from population covariance, and dependence margins are estimated from 16 independent reference samples of 8,192 observations.

External-edge strength. We vary the external-edge scale in a six-variable construction with SCCs $\{1, 2, 3\}$, $\{4, 5\}$, and $\{6\}$. Before rescaling, the coefficients within the cyclic SCCs are

$(B_{21}, B_{32}, B_{13}, B_{31}) = (0.60, -0.55, 0.58, 0.28)$ and $(B_{54}, B_{45}) = (0.55, -0.50)$. The coefficients on edges between SCCs are $(B_{41}, B_{43}, B_{52}, B_{65}) = a(1, 1/2, -3/4, 1)$, with $a \in \{0.10, 0.20, 0.35, 0.50, 0.70, 0.90\}$; all other entries are zero. The two cyclic blocks are rescaled so that their entrywise absolute coefficient matrices have spectral radii 0.55 and 0.50, respectively. We also consider three variants obtained by multiplying nonzero coefficients independently by $\mathrm{Unif}(0.94, 1.06)$ factors and rescaling the cyclic blocks again, with the same factors across scales. Errors are independent $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$ variables.

Figure 6(a) shows median margins across the four coefficient variants at each scale. The parent margin increases with external-edge strength, whereas the estimated dependence margin does not. Thus stronger external edges do not necessarily give greater dependence separation.

Noise distribution. We vary the noise distribution while keeping the graph, coefficients, and covariance fixed in the one-SCC construction with $p = 10$, $s_{\max} = 3$, and $d_B = 2$. The four noise laws are uniform, $\mathrm{Beta}(2, 5)$, symmetric mixture $\frac{1}{2}N(-0.95, 1 - 0.95^2) + \frac{1}{2}N(0.95, 1 - 0.95^2)$, and the skewed mixture in Appendix F.1, all standardized to zero mean and unit variance. The dependence boxplots in Figure 6(b) summarize the Monte Carlo variability of the margin estimates computed separately from each reference sample.

The parent margin remains 0.223, while the estimated dependence margin ranges from 0.00092 for uniform noise to 0.0471 for the skewed Gaussian mixture, a roughly 51-fold difference. Thus the noise distribution can substantially change dependence separation even when the graph and covariance are unchanged.

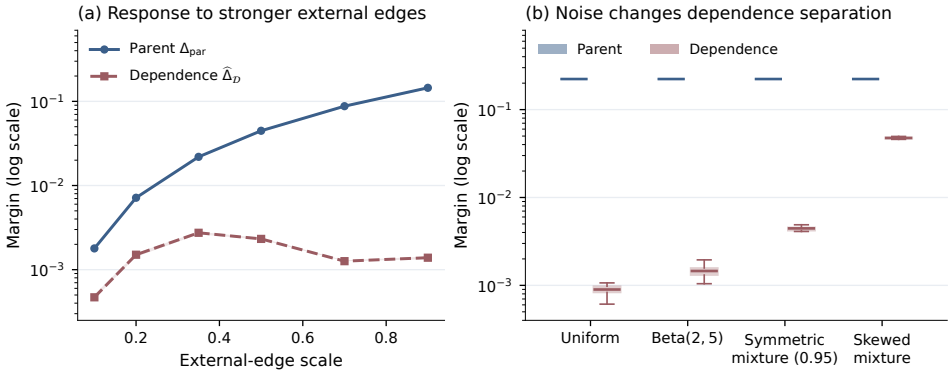

Figure 6: Parent and dependence margins under varying external-edge strength and noise distributions.

F.5 Additional sample-efficiency comparisons

We extend the method comparison in Figure 4 to different noise distributions and DAGs, using $(s, d) = (3, 2)$ for BlockExo throughout.

Noise distributions. We use the four-SCC construction at $p = 50$ with $s_{\max} \in \{2, 3\}$ and $d_B \in \{1, 2\}$, replacing the noise law by either the symmetric Gaussian mixture $\frac{1}{2}N(-0.95, 1 - 0.95^2) + \frac{1}{2}N(0.95, 1 - 0.95^2)$ or $\mathrm{Unif}(-\sqrt{3}, \sqrt{3})$. The graph and coefficients remain fixed; both noise laws have zero mean and unit variance, preserving the population covariance.

