LOCAL FISHER INFORMATION ENABLES SPARSE CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse causal discovery calls for methods that exploit graph structure without estimating high-dimensional densities. We introduce Fisher Information Completion Search (FiCS), a source-first algorithm for additive noise models that uses one local Fisher score for both ordering and parent selection. Under regularity and nonconstant-parent conditions, we prove that a node's local Fisher information equals the noise Fisher information exactly when the conditioning set contains all parents, provided that it contains no descendants. This Fisher parent completion identifies the parent set as the unique minimal Fisher completion. With a maximum conditioning set size q at least the maximum indegree d, population FiCS queries marginals of at most $q + 1$ variables and recovers the true directed acyclic graph under a positive ordering margin. Bounded conditioning also has a population advantage: reducing q toward d cannot decrease, and can strictly increase, the ordering margin. A growing non-Gaussian family separates local Fisher selection from conditional-variance and leaf-first Fisher ordering. For the regularized kernel Stein estimator, we establish high-dimensional DAG consistency under $q\{1 + \log(p/q)\} + \log p = o(n)$, uniform Fisher separation, local approximation, and compatible ridge and parent penalty parameters. Experiments show the strongest gains when n is small relative to p, quantify the effect of the conditioning size, and demonstrate competitive reference-graph recovery on three real-data benchmarks.

1 INTRODUCTION

Graph sparsity is central to statistically efficient high-dimensional directed acyclic graph (DAG) learning (Kalisch & Bühlmann, 2007; Nandy et al., 2018; Aragam et al., 2019). For identifiable subclasses of linear additive noise models (ANMs), an indegree bound reduces structure recovery to local regressions or subset comparisons. This yields minimax-optimal dependence on indegree and ambient dimension for equal-variance Gaussian models (Gao et al., 2022) and, under separate assumptions, sub-Gaussian non-Gaussian models (Oh et al., 2025). Can sparsity similarly reduce distribution-sensitive recovery of both the order and parent sets in nonlinear ANMs to bounded-dimensional estimation?

Source-first search is naturally aligned with sparsity: a node becomes eligible once its parents have been selected. Conditional-variance criteria extend this strategy beyond linear-Gaussian models (Park, 2020). Prominent density-derivative methods instead exploit distributional shape through leaf-first queries on the joint distribution of the remaining variables (Rolland et al., 2022; Montagna et al., 2023a;b; Xu et al., 2024; Chen et al., 2026). Graph sparsity alone does not make these ordering queries low-dimensional. We resolve this sparsity–query mismatch through Fisher parent completion, which characterizes the parent set as the minimal Fisher completion for source-first ordering and parent recovery.

Our key result is *Fisher parent completion*: among nondescendant conditioning sets, local Fisher information attains the noise Fisher information if and only if all parents are included. Conditioning on the parents fixes the structural signal; omitting a parent yields a positive parent-omission margin. This yields Fisher Information Completion Search (FiCS), which uses the same local score for source-first ordering and size-penalized parent selection. Node ordering additionally requires a positive ordering margin because noise Fisher information may differ across nodes.

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.

Our contributions are threefold.

- **Fisher parent completion.** We characterize the parent set as the unique minimal Fisher completion, yielding FiCS, which uses one local score for both source-first ordering and parent recovery from marginals of dimension at most $q + 1$ when $q \geq d$ and the ordering margin is positive.
- **Two benefits of bounded conditioning.** Restricting q limits estimation dimension and, once $q \geq d$, reducing q cannot decrease and can strictly increase the population ordering margin. A growing non-Gaussian family satisfies uniform Fisher separation while the CaPS leaf rule and conditional-variance ordering make incorrect first population selections.
- **High-dimensional consistency.** We establish DAG consistency using a uniform local score error bound and derive the corresponding error bound for the regularized kernel Stein estimator, with complexity $q\{1 + \log(p/q)\} + \log p$. Consistency holds under uniform Fisher separation, local approximation, and compatible ridge and parent penalty parameters.

2 Related Work

Source-first recovery from conditional uncertainty. In linear equal-variance models, Chen et al. (2019) introduced indegree-bounded conditional-variance ordering. Gao et al. (2022, Algorithm 1, LearnDAG) combines this ordering with parent recovery and establishes minimax-optimal sample complexity. Conditional-variance conditions also cover nonlinear ANMs with heterogeneous noise variances (Park, 2020). For jointly Gaussian variables, FiCS ordering reduces to bounded-subset conditional-variance ordering; Appendix C gives the corresponding score identity and estimation bound. Fisher parent completion extends this local approach to distribution-sensitive recovery in nonlinear ANMs. Unlike CAM's additive regression model (Bühlmann et al., 2014), it permits structural functions that are not additive across individual parents.

Leaf-first recovery from density derivatives. SCORE, DAS, NoGAM, CaPS, and FICO infer leaf-first orderings from density scores, Hessians, or Fisher information (Rolland et al., 2022; Montagna et al., 2023a;b; Xu et al., 2024; Chen et al., 2026). Under Stein integration conditions, the expected-curvature and Fisher statistics used by CaPS and FICO give equivalent population leaf rankings. Their queries use the remaining-variable distribution, whereas FiCS conditions on selected predecessors to identify completed parent mechanisms. Section 3.4 compares the resulting ordering rules; existing comparisons of sufficient identifiability conditions address a different question (Park, 2026).

Other uses of locality. AdaScore extends score-based discovery across linear, nonlinear, and latent-variable models; HiTOC localizes leaf search to Markov-blanket subsets; SSTS reuses score geometry through Schur-complement updates (Montagna et al., 2025; Sun et al., 2026; Wu & Xie, 2026). FiCS instead bounds each queried marginal by the subset size and uses the same local score for ordering and parent selection. Section 5 analyzes regularized kernel Stein estimation uniformly over those local queries.

3 ANM Identifiability via Fisher Parent Completion

Fisher parent completion characterizes when a local conditioning set contains all parents. We use this principle to recover an ordering and parent sets, establish the population advantage of locality, and compare the resulting ordering regimes.

3.1 Fisher Parent Completion in Additive Noise Models

Let $G = (V, E)$ be a directed acyclic graph on $V = [p] := \{1, \ldots, p\}$, let $\mathrm{Pa}(i)$ denote the parents of node i, and let $d = \max_i |\mathrm{Pa}(i)|$ be the maximum indegree. Let $\mathrm{Nd}(i)$ denote the nondescendants of i, excluding i. We consider the additive noise model over G:

$$X_i = f_i(X_{\mathrm{Pa}(i)}) + \varepsilon_i, \qquad \text{independent } \varepsilon_1, \ldots, \varepsilon_p. \tag{1}$$

The structural function f_i may be nonlinear and need not be additive across its parent arguments. We write $\mathbb{E}$, Var, and $\mathbb{P}$ for expectation, variance, and probability, respectively.

For the Fisher parent completion result below, we assume that each noise density is everywhere positive and C^1, with finite location Fisher information, and that each f_i is continuous and nonconstant

in every parent coordinate. The remaining score-projection and integration-by-parts conditions are stated in Appendix A.1.

Let $s_{\varepsilon_i}(e) := \partial_e \log p_{\varepsilon_i}(e)$ and $I_i^\varepsilon := \mathbb{E}[s_{\varepsilon_i}(\varepsilon_i)^2]$ denote the noise score and noise Fisher information, respectively. For $i \notin T$, let $p_{i,T}$ denote the joint density of (X_i, X_T), and define

$$I_i(T) := \mathbb{E}\Big[\{\partial_{x_i} \log p(X_i \mid X_T)\}^2\Big], \qquad J_i(T) := -I_i(T) = \mathbb{E}\big[\partial_{x_i}^2 \log p_{i,T}(X_i, X_T)\big]. \tag{2}$$

We call $I_i(T)$ the local Fisher information and $J_i(T)$ the local Fisher score. A nondescendant set T satisfying $I_i(T) = I_i^\varepsilon$ is called a Fisher completion for node i.

Theorem 1 (Fisher parent completion)**.** *Consider an ANM in Equation (1). Under the stated regularity and nonconstant-parent conditions, for every node i and every $T \subseteq \mathrm{Nd}(i)$,*

$$\Gamma_i(T) := I_i^\varepsilon - I_i(T) \geq 0, \qquad \Gamma_i(T) = 0 \iff \mathrm{Pa}(i) \subseteq T. \tag{3}$$

Consequently, $\mathrm{Pa}(i)$ is the unique minimal Fisher completion.

Fisher parent completion has a simple projection interpretation:

$$\Gamma_i(T) = \mathbb{E}[\mathrm{Var}\{s_{\varepsilon_i}(\varepsilon_i) \mid X_i, X_T\}] = \min_g \mathbb{E}\Big[\{s_{\varepsilon_i}(\varepsilon_i) - g(X_i, X_T)\}^2\Big].$$

Thus, $\Gamma_i(T)$ is the minimum prediction error of the noise score from (X_i, X_T). If T contains all parents, then the noise score $s_{\varepsilon_i}(\varepsilon_i)$ is exactly recovered by (X_i, X_T) and the error is zero; if a parent is omitted, the error is strictly positive. Figure 1 illustrates this distinction through the conditional density of X_i given X_T.

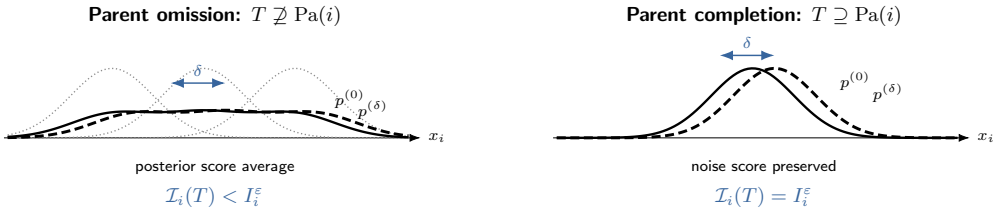


Figure 1: Fisher parent completion. Black curves show the conditional density of X_i given X_T, while gray curves show translated noise densities $p_{\varepsilon_i}(x_i - m)$ for possible values of $m = f_i(X_{\mathrm{Pa}(i)})$. Parent omission averages over multiple translations, giving $I_i(T) < I_i^\varepsilon$ (left); parent completion leaves a single translation, giving $I_i(T) = I_i^\varepsilon$ (right). Solid and dashed black curves denote the conditional density before and after a small location shift δ.

3.2 Identifiability via Local Fisher Information

A topological ordering is a bijection $\pi : V \to [p]$ such that $\pi(i) < \pi(j)$ for every edge $(i, j) \in E$, where $\pi(i)$ denotes the position of node i. A set $S \subseteq V$ is *ancestral* if it contains every ancestor of each of its nodes. For an ancestral set S of already selected variables, write $\mathrm{Src}(S) := \{i \notin S : \mathrm{Pa}(i) \subseteq S\}$ for the current sources.

For an ancestral set S and maximum conditioning set size q, let $q_S := \min\{q, |S|\}$ and define the best local Fisher score

$$Q_i^{(q)}(S) := \min_{T \subseteq S,\ |T| = q_S} J_i(T). \tag{4}$$

By monotonicity of local Fisher information under additional conditioning, the minimum over subsets of cardinality q_S in Equation (4) equals the minimum of $J_i(T)$ over all $T \subseteq S$ with $|T| \leq q$. We call the rule that selects a minimizer of $Q_i^{(q)}(S)$ over $i \notin S$ *local Fisher selection*.

Whenever a nonsource remains, define the ordering margin

$$g_q(S) := \min_{i \notin S \cup \mathrm{Src}(S)} Q_i^{(q)}(S) - \min_{s \in \mathrm{Src}(S)} Q_s^{(q)}(S). \tag{5}$$

Differences in noise Fisher information across nodes mean that attainability of a Fisher completion alone does not rank the remaining candidates. The margin $g_q(S)$ measures whether the best attainable score among current sources is separated from that among nonsources.

Theorem 2 (Sparse population DAG recovery)**.** *Consider an ANM in Equation (1) satisfying the conditions of Theorem 1, and suppose that $q \geq d$. Then, at every ancestral set S, the following statements hold.*

(a) ***Ordering.*** *For every $i \notin S$, we have that $Q_i^{(q)}(S) = -I_i^{\varepsilon} + \min_{T \subseteq S,\, |T| = q_S} \Gamma_i(T)$, and that*

$$i \in \mathrm{Src}(S) \quad \Longleftrightarrow \quad Q_i^{(q)}(S) = -I_i^{\varepsilon}. \tag{6}$$

If a nonsource remains and $g_q(S) > 0$, every minimizer of $Q_i^{(q)}(S)$ over $i \notin S$ is a current source.

(b) ***Parent recovery.*** *For every $i \in \mathrm{Src}(S)$ and every $T \subseteq S$ satisfying $|T| \leq q$,*

$$T = \mathrm{Pa}(i) \quad \Longleftrightarrow \quad J_i(T) = Q_i^{(q)}(S) \quad \textit{and} \quad J_i(U) > Q_i^{(q)}(S) \ \textit{ for every } \ U \subsetneq T. \tag{7}$$

Consequently, if $g_q(S) > 0$ at every ancestral set S for which a nonsource remains, repeated application of (a) *and* (b) *recovers the true DAG using marginals involving at most $q + 1$ variables.*

A transparent sufficient case is equal noise Fisher information across nodes; the same conclusion holds when the noise Fisher information does not increase along directed paths (Corollary 2).

3.3 Population Advantages of Locality

Once $q \geq d$, increasing q leaves current-source scores unchanged but can lower nonsource scores by admitting more proxies for unselected parents.

Proposition 1 (Population advantage of locality)**.** *For $d \leq q_1 \leq q_2$ and every ancestral set S,*

$$\begin{aligned} &Q_s^{(q_1)}(S) = Q_s^{(q_2)}(S) = -I_s^{\varepsilon} \quad \textit{for every } s \in \mathrm{Src}(S), \\ &Q_i^{(q_1)}(S) \geq Q_i^{(q_2)}(S) \qquad\quad \textit{for every } i \notin S \cup \mathrm{Src}(S). \end{aligned} \tag{8}$$

Consequently, $g_{q_1}(S) \geq g_{q_2}(S)$ whenever a nonsource remains, and this inequality can be strict.

Appendix A.3 gives the proof and an example in which $q = d$ yields a positive ordering margin, whereas conditioning on the entire selected set yields a negative margin.

3.4 Comparison with Other Ordering Criteria

We compare local Fisher selection with the CaPS leaf rule and conditional-variance ordering on a growing non-Gaussian family with uniform Fisher separation.

Definition 1 (Uniform Fisher separation)**.** A family of ANMs, each equipped with a subset size q, is *uniformly Fisher-separated* if there exist constants $\Delta_{\mathrm{ord}}, \Delta_{\mathrm{par}} > 0$, common to the entire family, such that, for every model in the family, $g_q(S) \geq \Delta_{\mathrm{ord}}$ for every ancestral set S with $V \setminus (S \cup \mathrm{Src}(S)) \neq \varnothing$, and $\Gamma_i(T) \geq \Delta_{\mathrm{par}}$ for every nonroot node i and every parent-omitting set $T \subseteq \mathrm{Nd}(i)$ with $|T| \leq q$. Here, Δ_{ord} and Δ_{par} are uniform lower bounds for the ordering and parent-omission margins, respectively.

Proposition 2 (Uniform Fisher separation in non-Gaussian models)**.** *There exists a family of non-Gaussian chain ANMs $\{\mathcal{M}_p\}_{p \in 2\mathbb{N}}$ with maximum indegree one that is uniformly Fisher-separated at $q = 1$. Consequently, Theorem 2 guarantees population DAG recovery for every $\mathcal{M}_p$ using marginals of dimension at most two. In contrast, on every $\mathcal{M}_p$, the CaPS leaf-selection rule and the conditional-variance ordering rule in Gao et al. (2022, Algorithm 1) choose a nonleaf and a nonsource, respectively, at their first population steps.*

Appendix A.4 constructs the family and verifies both margin lower bounds. The conditional-variance comparison evaluates the ordering criterion beyond its equal-variance model class.

