CAUSAL DISCOVERY FROM QUANTILE SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery exploits asymmetries in uncertainty measures such as variance. However, many uncertainty-based methods rely on additive-noise assumptions. We extend this principle to structural causal models beyond additive noise using the Pairwise Average Separation Score (PASS), which measures uncertainty through quantile separation. Under a quantile-separation condition, our algorithm uses the same nodewise PASS score for both topological ordering and parent recovery. Under stated conditions, we establish graph-recovery consistency for a growing tensor-spline estimator. The criterion accommodates heavy-tailed distributions, including those with infinite variance, under suitable fractional-moment conditions. Experiments demonstrate robust graph recovery across additive-noise, heteroskedastic-noise, and heavy-tailed settings, with evaluations on three real-data benchmarks.

1 INTRODUCTION

Causal discovery from observational data includes constraint-based methods (Spirtes et al., 2000; Tsamardinos et al., 2003), continuous optimization approaches (Zheng et al., 2018; Bello et al., 2022), and asymmetry-based methods. The latter exploit non-Gaussianity in linear models (Shimizu et al., 2006), residual independence in nonlinear additive-noise models (ANMs) (Hoyer et al., 2008; Peters et al., 2014), or derivatives of the joint log-density (Rolland et al., 2022).

One line of asymmetry-based causal discovery exploits conditional uncertainty, measured through quantities such as variance. This approach can identify the graph even in linear Gaussian models when the error variances are equal (Peters & Bühlmann, 2014). Subsequent variance-based methods relax the equal-error-variance assumption in linear ANMs under suitable ordering conditions (Ghoshal & Honorio, 2018; Park, 2020), while CaPS develops a related ordering criterion for both linear and nonlinear ANMs with Gaussian noise (Xu et al., 2024). These methods rely on additive-noise models, where conditional uncertainty does not depend on parent values. We extend this uncertainty-based ordering principle to SCMs where uncertainty can depend on parent values, using quantile separation.

We introduce the *Pairwise Average Separation Score (PASS)*, which measures conditional uncertainty through average β-powered distances between quantile values. Under a quantile-separation condition, our algorithm recovers a topological ordering and the parents of each node in SCMs using the same score. This condition encompasses equal-error Gaussian models and the Gaussian variance-ordering conditions used in CaPS.

Our main contributions are as follows:

- **Uncertainty-based causal discovery beyond ANMs.** We establish recovery of a topological ordering and the parents of each node in SCMs beyond additive noise under an edgewise quantile-separation condition, using the same nodewise PASS score for both tasks.
- **Consistent graph recovery.** We establish graph-recovery consistency for a growing tensor-spline estimator under stated conditions.
- **Heavy-tailed extension.** Our consistency result accommodates infinite-variance distributions under the stated assumptions. Experiments show robust recovery on heavy-tailed synthetic data and competitive performance on the 12-station Danube and CausalRivers datasets.

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.

2 Related Work

Variance-based causal discovery. Variance-based methods and CaPS establish graph identifiability under restrictions on noise variances or causal mechanisms (Peters & Bühlmann, 2014; Ghoshal & Honorio, 2018; Park, 2020; Xu et al., 2024). Variance and R^2 sortability measure how consistently marginal variance and R^2, respectively, increase along directed paths (Reisach et al., 2021; 2023). Var-SortnRegress and R^2-SortnRegress sort once by increasing marginal variance and R^2, respectively. PASS recomputes conditional uncertainty after each removal, and its recursive sink-separation condition requires maxima only at sinks rather than monotonic increases along edges.

Uncertainty-based causal discovery. GFBS generalizes conditional-variance-based ordering and pruning through Bregman scores estimated via conditional-mean regression; under squared loss, these scores reduce to expected conditional variance (Rajendran et al., 2021). PASS measures uncertainty across the full conditional distribution through pairwise quantile separation, using the resulting energy-score generalized entropy (Gneiting & Raftery, 2007) for recursive sink selection and parent recovery. We establish quantile-separation conditions for graph identification and consistency of a tensor-spline quantile estimator under fractional-moment and regularity assumptions that permit infinite-variance distributions. Conditional-entropy methods also recover causal structure in discrete SCMs (Feigenbaum et al., 2024).

Quantile-based causal discovery. bQCD compares integrated quantile-regression losses between the two causal directions in bivariate settings (Tagasovska et al., 2020). At $\beta = 1$, PASS is proportional to the corresponding expected integrated pinball loss. Quantile Partial Effect (QPE) uses conditional-quantile derivatives for bivariate identification, while FICO uses Fisher information for multivariate sink ordering; under integration by parts, the latter equals the negative expected score-Jacobian diagonal used by CaPS (Chen et al., 2026; Xu et al., 2024). These approaches differ in how quantiles inform causal recovery: PASS uses the same quantile-separation score directly for both topological ordering and parent recovery.

3 Structural Causal Models and Notation

We work with real-valued structural causal models (SCMs) on a directed acyclic graph (DAG) $G = (V, E)$, with node set $V = \{1, \dots, p\}$ and directed edge set $E \subseteq V \times V$; $(k, j) \in E$ denotes $k \to j$. An SCM consists of endogenous random variables $X_1, \dots, X_p$ and exogenous random variables $\varepsilon_1, \dots, \varepsilon_p$. For $A \subseteq V$, write $X_A = (X_k)_{k \in A}$. The measurable structural assignments are

$$X_j = m_j(X_{\mathrm{Pa}(j)}, \varepsilon_j), \qquad \text{for all } j \in V, \tag{1}$$

where $\mathrm{Pa}(j)$ is the parent set of node j, m_j is a measurable function, and the exogenous variables $\varepsilon_1, \dots, \varepsilon_p$ are mutually independent.

Let $\mathcal{T}(G)$ denote the set of topological orderings of G. For a topological ordering π, write $\mathrm{Pred}_\pi(j) := \{k \in V : \pi(k) < \pi(j)\}$ for the set of predecessors of j under π.

We write $\mathbb{E}$, $\mathbb{E}_n$, and $\mathbb{P}$ for population expectation, empirical mean, and probability. For a random variable W, P_W denotes its distribution and F_W its marginal cumulative distribution function (CDF). For a random variable Y and a random vector Z, $P_{Y|Z=z}$ and $F_{Y|Z=z}$ denote a regular conditional distribution of Y given $Z = z$ and its CDF. We also write $P_{Y|Z}$ and $F_{Y|Z}$ for the corresponding random conditional distribution and random conditional CDF evaluated at Z.

For such Y and Z and $\tau \in (0, 1)$, define the conditional τ-quantile

$$q_{Y|Z}(\tau \mid z) := \inf\{t \in \mathbb{R} : F_{Y|Z=z}(t) \geq \tau\}.$$

For a realization $x = (x_1, \dots, x_p)$, write $x_A = (x_k)_{k \in A}$. For $A \subseteq V \setminus \{j\}$, write $q_{j|A}(\tau \mid x_A)$ for the conditional τ-quantile of X_j given $X_A = x_A$. When $Z = \varnothing$, we write $q_Y(\tau) := q_{Y|\varnothing}(\tau)$ for the marginal quantile.

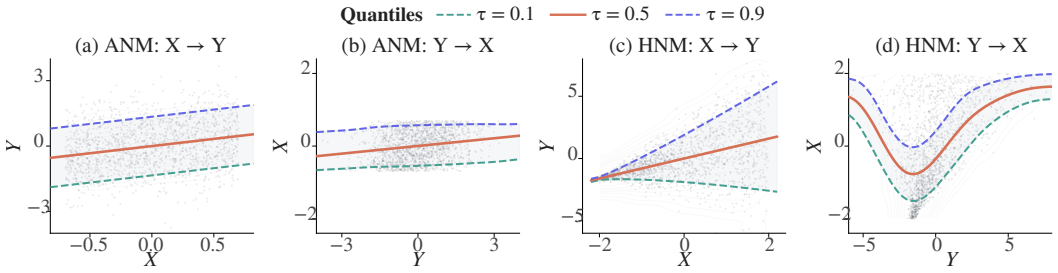


Figure 1: Illustration of PASS. PASS measures conditional uncertainty through quantile separation. Plots (a) and (c) show conditional quantiles for $Y \mid X$, while plots (b) and (d) show conditional quantiles for $X \mid Y$ using the same samples. Switching from $Y \mid X$ to $X \mid Y$ changes the conditional quantile curves, illustrating the directional asymmetry exploited by PASS. The ANM follows $Y = f(X) + \varepsilon$, while the heteroskedastic-noise model (HNM) follows $Y = f(X) + \sigma(X)\varepsilon$. The displayed quantiles correspond to $\tau = 0.1, 0.5, 0.9$.

4 PASS: Pairwise Average Separation Score

4.1 PASS and Detection of Conditional Distribution Changes

Pairwise quantile separation measures uncertainty in a conditional distribution. PASS averages this quantity over the distribution of the conditioning variables.

The *pairwise quantile separation* of Y given Z is defined, for $0 < \beta \leq 1$, by

$$\Gamma_\beta(Y \mid Z) := \mathbb{E}_{T,T'}\left[\left|q_{Y|Z}(T \mid Z) - q_{Y|Z}(T' \mid Z)\right|^\beta \middle| Z\right], \quad T, T' \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1),$$

where T, T' are independent of (Y, Z). When $Z = \varnothing$, this reduces to the marginal pairwise quantile separation $\Gamma_\beta(Y) := \Gamma_\beta(Y \mid \varnothing)$. Thus, $\Gamma_\beta(Y \mid Z)$ is the conditional expected β-powered distance between two independently sampled quantile values from $P_{Y|Z}$.

PASS for Y given Z at exponent β is then defined by

$$S_\beta(Y \mid Z) := \mathbb{E}[\Gamma_\beta(Y \mid Z)].$$

Thus, $\Gamma_\beta(Y \mid Z)$ can vary with Z, while $S_\beta(Y \mid Z)$ is its average. Figure 1 illustrates the quantile gaps in both conditioning directions. In the forward ANM, the gaps are constant across parent values; in the HNM, they vary with the conditional noise scale. In the ANM example, the conditional quantile gaps are wider for $Y \mid X$ than for $X \mid Y$. PASS aggregates these gaps over quantile levels and conditioning values, yielding a score that can differ between the two directions.

Additional conditioning cannot increase PASS. The following proposition characterizes when the decrease is strict.

Proposition 4.1 (Detection of conditional distribution changes)**.** *Fix $0 < \beta \leq 1$, and assume $\mathbb{E}|Y|^\beta < \infty$. For any node subsets $A \subseteq B \subseteq V$,*

$$S_\beta(Y \mid X_A) - S_\beta(Y \mid X_B) \geq 0.$$

Moreover, equality holds if and only if $P_{Y|X_B} = P_{Y|X_A}$ almost surely.

Thus, equality holds exactly when Y is conditionally independent of the added variables $X_{B\setminus A}$ given X_A. This behavior under conditioning has an entropy interpretation: $S_\beta(Y \mid Z)/2$ equals the generalized entropy of $P_{Y|Z}$ induced by the energy score, averaged over Z (Gneiting & Raftery, 2007).

4.2 Equivalent Representations of PASS

The following proposition gives two equivalent representations of PASS. The first representation rewrites pairwise quantile separation as an average β-powered pairwise distance between conditionally

independent copies from the same conditional distribution. The second expresses PASS through integrals of conditional quantiles.

Proposition 4.2 (Equivalent pairwise-distance and quantile-integral representations)**.** *Fix $0 < \beta \leq 1$, and assume $\mathbb{E}|Y|^\beta < \infty$. Let Y' be a conditionally independent copy of Y given Z, so that $Y, Y' \mid Z$ are conditionally i.i.d. from $P_{Y|Z}$. Then*

$$\Gamma_\beta(Y \mid Z) = \mathbb{E}\big(|Y - Y'|^\beta \mid Z\big), \qquad S_\beta(Y \mid Z) = \mathbb{E}|Y - Y'|^\beta.$$

Moreover,

$$S_\beta(Y \mid Z) = \mathbb{E}\left[2\int_0^1 \big|q_{Y|Z}(\tau \mid Z) - Y\big|^\beta \, d\tau - \int_0^1 \int_0^1 \big|q_{Y|Z}(\tau \mid Z) - q_{Y|Z}(\tau' \mid Z)\big|^\beta \, d\tau \, d\tau'\right]. \quad (2)$$

The pairwise-distance representation shows how the exponent β controls the moment requirement. PASS is finite under finite β-moments, so smaller values of β allow heavier-tailed distributions. We use the quantile-integral representation to construct the estimator in Section 6.

5 Identifiability via PASS

We establish conditions for graph recovery in two stages: recursive sink removal to obtain a topological order, followed by parent selection within that order.