Under the symmetric Gaussian mixture, BlockExo begins recovering disjoint cycles at smaller sample sizes than Coarsening (Figure 7). At $n = 512$, it achieves $15/20$ and $13/20$ exact recoveries for $d_B = 1, 2$, respectively, while both Coarsening thresholds achieve none. For overlapping cycles, the recovery transitions are similar at $d_B = 1$, but Coarsening 0.2 reaches high recovery accuracy earlier at $d_B = 2$. BlockExo attains $20/20$ exact recoveries in all four settings at $n = 1024$.

Under uniform noise, the relative sample cost depends on the cycle structure and external-parent count (Figure 8). BlockExo begins recovering disjoint cycles at smaller sample sizes, although Coarsening 0.2 reaches $20/20$ earlier at $d_B = 1$. For overlapping cycles, Coarsening 0.2 reaches high recovery

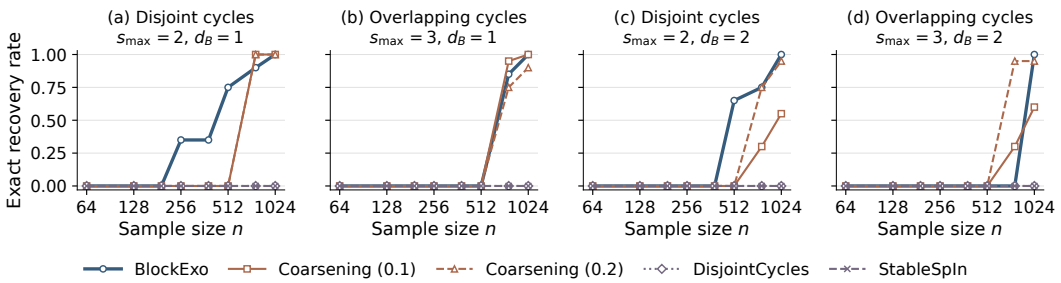

Figure 7: Exact condensation recovery under symmetric Gaussian mixture noise.

accuracy earlier at $d_B = 1$, while the two methods perform similarly at $d_B = 2$. BlockExo attains $20/20$ exact recoveries in all four settings at $n = 3072$.

Neither DisjointCycles nor StableSpIn achieves exact partition recovery under either noise law at the tested sample sizes. For DisjointCycles, both noise laws fall outside its generic-moment condition, and the overlapping construction also violates its cycle-disjointness assumption.

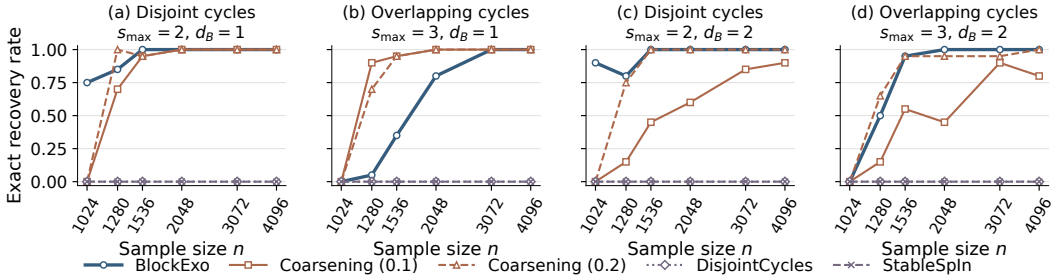

Figure 8: Exact condensation recovery under uniform noise.

The sample size needed for BlockExo to achieve high recovery accuracy increases from the skewed Gaussian mixture to the symmetric Gaussian mixture and then to uniform noise (Figures 4, 7, and 8), despite identical graphs, coefficients, and covariance. This ordering is consistent with the one-SCC margin analysis in Appendix F.4, where the estimated dependence margin decreases across these three noise laws in the same order. Together, these results highlight distribution-dependent separation as a source of recovery difficulty beyond SCC size and external-parent count.

DAG recovery. We consider DAGs with $p = 50$ and $d_B \in \{1, 2\}$, consisting of d_B roots, each connected to every remaining node. Every SCC is a singleton, so exact condensation recovery is exact DAG recovery. Even with search bounds $(s, d) = (3, 2)$ allowing cyclic blocks, BlockExo reaches $20/20$ exact recoveries at smaller sample sizes than either Coarsening threshold for $d_B = 1$, and at the same sample size as Coarsening 0.2 for $d_B = 2$ (Figure 9).