4 Fisher Information Completion Search (FiCS)

Suppose that we observe n independent copies $X^{(1)}, \ldots, X^{(n)}$ of X. FiCS operationalizes the Fisher parent completion principle of Section 3 by replacing each population score $J_i(T)$ with an estimate $\widehat{J}_i(T)$. We first construct the regularized kernel Stein estimator, then use it for ordering estimation and parent selection, and finally give the computational cost.

4.1 Local Fisher Score Estimation

We estimate local Fisher scores directly from observations using regularized kernel Stein equations (Zhou et al., 2020), extending the density-derivative estimation used by SCORE and CaPS (Rolland et al., 2022; Xu et al., 2024). The recovery guarantee applies to any estimator with sufficiently small uniform error over the local scores. The theory below analyzes this particular construction.

For $W := \{i\} \cup T$, let p_W be the density of X_W, and define $s_{W,i} := \partial_i \log p_W$ and $r_{W,i} := (\partial_i^2 p_W)/p_W$. The factorization $p_W(x_i, x_T) = p(x_i \mid x_T)p_T(x_T)$ yields

$$J_i(T) = \mathbb{E}\big[\partial_i^2 \log p_W(X_W)\big] = \mathbb{E}\big[r_{W,i}(X_W) - s_{W,i}(X_W)^2\big]. \tag{9}$$

We estimate the two derivative ratios with separate Stein equations. For observation $a \in [n]$, write $z_a = X_W^{(a)}$. Given a kernel k_W, define

$$(K_W)_{ab} := k_W(z_a, z_b), \qquad (\mathbf{D}_{W,i}^{(m)})_b := \sum_{a=1}^{n} \partial_i^m k_W(z_a, z_b), \qquad m \in \{1, 2\}, \tag{10}$$

where the derivatives act on the target coordinate of the first kernel argument. The regularized kernel Stein estimator is given by

$$\begin{aligned} \widehat{\mathbf{s}}_{W,i} &:= -(K_W + \eta_W I_n)^{-1}\mathbf{D}_{W,i}^{(1)}, \quad \widehat{\mathbf{r}}_{W,i} := (K_W + \eta_W I_n)^{-1}\mathbf{D}_{W,i}^{(2)}, \\ \widehat{J}_i(T) &:= \frac{1}{n}\mathbf{1}^\top \widehat{\mathbf{r}}_{W,i} - \frac{1}{n}\|\widehat{\mathbf{s}}_{W,i}\|_2^2. \end{aligned} \tag{11}$$

Here I_n is the $n \times n$ identity matrix, $\mathbf{1}$ is the vector of n ones, and $\eta_W > 0$ is the matrix ridge that regularizes the matrix inversion. The two Stein systems share the same kernel matrix and can be solved together. Appendix A.7 derives Equation (11), while Appendix E.1 specifies the kernels, bandwidths, and numerical safeguards used in the experiments.

4.2 FiCS Algorithm

Ordering estimation. For the current selected set S and q_S defined in Section 3.2, FiCS selects the next node by empirical local Fisher selection:

$$\widehat{i} \in \operatorname*{arg\,min}_{i \in V \setminus S} \min_{T \subseteq S,\ |T| = q_S} \widehat{J}_i(T). \tag{12}$$

At the population level, monotonicity of local Fisher information makes minimization over $|T| = q_S$ equivalent to minimization over $|T| \le q$. Estimated scores need not preserve this monotonicity, so Equation (12) uses the same cardinality q_S for every candidate. The guarantees in Section 5 require $q \ge d$ and positive lower bounds on the ordering and parent-omission margins; when d is unknown, q specifies the conditioning-size bound.

Parent selection. Given an estimated ordering $\widehat{\pi}$, let $\widehat{S}_i := \{j \in V : \widehat{\pi}(j) < \widehat{\pi}(i)\}$ denote the predecessors of node i. FiCS estimates its minimal Fisher completion by

$$\widehat{P}_i \in \operatorname*{arg\,min}_{T \subseteq \widehat{S}_i,\ |T| \le q} \Big\{ \widehat{J}_i(T) + \lambda |T| \Big\}, \qquad \lambda > 0. \tag{13}$$

Parent selection compares different subset sizes. Given a correct topological order, every parent-set superset has the same population score, so the parent penalty discards redundant predecessors; a parent-omitting set, by contrast, has a positive parent-omission margin. Section 5 specifies conditions under which λ separates these two cases. Restricting parent candidates to ordered predecessors makes the output acyclic, and Algorithm 1 summarizes both stages.

Algorithm 1 Fisher Information Completion Search (FiCS)

Input: Data $\mathbf{X} \in \mathbb{R}^{n \times p}$, conditioning size q, parent penalty λ, and local scores $\widehat{J}_i(T)$.

Stage 1: Ordering estimation

$S \leftarrow \varnothing$
while $|S| < p$ **do**
  $q_S \leftarrow \min\{q, |S|\}$
  $\widehat{i} \leftarrow$ solution of Eq. (12)
  $\widehat{\pi}(\widehat{i}) \leftarrow |S| + 1$
  $S \leftarrow S \cup \{\widehat{i}\}$
end while

Stage 2: Parent selection

$\widehat{E} \leftarrow \varnothing$
for $i \in V$ **do**
  $\widehat{S}_i \leftarrow \{j : \widehat{\pi}(j) < \widehat{\pi}(i)\}$
  $\widehat{P}_i \leftarrow$ solution of Eq. (13)
  $\widehat{E} \leftarrow \widehat{E} \cup (\widehat{P}_i \times \{i\})$
end for

Output: Ordering $\widehat{\pi}$ and DAG $\widehat{G} = (V, \widehat{E})$.

Population local Fisher scores are coordinate-scale dependent: replacing X_i by $a_i X_i$, with $a_i \neq 0$, multiplies $J_i(T)$ by a_i^{-2}. The reported FiCS fits mean-center variables without variance standardization; synthetic and real-data preprocessing are specified in Appendices F.1 and F.7, respectively.

4.3 Computational Complexity

Each queried marginal has at most $q + 1$ variables, while the size of the local query class is

$$N_{p,q} := p \sum_{r=0}^{q} \binom{p-1}{r} \leq p^{q+1} = O(p^{q+1}). \tag{14}$$

When several local queries share the same marginal W, their Stein systems use the same kernel matrix, so one factorization serves all target coordinates. Thus, at most $\sum_{k=1}^{q+1} \binom{p}{k}$ distinct factorizations are required. Appendix E.1 describes the incremental update used to avoid repeated subset comparisons.

A dense factorization costs $O(n^3)$ time and $O(n^2)$ memory. Locality bounds the estimation dimension; exhaustive subset search remains combinatorial in q.

5 Statistical Guarantees of FiCS

FiCS recovers the DAG when the uniform local-score error is sufficiently small relative to the ordering and parent-omission margins and the parent penalty. We first derive this recovery condition, then bound the uniform local-score error of the regularized kernel Stein estimator and combine the two results. Throughout, p, d, q, the parent penalty λ, and the ridge parameter ρ may depend on n, with this dependence suppressed.

Define the local query class $\mathcal{Q}_{p,q}$, its size $N_{p,q}$, and the logarithmic complexity term L by

$$\mathcal{Q}_{p,q} := \{(i, T) : i \in V,\ T \subseteq V \setminus \{i\},\ |T| \leq q\}, \qquad N_{p,q} := |\mathcal{Q}_{p,q}|,$$
$$L := q\left\{1 + \log\left(\frac{p}{q}\right)\right\} + \log p, \qquad \log N_{p,q} = O(L),$$

where $q\{1 + \log(p/q)\} := 0$ when $q = 0$; the last relation follows from the standard binomial-sum bound applied to Equation (14).

5.1 DAG Recovery under a Uniform Local Score Error Bound

For $\epsilon > 0$, define the uniform score-error event

$$\mathcal{E}(\epsilon) := \left\{ \max_{(i,T) \in \mathcal{Q}_{p,q}} \left| \widehat{J}_i(T) - J_i(T) \right| \leq \epsilon \right\}. \tag{15}$$

Theorem 3 (DAG recovery from a uniform local score error bound)**.** *Consider an ANM satisfying the conditions of Theorem 1, and let $d \leq q \leq p - 1$. Suppose $\Delta_{\mathrm{ord}}, \Delta_{\mathrm{par}} > 0$ are lower bounds on its ordering and parent-omission margins, respectively. On $\mathcal{E}(\epsilon)$, Algorithm 1 returns a topological ordering and the true DAG whenever*

$$2\epsilon < \min\{\Delta_{\mathrm{ord}}, \lambda, \Delta_{\mathrm{par}} - d\lambda\}. \tag{16}$$

The three terms in Equation (16) preserve source selection, exclude redundant parent supersets, and exclude parent-omitting sets, respectively.

5.2 Uniform Stein Estimation

The following condition allows $p > n$ while requiring the logarithmic complexity of the local query class to remain smaller than the sample size.

Assumption 1 (High-dimensional setting)**.** *The maximum indegree and subset size satisfy*

$$\max\{1, d\} \le q \le p - 1, \qquad L = o(n).$$

For $W = \{i\} \cup T$, let P_W be the law of X_W and let $\mathcal{H}_W$ be the reproducing kernel Hilbert space (RKHS) of k_W. Define the integral operator $\mathsf{T}_W : L^2(P_W) \to L^2(P_W)$ and the uniform approximation error by

$$\begin{aligned}(\mathsf{T}_W h)(x) &:= \mathbb{E}\{k_W(x, X_W)h(X_W)\},\\ A(\rho) &:= \sup_{n\ge 1} \max_{(i,T)\in\mathcal{Q}_{p,q}} \max_{u\in\{s,r\}} \left\|\rho(\mathsf{T}_W + \rho I)^{-1} u_{W,i}\right\|_{L^2(P_W)}.\end{aligned} \tag{17}$$

Assumption 2 (Uniform local L^2 Stein regularity)**.** *Uniformly over the model sequence and all $(i, T) \in \mathcal{Q}_{p,q}$, each k_W is deterministic, positive semidefinite, and C^4, with*

$$\sup_x \|\partial_i^m k_W(x, \cdot)\|_{\mathcal{H}_W} \le \kappa_m, \qquad m = 0, 1, 2.$$

The two Stein targets have $L^2(P_W)$ norms bounded by a common constant $M < \infty$, the corresponding first- and second-order integration-by-parts identities hold, and $A(\rho) \to 0$ as $\rho \downarrow 0$.

Assumption 2 extends regularized kernel Stein estimation (Zhou et al., 2020, Section 3.1 and Appendix B) and L^2 Tikhonov analysis (Smale & Zhou, 2007, Sections 2 and 6) uniformly over the local queries; it is an estimation condition, not a consequence of bounded indegree. Appendix A.6 gives its operator formulation. Uniform Fisher separation is imposed separately through Definition 1; Propositions 2 and 4 provide non-Gaussian and Gaussian examples.

For $\delta \in (0, 1)$, define

$$\ell_\delta := \log\left(\frac{16 N_{p,q}}{\delta}\right), \qquad b_\delta := C_B\left\{\sqrt{\frac{\ell_\delta}{n}} + \frac{\ell_\delta}{n}\right\}, \qquad e_\delta(\rho) := A(\rho) + \frac{b_\delta}{\rho},$$

where C_B depends only on the uniform bounds in Assumption 2.

Theorem 4 (Uniform local score error bound)**.** *Under Assumption 2, set the matrix ridge in Equation (11) to $\eta_W = n\rho$ for every queried W. For every $\delta \in (0, 1)$ and every deterministic $\rho \in [b_\delta, 1]$, with probability at least $1 - \delta$,*

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} \left|\widehat{J}_i(T) - J_i(T)\right| \le C\left\{e_\delta(\rho) + e_\delta(\rho)^2\right\}, \tag{18}$$

where C is independent of n, p, q, δ, and ρ.

The bound jointly controls the two Stein targets and their same-sample evaluation in the local Fisher functional. Its dependence on p and q enters through $\log N_{p,q} = O(L)$.

5.3 High-Dimensional Consistency

Assumption 1 gives $b := \sqrt{L/n} + L/n \to 0$. Choose a deterministic sequence $\rho \downarrow 0$ such that $b/\rho \to 0$, use the matrix ridge $\eta = n\rho$, and define

$$e := A(\rho) + \frac{b}{\rho}, \qquad R := e + e^2, \qquad m := \min\left\{\Delta_{\mathrm{ord}}, \lambda, \Delta_{\mathrm{par}} - d\lambda\right\}.$$

Corollary 1 (High-dimensional DAG consistency)**.** *Under Assumptions 1 and 2, suppose that the ANM sequence is uniformly Fisher-separated and that every model satisfies the conditions of Theorem 1. If $m > 0$ eventually and $R = o(m)$, then the output $\widehat{G}$ of Algorithm 1 satisfies $\mathbb{P}(\widehat{G} = G) \to 1$.*

Theorem 4 gives uniform local score error $O_{\mathbb{P}}(R)$, so $R = o(m)$ allows Theorem 3 to apply with probability tending to one. A sufficient parent penalty condition is $R = o(\lambda)$ and $d\lambda \le (1 - c_0)\Delta_{\mathrm{par}}$ for some fixed $c_0 \in (0, 1)$. Thus q controls L, whereas d enters the parent penalty constraint; when d grows, feasibility requires $dR \to 0$.

Proofs appear in Appendices A.5 and A.7. Appendix B gives explicit rates and balanced ridge tuning under a stronger RKHS source condition. Empirical bandwidth and synthetic calibration rules are specified in Appendices E.1 and E.3, respectively.

6 Numerical Experiments

We test two practical consequences of bounded conditioning: end-to-end recovery when p is large relative to n, and sensitivity to the conditioning size. We report directed-edge F1 and ordering accuracy A_{top}, the fraction of true edges consistent with the estimated order.

We generate 30 Erdős–Rényi (ER3) DAGs for each $p \in \{50, 100, 200\}$, with expected mean total degree three before capping the indegree at two and randomly permuting variable labels. Each nonroot mechanism is linear with probability 0.5 and otherwise a Gaussian-process function of its parents; errors are independent $N(0, 1)$. We use nested samples with $n \in \{50, 100, 200, 500, 1000\}$. We compare FiCS ($q = 2$) with CaPS (Xu et al., 2024), SCORE (Rolland et al., 2022), NoGAM (Montagna et al., 2023b), and nonlinear DAGMA (Bello et al., 2022). Each method uses one configuration selected on independent calibration DAGs and fixed across test settings. Appendices E–G provide implementation details, experimental designs, and additional results. CaPS can fail numerically during GAM pruning. Figures 2 and 3 include only valid DAG estimates; Appendix F.5 reports the failure counts.

6.1 Recovery on Random DAGs

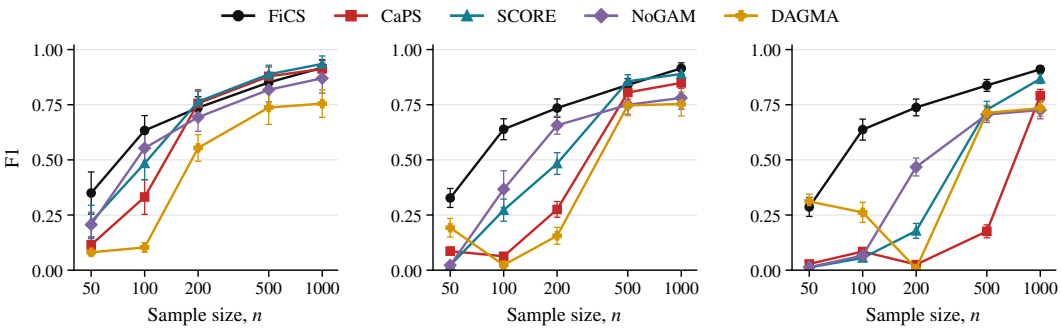


Figure 2: Mean graph-recovery F1 versus sample size for $p = 50$, 100, and 200 from left to right; error bars indicate one standard deviation.