For each node j and conditioning set $A \subseteq V \setminus \{j\}$, write $S_\beta(j \mid A) := S_\beta(X_j \mid X_A)$.

For a remaining node set $R \subseteq V$ and $j \in R$, the ordering score is

$$\Omega_{j,R} := S_\beta(j \mid R \setminus \{j\}). \quad (3)$$

Given a topological ordering π, the parent score of k for j is the reduction in PASS from adding X_k to the conditioning set:

$$\Psi_{j,\pi}(k) := S_\beta(j \mid \mathrm{Pred}_\pi(j) \setminus \{k\}) - S_\beta(j \mid \mathrm{Pred}_\pi(j)), \qquad k \in \mathrm{Pred}_\pi(j). \quad (4)$$

5.1 Ordering Recovery

For $R \subseteq V$, let $G[R]$ denote the induced subgraph and $\mathrm{sink}(G[R])$ its set of sinks. A set R is ancestral if $\mathrm{Pa}(j) \subseteq R$ for every $j \in R$; write $\mathcal{A}(G)$ for the family of ancestral sets. If every recursive step selects a sink, removing it and prepending it to the ordering recovers a topological ordering.

Theorem 5.1 (Ordering recovery under recursive sink separation)**.** *Consider an SCM of the form* (1)*. Fix $0 < \beta \leq 1$ and suppose all ordering scores $\Omega_{j,R}$ are finite for $R \in \mathcal{A}(G)$ and $j \in R$. Assume that, for every ancestral set R containing a non-sink,*

$$\max_{s \in \mathrm{sink}(G[R])} \Omega_{s,R} > \max_{u \in R \setminus \mathrm{sink}(G[R])} \Omega_{u,R}. \quad (5)$$

Then any maximizer of $\Omega_{j,R}$ over a nonempty ancestral set R is a sink of $G[R]$. Starting from $R = V$, recursively removing a maximizing node and prepending it to the ordering yields a topological ordering of G.

For an ancestral remaining set, conditioning on all other remaining nodes preserves a sink's parent-conditioned PASS and, under faithfulness, strictly reduces a non-sink's. The following edgewise condition then ensures recursive sink separation.

Corollary 5.1 (Ordering recovery under an edgewise quantile-separation condition)**.** *Consider an SCM of the form* (1) *whose observational distribution is faithful to G. Fix $0 < \beta \leq 1$ and assume $\max_{j \in V} \mathbb{E}|X_j|^\beta < \infty$. Suppose that, for every edge $u \to v$,*

$$S_\beta(u \,|\, \mathrm{Pa}(u)) \leq S_\beta(v \,|\, \mathrm{Pa}(v)).$$

Then recursive sink separation (5) *holds. Consequently, Theorem 5.1 yields recovery of a topological ordering by recursive sink removal.*

This includes equal-error Gaussian models and CaPS's Gaussian variance-ordering conditions (Peters & Bühlmann, 2014; Xu et al., 2024). For HNMs whose errors have unit pairwise quantile separation, the condition compares the expected βth powers of the conditional noise scales (Appendix E.2). Because PASS identifies order through comparisons of conditional uncertainty across nodes, relative measurement scales are part of its identifying assumptions. Common rescaling preserves these comparisons, whereas variable-specific rescaling may invalidate the separation condition (Proposition B.1).

5.2 Parent Recovery

Given a correct order, removing a non-parent predecessor leaves all parents in the conditioning set, so the local Markov property preserves the conditional distribution of X_j. Removing a parent changes this distribution under faithfulness. By Proposition 4.1, PASS therefore remains unchanged when a non-parent is removed and increases strictly when a parent is removed. The parent score $\Psi_{j,\pi}(k)$ measures this increase.

Theorem 5.2 (Parent recovery)**.** *Fix $0 < \beta \leq 1$ and a correct topological ordering $\pi \in \mathcal{T}(G)$. Assume that the induced observational distribution is faithful to G, and $\max_{j \in V} \mathbb{E}|X_j|^{\beta} < \infty$. Then, for every node j and predecessor $k \in \mathrm{Pred}_{\pi}(j)$,*

$$k \in \mathrm{Pa}(j) \quad \Longleftrightarrow \quad \Psi_{j,\pi}(k) > 0.$$

Consequently, there exists a constant $c_{\mathrm{par}} > 0$, common to all $\pi \in \mathcal{T}(G)$, such that every true parent satisfies $\Psi_{j,\pi}(k) \geq c_{\mathrm{par}}$, while every non-parent predecessor has score zero.

For a given correct order, variable-specific affine transformations preserve the zero/positive distinction in the population parent scores (Proposition B.1). Combining ordering recovery and parent recovery gives graph identifiability.

Corollary 5.2 (Graph identifiability via PASS)**.** *Under the conditions of Corollary 5.1 and Theorem 5.2, the graph structure of the SCM is identifiable via PASS.*

6 Algorithm

We construct a PASS estimator from spline quantile estimates and establish graph-recovery consistency under stated conditions.

6.1 PASS Score Estimator

We map each predictor to $[0, 1]$ for spline fitting, leaving the response unchanged. For $A \subseteq V \setminus \{j\}$, write $F_\ell := F_{X_\ell}$ and $U_A = (F_\ell(X_\ell))_{\ell \in A}$. With continuous marginals, conditioning on U_A is equivalent to conditioning on X_A, up to null sets. Let $q^U_{j|A}(\tau, u)$ be the conditional quantile of X_j given $U_A = u$. We use the empirical ranks:

$$\widehat{U}_{i\ell} = \frac{R_{i\ell}}{n+1}, \qquad R_{i\ell} = \mathrm{rank}(X_\ell^{(i)}; X_\ell^{(1:n)}).$$

For nonempty A, let $\mathcal{S}_{A,k}$ be deterministic full tensor B-spline spaces on $[0, 1]^{|A|}$ containing constants, with bases $\Phi_{A,k}$ and dimensions $m_{A,k}$. We use equally spaced knots, with spacing tending to zero as $k \to \infty$. The spline order is fixed and at least two. Appendix D.1 gives the construction. At sample size n, write $\Phi_A = \Phi_{A,k(n)}$ and $m_{n,A} = m_{A,k(n)}$ for a deterministic index $k(n)$. Define the spline quantile estimator by

$$\begin{aligned} \widehat{\theta}_{j,A}(\tau) &\in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{m_{n,A}}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\Big(X_j^{(i)} - \Phi_A(\widehat{U}_{i,A})^\top \theta\Big), \\ \widehat{q}^{\,E}_{j|A}(\tau, u) &= \Phi_A(u)^\top \widehat{\theta}_{j,A}(\tau), \qquad \rho_\tau(v) = v\{\tau - \mathbf{1}(v < 0)\}. \end{aligned} \tag{6}$$

Here ρ_τ is the pinball loss, and the superscript E denotes fitting with empirical predictor ranks. For $A = \varnothing$, use the marginal empirical quantile. Choose exact minimizers jointly measurably in the data and $\tau \in (0, 1)$.

To approximate the quantile integrals in (2), draw $T_1, \ldots, T_{K_n} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$ independently of the observations, with $K_n \geq 2$, and share them across all nodewise fits. For a fixed pair (j, A), write $\widehat{q}_{i,k} = \widehat{q}^E_{j|A}(T_k, \widehat{U}_{i,A})$.

To control tail contributions in the consistency analysis, we clip all nK_n fitted values for this pair at a common bound B: $\overline{q}_{i,k} = \max\{-B, \min(\widehat{q}_{i,k}, B)\}$. We choose B so that the average of $|\overline{q}_{i,k}|^{2\beta}$ over these values is at most twice the sample response moment $n^{-1}\sum_i |X_j^{(i)}|^{2\beta}$. We leave the predictions unchanged if they satisfy this moment bound before clipping. Appendix D.2 specifies the choice of B.

Using the clipped quantiles, we estimate PASS by

$$\widehat{S}_\beta(j \mid A) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2}{K_n}\sum_{k=1}^{K_n}|\overline{q}_{i,k} - X_j^{(i)}|^\beta - \frac{1}{K_n(K_n-1)}\sum_{k\neq l}|\overline{q}_{i,k} - \overline{q}_{i,l}|^\beta\right]. \tag{7}$$

The second term averages over the $K_n(K_n-1)$ ordered pairs of distinct quantile levels.

6.2 The PASS(β) Algorithm

Algorithm 1 takes a nodewise PASS estimator $\widehat{S}_\beta$ and parent thresholds $t_{j,k}$ as inputs. Its ordering and parent scores are

$$\widehat{\Omega}_{j,R} := \widehat{S}_\beta(j \mid R \setminus \{j\}), \qquad \widehat{\Psi}_{j,\pi}(k) := \widehat{S}_\beta(j \mid \text{Pred}_\pi(j) \setminus \{k\}) - \widehat{S}_\beta(j \mid \text{Pred}_\pi(j)). \tag{8}$$

Stage 1 removes a maximizing node as the next sink and prepends it to the order. Stage 2 retains $k \to j$ when $\widehat{\Psi}_{j,\hat{\pi}}(k) > t_{j,k}$.

Algorithm 1 PASS(β)

Given: data $X^{(1:n)}$, nodewise score routine $\widehat{S}_\beta$, and parent-threshold rule $t_{j,k}$.

Stage 1: Ordering recovery
- $R \leftarrow V$, $\hat{\pi} \leftarrow ()$
- **while** $R \neq \varnothing$ **do**
  - compute $\widehat{\Omega}_{j,R}$ for all $j \in R$
  - choose $\hat{s}_R \in \arg\max_{j\in R}\widehat{\Omega}_{j,R}$
  - prepend $\hat{s}_R$ to $\hat{\pi}$; update $R \leftarrow R \setminus \{\hat{s}_R\}$
- **end while**

Stage 2: Parent recovery
- **for** each node $j \in V$ **do**
  - $P_j \leftarrow \text{Pred}_{\hat{\pi}}(j)$
  - score P_j and each $P_j \setminus \{k\}$
  - compute $\widehat{\Psi}_{j,\hat{\pi}}(k)$ for $k \in P_j$
  - set $t_{j,k}$ using the threshold rule
  - $\widehat{\text{Pa}}(j) \leftarrow \{k \in P_j : \widehat{\Psi}_{j,\hat{\pi}}(k) > t_{j,k}\}$
- **end for**

Return: $\hat{\pi}$ and $\widehat{G} = (V, \{k \to j : k \in \widehat{\text{Pa}}(j)\})$.

Score implementation. Algorithm 1 requires $O(p^2)$ nodewise score evaluations across both stages. Experiments use additive spline quantile regressions with $O(p)$ basis functions per fit. Each quantile fit excludes the fold on which its score is evaluated, separating model fitting from score evaluation. The held-out predictions enter score contributions of the form in (7).

For parent recovery, fits with all predecessors and with one candidate removed share folds and quantile levels. Let $\widehat{\psi}^{(i)}_{j,k}$ be the score with candidate k omitted minus that with k included, evaluated on observation i. The parent score is $\widehat{\Psi}_{j,\hat{\pi}}(k) = n^{-1}\sum_{i=1}^n \widehat{\psi}^{(i)}_{j,k}$. Its estimated standard error is $\widehat{\text{se}}_{j,k} = \text{sd}\{\widehat{\psi}^{(i)}_{j,k}\}_{i=1}^n/\sqrt{n}$. The implementation uses the threshold $t_{j,k} = z_{1-\alpha}\widehat{\text{se}}_{j,k}$. Here $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. Fitting details and numerical settings are in Appendix F.

6.3 Consistency of PASS

We analyze the tensor-spline estimator (6) uniformly over all pairs (j, A), including adaptively selected sets. Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d., with $p \geq 2$, continuous marginals, and $\max_j \mathbb{E}|X_j|^{2\beta} < \infty$ for a fixed $0 < \beta \leq 1$.

Assumption 6.1 (Quantile-regression regularity). *For each pair $a = (j, A)$, let $F_a(\cdot \mid u)$ and $q_{a,\tau}(u) = q^U_{j|A}(\tau, u)$ denote the conditional CDF and quantile functions. Write $I_\epsilon = [\epsilon, 1-\epsilon]$ for $0 < \epsilon < 1/2$. Assume:*

(i) For nonempty A, P_{U_A} gives positive probability to every nonempty open subset of $(0,1)^{|A|}$.

(ii) For P_{U_A}-almost every u, $F_a(\cdot \mid u)$ is continuous and, for all $\tau \in (0,1)$ and $t > 0$, satisfies $F_a(q_{a,\tau}(u) - t \mid u) < \tau < F_a(q_{a,\tau}(u) + t \mid u)$.

(iii) For nonempty A and every $0 < \epsilon < 1/2$, the envelope $M_{a,\epsilon}(u) = \sup_{\tau \in I_\epsilon} |q_{a,\tau}(u)|$ satisfies $\mathbb{E}M_{a,\epsilon}(U_A) < \infty$.