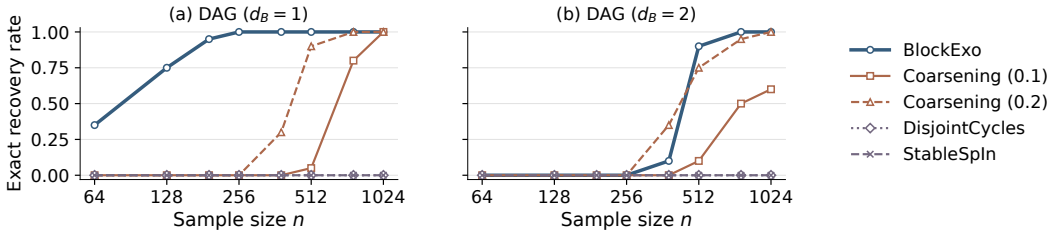

Figure 9: Exact DAG recovery across methods.

F.6 SENSITIVITY TO SEARCH BOUNDS

We compare the common search bounds $(s, d) = (3, 2)$ with the exact bounds $(s, d) = (s_{\max}, d_B)$ on the same four-SCC datasets from Figure 4, keeping the search order, threshold rule, and multipliers fixed. Even without the exact search bounds, BlockExo attains $20/20$ exact recoveries in every setting, although recovery requires larger sample sizes at $d_B = 1$. For disjoint cycles at $n = 192$, the common bounds reduce the exact-recovery rate from $18/20$ to $1/20$; for overlapping cycles at $n = 384$, the rate decreases from $20/20$ to $3/20$. At $(s_{\max}, d_B) = (2, 2)$, the two choices agree in exact-recovery outcome for every replicate despite the different bounds.

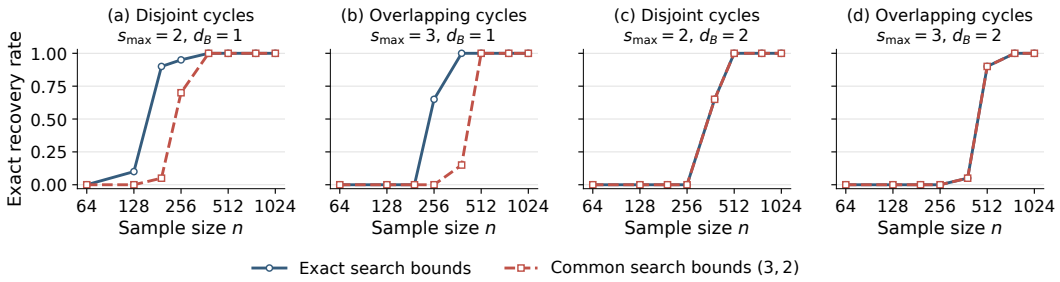

Figure 10: Sensitivity of exact condensation recovery to search bounds.

F.7 RUNTIME

BlockExo's median runtime ranges from 8 to 34 seconds across the four settings in Figure 4. BlockExo is slower than Coarsening; on disjoint cycles, its median runtime is comparable to DisjointCycles and shorter than StableSpIn. Table 2 reports median runtimes pooled across sample sizes on the same inputs for which BlockExo completed, regardless of recovery accuracy.

Table 2: Median runtimes in seconds for the Figure 4 experiments.

$(s_{\max}, d_B)$	BlockExo	Coarsening 0.1	Coarsening 0.2	DisjointCycles	StableSpIn
$(2, 1)$	8.12	1.87	1.86	8.09	27.14
$(2, 2)$	11.14	1.85	1.87	8.26	27.09
$(3, 1)$	34.11	1.86	1.84	8.71	26.88
$(3, 2)$	27.88	1.87	1.84	8.54	27.12

Timing covers execution from initialization through recovery evaluation, including subprocess overhead. CPU fits ran on an Intel Xeon Gold 6258R with one BLAS/OpenMP thread per fit. StableSpIn ran on an NVIDIA RTX A5000 GPU.