Figure 2 shows FiCS's largest relative gains in the considered small-sample, high-dimensional settings. At $p = 200$ and $n = 100$, FiCS attains mean F1 0.637, compared with 0.262 for DAGMA, 0.055 for SCORE, and 0.066 for NoGAM. Hence, FiCS recovers substantial graph structure even when $p > n$, supporting the benefit of bounding the dimension of local conditioning. Across the three dimensions, the performance gap generally narrows as n increases, suggesting that FiCS's main empirical advantage in these experiments lies in sample-efficient recovery when data are limited rather than a performance advantage that persists across sample sizes.

6.2 Sensitivity to the Conditioning Size

This ablation examines the effect of setting the conditioning size for ordering below, at, or above the maximum indegree. At $n = 100$ and $p \in \{50, 100, 150, 200\}$, we vary only FiCS's conditioning size for ordering, using $q \in \{1, 2, 3\}$ or the entire selected set, while fixing the parent-selection subset size at two and all other components. Because the design caps indegree at two, $q = 1$ can omit a parent at two-parent nodes and therefore falls outside the uniform DAG recovery guarantee, whereas $q \geq 2$ satisfies the indegree condition. CaPS is shown as an external reference.

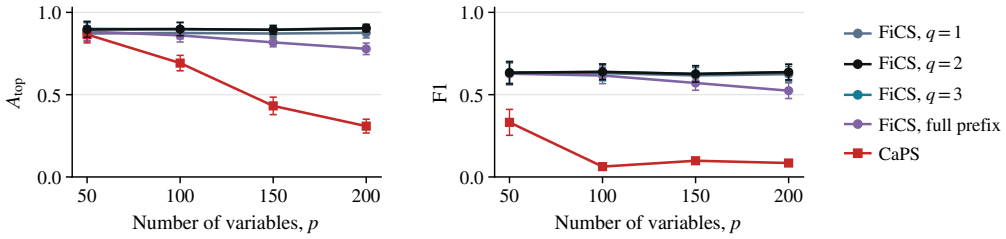


Figure 3: Mean ordering accuracy (left) and graph-recovery F1 (right) versus dimension at $n = 100$.

Figure 3 reveals asymmetric effects of the conditioning size. At $p = 200$, $q = 1$ attains $A_{\text{top}} = 0.876$ and F1 0.625, only modestly below 0.904 and 0.637 for $q = 2$. This small empirical loss is consistent with $q = 1$ remaining adequate for roots and one-parent nodes, but it does not restore the DAG recovery guarantee. In contrast, conditioning on the entire selected set yields lower values of 0.779 and 0.524, while $q = 3$ performs similarly to $q = 2$. The performance difference widens over the displayed dimensions, consistent with the increasing estimation burden of conditioning on the entire selected set. Because its dimension grows with p whereas fixed-q conditioning remains bounded, this trend suggests a larger disadvantage at still higher dimensions, although those settings are not simulated. Proposition 1 shows that locality can also improve the population ordering margin for $q \geq d$.

Appendices G.1–G.2 assess sensitivity to graph structure, nonlinear mechanisms, noise distributions, and the linear-mechanism proportion. Appendix G.4 compares population and finite-sample ordering with VarSort and R^2Sort across six Gaussian generator regimes.

7 Real-Data Benchmarks

We evaluate whether FiCS ordering transfers beyond synthetic designs and supports reference-graph recovery. Table 1 compares Danube (Gnecco et al., 2021), Sachs (Sachs et al., 2005), and 30 CausalRivers subnetworks (Stein et al., 2025), using one configuration per method across datasets and $q = 2$ for FiCS. We compare the same five methods as in Section 6. Because the river data are temporally dependent, these experiments assess reference-graph agreement outside the i.i.d. ANM setting.

Table 1: Reference-graph recovery on Danube ($n = 500, p = 12$), Sachs ($n = 853, p = 11$), and CausalRivers ($n = 500, p = 5$). Parentheses give SD over 30 subgraphs. Bold marks the best reported mean per metric; – denotes no separate ordering stage.

	Danube			Sachs			CausalRivers		
Method	$A_{\text{top}}\uparrow$	F1↑	SHD↓	$A_{\text{top}}\uparrow$	F1↑	SHD↓	$A_{\text{top}}\uparrow$	F1↑	SHD↓
FiCS	**1.000**	**0.452**	17	**0.529**	**0.538**	**11**	**0.975** (0.076)	**0.563** (0.165)	4.53 (1.63)
CaPS	0.909	0.351	36	**0.529**	0.500	**11**	0.733 (0.314)	0.408 (0.166)	6.03 (1.43)
SCORE	**1.000**	0.400	33	**0.529**	0.444	12	0.717 (0.292)	0.401 (0.175)	6.20 (1.40)
NoGAM	0.727	0.225	52	**0.529**	0.452	14	0.742 (0.304)	0.416 (0.188)	6.83 (1.66)
DAGMA	–	0.000	**11**	–	0.000	17	–	0.286 (0.337)	**3.53** (1.25)

Table 1 shows that FiCS achieves the highest F1 on Danube and Sachs and the best or tied-best ordering accuracy across all three benchmarks. On Danube, FiCS orders all 11 reference edges forward but selects seven true positives and 13 false positives, showing that correct ordering does not ensure accurate parent selection. DAGMA's lower SHD on Danube and CausalRivers accompanies empty or near-empty graphs with lower F1, so SHD should be interpreted alongside edge recovery.

Appendix F.7 describes the real-data benchmarks, and Appendix G.3 reports the sensitivity analysis. Overall, while local Fisher information is especially advantageous in sparse, high-dimensional settings, these results suggest that it also supports effective recovery on low-dimensional benchmarks.

8 Discussion

FiCS shows that sparse causal discovery can exploit distributional information without estimating full-dimensional scores. Fisher parent completion characterizes each parent set as the unique minimal Fisher completion, enabling one bounded-dimensional score to drive both source-first ordering and parent recovery. The theory and experiments further show that locality is not merely computational: it can improve the population ordering margin and finite-sample accuracy.

Its DAG recovery guarantee requires $q \geq d$; the method remains sensitive to coordinate scaling and relies on subset search and dense kernel solves. The estimator-agnostic recovery result suggests a clear path forward: combine screening that preserves minimal Fisher completions with faster local score estimators and data-driven, scale-aware tuning. More broadly, Fisher completions may provide a reusable principle for distribution-sensitive causal discovery beyond additive noise models.

REPRODUCIBILITY STATEMENT

Synthetic tuning and test DAGs are generated independently, and all methods are evaluated on the same datasets within each setting. The appendices provide the assumptions and proofs; implementation and hyperparameter selection, experimental design and evaluation metrics, and additional results are detailed in Appendices E, F, and G, respectively.

AI USE STATEMENT

We used ChatGPT (OpenAI) to assist literature retrieval, manuscript drafting and editing, experiment implementation and analysis, and mathematical derivations and proofs. The authors are responsible for verifying the cited sources, code, results, and arguments, and for all claims and content in the final paper.

REFERENCES

Bryon Aragam, Arash Amini, and Qing Zhou. Globally optimal score-based learning of directed acyclic graphs in high-dimensions. *Advances in Neural Information Processing Systems*, 32, 2019.

Kevin Bello, Bryon Aragam, and Pradeep K. Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Yikang Chen, Xingzhe Sun, and Dehui Du. Causal discovery via quantile partial effect. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.12981.

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.

Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778, 2021. doi: 10.1214/20-AOS2021.

Rebecca J. Herman, Jonas Wahl, Urmi Ninad, and Jakob Runge. Unitless Unrestricted Markov-Consistent SCM Generation: Better benchmark datasets for causal discovery. In Biwei Huang and Mathias Drton (eds.), *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 1506–1531. PMLR, 2025. URL https://proceedings.mlr.press/v275/herman25a.html.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007. URL https://www.jmlr.org/papers/v8/kalisch07a.html.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pp. 752–771, 2023a.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751, 2023b.

Francesco Montagna, Philipp Michael Faller, Patrick Blöbaum, Elke Kirschbaum, and Francesco Locatello. Score matching through the roof: Linear, nonlinear, and latent variables causal discovery. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 552–605. PMLR, 2025. URL https://proceedings.mlr.press/v275/montagna25a.html.

Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018. doi: 10.1214/17-AOS1654.

Sunmin Oh, Seungsu Han, and Gunwoong Park. Optimal estimation of linear non-gaussian structural 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. PMLR, 2025. URL https://proceedings.mlr.press/v258/oh25a.html.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Gunwoong Park. Comparison of identifiability conditions in linear additive noise models. *Journal of the Korean Statistical Society*, 2026. doi: 10.1007/s42952-026-00379-x.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. doi: 10.1214/aop/1176988477.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753, 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007. doi: 10.1007/s00365-006-0659-y.

A. J. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959. doi: 10.1016/S0019-9958(59)90348-1.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers – scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wmV4cIbgl6.

Haixiang Sun, Pengchao Tian, Zihan Zhou, and Andrew L. Liu. Hierarchical causal structure discovery from layered topological orderings. *International Journal of Data Science and Analytics*, 22:243, 2026. doi: 10.1007/s41060-026-01211-9.

Rui Wu and Hong Xie. Optimization-free topological sort for causal discovery via the Schur complement of score Jacobians, 2026. URL https://arxiv.org/abs/2604.25295.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0141.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11513–11522. PMLR, 2020.

A Proofs of the Main Results

This section proves the population and statistical results in the order in which they appear in the main text. Appendix A.7 also proves the query-uniform concentration lemma used by both the main L^2 analysis and Appendix B. Throughout, the notation agrees with Sections 3–5; in particular, $g_q(S)$ and $\Gamma_i(T)$ denote the ordering margin and parent-omission margin, respectively.

A.1 Proof of Theorem 1

For each local query (i, T), assume that $p(x_i \mid x_T)$ is twice continuously differentiable in x_i, its conditional score is square-integrable, and differentiation may pass through the conditional integrals below. Assume that the first and second derivatives are integrable where used and that the boundary terms in integration by parts vanish. Then $\mathbb{E}[\partial^2_{x_i} \log p(X_i \mid X_T)] = -\mathbb{E}[(\partial_{x_i} \log p(X_i \mid X_T))^2]$. These are the score-projection and integration-by-parts conditions used in Section 3.

Proof of Theorem 1. Fix i and $T \subseteq \mathrm{Nd}(i)$, put $Z := X_T$, $M := f_i(X_{\mathrm{Pa}(i)})$, and write p_ε and s_ε for the density and score of ε_i. Since T contains no descendant of i, ε_i is independent of (M, Z). Let μ_z be a regular conditional distribution of M given $Z = z$. Then

$$p(x \mid z) = \int p_\varepsilon(x - m)\, \mu_z(dm).$$

Differentiating under the integral and applying Bayes' formula gives, for almost every (x, z),

$$\partial_x \log p(x \mid z) = \int s_\varepsilon(x - m)\, \pi_{x,z}(dm), \qquad \pi_{x,z}(dm) := \frac{p_\varepsilon(x - m)\mu_z(dm)}{\int p_\varepsilon(x - v)\mu_z(dv)}.$$

Because p_ε is everywhere positive, $\pi_{x,z}$ and μ_z are equivalent measures. Evaluating at $x = X_i = M + \varepsilon_i$ yields

$$\partial_{x_i} \log p(X_i \mid Z) = \mathbb{E}\{s_\varepsilon(\varepsilon_i) \mid X_i, Z\}.$$

The conditional variance decomposition therefore gives

$$\Gamma_i(T) = I_i^\varepsilon - I_i(T) = \mathbb{E}[\mathrm{Var}\{s_\varepsilon(\varepsilon_i) \mid X_i, Z\}] \geq 0,$$

with equality if and only if $s_\varepsilon(\varepsilon_i)$ is $\sigma(X_i, Z)$-measurable.

We next characterize the equality case. If M is $\sigma(Z)$-measurable, then $\varepsilon_i = X_i - M$ is $\sigma(X_i, Z)$-measurable, so $\Gamma_i(T) = 0$. Conversely, suppose that $\Gamma_i(T) = 0$. For P_Z-almost every z, the conditional variance under $\pi_{x,z}$ is zero for $P_{X_i|Z=z}$-almost every x. The mixture density $p(x \mid z)$ is everywhere positive, so this holds for Lebesgue-almost every x. Equivalence of $\pi_{x,z}$ and μ_z, followed by Fubini's theorem, gives

$$s_\varepsilon(x - m) = s_\varepsilon(x - m')$$

for $\mu_z \otimes \mu_z$-almost every (m, m') and Lebesgue-almost every x. If μ_z were not a point mass, two independent draws from μ_z would differ with positive probability; hence there would exist $m \neq m'$ for which the displayed equality holds for almost every x. With $t := m - m' \neq 0$, this implies $s_\varepsilon(y + t) = s_\varepsilon(y)$ almost everywhere. Since p_ε is positive and C^1, the function $y \mapsto \log p_\varepsilon(y + t) - \log p_\varepsilon(y)$ is continuously differentiable with derivative zero almost everywhere and hence everywhere. It is therefore constant, so $p_\varepsilon(y + t) = C p_\varepsilon(y)$ for some $C > 0$. Integrating both sides over y forces $C = 1$, which would make the positive density p_ε periodic with nonzero period and therefore nonintegrable. This contradiction shows that μ_z is a point mass for P_Z-almost every z. Because M is real-valued and regular conditional laws may be chosen as measurable probability kernels, this is equivalent to the existence of a measurable m_0 such that $M = m_0(Z)$ almost surely. Thus

$$\Gamma_i(T) = 0 \quad \Longleftrightarrow \quad M \text{ is } \sigma(Z)\text{-measurable.}$$

It remains to determine when the structural shift is measurable from X_T. If $\mathrm{Pa}(i) \subseteq T$, then M is directly determined by X_T. Conversely, suppose that T omits a parent and write

$$A := \mathrm{Pa}(i) \cap T, \qquad B := \mathrm{Pa}(i) \setminus T, \qquad C := T \setminus \mathrm{Pa}(i).$$

Then $B \neq \varnothing$, $X_{\mathrm{Pa}(i)} = (X_A, X_B)$, and $X_T = (X_A, X_C)$. If M were $\sigma(X_T)$-measurable, there would exist a measurable function h such that

$$f_i(X_A, X_B) = h(X_A, X_C) \quad \text{a.s.}$$

The positive noise densities imply that (X_A, X_B, X_C) has an everywhere positive joint density, so this equality holds for Lebesgue-almost every (a, b, c). By Fubini's theorem, for almost every a and almost every c, the equality holds for almost every b. Fix such an a and choose one c from the corresponding full-measure set. Then $f_i(a, b) = h(a, c)$ for almost every b, and continuity in b makes $f_i(a, b)$ constant for every b. The full-measure set of such a is dense. Hence, for arbitrary a and any b, b', choose a sequence of good points $a_k \to a$; continuity gives $f_i(a, b) = \lim_k f_i(a_k, b) = \lim_k f_i(a_k, b') = f_i(a, b')$. Thus f_i is independent of all coordinates in B, contradicting its nonconstant dependence on every parent coordinate. Hence

$$\Gamma_i(T) = 0 \quad \Longleftrightarrow \quad \mathrm{Pa}(i) \subseteq T.$$

The Fisher completions are therefore precisely the nondescendant supersets of $\mathrm{Pa}(i)$, whose unique inclusion-minimal and unique minimum-cardinality member is $\mathrm{Pa}(i)$. □

A.2 Proof of Theorem 2 and Additional Corollary

Proof of Theorem 2. Fix an ancestral set S. Because every $T \subseteq S$ contains only nondescendants of an unselected node i, Theorem 1 gives $J_i(T) = -I_i^\varepsilon + \Gamma_i(T)$ and hence

$$Q_i^{(q)}(S) = -I_i^\varepsilon + \min_{T \subseteq S,\, |T| = q_S} \Gamma_i(T).$$

If $i \in \mathrm{Src}(S)$, then $\mathrm{Pa}(i) \subseteq S$ and $|\mathrm{Pa}(i)| \le d \le q$. Hence $\mathrm{Pa}(i)$ can be extended to a set $T \subseteq S$ with $|T| = q_S$, and Theorem 1 gives $Q_i^{(q)}(S) = -I_i^\varepsilon$. Conversely, if $Q_i^{(q)}(S) = -I_i^\varepsilon$, some admissible $T \subseteq S$ satisfies $\Gamma_i(T) = 0$, so Theorem 1 gives $\mathrm{Pa}(i) \subseteq T \subseteq S$. Therefore

$$i \in \mathrm{Src}(S) \quad \Longleftrightarrow \quad Q_i^{(q)}(S) = -I_i^\varepsilon.$$

If a nonsource remains and $g_q(S) > 0$, the definition of the ordering margin gives

$$\min_{i \notin S \cup \mathrm{Src}(S)} Q_i^{(q)}(S) > \min_{s \in \mathrm{Src}(S)} Q_s^{(q)}(S),$$

so every minimizer over $i \notin S$ is a current source.