(iv) For nonempty A, use the spline spaces $\mathcal{S}_{A,k}$ and bases $\Phi_{A,k}$ specified in Section 6.1.

For $A = \varnothing$, (ii) applies to the marginal distribution, and (iii) holds automatically because central marginal quantiles are bounded.

Using the quantile-regression series formulation of Belloni et al. (2019), we establish uniform consistency of the estimated PASS scores under the above conditions. Combining this consistency with the population separation margins yields the following graph-recovery result.

Theorem 6.1 (Graph recovery). *Under Assumption 6.1 and the conditions of Corollary 5.1 and Theorem 5.2, there exists a deterministic nondecreasing sequence of indices $k(n) \to \infty$, common to all pairs (j, A), for which the following holds. Use the moment-clipped estimator* (7) *with this choice of $k(n)$ and any deterministic $K_n \geq 2$ tending to infinity. With a fixed measurable tie-breaking convention among score maximizers, Algorithm 1 with a fixed parent threshold $0 < t < c_{\mathrm{par}}$ gives $\widehat{G}_t$ with*

$$\mathbb{P}(\widehat{G}_t = G) \longrightarrow 1.$$

The proof is given in Appendix C.1.

7 Numerical Experiments

We evaluate PASS using structural Hamming distance (SHD), directed-edge F1, and ordering accuracy A_{top}, the fraction of true edges consistent with the estimated order. PASS uses shared random quantile levels and off-diagonal energy estimates. PASS settings are given in Appendix F; protocols, full metrics, standard deviations where available, and runtimes for the simulation and real-data benchmarks appear in Appendix G.

7.1 ANM/HNM Simulations

We use ER1 and ER3 Erdős–Rényi graphs with expected edge counts p and $3p$, respectively. We compare PASS with FICO (Chen et al., 2026), TOPIC (Xu et al., 2025), CAM (Bühlmann et al., 2014), SCORE (Rolland et al., 2022), NoGAM (Montagna et al., 2023), DirectLiNGAM (Shimizu et al., 2011), HOST (Duong & Nguyen, 2023), and SkewScore (Lin et al., 2025).

ANM and HNM recovery. Under ER1, PASS has the lowest mean SHD in all six settings and the highest F1 except in Gaussian ANMs, where FICO leads (Table 1). Under ER3, PASS has the lowest SHD but not the highest F1 (Appendix G.2.2). Depth-based scale calibration favors PASS's ordering condition in these HNMs, without guaranteeing it. Without this calibration, PASS's SHD and F1 improve with sample size in RBF-GP/sigmoid HNMs (Appendix G.2.7). Given the true order, PASS performs best under Cauchy noise and CAM/GAM under Gaussian noise (Appendix G.3.3).

Sample-size trends. Figure 2 compares sample-size trends under ER1 and ER3 (Appendices G.2.4 and G.2.6). PASS's mean SHD generally decreases with n, but it is less competitive at small n under ER3; FICO retains lower SHD in Gaussian ANMs under ER3. Appendix G.3.6 compares additive and pairwise spline bases in the same ER1 experiment.

Table 1: Graph recovery at $p = 50$, $n = 1000$ (ER1): mean performance. Bold marks columnwise best means, including ties (SHD lower, F1 higher).

Method	ANM						HNM					
	Gaussian		Pareto(2)		Cauchy		Gaussian		Pareto(2)		Cauchy	
	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑
PASS	**32.4**	0.682	**24.0**	**0.753**	**23.1**	**0.759**	**24.2**	**0.751**	**18.6**	**0.807**	**19.8**	**0.790**
FICO	37.0	**0.704**	69.3	0.557	90.4	0.388	418.7	0.186	180.1	0.315	67.4	0.424
TOPIC	62.3	0.483	103.9	0.184	108.4	0.055	77.9	0.351	104.6	0.158	104.6	0.053
CAM	50.6	0.578	105.3	0.269	128.4	0.100	67.8	0.440	116.8	0.218	128.1	0.082
SCORE	208.9	0.186	607.2	0.104	678.2	0.071	204.7	0.151	601.4	0.102	683.4	0.063
NoGAM	217.8	0.181	613.4	0.106	683.9	0.072	202.1	0.150	598.1	0.110	684.4	0.065
DirectLiNGAM	104.3	0.351	55.1	0.587	48.2	0.525	85.7	0.411	68.3	0.523	49.6	0.496
HOST	134.0	0.338	164.0	0.113	150.1	0.208	115.3	0.402	152.5	0.123	116.0	0.243
SkewScore	108.0	0.295	114.0	0.359	139.6	0.199	92.3	0.329	120.7	0.313	145.4	0.151

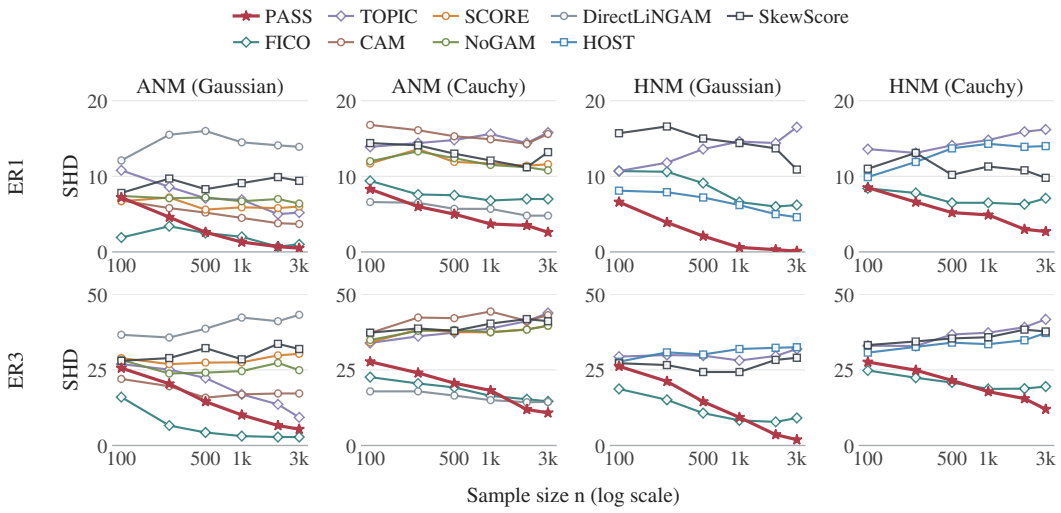


Figure 2: Sample-size trends at $p = 10$ with additive Gaussian-process (GP) mechanisms: mean SHD. Rows show ER1 (top) and ER3 (bottom), with different SHD ranges; lower is better. Panel titles identify model and noise. The n-axis is logarithmic. CAM, SCORE, NoGAM, and DirectLiNGAM appear only under ANM; HOST only under HNM.

7.2 Scale Dependence and Sortability

We evaluate PASS ordering under changes in measurement scale and compare it with variance- and R^2-based ordering across the UUMC generator regimes.

Sensitivity to variable scaling. We multiply each variable by an independently drawn random factor, with distributions specified in Table 2. The results confirm the scale dependence discussed in Section 5.1: variable-specific scaling materially reduces ordering accuracy and increases SHD, although the degradation is gradual over the tested ranges. In contrast, common tenfold scaling leaves performance nearly unchanged.

UUMC Benchmark. UUMC is designed to evaluate causal discovery methods under generator-induced variance and R^2 sortability. Its six linear Gaussian settings from UUMCdata (Herman et al., 2025) vary coefficients and noise scales to produce informative, weak, and reversed sortability regimes. Table 3 compares the ordering accuracy of PASS, Var-SortnRegress, and R^2-SortnRegress

over 100 matched seeds per setting. Appendices E.1 and G.3.1 give a two-node population comparison and the full benchmark results, respectively.

Table 2: PASS under scaling $X_j \mapsto s_j X_j$: linear Gaussian ANMs, $p = 10$, $n = 1000$, ER1. Entries are mean (SD).

Scaling	Factor s_j	$A_{\text{top}} \uparrow$	SHD$\downarrow$
None	1	1.000 (0.000)	1.2 (1.0)
Variable-specific	$\text{Uniform}(1, 2)$	0.853 (0.139)	4.7 (4.4)
	$\text{Uniform}(1, 5)$	0.811 (0.162)	5.2 (4.6)
	$\text{Uniform}(1, 10)$	0.786 (0.168)	5.9 (5.1)
Common	10	0.991 (0.029)	1.3 (1.2)

Table 3: Mean ordering accuracy A_{top} on the UUMC generator family ($p = 20$, $n = 3000$). Higher is better.

Generator setting	PASS	VarSort	R^2Sort
50–50	**0.911**	0.485	0.127
DaO	0.365	0.148	**0.566**
IPA	**0.497**	0.434	0.436
UUMC	**0.604**	0.505	0.423
iSCM	**0.591**	0.509	0.451
Unit-variance noise	**0.977**	0.934	0.870

PASS has the highest ordering accuracy in five of the six settings, but accuracy is low on DaO and IPA ($0.365, 0.497$). These results demonstrate that PASS does not rely solely on favorable variance or R^2 sortability induced by synthetic data generation. Table 8 reports the corresponding graph-recovery results.

8 Real Data Benchmarks

We evaluate the 12-station Danube daily-discharge network ($n = 4600$) (Gnecco et al., 2021), the Sachs cell-signaling dataset ($p = 11$, $n = 853$) (Sachs et al., 2005), and connected five-station river-discharge subnetworks from CausalRivers (Stein et al., 2025). Danube has positive generalized-Pareto tail-shape estimates at all stations using the top 10% of observations. CausalRivers preprocessing retains $7{,}486$ of the $7{,}521$ Random-5 subgraphs: six-hour aggregation and complete-case filtering leave fewer than two observations in the other 35 (Appendix G.4.2).

In Table 4, PASS attains F1 0.490 on Danube and 0.389 on Sachs; on Sachs, FICO and TOPIC each attain 0.462. CausalRivers subgraphs overlap, so small F1 differences are inconclusive; Appendix G.4.3 reports the official-protocol comparison.

Table 4: Real-data SHD and F1; parentheses give sample SDs. Danube/Sachs report ten algorithmic seeds for PASS and single runs for comparators. CausalRivers summarizes $7{,}486$ subgraphs. Lower SHD and higher F1 are better; bold marks columnwise best means, including ties.

	Danube		Sachs		CausalRivers	
Method	SHD$\downarrow$	F1$\uparrow$	SHD$\downarrow$	F1$\uparrow$	SHD$\downarrow$	F1$\uparrow$
PASS	**22.5 (1.4)**	**0.490 (0.026)**	15.4 (0.5)	0.389 (0.016)	**6.3 (2.0)**	**0.458 (0.193)**
FICO	51	0.301	**14**	**0.462**	7.6 (1.9)	0.453 (0.142)
TOPIC	41	0.281	**14**	**0.462**	9.5 (2.0)	0.290 (0.149)
CAM	43	0.189	15	0.444	9.0 (2.2)	0.304 (0.172)
SCORE	73	0.052	24	0.200	10.4 (2.2)	0.253 (0.155)
NoGAM	73	0.052	24	0.200	10.3 (2.2)	0.258 (0.157)
DirectLiNGAM	49	0.269	18	0.357	8.9 (2.2)	0.306 (0.159)
HOST	64	0.030	15	0.400	10.3 (2.0)	0.170 (0.140)
SkewScore	67	0.130	21	0.222	10.1 (2.0)	0.272 (0.141)

9 Conclusion

PASS uses one quantile-separation score for ordering and parent recovery. We characterize its asymmetry in ANMs and HNMs and prove graph-recovery consistency for a tensor-spline estimator under stated conditions. Identification is scale-sensitive; guarantees for the experimental additive-spline estimator and data-dependent parent threshold remain open. Future work includes studying separation in other SCMs and developing scalable quantile estimators.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, mathematical proof development and checking, experimental design, implementation, and interpretation of results. The authors take responsibility for the final content, including all claims and results.

REPRODUCIBILITY STATEMENT

Proofs and supporting regularity arguments are provided in Appendices A–D. Appendices F and G.1 specify the score estimators, fitting procedures, hyperparameters, simulation protocol, evaluation metrics, and baseline implementations. The real-data protocols and additional results are described in Section 8 and Appendix G.4. The supplementary archive `PASS_submission_code.zip` contains the experimental PASS implementation, synthetic data generators, and scripts to run the PASS settings in Table 1 and Figure 2. Data and precomputed results are not included.

REFERENCES

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, pp. 8226–8239. Curran Associates, Inc., 2022.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019. doi: 10.1016/j.jeconom.2019.04.003.

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. doi: 10.1214/14-AOS1260.