Now fix $i \in \mathrm{Src}(S)$. The source characterization gives $Q_i^{(q)}(S) = -I_i^\varepsilon$, while Theorem 1 gives, for every $T \subseteq S$ with $|T| \le q$,

$$J_i(T) = Q_i^{(q)}(S) \quad \Longleftrightarrow \quad \Gamma_i(T) = 0 \quad \Longleftrightarrow \quad \mathrm{Pa}(i) \subseteq T.$$

If $T = \mathrm{Pa}(i)$, then $J_i(T) = Q_i^{(q)}(S)$, while every proper subset $U \subsetneq T$ omits a parent and satisfies $J_i(U) > Q_i^{(q)}(S)$. Conversely, suppose $J_i(T) = Q_i^{(q)}(S)$ and every proper subset $U \subsetneq T$ has $J_i(U) > Q_i^{(q)}(S)$. The displayed equivalence gives $\mathrm{Pa}(i) \subseteq T$. If the inclusion were strict, $U = \mathrm{Pa}(i)$ would be a proper subset with $J_i(U) = Q_i^{(q)}(S)$, a contradiction. Hence $T = \mathrm{Pa}(i)$, which also proves the two uniqueness claims.

Begin the recursion with $S = \varnothing$, which is ancestral. At any stage, if a nonsource remains, the positive ordering margin forces every minimizer to be a current source. If no nonsource remains, every unselected node is already a current source, so any minimizer is valid. Selecting and adding a current source preserves ancestrality, and the recursion therefore produces a topological ordering. The parent characterization then recovers each selected node's true parent set from its predecessors. Thus the procedure recovers the true DAG, and every queried marginal contains at most one target node and q conditioning variables. □

Theorem 2 reduces population recovery to positivity of the ordering margin, but checking this margin over every ancestral set may be difficult. A transparent sufficient condition can instead be stated directly in terms of the noise Fisher information along causal paths. The following corollary shows that pathwise nonincrease, including equal noise Fisher information as a special case, guarantees a positive ordering margin.

Corollary 2 (A sufficient ordering condition)**.** *Under the conditions of Theorem 1, suppose $q \ge d$ and the noise Fisher information does not increase along any directed path. Then $g_q(S) > 0$ at every ancestral set with a remaining nonsource, and population FiCS recovers the true DAG. Equal noise Fisher information is a special case.*

Proof. Fix an ancestral set S. If no nonsource remains, every unselected node is a current source, and there is nothing to prove. Otherwise, fix a remaining nonsource i. Following incoming edges from i through unselected parents must terminate at a node $s \in \mathrm{Src}(S)$, and s is an ancestor of i. The pathwise nonincrease assumption gives $I_s^\varepsilon \geq I_i^\varepsilon$. Every admissible subset $T \subseteq S$ for node i omits at least one parent, so Theorem 1 gives $\Gamma_i(T) > 0$. Because the admissible subset collection is finite,

$$Q_i^{(q)}(S) > -I_i^\varepsilon \geq -I_s^\varepsilon = Q_s^{(q)}(S) \geq \min_{t \in \mathrm{Src}(S)} Q_t^{(q)}(S).$$

This holds for every remaining nonsource i. Taking the minimum over them yields $g_q(S) > 0$. Theorem 2 therefore gives population DAG recovery. □

A.3 Proof of Proposition 1

Proof of Proposition 1. Let $T \subseteq U \subseteq V \setminus \{i\}$ and define $s_{i,A} := \partial_{x_i} \log p(X_i \mid X_A)$ for $A \in \{T, U\}$. Marginalizing the conditional distribution given X_U over $X_{U \setminus T}$ and differentiating with respect to x_i gives

$$s_{i,T} = \mathbb{E}[s_{i,U} \mid X_i, X_T].$$

The conditional variance decomposition and $J_i(A) = -\mathbb{E}[s_{i,A}^2]$ therefore yield

$$J_i(T) - J_i(U) = \mathbb{E}[s_{i,U}^2] - \mathbb{E}[s_{i,T}^2] = \mathbb{E}[(s_{i,U} - s_{i,T})^2] \geq 0.$$

Every subset of S of size $\min\{q_1, |S|\}$ can be extended to a subset of size $\min\{q_2, |S|\}$. Applying the preceding monotonicity identity gives

$$Q_i^{(q_1)}(S) \geq Q_i^{(q_2)}(S)$$

for every $i \notin S$. If $s \in \mathrm{Src}(S)$, both subset sizes admit a set containing $\mathrm{Pa}(s)$ because $q_1 \geq d$. Theorem 1 therefore gives

$$Q_s^{(q_1)}(S) = Q_s^{(q_2)}(S) = -I_s^\varepsilon.$$

Subtracting the common minimum source score from the minimum nonsource score gives $g_{q_1}(S) \geq g_{q_2}(S)$ whenever a nonsource remains.

Strictness. The nonsource score $Q_i^{(q)}(S)$ can strictly decrease as q increases. Consider

$$X_1 = \varepsilon_1, \qquad X_2 = \varepsilon_2, \qquad X_3 = \varepsilon_3, \qquad X_4 = 3X_1 + 3X_2 + \varepsilon_4, \qquad X_5 = \tfrac{1}{2}X_4 + 2X_3 + \varepsilon_5,$$

where the errors are independent centered Gaussians with variances $(1, 1, 1, 2, 1/2)$. At the ancestral set $S = \{1, 2, 3\}$, node 4 is the only current source and node 5 is a nonsource. The corresponding population scores are

Subset size	$Q_4^{(q)}(S)$	$Q_5^{(q)}(S)$	$g_q(S)$
$q = 2$	$-1/2$	$-4/13$	$5/26$
$q = 3$	$-1/2$	-1	$-1/2$

Conditioning node 5 on its best two roots leaves conditional variance $13/4$, whereas conditioning on all three roots leaves conditional variance one. Hence $g_2(S) > g_3(S)$, and increasing the subset size can reverse the population ordering decision. □

A.4 Proof of Proposition 2

For a scalar random variable Y with differentiable density, write $I(Y) := \mathbb{E}[(\partial_y \log p_Y(Y))^2]$ for its location Fisher information. For each even p, consider the chain ANM $\mathcal{M}_p$ given by

$$X_1 = \varepsilon_1, \qquad X_j = a_j X_{j-1} + \varepsilon_j, \quad 2 \leq j \leq p,$$

where the errors are mutually independent, ε_j follows the standard Student t_3 distribution for odd j, and $\varepsilon_j \sim N(0, 1)$ for even j. Choose the coefficients to satisfy

$$\frac{13}{40} \leq a_j^2 \leq \frac{33}{100} \quad (j \text{ even}), \qquad \frac{1}{4} \leq |a_j| \leq \frac{1}{3} \quad (j \geq 3 \text{ odd}).$$

The standard Student t_3 distribution has variance 3 and location Fisher information $2/3$, whereas the standard Gaussian distribution has variance and location Fisher information one.

Proof. We verify lower bounds on the ordering and parent-omission margins and then the failures of the competing ordering rules. First, for every nonroot node j and every parent-omitting set $T \subseteq \mathrm{Nd}(j)$ satisfying $|T| \leq 1$, we claim that

$$\Gamma_j(T) \geq \frac{2}{75}.$$

Lower bound on the parent-omission margin. Since $\mathrm{Pa}(j) = \{j-1\}$, every parent-omitting set $T \subseteq \mathrm{Nd}(j)$ contains only variables among $X_1, \ldots, X_{j-2}$. For $j \geq 3$, write

$$X_j = a_j a_{j-1} X_{j-2} + V_j, \qquad V_j := a_j \varepsilon_{j-1} + \varepsilon_j.$$

The random variable V_j is independent of (X_{j-2}, X_T). Applying the conditional-score projection identity from the proof of Theorem 1 to the structural shift $a_j a_{j-1} X_{j-2}$ and noise V_j gives

$$I_j(T) \leq I(V_j).$$

For $j = 2$, a parent-omitting set is empty and $X_2 = V_2$, so the same inequality holds.

Suppose first that j is odd. Then $\varepsilon_{j-1} \sim N(0,1)$ and $I(\varepsilon_j) = 2/3$. Stam's inequality (Stam, 1959) gives

$$\frac{1}{I(V_j)} \geq \frac{1}{I(a_j \varepsilon_{j-1})} + \frac{1}{I(\varepsilon_j)} = a_j^2 + \frac{3}{2}.$$

Since $|a_j| \geq 1/4$,

$$I(V_j) \leq \left(a_j^2 + \frac{3}{2}\right)^{-1} \leq \left(\frac{1}{16} + \frac{3}{2}\right)^{-1} = \frac{16}{25}.$$

Consequently,

$$\Gamma_j(T) = I_j^{\varepsilon} - I_j(T) \geq \frac{2}{3} - \frac{16}{25} = \frac{2}{75}.$$

Now suppose that j is even. Write $\varepsilon_{j-1} = \sqrt{3/U}\, G$, where $U \sim \chi_3^2$ and $G \sim N(0,1)$ are independent. Conditional on U, the variable V_j is Gaussian with variance $1 + c/U$, where $c := 3a_j^2 \geq 39/40$. The marginal score of V_j is the conditional expectation of its Gaussian score given V_j. Conditional Jensen's inequality therefore yields

$$I(V_j) \leq \mathbb{E}\left[\frac{1}{1 + c/U}\right] = \mathbb{E}\left[\frac{U}{U + c}\right] \leq \mathbb{E}\left[\frac{U}{U + 39/40}\right].$$

We next verify a convenient rational upper bound for the last expectation. Let

$$R(u) := 25000 - 8000u + 1050u^2 - 56u^3 + u^4.$$

Cauchy–Schwarz gives

$$\mathbb{E}\left[\frac{1}{U + 39/40}\right] \geq \frac{\{\mathbb{E}[R(U)]\}^2}{\mathbb{E}[(U + 39/40)R(U)^2]}.$$

Using $\mathbb{E}[U^k] = \prod_{\ell=0}^{k}(2\ell + 1)$ for $U \sim \chi_3^2$ gives

$$\mathbb{E}[R(U)] = 11815, \qquad \mathbb{E}[(U + 39/40)R(U)^2] = \frac{3252350883}{8}.$$

It follows that

$$\mathbb{E}\left[\frac{U}{U + 39/40}\right] = 1 - \frac{39}{40}\mathbb{E}\left[\frac{1}{U + 39/40}\right] \leq \frac{721171976}{1084116961} < \frac{2}{3} - \frac{1}{1000}.$$

Therefore,

$$\Gamma_j(T) = 1 - I_j(T) \geq 1 - I(V_j) > \frac{1}{3} + \frac{1}{1000} > \frac{2}{75}.$$

The two parity cases establish $\Delta_{\mathrm{par}} = 2/75$ as a lower bound on the parent-omission margin.

Lower bound on the ordering margin. At every ancestral set S for which a nonsource remains, we have

$$g_1(S) \geq \frac{1}{1000}.$$

An ancestral set in a directed chain consists of an initial segment of the chain. Its unique current source s has $Q_s^{(1)}(S) = -I_s^\varepsilon$. For every remaining nonsource j, every admissible subset omits its parent, so the lower bounds on the parent-omission margin above give

$$Q_j^{(1)}(S) - Q_s^{(1)}(S) = I_s^\varepsilon - I_j^\varepsilon + \min_{T \subseteq S,\ |T| = \min\{1,|S|\}} \Gamma_j(T).$$

If j is odd, the right-hand side is at least $2/75$ because $I_s^\varepsilon \geq 2/3 = I_j^\varepsilon$. If j is even, it is greater than $2/3 - 1 + (1/3 + 1/1000) = 1/1000$. Taking the minimum over nonsources establishes $\Delta_{\mathrm{ord}} = 1/1000$.

Failure of the competing ordering rules. For a nonleaf j, independence and the zero means of the noise scores give

$$\mathbb{E}[(\partial_j \log p(X))^2] = I_j^\varepsilon + a_{j+1}^2 I_{j+1}^\varepsilon.$$

Every odd nonleaf has joint Fisher information at most $2/3 + 33/100 = 299/300 < 1$, whereas the unique leaf p has joint Fisher information one. Thus the first CaPS minimizer is a nonleaf. For LearnDAG, the unique initial source has variance $\mathrm{Var}(X_1) = 3$, whereas

$$\mathrm{Var}(X_2) = 3a_2^2 + 1 \leq 199/100 < 3.$$

Hence X_1 cannot be a first-step minimizer; because it is the unique source, the first selected node is necessarily a nonsource. The two margin lower bounds are independent of p, so the family is uniformly Fisher-separated at $q = 1$. Theorem 2 gives population DAG recovery, and the preceding calculations give the two claimed first-step failures. □

A.5 Proof of Theorem 3

Proof of Theorem 3. On $\mathcal{E}(\epsilon)$, the uniform error

$$U := \max_{(i,T) \in \mathcal{Q}_{p,q}} |\widehat{J}_i(T) - J_i(T)|$$

satisfies $U \leq \epsilon$. For an ancestral set S, define

$$\widehat{Q}_i^{(q)}(S) := \min_{T \subseteq S,\ |T| = q_S} \widehat{J}_i(T).$$

The uniform local score error bound implies

$$|\widehat{Q}_i^{(q)}(S) - Q_i^{(q)}(S)| \leq \epsilon$$

for every $i \notin S$. Hence the empirical difference between the best nonsource score and the best current-source score is at least $\Delta_{\mathrm{ord}} - 2\epsilon > 0$. The ordering rule therefore selects a current source. Adding it to S preserves ancestrality, so Stage 1 recursively returns a topological ordering.

Fix a node i and let S be its predecessors in this order, so $\mathrm{Pa}(i) \subseteq S \subseteq \mathrm{Nd}(i)$. Every parent-containing set has population score $-I_i^\varepsilon$. Thus, for any strict parent-containing superset $T \supsetneq \mathrm{Pa}(i)$,

$$\{\widehat{J}_i(T) + \lambda|T|\} - \{\widehat{J}_i(\mathrm{Pa}(i)) + \lambda|\,\mathrm{Pa}(i)|\} \geq \lambda - 2\epsilon > 0.$$

If T omits a parent, Theorem 1 and the lower bound on the parent-omission margin give $J_i(T) - J_i(\mathrm{Pa}(i)) \geq \Delta_{\mathrm{par}}$. Since $|\,\mathrm{Pa}(i)| \leq d$, the largest possible reduction in the parent penalty is $d\lambda$, and hence

$$\{\widehat{J}_i(T) + \lambda|T|\} - \{\widehat{J}_i(\mathrm{Pa}(i)) + \lambda|\,\mathrm{Pa}(i)|\} \geq \Delta_{\mathrm{par}} - d\lambda - 2\epsilon > 0.$$

Therefore $\mathrm{Pa}(i)$ is the unique penalized minimizer. The same argument applies to every node, so Stage 2 recovers all parent sets and Algorithm 1 returns the true DAG. □

For the covariance-estimation result in Appendix C, write the recovery tolerance of Theorem 3 as

$$\epsilon_{\mathrm{alg}} := \frac{1}{2} \min\{\Delta_{\mathrm{ord}}, \lambda, \Delta_{\mathrm{par}} - d\lambda\}. \tag{19}$$

Uniform local score error strictly below $\epsilon_{\mathrm{alg}} > 0$ ensures DAG recovery by Algorithm 1.

A.6 Detailed Operator Formulation of Assumption 2

We use the notation of Section 5 and introduce the operators needed for the proofs below.

Let $\phi_W(x) := k_W(x, \cdot)$, and let $S_W : \mathcal{H}_W \to L^2(P_W)$ be the canonical inclusion operator. Define

$$L_W := S_W^* S_W = \mathbb{E}\{\phi_W(X_W) \otimes \phi_W(X_W)\},$$

where $\otimes$ denotes the rank-one operator, $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm, and $\|\cdot\|_{\mathrm{op}}$ denotes the induced operator norm. The integral operator T_W in Equation (17) admits the factorization $\mathsf{T}_W = S_W S_W^*$ because, for every $h \in L^2(P_W)$,

$$(S_W S_W^* h)(x) = \mathbb{E}\{k_W(x, X_W)h(X_W)\} = (\mathsf{T}_W h)(x).$$

Thus, the integral and factorized representations give the same operator T_W and the same value of $A(\rho)$. Notice that L_W and T_W are distinct: the former acts on $\mathcal{H}_W$, whereas the latter acts on $L^2(P_W)$.

Define

$$\zeta^s_{W,i} := \mathbb{E}\{\partial_i k_W(X_W, \cdot)\}, \qquad \zeta^r_{W,i} := \mathbb{E}\{\partial_i^2 k_W(X_W, \cdot)\}.$$

The first- and second-order Stein identities in Assumption 2 take the operator form

$$S_W^* s_{W,i} = -\zeta^s_{W,i}, \qquad S_W^* r_{W,i} = \zeta^r_{W,i}. \tag{20}$$

All regularity, target-norm, and approximation bounds are those of Assumption 2, uniformly over the local query class.

A.7 Proof of Theorem 4 and Corollary 1

Write $\rho := \eta/n$, let $\sigma_s := -1$ and $\sigma_r := 1$, and define

$$\mu_W := \mathbb{E}\{\phi_W(X_W)\}, \qquad \widehat{\mu}_W := \frac{1}{n}\sum_{a=1}^n \phi_W(X_W^{(a)}),$$

$$\widehat{L}_W := \frac{1}{n}\sum_{a=1}^n \phi_W(X_W^{(a)}) \otimes \phi_W(X_W^{(a)}), \quad \widehat{\zeta}^s_{W,i} := \frac{1}{n}\sum_{a=1}^n \partial_i k_W(X_W^{(a)}, \cdot),$$

$$\widehat{\zeta}^r_{W,i} := \frac{1}{n}\sum_{a=1}^n \partial_i^2 k_W(X_W^{(a)}, \cdot).$$

For $u \in \{s, r\}$, define the corresponding operator estimator by

$$\widehat{u}_{W,i} := \sigma_u(\widehat{L}_W + \rho I)^{-1}\widehat{\zeta}^u_{W,i}.$$

The representer form of the empirical normal equations gives

$$(K_W + \eta I_n)\widehat{\mathbf{s}}_{W,i} = -\mathbf{D}^{(1)}_{W,i}, \qquad (K_W + \eta I_n)\widehat{\mathbf{r}}_{W,i} = \mathbf{D}^{(2)}_{W,i}.$$

Hence the sample-value vectors of these operator estimators coincide with the matrix estimators in Equation (11).

Lemma 1 (Uniform concentration of the empirical Stein quantities)**.** *Under Assumption 2, with probability at least* $1 - \delta$*,*

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} \left\{ \|\widehat{L}_W - L_W\|_{\mathrm{HS}}, \|\widehat{\mu}_W - \mu_W\|_{\mathcal{H}_W}, \|\widehat{\zeta}^s_{W,i} - \zeta^s_{W,i}\|_{\mathcal{H}_W}, \|\widehat{\zeta}^r_{W,i} - \zeta^r_{W,i}\|_{\mathcal{H}_W} \right\} \le b_\delta.$$

Proof. Fix a local query (i, T) and abbreviate $\phi := \phi_W(X_W)$. Define the centered Hilbert-valued variables

$$Y^L := \phi \otimes \phi - L_W, \qquad Y^\mu := \phi - \mu_W,$$
$$Y^s := \partial_i k_W(X_W, \cdot) - \zeta^s_{W,i}, \qquad Y^r := \partial_i^2 k_W(X_W, \cdot) - \zeta^r_{W,i}.$$

The first variable takes values in the Hilbert space of Hilbert–Schmidt operators, and the remaining three take values in $\mathcal{H}_W$. All four variables are mean zero. Assumption 2, in the detailed notation of Appendix A.6, gives

$$\|Y^L\|_{\mathrm{HS}} \le 2\kappa_0^2, \qquad \|Y^\mu\|_{\mathcal{H}_W} \le 2\kappa_0, \qquad \|Y^s\|_{\mathcal{H}_W} \le 2\kappa_1, \qquad \|Y^r\|_{\mathcal{H}_W} \le 2\kappa_2$$

almost surely, uniformly over all local queries. Indeed, $\|\phi \otimes \phi\|_{\mathrm{HS}} = \|\phi\|^2_{\mathcal{H}_W}$, while the norm of each expectation is bounded by the corresponding expected norm.

The Bernstein inequality for sums of independent, mean-zero Hilbert-space-valued random variables (Pinelis, 1994) therefore implies that, for each of the four empirical averages and every $x > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{a=1}^{n} Y_a\right\| > C_0\left\{\sqrt{\frac{x}{n}} + \frac{x}{n}\right\}\right) \le 2e^{-x},$$

where C_0 depends only on κ_0, κ_1, and κ_2. Apply this inequality with $x = \ell_\delta := \log(16N_{p,q}/\delta)$. A union bound over the four quantities and at most $N_{p,q}$ local queries gives total failure probability at most $8N_{p,q}e^{-\ell_\delta} \le \delta$. Enlarging C_B if necessary makes the common right-hand side equal to $b_\delta = C_B\{\sqrt{\ell_\delta/n} + \ell_\delta/n\}$, proving the lemma. □

Proof of Theorem 4. Work on the event in Lemma 1, and suppose that $b_\delta \le \rho \le 1$.

Population ridge approximation and empirical target error. Fix a local query and $u \in \{s, r\}$, write $g := u_{W,i}$, and define the population-regularized target

$$u_\rho := (L_W + \rho I)^{-1} S_W^* g = \sigma_u (L_W + \rho I)^{-1} \zeta^u_{W,i}.$$

The identities $L_W = S_W^* S_W$ and $\mathsf{T}_W = S_W S_W^*$ give

$$S_W u_\rho = \mathsf{T}_W(\mathsf{T}_W + \rho I)^{-1} g, \qquad \|g - S_W u_\rho\|_{L^2(P_W)} \le A(\rho).$$

Spectral calculus and the uniform target bound in Assumption 2 also give

$$\|u_\rho\|_{\mathcal{H}_W} \le \frac{M}{2\sqrt{\rho}},$$

because the singular values of $(L_W + \rho I)^{-1} S_W^*$ are $\sqrt{t}/(t+\rho)$ and $\sup_{t \ge 0} \sqrt{t}/(t+\rho) = 1/(2\sqrt{\rho})$. Subtracting the empirical and population normal equations yields

$$\widehat{u}_{W,i} - u_\rho = (\widehat{L}_W + \rho I)^{-1}\left[\sigma_u(\widehat{\zeta}^u_{W,i} - \zeta^u_{W,i}) + (L_W - \widehat{L}_W)u_\rho\right].$$

Put $b := b_\delta$. Since $\|\widehat{L}_W - L_W\|_{\mathrm{op}} \le b \le \rho$, we have $L_W \preceq \widehat{L}_W + \rho I$ and hence

$$\|S_W(\widehat{L}_W + \rho I)^{-1}\|_{\mathrm{op}} \le \rho^{-1/2}.$$

Indeed, with $A_W := \widehat{L}_W + \rho I$, the operator inequality $L_W \preceq A_W$ gives $A_W^{-1} L_W A_W^{-1} \preceq A_W^{-1} \preceq \rho^{-1} I$, and the squared norm on the preceding display equals $\|A_W^{-1} L_W A_W^{-1}\|_{\mathrm{op}}$. The concentration lemma and the bound on $\|u_\rho\|_{\mathcal{H}_W}$ therefore imply

$$\|S_W(\widehat{u}_{W,i} - u_\rho)\|_{L^2(P_W)} \le \rho^{-1/2}\left(b + \frac{bM}{2\sqrt{\rho}}\right) \le C\frac{b}{\rho}.$$

Combining the last display with the population approximation bound gives, uniformly over all local queries and $u \in \{s, r\}$,

$$\|S_W \widehat{u}_{W,i} - u_{W,i}\|_{L^2(P_W)} \le A(\rho) + C\frac{b_\delta}{\rho}.$$

Using $\|A_W^{-1}\|_{\mathrm{op}} \le 1/\rho$ in the same inverse identity gives

$$\|\widehat{u}_{W,i} - u_\rho\|_{\mathcal{H}_W} \le \frac{b}{\rho} + \frac{bM}{2\rho^{3/2}} \le 1 + \frac{M}{2\sqrt{\rho}}.$$

Together with the bound on $\|u_\rho\|_{\mathcal{H}_W}$ and $\rho \le 1$, this proves

$$\|\widehat{u}_{W,i}\|_{\mathcal{H}_W} \le \frac{C}{\sqrt{\rho}}.$$

Propagation to the local Fisher scores. Fix a local query and abbreviate $s := s_{W,i}$, $r := r_{W,i}$, $\widehat{s} := \widehat{s}_{W,i}$, and $\widehat{r} := \widehat{r}_{W,i}$. The empirical and population scores satisfy

$$\widehat{J}_i(T) = \langle \widehat{\mu}_W, \widehat{r}\rangle_{\mathcal{H}_W} - \langle \widehat{s}, \widehat{L}_W \widehat{s}\rangle_{\mathcal{H}_W}, \qquad J_i(T) = \mathbb{E}\{r_{W,i}(X_W)\} - \|s_{W,i}\|^2_{L^2(P_W)}.$$

Let $e_u := \|S_W \widehat{u} - u\|_{L^2(P_W)}$ for $u \in \{s, r\}$. Adding and subtracting the corresponding population expectations gives

$$|\widehat{J}_i(T) - J_i(T)| \le b_\delta \|\widehat{r}\|_{\mathcal{H}_W} + e_r + b_\delta \|\widehat{s}\|^2_{\mathcal{H}_W} + e_s(2M + e_s).$$

Here the linear term follows from $|\langle \widehat{\mu}_W - \mu_W, \widehat{r}\rangle| \le b_\delta \|\widehat{r}\|$ and $|\mathbb{E}\{S_W \widehat{r} - r\}| \le e_r$. For the quadratic term,

$$\left| \langle \widehat{s}, \widehat{L}_W \widehat{s}\rangle - \|s\|^2_{L^2(P_W)} \right| \le b_\delta \|\widehat{s}\|^2_{\mathcal{H}_W} + \left| \|S_W \widehat{s}\|^2_{L^2(P_W)} - \|s\|^2_{L^2(P_W)} \right|,$$

and the last absolute value is at most $e_s(2M + e_s)$ because $\|S_W \widehat{s}\|_{L^2(P_W)} \le M + e_s$. The target bounds above imply $e_s, e_r \le C e_\delta(\rho)$ and $\|\widehat{s}\|_{\mathcal{H}_W}, \|\widehat{r}\|_{\mathcal{H}_W} \le C/\sqrt{\rho}$. Since $\rho \le 1$, the empirical-evaluation terms are bounded by $C b_\delta/\rho$, which is absorbed into $C e_\delta(\rho)$ because $b_\delta/\rho \le e_\delta(\rho)$. Taking the maximum over all local queries therefore gives

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} |\widehat{J}_i(T) - J_i(T)| \le C \left\{ e_\delta(\rho) + e_\delta(\rho)^2 \right\}.$$

This is Equation (18). Lemma 1 holds with probability at least $1 - \delta$, completing the proof. □

Proof of Corollary 1. Assumption 1 gives $L = o(n)$ and hence $b \to 0$. Fix any $\delta \in (0, 1)$. Since $\log N_{p,q} = O(L)$ and $\log(1/\delta)$ is constant, $b_\delta \le C_\delta b$ for all sufficiently large n. Moreover, $b/\rho \to 0$ implies $b_\delta \le \rho$ eventually, so Theorem 4 is applicable at this fixed confidence level. The definitions of e and R therefore give a constant $M_\delta < \infty$ such that $\mathbb{P}(U > M_\delta R) \le \delta$ eventually, where

$$U := \max_{(i,T)\in\mathcal{Q}_{p,q}} |\widehat{J}_i(T) - J_i(T)| = O_{\mathbb{P}}(R).$$

Because δ is arbitrary, this is precisely $U = O_{\mathbb{P}}(R)$. Since $R = o(m)$, Slutsky's theorem gives $U/m = (U/R)(R/m) = o_{\mathbb{P}}(1)$ and therefore $\mathbb{P}(2U < m) \to 1$. On the event $2U < m$, Theorem 3 applies with $\epsilon = U$, so Algorithm 1 returns the true DAG. Hence $\mathbb{P}(\widehat{G} = G) \to 1$. □

B RKHS Source Analysis for Regularized Kernel Stein Estimation

The main analysis requires only uniform L^2 Stein regularity and approximation, allowing the first- and second-order Stein targets to lie outside their kernel RKHSs. This appendix replaces the abstract approximation condition with the stronger requirement that both targets belong uniformly to source spaces of the RKHS covariance operators. In return, it yields convergence in RKHS norm and an explicit regularization-bias rate. These stronger conclusions are not needed for the main DAG recovery result.

B.1 Uniform RKHS source regularity and strengthened rates

For $u \in \{s, r\}$, let $\beta_u \in (0, 1]$ denote its source exponent and define $\beta_* := \min\{\beta_s, \beta_r\}$. For a positive self-adjoint operator L_W, let $L_W^{\beta_u}$ denote its spectral power.

Assumption 3 (Uniform local RKHS source regularity)**.** *Retain the uniform kernel regularity, target-norm bounds, and Stein identities of Assumption 2. For each $u \in \{s, r\}$, suppose that the $L^2(P_W)$ target $u_{W,i}$ admits an RKHS representative, denoted by the same symbol, and that there exist a fixed exponent $\beta_u \in (0, 1]$ and an element $f^u_{W,i} \in \mathcal{H}_W$ such that*

$$u_{W,i} = L_W^{\beta_u} f^u_{W,i}, \qquad \sup_{n\ge 1} \max_{(i,T)\in\mathcal{Q}_{p,q}} \max_{u\in\{s,r\}} \|f^u_{W,i}\|_{\mathcal{H}_W} \le F$$

for some $F < \infty$.

The source condition implies the uniform approximation requirement $A(\rho) \to 0$ in Assumption 2. Indeed, the intertwining identity $(\mathsf{T}_W + \rho I)^{-1} S_W = S_W (L_W + \rho I)^{-1}$ implies $A(\rho) \le C\rho^{\beta_*}$, so the original approximation condition follows automatically. Define

$$\omega_\delta(\rho) := \rho^{\beta_*} + \frac{b_\delta}{\rho}.$$

Theorem 5 (Uniform RKHS error bounds and explicit source rate)**.** *Under Assumption 3, set the matrix ridge in Equation (11) to $\eta_W = n\rho$ for every queried W. For every $\delta \in (0,1)$ and $b_\delta \le \rho \le 1$, with probability at least $1-\delta$,*

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} \|\widehat{u}_{W,i} - u_{W,i}\|_{\mathcal{H}_W} \le C\left\{\rho^{\beta_u} + \frac{b_\delta}{\rho}\right\}, \qquad u \in \{s,r\},$$

and

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} \left|\widehat{J}_i(T) - J_i(T)\right| \le C\left\{\omega_\delta(\rho) + \omega_\delta(\rho)^2\right\}. \tag{21}$$

Here C is independent of n, p, q, δ, and ρ.