Yikang Chen, Xingzhe Sun, and Dehui Du. Causal discovery via quantile partial effect. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bao Duong and Thin Nguyen. Heteroscedastic causal structure learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pp. 598–605. IOS Press, 2023. doi: 10.3233/FAIA230321.

Itai Feigenbaum, Devansh Arpit, Shelby Heinecke, Juan Carlos Niebles, Weiran Yao, Caiming Xiong, Silvio Savarese, and Huan Wang. Causal layering via conditional entropy. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1176–1191. PMLR, 2024.

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

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.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Rebecca J. Herman, Jonas Wahl, Urmi Ninad, and Jakob Runge. Unitless unrestricted Markov-consistent SCM generation: Better benchmark datasets for causal discovery. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 1506–1531. PMLR, 2025.

Sujai Hiremath, Promit Ghosal, and Kyra Gan. LoSAM: Local search in additive noise models with mixed mechanisms and general noise for global causal discovery. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1677–1709. PMLR, 21–25 Jul 2025.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pp. 689–696. Curran Associates, Inc., 2008.

Yingyu Lin, Yuxing Huang, Wenqin Liu, Haoran Deng, Ignavier Ng, Kun Zhang, Mingming Gong, Yi-An Ma, and Biwei Huang. A skewness-based criterion for addressing heteroscedastic noise in causal discovery. In *International Conference on Learning Representations*, 2025.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 726–751. PMLR, 2023.

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

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. doi: 10.1093/biomet/ast043.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Goutham Rajendran, Bohdan Kivva, Ming Gao, and Bryon Aragam. Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18660–18672. Curran Associates, Inc., 2021.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784. Curran Associates, Inc., 2021.

Alexander Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. A scale-invariant sorting criterion to find a causal order in additive noise models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 785–807. Curran Associates, Inc., 2023. doi: 10.52202/075280-0036.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753. PMLR, 2022.

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. doi: 10.1126/science.1105809.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear Non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear Non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2 edition, 2000.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers - scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9311–9323. PMLR, 2020.

Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 673–678, 2003. doi: 10.1145/956750.956838.

Sascha Xu, Sarah Mameche, and Jilles Vreeken. Information-theoretic causal discovery in topological order. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 03–05 May 2025.

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, pp. 4315–4340. Curran Associates, Inc., 2024.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9472–9483. Curran Associates, Inc., 2018.

APPENDIX

A PROPERTIES OF PASS

A.1 PROOF OF PROPOSITION 4.2

Proof. Let $T, T' \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1)$ be independent of (Y, Z). Since $q_Y(T) \sim P_Y$ and $q_Y(T') \sim P_Y$ by the inverse probability-integral transform, and T and T' are independent, $q_Y(T)$ and $q_Y(T')$ are independent copies of Y. Therefore, $\Gamma_\beta(Y) = \mathbb{E}_{T,T'}[|q_Y(T) - q_Y(T')|^\beta] = \mathbb{E}|Y - Y'|^\beta$, where Y' is an independent copy of Y. Applying the same identity conditionally to $P_{Y|Z}$ gives $\Gamma_\beta(Y \mid Z) = \mathbb{E}[|Y - Y'|^\beta \mid Z]$, where $Y, Y' \mid Z \overset{\text{i.i.d.}}{\sim} P_{Y|Z}$. Taking expectation over Z yields $S_\beta(Y \mid Z) = \mathbb{E}[\Gamma_\beta(Y \mid Z)] = \mathbb{E}_Z[\mathbb{E}(|Y - Y'|^\beta \mid Z)]$.

It remains to verify the quantile-integral representation in Proposition 4.2. Conditional on $Z = z$, let $\widetilde{Y} := q_{Y|Z}(T \mid z)$ and $\widetilde{Y}' := q_{Y|Z}(T' \mid z)$. Then $\widetilde{Y}, \widetilde{Y}'$, and Y are conditionally i.i.d. from $P_{Y|Z=z}$. The bracketed expression in (2) therefore has conditional expectation $2\,\mathbb{E}(|\widetilde{Y} - Y|^\beta \mid Z = z) - \mathbb{E}(|\widetilde{Y} - \widetilde{Y}'|^\beta \mid Z = z) = \Gamma_\beta(Y \mid Z = z)$. Taking expectations over Z proves (2). When $Z = \varnothing$, the same argument applies with the conditioning argument omitted. □

A.2 PROOF OF PROPOSITION 4.1

For the proof, define the two-distribution discrepancy

$$\mathcal{D}_\beta(Y \mid Z_1, Y \mid Z_2) := 2\,\mathbb{E}_{T,T'}\Big[\big|q_{Y|Z_1}(T \mid Z_1) - q_{Y|Z_2}(T' \mid Z_2)\big|^\beta \Big| Z_1, Z_2\Big] - \Gamma_\beta(Y \mid Z_1) - \Gamma_\beta(Y \mid Z_2),$$

where $T, T' \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0,1)$ are independent of all observed variables.

Proof. Since $A \subseteq B$, X_A is a subvector of X_B. On an extended probability space, condition on X_B and draw Y'_A from $P_{Y|X_A}$ and Y'_B from $P_{Y|X_B}$, independently of Y and of each other given X_B. For $y \in \mathbb{R}$, define $U_A(y) := \mathbb{E}(|Y'_A - y|^\beta \mid X_A) - \frac{1}{2}\Gamma_\beta(Y \mid X_A)$ and $U_B(y) := \mathbb{E}(|Y'_B - y|^\beta \mid X_B) - \frac{1}{2}\Gamma_\beta(Y \mid X_B)$.

Since $Y \mid X_A \sim P_{Y|X_A}$ and $Y \mid X_B \sim P_{Y|X_B}$,

$$\mathbb{E}[U_A(Y) \mid X_A] = \frac{1}{2}\Gamma_\beta(Y \mid X_A), \qquad \mathbb{E}[U_B(Y) \mid X_B] = \frac{1}{2}\Gamma_\beta(Y \mid X_B).$$

Taking expectations gives

$$\frac{1}{2}\{S_\beta(Y \mid X_A) - S_\beta(Y \mid X_B)\} = \mathbb{E}[U_A(Y) - U_B(Y)].$$

Conditioning on X_B gives

$$\begin{aligned}\mathbb{E}[U_A(Y) - U_B(Y) \mid X_B] &= \mathbb{E}(|Y'_A - Y|^\beta \mid X_B) - \frac{1}{2}\Gamma_\beta(Y \mid X_A) - \frac{1}{2}\Gamma_\beta(Y \mid X_B) \\ &= \frac{1}{2}\mathcal{D}_\beta(Y \mid X_A, Y \mid X_B).\end{aligned}$$

Here $Y \mid X_B$ and $Y'_B \mid X_B$ are conditionally i.i.d., so $\mathbb{E}(|Y'_B - Y|^\beta \mid X_B) = \Gamma_\beta(Y \mid X_B)$. Taking expectations gives the exact identity $S_\beta(Y \mid X_A) - S_\beta(Y \mid X_B) = \mathbb{E}[\mathcal{D}_\beta(Y \mid X_A, Y \mid X_B)]$. By Gneiting & Raftery (2007, Theorem 5 and Example 9), the discrepancy for the kernel $g(x, y) = |x - y|^\beta$ is nonnegative and vanishes exactly when the two distributions agree. Its expectation is therefore zero if and only if $P_{Y|X_B} = P_{Y|X_A}$ almost surely, proving both claims. □

B Proofs of Structural Identifiability via PASS in Section 5

B.1 Proof of Theorem 5.1

Proof. If R contains a non-sink, (5) ensures that every maximizer is a sink; otherwise every node in R is a sink. Removing a sink preserves ancestralness, since it is not a parent of any remaining node. Induction from $R = V$ therefore selects only sinks, and reversing their removal order gives a topological ordering. □

B.2 Proof of Corollary 5.1

Throughout the appendix, we use $G[R]$, $\mathrm{sink}(G[R])$, and $\mathcal{A}(G)$ as defined in Section 5.1.

Proof. Sink baseline. Fix $R \in \mathcal{A}(G)$. Let $s \in \mathrm{sink}(G[R])$. Because R is ancestral, every node in $R \setminus (\{s\} \cup \mathrm{Pa}(s))$ is a nondescendant of s: a directed path starting from s cannot leave R and later re-enter R, because that would give some node of R a parent outside R, contradicting ancestralness. By the local Markov property of the distribution induced by the SCM with independent exogenous variables, $X_s \perp X_{R\setminus(\{s\}\cup\mathrm{Pa}(s))} \mid X_{\mathrm{Pa}(s)}$. Since $R \setminus \{s\}$ is the union of $\mathrm{Pa}(s)$ and $R \setminus (\{s\} \cup \mathrm{Pa}(s))$, this conditional independence implies $P_{X_s|X_{R\setminus\{s\}}} = P_{X_s|X_{\mathrm{Pa}(s)}}$ a.s. The nodewise PASS value depends on the conditioning set only through this conditional distribution, and therefore $\Omega_{s,R} = S_\beta(s \,|\, R \setminus \{s\}) = S_\beta(s \,|\, \mathrm{Pa}(s))$.

Non-sink strictness. Let $u \in R \setminus \mathrm{sink}(G[R])$. Since u has a child in $G[R]$, faithfulness implies

$$X_u \not\perp X_{R\setminus(\{u\}\cup\mathrm{Pa}(u))} \mid X_{\mathrm{Pa}(u)}.$$

Equivalently, conditioning on $X_{R\setminus\{u\}}$ changes the conditional distribution of X_u relative to conditioning only on $X_{\mathrm{Pa}(u)}$. Apply Proposition 4.1 with $Y = X_u$, $A = \mathrm{Pa}(u)$, and $B = R \setminus \{u\}$. Because R is ancestral, $\mathrm{Pa}(u) \subseteq R \setminus \{u\}$, so

$$S_\beta(u \,|\, \mathrm{Pa}(u)) - \Omega_{u,R} > 0.$$

Cross-node comparison. Since u is not a sink of $G[R]$, it has a descendant sink $s_u \in \mathrm{sink}(G[R])$. The edgewise quantile-separation condition propagates along directed paths, so $S_\beta(u \,|\, \mathrm{Pa}(u)) \le S_\beta(s_u \,|\, \mathrm{Pa}(s_u))$. Combining this pathwise comparison with non-sink strictness gives

$$\Omega_{u,R} < S_\beta(u \,|\, \mathrm{Pa}(u)) \le S_\beta(s_u \,|\, \mathrm{Pa}(s_u)) = \Omega_{s_u,R}.$$

Every non-sink therefore has a smaller score than some sink. Since R is finite, this implies (5) whenever R contains a non-sink. The finite β-moment assumption makes all ordering scores finite. Theorem 5.1 now gives the claimed topological ordering. □

B.3 Proof of Theorem 5.2

Proof. Write $\Psi_j(k) := \Psi_{j,\pi}(k)$. First, we relate a zero parent score to conditional independence. Fix a node j and a predecessor $k \in \mathrm{Pred}_\pi(j)$. Set $A_0 := \mathrm{Pred}_\pi(j) \setminus \{k\}$ and $B_0 := \mathrm{Pred}_\pi(j)$. The inclusion $A_0 \subseteq B_0$ ensures that the A_0-conditioning variables are a subvector of the B_0-conditioning variables. The moment assumption also makes the nodewise score conditioned on X_{A_0} finite, since Proposition 4.2 and $0 < \beta \le 1$ give $S_\beta(X_j \mid X_{A_0}) \le 2\mathbb{E}|X_j|^\beta < \infty$ after taking expectations.

Since $\mathbb{E}|X_j|^\beta < \infty$, Proposition 4.1 applies with $Y = X_j$, smaller conditioning set A_0, and larger conditioning set B_0. It gives $\Psi_{j,\pi}(k) \ge 0$, with equality if and only if $P_{X_j|X_{\mathrm{Pred}_\pi(j)}} = P_{X_j|X_{\mathrm{Pred}_\pi(j)\setminus\{k\}}}$ almost surely. This equality is equivalent to $X_j \perp X_k \mid X_{\mathrm{Pred}_\pi(j)\setminus\{k\}}$.

Now consider any node j and predecessor $k \in \mathrm{Pred}_\pi(j)$. If $k \notin \mathrm{Pa}(j)$, then k is a non-parent predecessor of j, hence a nondescendant of j. Because π is a topological ordering, every predecessor of j is a nondescendant of j. Because the SCM with independent exogenous variables induces a distribution that is Markov with respect to G, the local Markov property gives $X_j \perp X_{\mathrm{Pred}_\pi(j)\setminus\mathrm{Pa}(j)} \mid X_{\mathrm{Pa}(j)}$. To isolate the single non-parent predecessor k, write $N_{j,k} := \mathrm{Pred}_\pi(j) \setminus (\mathrm{Pa}(j) \cup \{k\})$. The local Markov conditional independence can then be written as $X_j \perp (X_k, X_{N_{j,k}}) \mid X_{\mathrm{Pa}(j)}$. By the weak-union property of conditional independence, this implies $X_j \perp X_k \mid (X_{\mathrm{Pa}(j)}, X_{N_{j,k}})$,

equivalently $X_j \perp X_k \mid X_{\mathrm{Pred}_\pi(j)\setminus\{k\}}$. By the zero-score equivalence just proved, $\Psi_j(k) = 0$ for every $k \in \mathrm{Pred}_\pi(j) \setminus \mathrm{Pa}(j)$.