When $b_\delta \to 0$, balancing ρ^{β_*} and b_δ/ρ gives $\rho \asymp b_\delta^{1/(1+\beta_*)}$ and the leading rate $b_\delta^{\beta_*/(1+\beta_*)}$. Thus, Theorem 5 sharpens Theorem 4 when the abstract approximation condition is replaced by the stronger RKHS source condition. In the asymptotic notation of Corollary 1, put $R_{\mathrm{src}} := b^{\beta_*/(1+\beta_*)}$. If $d \ge 1$ and $dR_{\mathrm{src}} \to 0$, the balanced ridge parameter $\rho \asymp b^{1/(1+\beta_*)}$ and, for example, $\lambda := \sqrt{R_{\mathrm{src}}/d}$ satisfy $R_{\mathrm{src}} = o(\lambda)$ and $d\lambda \to 0$. If $d = 0$, taking $\lambda := \sqrt{R_{\mathrm{src}}}$ gives $R_{\mathrm{src}} = o(\lambda)$ and $\lambda \to 0$, so redundant nonempty parent sets are excluded. Hence the source condition supplies compatible ridge-parameter and parent penalty sequences whenever $dR_{\mathrm{src}} \to 0$.

B.2 RKHS STABILITY UNDER THE SOURCE CONDITION

For $u \in \{s,r\}$, define

$$\widehat{u}_{W,i} := \sigma_u(\widehat{L}_W + \rho I)^{-1}\widehat{\zeta}^u_{W,i}, \qquad u_{W,i,\rho} := (L_W + \rho I)^{-1} L_W u_{W,i},$$

where $\sigma_s := -1$ and $\sigma_r := 1$.

Lemma 2 (RKHS stability under a source condition)**.** *On the event of Lemma 1, Assumption 3 implies, uniformly over $(i,T) \in \mathcal{Q}_{p,q}$ and $u \in \{s,r\}$,*

$$\|u_{W,i,\rho} - u_{W,i}\|_{\mathcal{H}_W} \le C\rho^{\beta_u}, \qquad \|\widehat{u}_{W,i} - u_{W,i,\rho}\|_{\mathcal{H}_W} \le C\frac{b_\delta}{\rho}.$$

This lemma isolates the only nontrivial operator-stability step needed for Theorem 5. Appendix D gives its proof and integrates the remaining score-propagation argument directly into the proof of the theorem.

B.3 EXAMPLES INSIDE AND OUTSIDE THE RKHS SOURCE CLASS

The following propositions consider nonlinear Cauchy and linear Gaussian families, respectively.

Proposition 3 (A connected nonlinear family satisfying the RKHS source condition)**.** *Fix $r_0 \in \{1,2\}$ and set $q = r_0$. For any $p > r_0$, let the first r_0 variables be independent standard Cauchy roots and let*

$$X_k = \varepsilon_k \quad (k \le r_0), \qquad X_i = \sum_{k=1}^{r_0} b_{ki}\tanh(X_k) + \varepsilon_i \quad (i > r_0), \qquad b_{ki} \in \{-1,1\}, \tag{22}$$

where all errors are independent standard Cauchy variables. Arbitrary node-label permutations are allowed. This connected nonlinear family is uniformly Fisher-separated at $q = r_0$, with positive lower bounds on the ordering and parent-omission margins independent of p and the coefficient signs. Moreover, there is a deterministic finite-dimensional kernel family satisfying Assumption 3 uniformly, and hence also Assumption 2, with $\beta_s = \beta_r = 1$.

Proposition 4 (An L^2-admissible family outside the Gaussian-RBF source class)**.** *Fix $r_0 \in \{1,2\}$ and set $q = r_0$. Let $X_1, \dots, X_{r_0}$ be independent standard Gaussian roots and, for every $i > r_0$, let*

$$X_i = \sum_{k=1}^{r_0} b_{ki}X_k + \varepsilon_i, \qquad b_{ki} \in \{-1,1\}, \qquad \varepsilon_i \overset{\text{ind}}{\sim} N(0,1).$$

Use a deterministic Gaussian RBF kernel with fixed positive bandwidth in each local dimension. This family satisfies Assumption 2 uniformly and is uniformly Fisher-separated at $q = r_0$. However, it does not satisfy Assumption 3 for this fixed-bandwidth Gaussian RBF kernel family.

The constructed kernel in Proposition 3 is not the kernel used in the experiments. Proofs of both propositions are given in Appendix D.

C Linear Gaussian Specialization

This appendix specializes the estimator-agnostic recovery result in Theorem 3 to a Gaussian covariance plug-in estimator; it does not repeat the regularized kernel Stein analysis of Theorem 4. For a Gaussian marginal on $W = \{i\} \cup T$, with covariance matrix $\Sigma_W := \mathrm{Cov}(X_W)$,

$$J_i(T) = -\frac{1}{\mathrm{Var}(X_i \mid X_T)} = -(\Sigma_W^{-1})_{ii}.$$

Because $v \mapsto -1/v$ is increasing, FiCS then agrees with the bounded-subset conditional-variance criterion of Chen et al. (2019, Eq. (5)). Here $\|B\|_{\mathrm{op}} := \sup_{\|v\|_2=1} \|Bv\|_2$ is the induced matrix operator norm, consistent with the operator-norm notation used above, and $\|B\|_{\mathrm{F}}$ is the Frobenius norm.

Corollary 3 (Finite-sample Gaussian DAG recovery)**.** *Consider a linear Gaussian ANM satisfying the conditions of Theorem 1, with* $\max\{1, d\} \le q \le p-1$ *and positive lower bounds* $\Delta_{\mathrm{ord}}, \Delta_{\mathrm{par}}$ *on the ordering and parent-omission margins. Use the parent selection rule with* $\lambda > 0$ *and* $d\lambda < \Delta_{\mathrm{par}}$*, and let* $\epsilon_{\mathrm{alg}} > 0$ *be the recovery tolerance in Equation (19). For every queried* W*, define*

$$\bar{X}_W := \frac{1}{n}\sum_{a=1}^{n} X_W^{(a)}, \qquad \widehat{\Sigma}_W := \frac{1}{n}\sum_{a=1}^{n}(X_W^{(a)} - \bar{X}_W)(X_W^{(a)} - \bar{X}_W)^\top, \qquad \widehat{J}_i^{\mathrm{G}}(T) := -(\widehat{\Sigma}_W^{-1})_{ii}.$$

Here the superscript distinguishes the Gaussian covariance plug-in estimator from the regularized kernel Stein estimator in Equation (11); both estimate the same population score $J_i(T)$*. Suppose that every* Σ_W *has eigenvalues in* $[c_{\min}, c_{\max}]$*, where* $0 < c_{\min} \le c_{\max} < \infty$*. Then there is a universal constant* C *such that, for every* $\delta \in (0,1)$*, FiCS using the scores* $\widehat{J}_i^{\mathrm{G}}(T)$ *recovers the true DAG with probability at least* $1-\delta$ *whenever*

$$n \ge C\left[q\left\{1 + \log\left(\frac{p}{q}\right)\right\} + \log\left(\frac{p}{\delta}\right)\right]\left(\frac{c_{\max}}{c_{\min}}\right)^2\left(1 + \frac{1}{c_{\min}^2 \epsilon_{\mathrm{alg}}^2}\right). \tag{23}$$

Equation (23) separates the combinatorial term from the condition-number and recovery-tolerance factors. The tolerance ϵ_{alg} also depends on the indegree through $d\lambda$; the displayed combinatorial term alone therefore does not establish a minimax rate for growing d. Appendix D.5 proves the bound.

D Proofs of Auxiliary Results

This appendix proves the strengthened RKHS results and examples in Appendix B, followed by the Gaussian corollary in Appendix C.

D.1 Proof of Lemma 2

Proof. Fix a local query, suppress (W, i), and write $u = L^{\beta_u} f^u$. Spectral calculus gives

$$u_\rho - u = -\rho(L + \rho I)^{-1} L^{\beta_u} f^u.$$

Since $\sup_{t \ge 0} \rho t^{\beta_u}/(t+\rho) \le C\rho^{\beta_u}$ for $\beta_u \in (0, 1]$, the uniform bound on $\|f^u\|_{\mathcal{H}}$ proves the first claim. It also gives $\|u_\rho\|_{\mathcal{H}} \le \|u\|_{\mathcal{H}} \le C$ because $\|L\|_{\mathrm{op}} \le \kappa_0^2$.

The Stein identity gives $Lu = \sigma_u \zeta^u$. Subtracting the empirical and population regularized equations yields

$$\widehat{u} - u_\rho = (\widehat{L} + \rho I)^{-1}\left[\sigma_u(\widehat{\zeta}^u - \zeta^u) + (L - \widehat{L})u_\rho\right].$$

On the event of Lemma 1, both empirical deviations are at most b_δ, while $\|(\widehat{L} + \rho I)^{-1}\|_{\mathrm{op}} \le 1/\rho$. The uniform bound on $\|u_\rho\|_{\mathcal{H}}$ therefore proves the second claim. □

D.2 Proof of Theorem 5

Proof. Lemma 1 provides a common event of probability at least $1-\delta$. On this event, Lemma 2 and the triangle inequality give

$$d_u := \max_{(i,T)\in\mathcal{Q}_{p,q}} \|\widehat{u}_{W,i} - u_{W,i}\|_{\mathcal{H}_W} \le C\left\{\rho^{\beta_u} + \frac{b_\delta}{\rho}\right\}, \qquad u \in \{s, r\}.$$

It remains to propagate these target errors to the local Fisher scores. Fix a local query and abbreviate the corresponding targets by $s, r, \widehat{s}, \widehat{r}$. The empirical and population scores satisfy

$$\widehat{J}_i(T) = \langle \widehat{\mu}_W, \widehat{r} \rangle_{\mathcal{H}_W} - \langle \widehat{s}, \widehat{L}_W \widehat{s} \rangle_{\mathcal{H}_W}, \qquad J_i(T) = \langle \mu_W, r \rangle_{\mathcal{H}_W} - \langle s, L_W s \rangle_{\mathcal{H}_W}.$$

The first difference is bounded by

$$b_\delta \|\widehat{r}\|_{\mathcal{H}_W} + \|\mu_W\|_{\mathcal{H}_W} \|\widehat{r} - r\|_{\mathcal{H}_W},$$

and $\|\mu_W\|_{\mathcal{H}_W} \le \kappa_0$. For the quadratic term, adding and subtracting $\langle \widehat{s}, L_W \widehat{s} \rangle_{\mathcal{H}_W}$ gives

$$b_\delta \|\widehat{s}\|^2_{\mathcal{H}_W} + \|L_W\|_{\mathrm{op}} \|\widehat{s} - s\|_{\mathcal{H}_W} \left(\|\widehat{s}\|_{\mathcal{H}_W} + \|s\|_{\mathcal{H}_W} \right).$$

Assumption 3 and $\|L_W\|_{\mathrm{op}} \le \kappa_0^2$ give $\|u_{W,i}\|_{\mathcal{H}_W} \le \kappa_0^{2\beta_u} F$, so both $\|s\|_{\mathcal{H}_W}$ and $\|r\|_{\mathcal{H}_W}$ are uniformly bounded. Moreover, $\|\widehat{u}\|_{\mathcal{H}_W} \le \|u\|_{\mathcal{H}_W} + d_u$. Hence

$$\max_{(i,T) \in \mathcal{Q}_{p,q}} |\widehat{J}_i(T) - J_i(T)| \le C \left\{ d_s + d_r + d_s^2 + b_\delta (1 + d_s + d_r)^2 \right\}.$$

Since $b_\delta \le \rho \le 1$, we have $\omega_\delta(\rho) \le 2$, $d_s + d_r \le C\omega_\delta(\rho)$, and $b_\delta \le \omega_\delta(\rho)$. Substitution into the preceding display bounds its right-hand side by $C\{\omega_\delta(\rho) + \omega_\delta(\rho)^2\}$, proving Equation (21). □

D.3 Proof of Proposition 3

Proof. All errors have Fisher information $1/2$. We first verify uniform Fisher separation. Let $i > r_0$ be a leaf and let $T \subseteq \mathrm{Nd}(i)$, $|T| \le r_0$, omit at least one parent. The joint density of the roots and any finite collection of leaves is everywhere positive because it is a product of positive Cauchy error densities under the triangular structural transformation. Consequently, conditional on X_T, every omitted root has support $\mathbb{R}$ almost surely. Holding the other roots fixed and varying an omitted root changes $M_i = \sum_{k=1}^{r_0} b_{ki} \tanh(X_k)$ because $b_{ki} \neq 0$ and $\tanh$ is strictly monotone. Hence M_i is not $\sigma(X_T)$-measurable. The equality characterization in Theorem 1 gives $\Gamma_i(T) > 0$.

For fixed r_0, a local query contains at most $r_0 + 1$ variables. Up to coordinate permutations, its law is determined by finitely many root–leaf roles and coefficient-sign patterns, independently of p. The resulting parent-omission margins range over a finite collection of positive numbers, whose minimum gives a common lower bound $\Delta_{\mathrm{par}} > 0$ on the parent-omission margin. At any ancestral set with a remaining nonsource, every current source is an unselected root and every nonsource is a leaf missing at least one root parent. Since the noise Fisher information is $1/2$ at every node, Theorem 2 then gives $g_q(S) \ge \Delta_{\mathrm{par}}$. Thus the family is uniformly Fisher-separated.

We next construct an RKHS family and verify all parts of Assumption 3. For each local dimension $k \le r_0 + 1$, let $\mathcal{D}_k$ be the finite collection of marginal densities arising from the local types above. The standard Cauchy density is positive and C^∞, and each fixed-order derivative divided by the density is bounded. All derivatives of $\tanh$ are also bounded. Repeated differentiation of the joint density therefore produces the joint density multiplied by a bounded function. Dominated differentiation under the integral over unobserved roots shows that every $f \in \mathcal{D}_k$ is positive and C^∞ and that the first- and second-order targets $\partial_j \log f$ and $(\partial_j^2 f)/f$, together with all derivatives needed below, are bounded. The same domination shows that the marginal density and its relevant derivatives vanish at infinity, so both Stein integration-by-parts identities hold.

Form a finite feature vector ψ_k from the constant function and all distinct target functions $\partial_j \log f$ and $(\partial_j^2 f)/f$ with $f \in \mathcal{D}_k$ and $j \in [k]$, removing pointwise linear redundancies, and set $k_k(x, y) := \langle \psi_k(x), \psi_k(y) \rangle$. For every queried W, use $k_W := k_{|W|}$ and $\mathcal{H}_W := \mathrm{span}(\psi_{|W|})$. This kernel family is deterministic, positive semidefinite, and C^4. Bounded derivatives of the finite feature vectors give constants $\kappa_0, \kappa_1, \kappa_2$ that are common to all local types. Every local target belongs to $\mathcal{H}_W$ and has a uniformly bounded $L^2(P_W)$ norm.