Now suppose $k \in \mathrm{Pa}(j)$, and set $A := \mathrm{Pred}_\pi(j) \setminus \{k\}$. If $\Psi_j(k) = 0$, the zero-score equivalence gives $X_j \perp X_k \mid X_A$. But the direct edge $k \to j$ remains graphically active after conditioning on the other predecessors X_A, since the one-edge path from k to j has no intermediate node to block. Faithfulness therefore rules out $X_j \perp X_k \mid X_A$. Since Proposition 4.1 also gives $\Psi_j(k) \geq 0$, it follows that $\Psi_j(k) > 0$ for every $k \in \mathrm{Pa}(j)$. Combining the two cases gives zero scores for non-parent predecessors and strictly positive scores for true parents. There are finitely many valid orders and true-parent pairs, so these scores admit a common positive lower bound c_{par} over all valid orders. This proves the margin separation in Theorem 5.2. □

B.4 Proof of Corollary 5.2

Proof. Corollary 5.1 recovers a correct topological ordering by recursive sink removal. Conditional on any such ordering, Theorem 5.2 separates each true parent from each non-parent predecessor. Applying these two stages gives the edge set of G, so the population PASS procedure recovers the graph. □

B.5 Affine Transformations

Proposition B.1 (Affine equivariance)**.** *Fix $0 < \beta \leq 1$ and suppose $\max_{j\in V} \mathbb{E}|X_j|^\beta < \infty$. Let $X_j^\star = a_j X_j + b_j$, where $a_j \neq 0$ and b_j are constants. For every $j \in V$ and $A \subseteq V \setminus \{j\}$,*

$$S_\beta(X_j^\star \mid X_A^\star) = |a_j|^\beta S_\beta(X_j \mid X_A).$$

Consequently, the population ordering and parent scores for $X^\star$ satisfy

$$\Omega^\star_{j,R} = |a_j|^\beta \Omega_{j,R}, \qquad \Psi^\star_{j,\pi}(k) = |a_j|^\beta \Psi_{j,\pi}(k)$$

for every $j \in R$ and fixed order π with $k \in \mathrm{Pred}_\pi(j)$. Common absolute scales preserve all population ordering comparisons; unequal scales need not. For a given correct order, the zero/positive parent characterization in Theorem 5.2 is unchanged.

Proof. The coordinatewise transformations are bijective, so $\sigma(X_A^\star) = \sigma(X_A)$. If X_j' is a conditionally independent copy of X_j given X_A, then $a_j X_j' + b_j$ is a conditionally independent copy of $X_j^\star$ given $X_A^\star$. Proposition 4.2 therefore gives $S_\beta(X_j^\star \mid X_A^\star) = \mathbb{E}|a_j(X_j - X_j')|^\beta = |a_j|^\beta S_\beta(X_j \mid X_A)$. Substitution into (3) and (4) proves the remaining identities and conclusions. □

C Proofs of Consistency Results

C.1 Proof of Theorem 6.1

Proof. We first prove uniform score consistency and then use the population margins to recover the graph. By Proposition D.1, we can choose a deterministic $k(n) \uparrow \infty$, common to all pairs (j, A), such that, for every fixed $\epsilon \in (0, 1/2)$,

$$\mathcal{R}_n(\epsilon) := \max_{j,A} \sup_{\tau\in[\epsilon,1-\epsilon]} \frac{1}{n}\sum_{i=1}^n \left|\widehat{q}^{E}_{j|A}(\tau, \widehat{U}_{i,A}) - q^{U}_{j|A}(\tau, U_{i,A})\right|^\beta = o_P(1). \tag{9}$$

Define the uniform score error by

$$e_n := \max_{j,A} |\widehat{S}_\beta(j \mid A) - S_\beta(j \mid A)|. \tag{10}$$

For any deterministic $K_n \geq 2$ tending to infinity, we compare each estimated score with the same score evaluated at the true conditional quantiles, separating quantile-estimation error from the error due to sampling observations and quantile levels.

For a pair $a = (j, A)$, write $q_a(\tau \mid x_A) := q_{j|A}(\tau \mid x_A)$ and $S_a := S_\beta(X_j \mid X_A)$.

Let $T_1, \dots, T_K \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$ be independent of the observations and shared by all pairs (j, A). For this comparison, define the score using the true conditional quantiles and the same off-diagonal normalization:

$$\widetilde{S}^U_{a,n,K} := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{2}{K}\sum_{k=1}^{K}|q_a(T_k \mid X_A^{(i)}) - X_j^{(i)}|^\beta - \frac{1}{K(K-1)}\sum_{k\neq\ell}|q_a(T_k \mid X_A^{(i)}) - q_a(T_\ell \mid X_A^{(i)})|^\beta\right\}. \tag{11}$$

The superscript U denotes the off-diagonal normalization in (7), which requires $K \geq 2$.

Let $\widehat{S}^U_{a,n,K}$ be the corresponding fitted score, using the moment-clipped predictions $\overline{q}_{i,k}$ from (7) in place of the true quantiles. Let $\mathcal{I} = \{(j, A) : j \in V,\ A \subseteq V \setminus \{j\}\}$ be the finite collection of pairs. The triangle inequality gives

$$e_n \leq \underbrace{\max_{a\in\mathcal{I}}|\widehat{S}^U_{a,n,K_n} - \widetilde{S}^U_{a,n,K_n}|}_{\text{quantile-estimation error}} + \underbrace{\max_{a\in\mathcal{I}}|\widetilde{S}^U_{a,n,K_n} - S_a|}_{\text{sampling and random-quantile approximation error}}. \tag{12}$$

The following two lemmas control the two terms; their proofs are given in Appendices D.2 and D.3, respectively.

Lemma C.1 (Quantile-estimation gap)**.** *Under the sampling and moment conditions of Section 6.3 and Assumption 6.1, use $k(n)$ from Proposition D.1 and the moment-clipped scores* (7)*, with shared i.i.d.* $\text{Uniform}(0,1)$ *levels independent of the observations. For every deterministic $K_n \geq 2$ tending to infinity,*

$$\max_{a\in\mathcal{I}}|\widehat{S}^U_{a,n,K_n} - \widetilde{S}^U_{a,n,K_n}| = o_P(1).$$

Lemma C.2 (Sampling and random-quantile error)**.** *Fix $0 < \beta \leq 1$ and a finite collection $\mathcal{I}$ of pairs (j, A). Suppose $X^{(1)}, \dots, X^{(n)}$ are i.i.d., $T_1, \dots, T_K \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$ are independent of the data, $K \geq 2$, and $\max_j \mathbb{E}|X_j|^{2\beta} < \infty$. Then, uniformly over $a \in \mathcal{I}$,*

$$\max_{a\in\mathcal{I}}|\widetilde{S}^U_{a,n,K} - S_a| = O_P(n^{-1/2} + K^{-1}).$$

Applying Lemmas C.1 and C.2 to (12) gives

$$e_n = o_P(1) + O_P(n^{-1/2} + K_n^{-1}) = o_P(1), \tag{13}$$

since $K_n \to \infty$.

We now use the uniform score convergence to preserve the population gaps that separate sinks from non-sinks and parents from non-parents. If G has no edges, every order is valid and every population parent score is zero. Then $2e_n < t$ guarantees exact recovery, with probability tending to one. Hence assume G has an edge.

Let $\mathcal{A}(G) = \{R \subseteq V : \text{Pa}(j) \subseteq R \text{ for every } j \in R\}$ denote the collection of ancestral sets and let $\text{sink}(R)$ denote the sinks of the induced graph on R. Define

$$c_{\text{ord}} := \min_{\substack{R\in\mathcal{A}(G)\\ R\setminus\text{sink}(R)\neq\varnothing}}\left\{\max_{s\in\text{sink}(R)}\Omega_{s,R} - \max_{j\in R\setminus\text{sink}(R)}\Omega_{j,R}\right\},$$

where the minimum is over a nonempty collection. Corollary 5.1 gives the strict separation in (5), so each difference is positive. Since the collection is finite, $c_{\text{ord}} > 0$. On $2e_n < c_{\text{ord}}$, every node

maximizing the fitted score is a true sink whenever a non-sink remains. Removing a sink preserves ancestrality, so induction yields a valid order.

For any valid order π and candidate $k \in \mathrm{Pred}_\pi(j)$, Theorem 5.2 gives $\Psi_{j,\pi}(k) = 0$ when $k \notin \mathrm{Pa}(j)$ and $\Psi_{j,\pi}(k) \geq c_{\mathrm{par}} > 0$ when $k \in \mathrm{Pa}(j)$. Each fitted contrast is a difference of two nodewise scores, hence

$$\max_{\pi \in \mathcal{T}(G)} \max_{j,\, k \in \mathrm{Pred}_\pi(j)} |\widehat{\Psi}_{j,\pi}(k) - \Psi_{j,\pi}(k)| \leq 2e_n.$$

The event $2e_n < \min\{c_{\mathrm{ord}}, t, c_{\mathrm{par}} - t\}$ therefore guarantees a valid order and exact parent recovery. With $\delta = \min\{c_{\mathrm{ord}}, t, c_{\mathrm{par}} - t\} > 0$,

$$\mathbb{P}(\widehat{G}_t \neq G) \leq \mathbb{P}(2e_n \geq \delta) \longrightarrow 0$$

by (13). The same argument gives $\mathbb{P}(\widehat{G}_{t_n} = G) \to 1$ for deterministic $t_n \downarrow 0$ satisfying $e_n = o_P(t_n)$. □

D Auxiliary Results for Consistency

We first establish the central-quantile prediction consistency needed for Lemma C.1. We then prove Lemmas C.1 and C.2.

D.1 Tensor-Spline Consistency with Empirical Predictor Ranks

The following proposition establishes the central-quantile prediction bound (9) under Assumption 6.1. Write $a = (j, A)$, $Y = X_j$, $F_a(\cdot \mid u)$ for the conditional CDF, and $q_{a,\tau}(u) = q^U_{j|A}(\tau, u)$. We take jointly measurable versions, with $q_{a,\tau}(u) = \inf\{y : F_a(y \mid u) \geq \tau\}$. For fixed k, $\widehat{f}_{a,n,k,\tau} = \Phi^\top_{A,k}\widehat{\theta}^E_{a,n,k}(\tau)$ denotes the unconstrained, unpenalized quantile-regression series fit from (6) (Belloni et al., 2019). For a generic function h, write $Ph = \mathbb{E}h$ and $P_n h = n^{-1}\sum_i h_i$.

Proposition D.1 (Quantile-regression consistency with empirical ranks)**.** *Assume i.i.d. sampling. For each distribution with continuous marginals satisfying Assumption 6.1 for all pairs (j, A), there is a deterministic nondecreasing sequence $k(n) \to \infty$ such that*

$$\mathcal{R}^{(1)}_n(\epsilon) := \max_{j,A} \sup_{\tau \in I_\epsilon} \frac{1}{n}\sum_i |\widehat{f}_{a,n,k(n),\tau}(\widehat{U}_{i,A}) - q_{a,\tau}(U_{i,A})| \xrightarrow{P} 0 \tag{14}$$

for every fixed $\epsilon \in (0, 1/2)$. For empty A use the marginal empirical quantile. The same sequence works for all central intervals and pairs. The bounds hold simultaneously for every minimizer on common measurable events whose probabilities tend to one. For every fixed $0 < \beta \leq 1$, the central error in (9) *satisfies $\mathcal{R}_n(\epsilon) \leq \{\mathcal{R}^{(1)}_n(\epsilon)\}^\beta = o_P(1)$.*

Suprema over arbitrary minimizer selections are understood in outer probability.

The four auxiliary lemmas below control rank perturbations, quantile continuity, spline approximation, and prediction error. For nonempty A, the deterministic full tensor B-spline spaces $\mathcal{S}_{A,k}$ in Assumption 6.1(iv) contain constants and have fixed order at least two, uniform grids, simple interior knots, clamped endpoints, and mesh size tending to zero. With $J_k \to \infty$ univariate basis functions and standard tensor basis $B_{A,k}$, use $\Phi_{A,k} = J_k^{|A|/2} B_{A,k}$, so $m_{A,k} \asymp J_k^{|A|}$. The deterministic scalar basis normalization changes coefficients but not the fitted-value space.

Lemma D.1 (Rank and tensor-basis perturbation)**.** *Under i.i.d. sampling with continuous marginals, use the full tensor B-splines and deterministic normalization of Assumption 6.1. Uniformly over the finite collection of nonempty conditioning sets, with $d = |A|$ and $m \asymp J^d$,*

$$\Delta_n := \max_{i,\ell} |\widehat{U}_{i\ell} - F_{X_\ell}(X^{(i)}_\ell)| = O_P(n^{-1/2}),$$

$$\sup_u \|\Phi_A(u)\|_2 \leq C\sqrt{m}, \tag{15}$$

$$\max_i \|\Phi_A(\widehat{U}_{i,A}) - \Phi_A(U_{i,A})\|_2 \leq C\sqrt{m}\, J\Delta_n. \tag{16}$$

Proof. Let $\widehat{F}_\ell$ be the empirical CDF of coordinate ℓ. Continuity of the marginals eliminates ties almost surely, and $\widehat{U}_{i\ell} = n\widehat{F}_\ell(X_\ell^{(i)})/(n+1)$. Thus

$$\Delta_n \le \max_\ell \|\widehat{F}_\ell - F_\ell\|_\infty + (n+1)^{-1}.$$

The Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) and a union bound give the stated order. A fixed-order unscaled univariate B-spline vector has bounded norm, a fixed number of active coordinates, and Lipschitz constant $O(J)$. A telescoping tensor-product expansion gives norm $O(1)$ and Lipschitz constant $O(J)$ before scaling. Multiplication by $J^{d/2} \asymp \sqrt{m}$ gives both bounds, with $\|u-v\|_1 \le d\|u-v\|_\infty$ and $d \le p-1$. □

Lemma D.2 (Continuity and first-moment envelope)**.** *Under Assumption 6.1(ii), $\tau \mapsto q_{a,\tau}(u)$ is continuous on $(0,1)$ on the common full-measure set. Under (iii) it is continuous into $L^1(P_{U_A})$ on every I_ϵ. In particular its image there is compact. A finite response absolute first moment is sufficient, but not necessary, for (iii):*

$$M_{a,\epsilon}(U_A) \le \epsilon^{-1}\mathbb{E}(|Y| \mid U_A) \quad \textit{almost surely}. \tag{17}$$

Proof. For every admissible u and $\tau \in (0,1)$, continuity of the CDF and the generalized-inverse definition give $F_a(q_{a,\tau}(u) \mid u) = \tau$. For a fixed $u, \tau, t > 0$, the two strict inequalities in Assumption 6.1(ii) persist when the level is changed sufficiently little. The generalized inverse is then trapped between $q_\tau(u) - t$ and $q_\tau(u) + t$. This proves continuity. Differences of central quantiles are dominated by $2M_{a,\epsilon}$, so dominated convergence gives L^1 continuity. The continuous image of I_ϵ is compact. Monotonicity gives $M_{a,\epsilon} = \max(|q_{a,\epsilon}|, |q_{a,1-\epsilon}|)$. If an endpoint quantile is nonnegative, its upper tail has probability at least ϵ; if negative, its lower tail does. Hence its magnitude is at most $\epsilon^{-1}\mathbb{E}(|Y| \mid U_A)$. Taking the larger endpoint proves (17). □

Lemma D.3 (L^1 spline comparators)**.** *Under Assumption 6.1, fix I_ϵ and a pair (j, A) with $A \neq \varnothing$. There exist comparators $s^0_{k,\tau} = \Phi_{A,k}^\top \theta^0_k(\tau)$ such that*

$$b_{k,\epsilon} := \sup_{\tau \in I_\epsilon} \mathbb{E}|s^0_{k,\tau}(U_A) - q_{a,\tau}(U_A)| \longrightarrow 0. \tag{18}$$

For each fixed k, $\theta^0_k(\tau)$ can be chosen continuous in τ and bounded over I_ϵ.

Proof. The standard nonnegative B-splines form a partition of unity, and the supports shrink with the mesh. Assigning coefficients from values of a continuous function in the corresponding supports gives uniform approximation of every continuous function on the cube as $k \to \infty$. Continuous functions are dense in L^1 of any Borel probability on this compact cube: for example, regularity approximates indicators by continuous functions, and simple functions approximate integrable functions. Thus $\mathrm{dist}_{L^1}(h, \mathcal{S}_{A,k}) \to 0$ for every $h \in L^1$. A finite net of the compact family in Lemma D.2, together with the triangle inequality for distance to a subspace, gives

$$a_{k,\epsilon} := \sup_{\tau \in I_\epsilon} \mathrm{dist}_{L^1}(q_{a,\tau}, \mathcal{S}_{A,k}) \longrightarrow 0.$$

Set $\eta_k = a_{k,\epsilon} + 1/k$. Choose a finite grid in I_ϵ fine enough that the L^1 oscillation of $q_{a,\tau}$ within each grid interval is at most η_k. At every grid point choose a spline whose L^1 error is at most η_k, and interpolate its coefficient vectors linearly between adjacent grid points. For an intermediate τ, the triangle inequality gives error at most $2\eta_k$. This constructs (18), continuous coefficients, and a finite fixed-k coefficient bound. □

Write $F_u = F_a(\cdot \mid u)$ and $q_\tau = q_{a,\tau}$. For integrable f, define the loss-difference risk and excess

$$Q_\tau(f) = \mathbb{E}[\rho_\tau(Y - f(U_A)) - \rho_\tau(Y)], \qquad D_\tau(f) = Q_\tau(f) - Q_\tau(q_\tau). \tag{19}$$

Both risks are finite because the absolute loss difference is bounded by $|f(U_A)|$ or $|q_\tau(U_A)|$. The expectation is taken after subtracting the losses, so their individual expectations need not be finite. Conditional on $U_A = u$, put

$$d_{\tau,u}(z) = \int_{q_\tau(u)}^{z} \{F_u(t) - \tau\}\, dt. \tag{20}$$

Conditional integration of the check-loss difference gives $D_\tau(f) = \mathbb{E}d_{\tau,U_A}(f(U_A))$. The integrand has the appropriate sign on either side of the unique quantile. Consequently

$$0 \le D_\tau(f) \le \mathbb{E}|f(U_A) - q_\tau(U_A)|. \tag{21}$$

Lemma D.4 (From excess loss to L^1 prediction error)**.** *Fix I_ϵ under Assumption 6.1. For every $r > 0$ there is a $\lambda_{a,\epsilon}(r) > 0$, independent of f, τ, and k, such that*

$$\tau \in I_\epsilon, \quad f \in L^1(P_{U_A}), \quad D_\tau(f) \le \lambda_{a,\epsilon}(r) \quad \Longrightarrow \quad \mathbb{E}|f - q_\tau| < r. \tag{22}$$

In particular $\mathbb{E}|f - q_\tau|^\beta \le (\mathbb{E}|f - q_\tau|)^\beta$ for $0 < \beta \le 1$.

Proof. Let $\alpha = \epsilon/2$, $M = M_{a,\alpha}$, and define the auxiliary interval projection

$$Tf(u) = \min\{q_{1-\alpha}(u), \max\{q_\alpha(u), f(u)\}\}.$$

Let $g = Tf$. Since q_τ is inside that interval, convexity and monotonicity of (20) give

$$\begin{aligned} d_{\tau,u}(f(u)) - d_{\tau,u}(g(u)) &\ge \alpha|f(u) - g(u)|, \\ d_{\tau,u}(g(u)) &\ge 0. \end{aligned} \tag{23}$$

Here g is integrable because $|g| \le M_{a,\alpha}$. For example, above $q_{1-\alpha}$ one has $F_u(t) - \tau \ge \alpha$; below q_α one has $\tau - F_u(t) \ge \alpha$. Integrating these bounds proves the first inequality; the second follows because $q_\tau(u)$ minimizes the conditional pinball risk. Hence $D_\tau(g) \le D_\tau(f), \mathbb{E}|f - g| \le D_\tau(f)/\alpha$, and $|g - q_\tau| \le 2M$.

For any $\eta > 0$, define a measurable positive pointwise separation

$$m_{\epsilon,\eta}(u) = \min_{\tau \in I_\epsilon} \min\{d_{\tau,u}(q_\tau(u) - \eta), d_{\tau,u}(q_\tau(u) + \eta)\} > 0. \tag{24}$$

Indeed, strict crossing makes each displayed conditional excess positive. Continuity of F_u and $q_\tau(u)$ makes it continuous in τ; compactness then gives a strictly positive minimum for each admissible u. Its value equals the infimum over a fixed countable dense subset of I_ϵ, proving measurability. Put $\pi_{\epsilon,\eta}(c) = \mathbb{P}\{m_{\epsilon,\eta}(U_A) \le c\}$. Then $\pi_{\epsilon,\eta}(c) \to 0$ as $c \downarrow 0$. Monotonicity of conditional excess away from the quantile gives

$$\mathbb{P}(|g - q_\tau| > \eta) \le \pi_{\epsilon,\eta}(c) + D_\tau(f)/c \qquad (c > 0).$$

For $T > 0$, split according to $|g - q_\tau| \le \eta$ and $M \le T$. Using (23), obtain the explicit bound

$$\begin{aligned} \mathbb{E}|f - q_\tau| \le \left(\frac{2}{\epsilon} + \frac{2T}{c}\right) D_\tau(f) + \eta + 2T\pi_{\epsilon,\eta}(c) \\ + 2\mathbb{E}[M\mathbf{1}\{M > T\}]. \end{aligned} \tag{25}$$

To make this less than r, first choose $\eta < r/4$, then T making the last term less than $r/4$, then c making $2T\pi_{\epsilon,\eta}(c)$ less than $r/4$, and finally a positive $\lambda_{a,\epsilon}(r)$ making the first term less than $r/4$. All choices are common to τ and f. This proves (22). The final assertion is Jensen's inequality. □

Proof of Proposition D.1. Fix a pair (j, A) with $A \ne \varnothing$, and fix k and I_ϵ. Write $\Phi = \Phi_{A,k}$. The basis is bounded and Lipschitz, with finite constants

$$\zeta_k = \sup_u \|\Phi(u)\|_2, \qquad \ell_k = \sup_{u \ne v} \frac{\|\Phi(u) - \Phi(v)\|_2}{\|u - v\|_\infty}.$$

Lemma D.1 gives $\Delta_n = O_P(n^{-1/2})$ and $\max_i \|\Phi(\widehat{U}_{i,A}) - \Phi(U_{i,A})\|_2 \le \ell_k\Delta_n = o_P(1)$ at fixed k. Full support and linear independence imply

$$\kappa_k := \inf_{\|v\|_2 = 1} \mathbb{E}|\Phi(U_A)^\top v| > 0. \tag{26}$$

For otherwise compactness of the coefficient sphere would produce a nonzero continuous spline vanishing almost surely, contradicting positive probability of every interior open set where it is nonzero.

We next compare the empirical and population objectives. Write $U = U_A$ and $U_i = U_{i,A}$. For $f_\theta = \Phi^\top \theta$, let $g_{\tau,\theta}(Y, U) = \rho_\tau(Y - f_\theta(U)) - \rho_\tau(Y)$, and define

$$Q^E_{n,\tau}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{\rho_\tau(Y_i - \Phi(\widehat{U}_{i,A})^\top \theta) - \rho_\tau(Y_i)\}.$$

Subtracting the second term, which is constant in θ, does not change any empirical minimizer. For a fixed radius L, these functions have bounded envelope $L\zeta_k$. The identity

$$g_{\tau,\theta} = \tfrac{1}{2}\{|Y - f_\theta(U)| - |Y|\} - (\tau - \tfrac{1}{2}) f_\theta(U) \tag{27}$$

also gives the deterministic continuity bound

$$|g_{\tau,\theta} - g_{\tau',\theta'}| \le \zeta_k \|\theta - \theta'\|_2 + L\zeta_k |\tau - \tau'|.$$

A finite parameter net and the ordinary i.i.d. law of large numbers yield $\sup_{\tau \in I_\epsilon, \|\theta\|_2 \le L} |(P_n - P) g_{\tau,\theta}| = o_P(1)$. Adding the pathwise rank error, at most $L\ell_k \Delta_n$, proves

$$Z_{n,k,\epsilon}(L) := \sup_{\tau \in I_\epsilon, \|\theta\|_2 \le L} |Q^E_{n,\tau}(\theta) - Q_\tau(f_\theta)| = o_P(1), \tag{28}$$

The i.i.d. law of large numbers above applies to (Y_i, U_i).