For any local law P_W and nonzero $h \in \mathcal{H}_W$, removal of pointwise redundancies makes h a nonzero continuous function. Since P_W has an everywhere positive density, $\langle h, L_W h \rangle_{\mathcal{H}_W} = \mathbb{E}\{h(X_W)^2\} > 0$. Thus L_W is positive definite and invertible on $\mathcal{H}_W$. There are only finitely many local types, so $\sup_W \|L_W^{-1}\|_{\mathrm{op}} < \infty$. Taking $f^u_{W,i} := L_W^{-1} u_{W,i}$ gives $u_{W,i} = L_W f^u_{W,i}$ with a common bound on $\|f^u_{W,i}\|_{\mathcal{H}_W}$. Hence Assumption 3 holds with $\beta_s = \beta_r = 1$. □

D.4 Proof of Proposition 4

Proof. Every queried marginal has dimension at most $r_0 + 1$. Up to coordinate permutation, its covariance is determined by finitely many root–leaf roles and coefficient-sign patterns. Each such covariance is positive definite because every observed leaf contains an independent unit-variance error. Finiteness of the type collection therefore gives common positive lower and finite upper eigenvalue bounds. Consequently, the Gaussian first-order targets are linear, the second-order targets are quadratic, and their $L^2(P_W)$ norms are uniformly bounded. Gaussian tails justify the first- and second-order Stein identities. For a fixed positive bandwidth in each of the finitely many local dimensions, the Gaussian RBF kernel is deterministic and C^∞, and its feature and first two derivative-feature norms are uniformly bounded by constants κ_0, κ_1, and κ_2.

It remains to verify the uniform approximation condition. Fix a local law P_W, denote its RBF bandwidth by $h_W > 0$, and let $v \in L^2(P_W)$. The finite signed measure $\nu(dx) := v(x)P_W(dx)$ has finite total variation. For the translation-invariant Gaussian kernel, Fourier inversion gives

$$\langle v, \mathsf{T}_W v\rangle_{L^2(P_W)} = c_W \int_{\mathbb{R}^{|W|}} |\widehat{\nu}(\omega)|^2 \exp\{-h_W^2\|\omega\|^2/2\}\, d\omega,$$

where $c_W > 0$ depends only on the bandwidth and dimension. The Gaussian spectral density is strictly positive, so $\mathsf{T}_W v = 0$ implies $\widehat{\nu} = 0$ and hence $v = 0$ in $L^2(P_W)$. Thus $\ker(\mathsf{T}_W) = \{0\}$ and, because T_W is self-adjoint, $\overline{\mathrm{Ran}(\mathsf{T}_W)} = L^2(P_W)$. Spectral calculus now gives $\|\rho(\mathsf{T}_W + \rho I)^{-1} u_{W,i}\|_{L^2(P_W)} \to 0$ for both targets. Only finitely many local distribution types occur, so taking their maximum preserves convergence and makes all constants uniform. This verifies Assumption 2.

The stronger RKHS source condition fails for this kernel family. The Gaussian RBF kernel has bounded diagonal, so every f in its RKHS is bounded by $|f(x)| \le \|f\|_{\mathcal{H}_W}\sqrt{k_W(x,x)}$. Already for a singleton standard Gaussian root, the targets are $s(x) = -x$ and $r(x) = x^2 - 1$, both unbounded. They therefore do not belong to this Gaussian RBF RKHS and cannot equal $L_W^{\beta_u} f_{W,i}^u$ for any $\beta_u > 0$ and $f_{W,i}^u \in \mathcal{H}_W$. Thus Assumption 3 fails specifically for the stated fixed-bandwidth Gaussian RBF kernel family.

Finally, we verify uniform Fisher separation. Suppose a leaf omits $r_{\mathrm{miss}} \ge 1$ root parents. Conditional on any included roots, the remaining roots form a standard Gaussian vector $R \in \mathbb{R}^{r_{\mathrm{miss}}}$. Because $|T| \le r_0$, the other variables in T provide at most r_{miss} noisy linear measurements $BR + \xi$, where every row of B has entries in $\{-1, 1\}$ and ξ has identity covariance. The conditional covariance of R is $(I + B^\top B)^{-1}$. If $a \in \{-1, 1\}^{r_{\mathrm{miss}}}$ is the target leaf's loading vector, then

$$\mathrm{Var}(a^\top R \mid BR + \xi) = a^\top (I + B^\top B)^{-1} a \ge \frac{\|a\|_2^2}{1 + \|B\|_{\mathrm{op}}^2} \ge \frac{r_{\mathrm{miss}}}{1 + r_{\mathrm{miss}}^2},$$

because B has at most r_{miss} rows and $\|B\|_{\mathrm{op}}^2 \le \|B\|_{\mathrm{F}}^2 \le r_{\mathrm{miss}}^2$. The conditional variance of the target leaf is therefore at least $1 + r_{\mathrm{miss}}/(1 + r_{\mathrm{miss}}^2)$. Since its noise Fisher information is one, the parent-omission margin is at least $1/3$ for $r_0 = 1$ and $2/7$ for $r_0 = 2$. Equal noise Fisher information across nodes then supplies the same positive lower bound on the ordering margin. □

D.5 Proof of Corollary 3

Proof. Let $k := |W| \le q + 1$. Standard Gaussian concentration for the centered sample covariance implies that, for every $x > 0$, with probability at least $1 - 2e^{-x}$,

$$\|\widehat{\Sigma}_W - \Sigma_W\|_{\mathrm{op}} \le C_0 c_{\max}\left\{\sqrt{\frac{k+x}{n}} + \frac{k+x}{n}\right\}. \tag{24}$$

Apply Equation (24) with $x = \log(2N_{p,q}/\delta)$ and take a union bound over all local queries. With $\ell_\delta := \log(16N_{p,q}/\delta)$, there is a universal constant C_2 such that, with probability at least $1 - \delta$,

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} \|\widehat{\Sigma}_W - \Sigma_W\|_{\mathrm{op}} \le C_2 c_{\max}\left\{\sqrt{\frac{\ell_\delta}{n}} + \frac{\ell_\delta}{n}\right\}. \tag{25}$$

Since $\ell_\delta \le C_3[q\{1 + \log(p/q)\} + \log(p/\delta)]$, the sample-size condition in Equation (23) implies $n \gtrsim \ell_\delta (c_{\max}/c_{\min})^2$. After increasing the universal constant if necessary, the right-hand side of

Equation (25) is at most $c_{\min}/2$. Weyl's inequality then gives $\lambda_{\min}(\widehat{\Sigma}_W) \geq c_{\min}/2$ simultaneously over all local queries, so every $\widehat{\Sigma}_W$ is invertible. The inverse perturbation identity yields

$$\|\widehat{\Sigma}_W^{-1} - \Sigma_W^{-1}\|_{\mathrm{op}} \leq \|\widehat{\Sigma}_W^{-1}\|_{\mathrm{op}}\|\widehat{\Sigma}_W - \Sigma_W\|_{\mathrm{op}}\|\Sigma_W^{-1}\|_{\mathrm{op}} \leq 2c_{\min}^{-2}\|\widehat{\Sigma}_W - \Sigma_W\|_{\mathrm{op}}.$$

Consequently,

$$\max_{(i,T)\in\mathcal{Q}_{p,q}} |\widehat{J}_i^{\mathrm{G}}(T) - J_i(T)| \leq 2c_{\min}^{-2} \max_{(i,T)\in\mathcal{Q}_{p,q}} \|\widehat{\Sigma}_W - \Sigma_W\|_{\mathrm{op}}.$$

The remaining factor in Equation (23) makes the last display at most $\epsilon_{\mathrm{alg}}/2$ after a further universal enlargement of C. Equation (19) and Theorem 3 therefore give DAG recovery with probability at least $1 - \delta$. □

E Implementation Details

E.1 FiCS Implementation

Score estimation and numerical solution. For W with k variables and observations $z_a = X_W^{(a)}$, Equation (11) uses

$$k_W(z, z') = h_W^{-1} \exp\{-\|z - z'\|^2/(2h_W^2)\}, \qquad (K_W)_{ab} = k_W(z_a, z_b),$$
$$\eta_W = \max\left\{10^{-10}, \frac{r\,\mathrm{tr}(K_W)}{nk}\right\}.$$

Here h_W is $u > 0$ times the median of all n^2 pairwise Euclidean distances, including the diagonal, and $r > 0$ is the ridge multiplier. One Cholesky factorization serves both derivative systems and all target coordinates of W. A nonpositive bandwidth or failed factorization raises an error. The regularized kernel Stein estimator uses all selected observations; synthetic and real-data preprocessing are specified in Appendices F.1 and F.7, respectively. Section 5 does not directly analyze this empirical bandwidth-and-ridge rule, but gives the local-score accuracy required for exact recovery.

Search and tie breaking. Ordering uses subsets of size $\min(q, |S|)$; parent selection uses all predecessor subsets of size at most q. Parent-objective ties prefer fewer parents, followed by lexicographic order. Parent selection uses parent penalty $\lambda = c\sqrt{\log(p)/n}$, with the parent penalty multiplier c specified in Appendix E.3.

Incremental score updates. Once $|S| \geq q$, retain each remaining candidate's minimum score. After adding node j, evaluate only new subsets $U \cup \{j\}$ with $U \subseteq S$ and $|U| = q - 1$. Together with the old subsets, these exhaust the q-subsets of $S \cup \{j\}$, so updating the minimum preserves the original criterion and deterministic tie breaking. Before $|S| = q$, use the entire selected set. For $A_{p,q} = \sum_{t=0}^{q-1}(p - t)$ and $1 \leq q < p$, this reduces ordering-score lookups from $A_{p,q} + \binom{p+1}{q+2}$ to $A_{p,q} + \binom{p}{q+1}$, or from $O(p^{q+2})$ to $O(p^{q+1})$ for fixed q. It changes neither the distinct kernel factorizations nor parent selection.

E.2 Baseline Implementations

CaPS and SCORE. CaPS (Xu et al., 2024) and SCORE (Rolland et al., 2022) use their implemented graph-estimation procedures. For the real-data benchmarks, CaPS uses the authors' `train_test` function with native R-CAM pruning, and SCORE uses the authors' implementation with CAM pruning. Hyperparameter settings are given in Appendix E.3.

NoGAM. For the synthetic experiments, NoGAM (Montagna et al., 2023b) uses five-fold kernel-ridge regression followed by CAM pruning. Its adapter passes the declared RBF parameter γ to the regressor; the public implementation mistakenly passes ridge α in that position. The correction is applied to both tuning and test runs. For the real-data benchmarks, NoGAM uses the public DoDiscover `NoGAM()` implementation with native PyGAM pruning. Numerical routines in these real-data runs are unchanged; only data and path interfaces are adapted. The corresponding settings, including the effective KRR RBF coefficient, are given in Appendix E.3.

DAGMA. Nonlinear DAGMA uses an MLP with architecture $[p, 10, 1]$, weight decay $\lambda_2 = 0.005$, and double precision. It starts with $T = 4$ optimization stages; cyclic outputs trigger fresh fits at $T = 6, 7, 8$, retaining the first acyclic output. The same nonlinear implementation and retry rule are used for the synthetic and real-data benchmarks.

E.3 Hyperparameter Selection and Configurations

Synthetic experiments. Tuning uses ten independent DAGs at each of $p = 20, 50$ and $n = 500, 1000$, giving 40 inputs disjoint from the test data. Each method has four predetermined configurations in Table 2. Candidates are ranked by successful-fit fraction, F1 on inputs completed by all tied candidates, SHD$/p$, and finally candidate order. The selected configuration is fixed across all synthetic test dimensions, sample sizes, and robustness conditions. Reference edges are used only to evaluate tuning candidates.

Table 2: Hyperparameter candidates evaluated on the independent synthetic tuning datasets.

Method	Candidate 1	Candidate 2	Candidate 3	Candidate 4
FiCS (u, r, c)	$(1.5, .5, .2)$	$(.75, .5, .2)$	$(1.5, .1, .2)$	$(1.5, .5, .8)$
CaPS (λ_1, λ_2)	$(50, 50)$	$(2, 50)$	$(2, 25)$	$(5, 50)$
SCORE (η, ξ)	$(.001, .001)$	$(.001, .05)$	$(.01, .001)$	$(.01, .05)$
NoGAM (α, γ)	$(.01, .1)$	$(.001, .1)$	$(.1, .1)$	$(.01, 1)$
DAGMA (λ_1, t)	$(.02, .3)$	$(.02, .1)$	$(.005, .3)$	$(.005, .1)$

The selected configurations are FiCS $(u, r, c) = (1.5, 0.5, 0.8)$, CaPS $(\lambda_1, \lambda_2) = (50, 50)$, SCORE $(\eta, \xi) = (0.001, 0.001)$, NoGAM $(\alpha, \gamma) = (0.001, 0.1)$, and DAGMA $(\lambda_1, t) = (0.02, 0.3)$. Here, ξ is the pruning level and t is the DAGMA edge threshold. CaPS uses λ_1 for initial parent-score pruning and λ_2 for subsequent edge addition; SCORE sets $\eta_G = \eta_H = \eta$; and NoGAM fixes its Stein regularization and pruning level at 0.001.

Real-data experiments. FiCS uses bandwidth multiplier $u = 0.5$, ridge multiplier $r = 1.0$, parent penalty multiplier $c = 0.0003$, and $q = \min\{2, p - 1\}$ in both stages. The kernel, bandwidth, and matrix ridge are defined in Appendix E.1; parent selection minimizes $\widehat{J}_i(T) + \lambda|T|$ with $\lambda = c\sqrt{\log(p)/n}$. The rule gives $q = 2$ on all three datasets; each search is also restricted to the available predecessors. Appendix G.3 reports sensitivity to this common configuration on the same benchmark inputs.

CaPS uses $\lambda_1 = \lambda_2 = 50$, pre-pruning and edge addition enabled, Stein regularization 0.001, and native R-CAM pruning at 0.001. SCORE uses $\eta_G = \eta_H = 0.001$, CAM cutoff 0.001, variance normalization disabled, and variance as the dispersion statistic. NoGAM uses the public DoDiscover `NoGAM()` defaults: five folds, KRR regularization 0.01, $\eta_G = \eta_H = 0.001$, and native PyGAM pruning at $\alpha = 0.05$ with ten cubic splines. Preliminary neighborhood selection is disabled. The effective KRR RBF coefficient in this implementation is 0.01. DAGMA uses the synthetic configuration $(\lambda_1, t) = (0.02, 0.3)$ and the architecture and retry rule in Appendix E.2.

Each method uses the same configuration and execution rule across the three real-data benchmarks.

F Experimental Design

F.1 Synthetic Data Generation

We generate ER3 DAGs by sampling a random topological order and including each admissible forward edge independently with probability $3/(p - 1)$. Thus the expected mean total degree is three before imposing the indegree cap of two. When necessary, two parents are retained using independent random priorities. Each nonroot mechanism is linear with probability 0.5, with a random coefficient sign and magnitude sampled uniformly from $[0.5, 1]$, and otherwise is sampled from a zero-mean Gaussian process on the full parent vector with kernel $k(u, v) = \exp(-\|u - v\|^2/2)$. The structural errors are independent $\mathcal{N}(0, 1)$.

F.2 Evaluation Metrics

For reference edges E and estimated ordering $\widehat{\pi}$, $A_{\text{top}} = |\{(i, j) \in E : \widehat{\pi}(i) < \widehat{\pi}(j)\}|/|E|$ is the fraction of reference edges ordered forward. F1 compares directed edges; a reversal counts as one

false positive and one false negative. SHD counts a reversal as one edit. Means and sample standard deviations are computed over valid DAG estimates; a missing graph is not assigned F1 zero. The instances included in these means can differ between methods.

F.3 Main Synthetic Benchmark

Figure 2 evaluates 30 independently generated DAGs at each $p \in \{50, 100, 200\}$, using the generator in Appendix F.1 and nested sample sizes $n \in \{50, 100, 200, 500, 1000\}$. We compare FiCS with CaPS, SCORE, NoGAM, and nonlinear DAGMA on the same datasets within each setting. FiCS uses subset size $q = 2$ in both ordering estimation and parent selection. Each method retains the configuration selected on independent tuning DAGs in Appendix E.3 across all test dimensions and sample sizes. Figure 2 reports mean directed-edge F1 and one sample standard deviation using the aggregation rule in Appendix F.2.