We now bound the coefficients of every empirical minimizer. Let $\alpha = \epsilon/2$ and let T be the projection onto $[q_\alpha, q_{1-\alpha}]$ used in the proof of Lemma D.4. Applying (23) gives $D_\tau(f_\theta) \ge \alpha \mathbb{E}|f_\theta - T f_\theta| \ge \alpha(\mathbb{E}|f_\theta| - \mathbb{E}M_{a,\alpha})$. Also $Q_\tau(q_\tau) \ge -\mathbb{E}M_{a,\epsilon}$, and therefore

$$Q_\tau(f_\theta) \ge \alpha \kappa_k \|\theta\|_2 - \alpha \mathbb{E}M_{a,\alpha} - \mathbb{E}M_{a,\epsilon}. \tag{29}$$

Choose finite $L = L_{a,k,\epsilon}$ large enough to contain all comparator coefficients from Lemma D.3, and so that the right side of (29) exceeds 1 on the radius-L sphere. On the measurable event $\mathcal{E}_{n,a,k,\epsilon} = \{Z_{n,k,\epsilon}(L) < 1/2\}$ the empirical objective is positive on that sphere, uniformly in τ, while $Q^E_{n,\tau}(0) = 0$.

A minimizer of the original finite-sample objective exists for every $\tau \in (0,1)$: minimize the continuous coercive unanchored loss on the closed linear space of fitted-value vectors, using $\rho_\tau(y - z) \ge \min(\tau, 1-\tau)|y - z|$. Represent a minimizing vector by coefficients. If any minimizer lay on or outside the radius-L sphere, its segment to zero and convexity would yield a sphere point with objective at most zero, a contradiction. Thus *every* minimizer is inside the ball on $\mathcal{E}_{n,a,k,\epsilon}$, whose probability tends to one at fixed k, ϵ.

A jointly measurable minimizer selection exists. With a fixed measurable rank tie rule, the empirical objective is Borel in the observations and τ, and continuous and convex in θ. For $\lambda > 0$, its sum with $\lambda\|\theta\|_2^2$ has a unique jointly measurable minimizer by the measurable argmin theorem. As $\lambda \downarrow 0$, these minimizers converge to the unique minimum-norm element of the unpenalized argmin set: comparison with that element bounds their norms, and every cluster point minimizes the original objective.

On this event, optimality and (21) give, simultaneously for all minimizers,

$$\sup_{\tau \in I_\epsilon} D_\tau(\widehat{f}_{a,n,k,\tau}) \le b_{k,\epsilon} + 2 Z_{n,k,\epsilon}(L). \tag{30}$$

Here fitted functions are evaluated at true coordinates inside the population risk.

For fixed k, ϵ, L the functions $h_{\tau,\theta}(u) = |\Phi(u)^\top \theta - q_\tau(u)|$ have integrable envelope $L\zeta_k + M_{a,\epsilon}$. Their empirical process satisfies

$$V_{n,k,\epsilon}(L) := \sup_{\tau \in I_\epsilon, \|\theta\|_2 \le L} |(P_n - P) h_{\tau,\theta}| = o_P(1). \tag{31}$$

Indeed, a finite parameter net bounds the oscillation by $\zeta_k \eta + \omega_\eta(u)$, where

$$\omega_\eta(u) = \sup_{\substack{t,t' \in I_\epsilon \\ |t - t'| \le \eta}} |q_t(u) - q_{t'}(u)|.$$

The modulus is measurable by taking the supremum over a countable dense set. It tends to zero almost surely by uniform continuity on I_ϵ and is dominated by $2M_{a,\epsilon} \in L^1$. If $\mathcal{N}_\eta$ is a finite product net for the compact parameter set, then

$$V_{n,k,\epsilon}(L) \le \max_{(t,v)\in\mathcal{N}_\eta} |(P_n - P)h_{t,v}| + 2\zeta_k\eta + P_n\omega_\eta + P\omega_\eta.$$

For fixed η, ordinary laws of large numbers apply to all net points and to ω_η. Next let $\eta \downarrow 0$ and use dominated convergence to obtain (31). All these bounding processes have measurable separable versions. Consequently, on the localization event,

$$\begin{aligned}\sup_{\tau\in I_\epsilon} P_n|\widehat{f}_{a,n,k,\tau}(\widehat{U}_A) - q_\tau(U_A)| \le \sup_{\tau\in I_\epsilon} \mathbb{E}_U|\widehat{f}_{a,n,k,\tau}(U_A) - q_\tau(U_A)| \\ + V_{n,k,\epsilon}(L) + L\ell_k\Delta_n.\end{aligned} \tag{32}$$

$\mathbb{E}_U$ integrates the fixed predictor law holding the fitted coefficients fixed. Uniformity over the coefficient ball permits substitution of the fitted coefficients. Lemma D.4 bounds the first term.

For $A = \varnothing$, the fit is the marginal empirical quantile. By Lemma D.2, the marginal quantile map is continuous on $I_{\epsilon/2}$. Let F and F_n denote the marginal CDF of Y and its empirical CDF. Put $h_n = \|F_n - F\|_\infty + 1/n$; the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) gives $h_n = o_P(1)$. On $h_n < \epsilon/2$, the generalized-inverse inequalities bound every empirical check-loss minimizer at level $\tau \in I_\epsilon$ by $q_{\tau-h_n}$ and $q_{\tau+h_n}$. The extra $1/n$ covers the possible interval of minimizers at empirical jump levels. Uniform continuity of q on $I_{\epsilon/2}$ yields $\sup_{\tau\in I_\epsilon} |\widehat{q}_\tau - q_\tau| = o_P(1)$.

We now choose a common $k(n)$ for all pairs (j, A) and central intervals. In this step, $b_{a,k,\epsilon}$, $Z_{n,a,k,\epsilon}$, $V_{n,a,k,\epsilon}$, and $\ell_{a,k}$ display the pair index a suppressed above. For fixed k, choose a minimizer selection that is jointly measurable in the data and all quantile levels, and write

$$\mathcal{R}^{(1)}_{n,k}(\epsilon) := \max_{a=(j,A)} \sup_{\tau\in I_\epsilon} \frac{1}{n}\sum_i |\widehat{f}_{a,n,k,\tau}(\widehat{U}_{i,A}) - q_{a,\tau}(U_{i,A})|,$$

using the marginal empirical quantile when $A = \varnothing$. Let $\epsilon_l = 1/(l+2)$. For each l, use Lemma D.4 for the finitely many pairs (j, A) with $A \neq \varnothing$ to choose $d_l > 0$ such that $D_\tau(f) \le d_l$ implies $\mathbb{E}|f - q_\tau| < 1/(4l)$, uniformly on I_{ϵ_l}. Choose strictly increasing $k_l \ge l$ such that $\max_{a=(j,A):\, A\neq\varnothing} b_{a,k_l,\epsilon_l} \le d_l/2$. For each such pair, put $L_a = L_{a,k_l,\epsilon_l}$, the localization radius chosen above. The bounds above allow strictly increasing deterministic $N_l \ge l$ such that, for every $n \ge N_l$, on a common measurable event $\mathcal{H}_{n,l}$ with $\mathbb{P}(\mathcal{H}_{n,l}) \ge 1 - 1/l$, all of the following hold simultaneously: localization for every pair with nonempty conditioning set; $\max_{a=(j,A):\, A\neq\varnothing} Z_{n,a,k_l,\epsilon_l}(L_a) \le d_l/4$; $\max_{a=(j,A):\, A\neq\varnothing}\{V_{n,a,k_l,\epsilon_l}(L_a) + L_a\ell_{a,k_l}\Delta_n\} \le 1/(4l)$; and the empty-set prediction errors are at most $1/(4l)$. On this event, (30), Lemma D.4, and (32) show that every central prediction error is less than $1/l$, for every minimizer. In particular,

$$\sup_{\text{empirical minimizers}} \mathcal{R}^{(1)}_{n,k_l}(\epsilon_l) < 1/l \quad \text{on } \mathcal{H}_{n,l}, \qquad n \ge N_l.$$

Put $l(n) = \max\{l : N_l \le n\}$ and $k(n) = k_{l(n)}$, assigning arbitrary initial values before N_1. For any fixed $\epsilon > 0$, eventually $I_\epsilon \subseteq I_{\epsilon_{l(n)}}$. Since $l(n) \to \infty$, for any fixed $r > 0$, eventually $1/l(n) < r$ and

$$\mathbb{P}^*\{\mathcal{R}^{(1)}_n(\epsilon) > r\} \le \mathbb{P}(\mathcal{H}^c_{n,l(n)}) \le 1/l(n) \longrightarrow 0.$$

Here $\mathbb{P}^*$ denotes outer probability. Thus (14) holds. Jensen's inequality for each finite sample gives the asserted β-error bound. □

D.2 Proof of Lemma C.1

Proof. We first bound score error by prediction error. For $0 < \beta \le 1$, $\big||x|^\beta - |y|^\beta\big| \le |x - y|^\beta$ and $|x + y|^\beta \le |x|^\beta + |y|^\beta$. Therefore,

$$|\widehat{S}^U_{a,n,K_n} - \widetilde{S}^U_{a,n,K_n}| \le \frac{4}{nK_n}\sum_{i=1}^n\sum_{k=1}^{K_n} |\overline{q}_{i,k} - q_a(T_k \mid X^{(i)}_A)|^\beta.$$

Indeed, the response-to-quantile term changes by at most twice the average prediction error. In the pairwise term, every prediction error appears $2(K_n - 1)$ times, giving another twice the same average. This bound holds pathwise for any fitted predictions in place of $\overline{q}_{i,k}$. It remains to show that the average error on the right tends to zero.

Use $k(n)$ from Proposition D.1 and any deterministic $K_n \geq 2$ tending to infinity. The empirical L^1 convergence in that proposition and Jensen's inequality give $\mathcal{R}_n(\epsilon) = o_P(1)$ for each fixed $\epsilon \in (0, 1/2)$; supremum bounds are understood in outer probability. Fix $a = (j, A)$, put $q_{i,k} = q^U_{j|A}(T_k, U_{i,A})$, and write $\mathbb{E}_{n,K_n} h = (nK_n)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K_n} h_{i,k}$ for the average over observations and quantile levels. Write $\widehat{q}_{i,k} = \widehat{q}^E_{j|A}(T_k, \widehat{U}_{i,A})$. Let $\widehat{M}_{j,n} = n^{-1} \sum_i |X_j^{(i)}|^{2\beta}$. Set $\widehat{B}_a = \infty$ if $\mathbb{E}_{n,K_n} |\widehat{q}|^{2\beta} \leq 2\widehat{M}_{j,n}$; otherwise choose $\widehat{B}_a \geq 0$ satisfying

$$\mathbb{E}_{n,K_n} \min\{|\widehat{q}|, \widehat{B}_a\}^{2\beta} = 2\widehat{M}_{j,n}.$$

Write $C_b(x) = \max\{-b, \min(x, b)\}$, with $C_\infty(x) = x$. Then $\overline{q}_{i,k} = C_{\widehat{B}_a}(\widehat{q}_{i,k})$ and $\mathbb{E}_{n,K_n} |\overline{q}|^{2\beta} \leq 2\widehat{M}_{j,n}$.

First, the bounded prediction error satisfies

$$D_{a,n} := \mathbb{E}_{n,K_n} \min\{1, |\widehat{q} - q|^\beta\} \leq \mathcal{R}_n(\epsilon) + \frac{1}{K_n} \sum_k \mathbf{1}\{T_k \notin [\epsilon, 1 - \epsilon]\}.$$

The last term has expectation 2ϵ. Sending $n \to \infty$ and then $\epsilon \downarrow 0$ gives $\max_a D_{a,n} = o_P(1)$.

Next, we show that the clipping boundary diverges. For fixed $0 \leq b < \infty$, $g_b(x) = \min\{|x|, b\}^{2\beta}$ is bounded and uniformly continuous. For any $\eta > 0$, choose $0 < \delta \leq 1$ so that $|g_b(x) - g_b(y)| \leq \eta$ whenever $|x - y| \leq \delta$. Splitting the average at $|\widehat{q} - q| = \delta$ gives

$$\mathbb{E}_{n,K_n} |g_b(\widehat{q}) - g_b(q)| \leq \eta + 2b^{2\beta} \delta^{-\beta} D_{a,n}.$$

Letting $n \to \infty$ and then $\eta \downarrow 0$ shows that $\mathbb{E}_{n,K_n} g_b(\widehat{q}) - \mathbb{E}_{n,K_n} g_b(q) = o_P(1)$. The distributional identity $q^U_{j|A}(T, U_A) \overset{d}{=} X_j$ gives $\mathbb{E}[\mathbb{E}_{n,K_n} g_b(q)] = \mathbb{E} g_b(X_j)$. Terms sharing neither an observation nor a level are independent. At most $nK_n^2 + n^2 K_n$ covariance terms share an index, and each is bounded, so $\mathrm{Var}\{\mathbb{E}_{n,K_n} g_b(q)\} = O(n^{-1} + K_n^{-1})$. Consequently,

$$\mathbb{E}_{n,K_n} g_b(\widehat{q}) \xrightarrow{P} \mathbb{E} g_b(X_j) \leq M_j, \qquad \widehat{M}_{j,n} \xrightarrow{P} M_j := \mathbb{E}|X_j|^{2\beta} \in (0, \infty).$$

Since $2M_j - \mathbb{E} g_b(X_j) \geq M_j > 0$, where positivity follows from the continuous marginal, $\mathbb{P}\{\mathbb{E}_{n,K_n} g_b(\widehat{q}) < 2\widehat{M}_{j,n}\} \to 1$ for every fixed b. Hence $\widehat{B}_a \xrightarrow{P} \infty$.