F.4 Effect of the Conditioning Size

Figure 3 uses the same 30 DAGs and observations per dimension, with $p \in \{50, 100, 150, 200\}$ and $n = 100$. FiCS varies only its conditioning size for ordering over $q = 1, 2, 3$ and the entire selected set; the parent-selection subset size remains two and the parent penalty is $0.8\sqrt{\log(p)/n}$, while the regularized kernel Stein estimator and its tuning remain fixed. CaPS provides an external reference with its native pruning and edge supplementation. For CaPS, A_{top} uses the order before edge supplementation, whereas F1 evaluates its final DAG.

F.5 CaPS Numerical Failures

CaPS returns no DAG when its native GAM pruning encounters a nonfinite covariance matrix and the eigendecomposition fails. Table 3 lists all 79 such numerical failures in Figure 2.

Table 3: CaPS numerical failures among 30 attempts per setting in Figure 2. Only settings with a recorded numerical error are listed.

p	n	Numerical failures
50	50	1
100	50	22
100	100	5
200	50	28
200	100	19
200	200	4

In Figure 3, the same numerical error occurs in 0, 5, 14, and 19 of 30 CaPS runs at $p = 50, 100, 150, 200$, respectively.

F.6 Runtime Measurements

For each setting, runtime is measured by refitting three of the 30 graph–sample inputs used in Figure 2, with the same inputs and fixed configurations for all methods. Reported wall-clock time covers fitting, including graph construction and pruning, on the same server; data generation and calibration are excluded.

Table 4: Wall-clock fitting time (minutes) on three paired graph–sample inputs per setting. Entries are medians of three completed fits, except CaPS at $(p, n) = (200, 200)$. Its marked value is the time of the single completed fit (1/3), not a median.

p	n	FiCS	CaPS	SCORE	NoGAM	DAGMA
50	200	0.45	0.37	0.30	0.64	5.61
50	500	2.84	1.90	0.51	1.39	13.48
100	200	3.65	16.72	1.15	2.72	14.52
100	500	23.18	8.77	1.91	5.55	42.07
200	200	58.75	1025.06†	11.85	23.27	173.34
200	500	307.32	630.60	20.36	50.09	239.59

Table 4 provides an empirical check of the complexity analysis in Section 4.3. With $q = 2$, the number of distinct marginal factorizations is bounded by $\sum_{k=1}^{3} \binom{p}{k} = (p^3 + 5p)/6$, which increases approximately eightfold when p doubles. Correspondingly, increasing p from 50 to 100 increases FiCS runtime by factors of 8.1 and 8.2 at $n = 200$ and $n = 500$, respectively, closely matching this prediction. The increase from $p = 100$ to 200 is steeper, at 13.3–16.1 times, so the agreement is strongest over the smaller dimensions. Increasing n from 200 to 500 raises runtime by 5.2–6.4 times, below the 15.6-fold increase associated with cubic factorization cost; thus, the asymptotic $O(n^3)$ cost does not manifest as cubic wall-clock scaling over this sample-size range.

F.7 Real-Data Benchmarks

All five methods in Table 1 receive the same observations in their original measurement units, with no external logarithm or variance standardization. We retain each implementation's internal transformations, including FiCS's mean-centering for score estimation. F1 and SHD are computed against the directed references below; SHD counts a reversal once (Appendix F.2).

Data and reference graphs. For Danube, we use the 12-station selection and 11-edge spatial reference from the official code of Gnecco et al. (2021). Intersecting dates over its prescribed 32-station pool and retaining June–August before 2010 gives $N = 4{,}600$ complete daily observations from 1960–2009. We take 500 equally spaced row indices, $\lfloor \mathrm{linspace}(0, N - 1, 500) \rfloor$, preserving chronological order. For Sachs, we use all 853 observations of the 11-protein benchmark matrix and its 17-edge reference DAG distributed with CaPS (Sachs et al., 2005; Xu et al., 2024).

For CausalRivers, we use the official East Germany release (Stein et al., 2025). Its sampler yields 7,521 connected five-station induced subnetworks. Candidates are ranked by SHA-256 hashes of ordered station identifiers with the prefix `FiCS-rivers-20260925:`. We apply the official six-hour rounding and mean aggregation, followed by linear interpolation, and take 500 equally spaced jointly finite rows per subnetwork. Eligibility requires at least 500 jointly finite rows and a nonconstant value in every coordinate. We use the first 30 eligible candidates: one constant-input candidate is excluded for every method and replaced by the next ranked candidate. Each reference has four edges. We transpose the release's effect-row labels to the cause-row adjacency convention used throughout this paper.

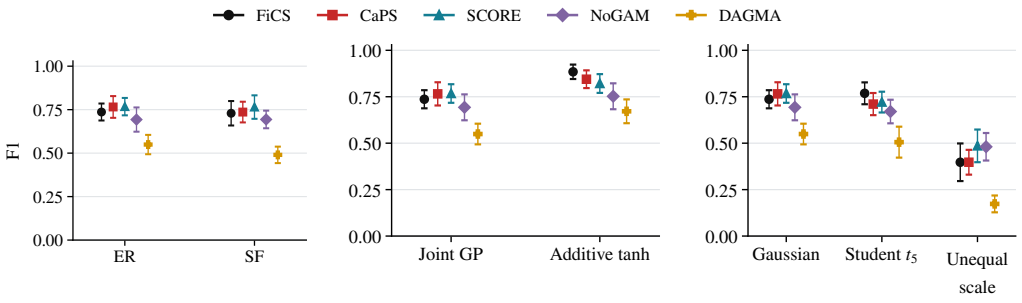
Scope of evaluation. FiCS uses $q = 2$ on all three datasets, with the settings in Appendix E.3. This is a computational cap, not an estimate of the true maximum indegree: the references have maximum indegree five on Danube, three on Sachs, and one to four across CausalRivers subnetworks. Thus, the cap can exclude reference parent sets.

These experiments assess agreement with biological or spatial reference graphs; they do not verify the theorem's i.i.d. ANM or recovery assumptions. River flows exhibit temporal dependence and may share rainfall drivers or unobserved upstream stations. The 30 CausalRivers subnetworks overlap in stations and observation times. Their reported SD describes variation across subnetworks, not uncertainty from 30 independent datasets.

G Additional Experiments

G.1 Sensitivity to Graph Structure, Mechanisms, and Noise

These tests examine whether the main random-DAG results depend on a particular graph family, nonlinear mechanism, or error distribution. Figure 4 changes one factor at a time at $p = 50, n = 200$, using 20 DAGs per condition and the fixed synthetic configurations. The scale-free (SF) graph family begins with an undirected preferential-attachment graph with two new edges per arriving node, orients its edges according to a random order, and caps the indegree at two. The additive alternative replaces nonlinear GP mechanisms by $\sum_{i \in \mathrm{Pa}(j)} b_{ij} \tanh(X_i)$ while retaining the 0.5 probability of a linear mechanism. Student-t_5 errors are scaled by $\sqrt{3/5}$ to have unit variance, and unequal Gaussian scales use noise standard deviation $\exp(U_j)$ with $U_j \sim \mathrm{Uniform}[-\log 2, \log 2]$. These DAGs are not screened for the sufficient recovery conditions.



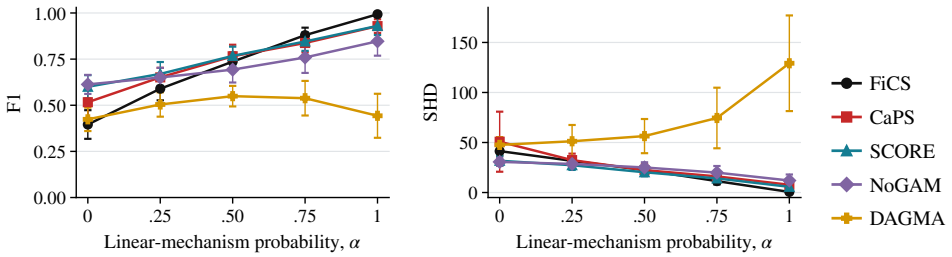
(a) Graph family **(b)** Nonlinear mechanism **(c)** Error distribution

Figure 4: Sensitivity to the data-generating model. Mean F1 and one-SD error bars over 20 DAGs per condition, at $p = 50, n = 200$. Each panel changes one factor from the ER3, joint-GP, unit-Gaussian reference condition. Horizontal offsets separate methods within each category.

FiCS's mean F1 is similar on ER3 and SF graphs (0.736 and 0.729), and increases to 0.884 for additive nonlinear mechanisms. The unequal-noise-scale condition reduces it to 0.397, compared with 0.768 for unit-variance Student-t_5 errors. Thus, the method tolerates these graph and mechanism changes better than heterogeneity in Gaussian noise scales; the results do not establish robustness to arbitrary error distributions.

G.2 Effect of the Linear-Mechanism Proportion

The main generator mixes linear and nonlinear mechanisms equally. To determine how this choice affects the comparison, we vary the probability of a linear nonroot mechanism over $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. The endpoints are all joint-parent GP mechanisms and all linear mechanisms. We retain $p = 50, n = 200$, maximum indegree two, unit Gaussian noise, and the fixed configurations of all five main methods, with $q = 2$ for FiCS. Graphs, coefficient draws, node-type uniforms, and primitive noise are paired across α; GP realizations are drawn separately across mechanism conditions. All 20 replicates per method and condition return DAGs (500 fits in total).



(a) F1 **(b)** SHD

Figure 5: Effect of the linear-mechanism proportion. Mean and one-SD error bars over all 20 DAGs per condition, at $p = 50, n = 200$. Increasing α replaces nonlinear GP mechanisms with linear ones; $\alpha = 0.5$ matches the main generator. All methods retain their fixed synthetic settings.

FiCS F1 increases from 0.396 at $\alpha = 0$ to 0.993 at $\alpha = 1$, while SHD decreases from 41.4 to 0.6. NoGAM and SCORE perform better at $\alpha = 0$ (F1 0.613 and 0.600); FiCS leads at $\alpha = 0.75$ and 1. Thus, performance depends on mechanism composition, and the main mixed-mechanism results should not be read as uniform superiority on purely nonlinear models.

G.3 Real-Data Sensitivity Analysis

We vary one parameter at a time around $(u, r, c) = (0.5, 1.0, 0.0003)$ and $q_0 = \min\{2, p - 1\}$, retaining all other parameters and the same 32 inputs. Table 5 varies the score-estimation parameters and subset size. Table 6 varies only the parent penalty multiplier.

Bandwidth multiplier, ridge multiplier, and subset size. We evaluate $u, r \in \{0.25, 0.5, 1\}$ and $q \in \{q_0, \min(3, p - 1), \min(5, p - 1), p - 1\}$ with $c = 0.0003$. Changing q changes both ordering

estimation and parent selection; it is not an isolated parent-selection comparison. For CausalRivers, $p = 5$, so $\min\{5, p-1\}$ and $p-1$ are the same setting. Changes in the bandwidth and ridge multipliers give mean CausalRivers F1 between 0.520 and 0.563. On Sachs, changing u from 0.5 to 0.25 or 1.0 reduces F1 from 0.538 to 0.429 or 0.417. Larger q improves recall on Danube and CausalRivers but adds false positives; F1 is nonmonotone and SHD increases relative to q_0.

Table 5: Sensitivity to bandwidth multiplier, ridge multiplier, and subset size on all three datasets. The common row uses $(u, r, c) = (0.5, 1.0, 0.0003)$ and $q = \min\{2, p-1\}$. Each other row changes only its named parameter. CausalRivers values are means over the same 30 subnetworks.

		Danube		Sachs		CausalRivers	
Parameter	Value	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓
Common configuration	–	0.452	17	0.538	11	0.563	4.53
Bandwidth multiplier u	0.25	0.467	16	0.429	12	0.548	4.77
	1.0	0.467	16	0.417	13	0.520	4.93
Ridge multiplier r	0.25	0.500	16	0.444	14	0.554	4.63
	0.5	0.452	17	0.500	13	0.551	4.67
Subset size q	$\min\{3, p-1\}$	0.474	20	0.500	13	0.579	5.03
	$\min\{5, p-1\}$	0.383	29	0.500	13	0.562	5.63
	$p-1$	0.353	33	0.500	13	0.562	5.63

Parent penalty multiplier. We vary $c \in \{0.0001, 0.0003, 0.0005, 0.001, 0.002, 0.003\}$ in $\lambda = c\sqrt{\log(p)/n}$, fixing $u = 0.5$, $r = 1.0$, and $q = q_0$. Changing c affects parent selection but leaves the ordering and A_{top} unchanged.

Table 6: Sensitivity to the parent penalty multiplier c in $\lambda = c\sqrt{\log(p)/n}$. Bandwidth and ridge multipliers remain 0.5 and 1.0, respectively, and $q = \min\{2, p-1\}$ in both stages. Bold c identifies the main-table setting. Edges is the number of selected edges; CausalRivers entries are means over the same 30 subnetworks.

	Danube			Sachs			CausalRivers		
c	F1↑	SHD↓	Edges	F1↑	SHD↓	Edges	F1↑	SHD↓	Edges
0.0001	0.438	18	21	0.500	11	11	0.567	4.57	6.80
0.0003	0.452	17	20	0.538	11	9	0.563	4.53	6.63
0.0005	0.452	17	20	0.538	11	9	0.534	4.67	6.37
0.001	0.308	18	15	0.538	11	9	0.514	4.73	6.17
0.002	0.308	18	15	0.560	10	8	0.510	4.77	6.13
0.003	0.240	19	14	0.560	10	8	0.512	4.73	6.10

Increasing c produces sparser graphs but has different effects across datasets (Table 6). From $c = 0.0001$ to 0.003, Danube F1 decreases from 0.438 to 0.240, whereas Sachs F1 increases from 0.500 to 0.560; CausalRivers F1 ranges from 0.510 to 0.567 across the grid. The common setting $c = 0.0003$ therefore does not maximize F1 on every dataset.

G.4 Gaussian Generator Regimes and Scope of Ordering Guarantees

To examine the scope of the population ordering guarantee beyond the main generator, we use the six Gaussian coefficient and noise-scale regimes of UUMCdata (Herman et al., 2025), version 0.0. Each regime contains 30 Erdős–Rényi DAGs with $p = 20$ and edge probability 0.1. We do not screen graphs for the theorem's recovery conditions. FiCS uses subset size $q = 5$, the maximum indegree in this cohort; this subset size makes every true parent set admissible but does not ensure a positive ordering margin.

Table 7 compares analytic population ordering with ordering at $n = 100$. FiCS uses $q = 5$ and its fixed synthetic estimator settings for the finite-sample results; VarSort and R^2Sort use empirical variance and linear-R^2 scores. Only ordering accuracy is compared. In 50–50 and iSCM, sample-size-dependent normalization changes the generated SCM, so the two columns are separate regime comparisons rather than a learning curve for a fixed SCM.

At the population level, FiCS has the highest A_{top} in four of the six regimes; R^2Sort leads in DaO and UUMC. FiCS orders all true edges correctly under unit-variance noise but does not do so in the other five regimes. Thus a subset size at least the maximum indegree does not imply the positive ordering margin required by Theorem 2.

Table 7: Mean ordering accuracy A_{top} across six linear Gaussian generator regimes ($p = 20$; 30 graphs per regime and setting). Population criteria use exact SCM covariances; finite-sample criteria use $n = 100$. Higher is better; bold marks the best method within each regime and setting.

	Population A_{top}			$n = 100$ A_{top}		
Generator	FiCS	VarSort	R^2Sort	FiCS	VarSort	R^2Sort
50–50	**0.482**	0.282	0.324	**0.565**	0.520	0.359
DaO	0.320	0.114	**0.535**	0.353	0.293	**0.543**
IPA	**0.478**	0.457	0.475	**0.510**	0.476	0.481
UUMC	0.465	0.468	**0.492**	**0.493**	0.471	0.482
iSCM	**0.499**	0.487	0.468	**0.526**	0.510	0.482
Unit-variance noise	**1.000**	0.944	0.611	**0.986**	0.945	0.609