The moment bound then converts the bounded error into convergence of the average β-powered prediction error. For fixed $L \geq 1$, on $\{\widehat{B}_a \geq L\}$ we have $C_L(\overline{q}) = C_L(\widehat{q})$. Contraction of C_L and subadditivity of $|\cdot|^\beta$ yield

$$\mathbb{E}_{n,K_n} |\overline{q} - q|^\beta \leq (2L)^\beta D_{a,n} + L^{-\beta} \{2\widehat{M}_{j,n} + \mathbb{E}_{n,K_n} |q|^{2\beta}\}.$$

The expression in braces has expectation $3M_j$. For fixed L, $\mathbb{P}(\widehat{B}_a < L) \to 0$ and $(2L)^\beta D_{a,n} = o_P(1)$. A union bound and Markov's inequality therefore give, for every $r > 0$,

$$\limsup_{n \to \infty} \mathbb{P}\{\mathbb{E}_{n,K_n} |\overline{q} - q|^\beta > r\} \leq \frac{6M_j}{rL^\beta}.$$

Letting $L \to \infty$ and taking a finite union over $a \in \mathcal{I}$ proves $\max_a \mathbb{E}_{n,K_n} |\overline{q} - q|^\beta = o_P(1)$. The rank and original conditional quantiles agree almost surely. The pathwise bound above now proves the result. □

D.3 Proof of Lemma C.2

Proof. We prove that, for each a there is a finite constant C_a, independent of n and K, such that

$$\mathbb{E}\big[(\widetilde{S}^U_{a,n,K} - S_a)^2\big] \leq C_a \left(\frac{1}{n} + \frac{1}{K(K-1)}\right).$$

The K^{-1} stochastic rate follows from cancellation for the true conditional quantiles under uniform sampling.

Fix $a = (j, A)$, abbreviate $X = (Y, Z) = (X_j, X_A)$, and set

$$h_a(X; u, v) := |q_a(u \mid Z) - Y|^\beta + |q_a(v \mid Z) - Y|^\beta - |q_a(u \mid Z) - q_a(v \mid Z)|^\beta.$$

Because each quantile level appears in exactly $K - 1$ unordered pairs,

$$\widetilde{S}^U_{a,n,K} = \frac{1}{n} \sum_{i=1}^{n} \binom{K}{2}^{-1} \sum_{1 \le k < \ell \le K} h_a(X^{(i)}; T_k, T_\ell).$$

Let $H_a(u, v) := \mathbb{E}_X h_a(X; u, v)$. If V is an independent uniform variable, then $q_a(V \mid Z) \mid Z$ has the conditional law of $Y \mid Z$. Consequently, for almost every u,

$$\begin{aligned} \mathbb{E}_V H_a(u, V) = \mathbb{E}|q_a(u \mid Z) - Y|^\beta + \mathbb{E}|q_a(V \mid Z) - Y|^\beta \\ - \mathbb{E}|q_a(u \mid Z) - q_a(V \mid Z)|^\beta = S_a. \end{aligned}$$

The first and third terms cancel conditionally on Z, while the middle term equals $\mathbb{E}|Y' - Y|^\beta = S_a$ for conditionally i.i.d. copies $Y, Y' \mid Z$. Thus the centered symmetric kernel $\overline{H}_a(u, v) := H_a(u, v) - S_a$ is first-order degenerate: $\mathbb{E}_V \overline{H}_a(u, V) = 0$ almost everywhere. It follows that

$$\mathrm{Var}\left[\binom{K}{2}^{-1} \sum_{k<\ell} H_a(T_k, T_\ell) \right] = \frac{2\,\mathbb{E}[\overline{H}_a(T, V)^2]}{K(K-1)}. \tag{33}$$

Pairs of kernel terms with disjoint indices are independent, and pairs sharing exactly one index have zero covariance by the first-order degeneracy. Only identical unordered pairs contribute to the variance, which gives the displayed identity.

The required second moments are finite. Indeed, $q_a(T \mid Z)$ has the marginal distribution of Y, and, for a constant $C_\beta < \infty$,

$$|h_a(X; T, V)|^2 \le C_\beta \big(|Y|^{2\beta} + |q_a(T \mid Z)|^{2\beta} + |q_a(V \mid Z)|^{2\beta} \big).$$

Conditional on all sampled quantile levels, the data summands are i.i.d. Jensen's inequality over the unordered pairs of quantile levels therefore gives

$$\begin{aligned} \mathbb{E}\Big[\mathrm{Var}(\widetilde{S}^U_{a,n,K} \mid T_1, \dots, T_K) \Big] &\le \frac{1}{n} \mathbb{E}\left[\binom{K}{2}^{-1} \sum_{k<\ell} h_a(X; T_k, T_\ell)^2 \right] \\ &\le \frac{C_a}{n}. \end{aligned}$$

Also $\mathbb{E}[\widetilde{S}^U_{a,n,K} \mid T_1, \dots, T_K]$ is the U-statistic in (33), and its mean is S_a. The law of total variance proves the mean-square bound. Chebyshev's inequality and a finite union bound give the stated $O_P(n^{-1/2} + K^{-1})$ rate uniformly over $\mathcal{I}$. □

E Additional Comparison of Identifying Conditions

This section compares PASS ordering with variance and R^2 sortability and with model-specific conditions for heteroskedastic ordering.

E.1 PASS AND SORTABILITY

Variance and R^2 sortability quantify how often marginal variance and R^2, respectively, increase along directed paths; they are graded diagnostics rather than requirements that every path be ordered correctly (Reisach et al., 2021; 2023). For a jointly Gaussian vector, write

$$v_j = \mathrm{Var}(X_j), \qquad d_{j,R} = \mathrm{Var}(X_j \mid X_{R\setminus\{j\}}).$$

The conditional variance is constant, and the pairwise-distance representation of PASS gives

$$\Omega_{j,R} = c_\beta d_{j,R}^{\beta/2}, \qquad R_j^2 = 1 - \frac{d_{j,V}}{v_j},$$

where $c_\beta = \mathbb{E}|Z - Z'|^\beta > 0$ for independent standard normal variables Z, Z'. Here R_j^2 uses regression on all other variables, not just the parents. Var-SortnRegress and R^2-SortnRegress sort once by v_j and R_j^2, whereas PASS recomputes $d_{j,R}$ after each sink removal.

Consider the two-node model

$$X = \varepsilon_X, \qquad Y = aX + \varepsilon_Y, \qquad \varepsilon_X, \varepsilon_Y \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad a \neq 0,\ \sigma > 0.$$

The regression scores coincide: $R_X^2 = R_Y^2 = a^2/(1 + a^2)$. In contrast, for every $0 < \beta \leq 1$,

$$S_\beta(Y \mid X) = c_\beta \sigma^\beta > c_\beta \sigma^\beta (1 + a^2)^{-\beta/2} = S_\beta(X \mid Y).$$

Thus R^2 sorting leaves the direction undetermined, whereas PASS selects Y as the sink and recovers $X \to Y$. Since $S_\beta(X) = S_\beta(Y \mid X)$, the edgewise condition in Corollary 5.1 holds with equality.

E.2 PASS AND HETEROSKEDASTIC ORDERING CONDITIONS

We compare PASS with two recent heteroskedastic ordering-based methods, HOST (Duong & Nguyen, 2023) and SkewScore (Lin et al., 2025). The comparison is not a strict set-inclusion statement: the methods impose different sufficient conditions and use different distributional properties to determine the order.

HOST. HOST is formalized for Gaussian heteroskedastic causal models

$$X_i = \mu_i(X_{\mathrm{Pa}(i)}) + \sigma_i(X_{\mathrm{Pa}(i)})E_i, \qquad E_i \sim N(0, 1).$$

Its ordering condition is conditional normality restoration: for a non-descendant conditioning set C in the current ordering step,

$$\frac{X_i - \mathbb{E}[X_i \mid X_C]}{\sqrt{\mathrm{Var}(X_i \mid X_C)}} \sim N(0, 1) \quad \Longleftrightarrow \quad \mathrm{Pa}(i) \subseteq C.$$

Thus, after an ancestral set C has been selected, a node in the reduced graph whose standardized conditional variable is Gaussian is identified as a source node. HOST also requires each parent to affect both the location and scale mechanisms nontrivially,

$$\partial_{X_j}\mu_i \not\equiv 0, \qquad \partial_{X_j}\sigma_i \not\equiv 0 \qquad (j \in \mathrm{Pa}(i)).$$

Therefore ANMs are outside the formal HOST condition, since their scale does not depend on the parents. Once an ordering is obtained, HOST recovers the DAG using conditional-independence tests.

SkewScore. SkewScore studies heteroskedastic symmetric-noise models

$$X_i = f_i(X_{\mathrm{Pa}(i)}) + \sigma_i(X_{\mathrm{Pa}(i)})N_i,$$

where the exogenous noises are mutually independent and symmetric. Its ordering condition is a sink characterization by density-score skewness. For the current node set R, with joint density p_R,

$$j \in \mathrm{sink}(G[R]) \quad \Longleftrightarrow \quad \mathrm{SkewScore}_{x_j}(p_R) = 0,$$

while non-sinks have nonzero score-coordinate skewness under the stated nondegeneracy condition. The algorithmic nodewise criterion uses the magnitude

$$\mathrm{SkewScore}_{x_j}(p) = \left|\mathbb{E}\left\{(\partial_{x_j} \log p(X))^3\right\}\right|.$$

This requires smooth positive densities and differentiable mechanisms so that $\nabla \log p(x)$ is well-defined. SkewScore covers some nonlinear ANMs, but the linear Gaussian ANM boundary is degenerate for the criterion: with a linear mechanism and a symmetric cause distribution, the bivariate nondegeneracy condition fails.

PASS. PASS uses the edgewise quantile-separation condition in Corollary 5.1 to recover the order by recursive sink removal, using the nodewise scores defined in Section 5. For the HNM $X_j = \mu_j(X_{\mathrm{Pa}(j)}) + \sigma_j(X_{\mathrm{Pa}(j)})\varepsilon_j$, with $\sigma_j(\cdot) > 0$, independent errors, and $0 < \Gamma_\beta(\varepsilon_j) < \infty$, normalize without loss of generality so that $\Gamma_\beta(\varepsilon_j) = 1$. Then $S_\beta(X_j \mid X_{\mathrm{Pa}(j)}) = \mathbb{E}[\sigma_j(X_{\mathrm{Pa}(j)})^\beta]$, so the PASS ordering condition becomes

$$u \to v \quad \Longrightarrow \quad \mathbb{E}[\sigma_u(X_{\mathrm{Pa}(u)})^\beta] \le \mathbb{E}[\sigma_v(X_{\mathrm{Pa}(v)})^\beta].$$

This condition depends on the conditional scale moment and not on the structural location function μ_j. A population construction motivating the scale-calibrated experiments is: write $\sigma_j(x) = a_j r_j(x)$, with $a_j > 0$, $r_j(x) > 0$, and $\mathbb{E}[r_j(X_{\mathrm{Pa}(j)})^\beta] = 1$. If the scale levels a_j are chosen as a nondecreasing function of graph depth, then every edge satisfies $u \to v \Rightarrow a_u \le a_v$, and the displayed edgewise inequality follows. This construction is only a simulation-level sufficient condition; the population assumption remains the edgewise scale-moment inequality.

For an ANM $X_j = \mu_j(X_{\mathrm{Pa}(j)}) + \varepsilon_j$, the same identity reduces to $S_\beta(X_j \mid X_{\mathrm{Pa}(j)}) = \Gamma_\beta(\varepsilon_j)$. Thus the PASS condition ranks the error-separation levels $\Gamma_\beta(\varepsilon_j)$, and identical error distributions make the condition hold with equality along every edge. More generally, for a common-shape location-scale family $\varepsilon_j = b_j + a_j Z$, translation invariance and scale equivariance give $\Gamma_\beta(\varepsilon_j) = |a_j|^\beta \Gamma_\beta(Z)$, so the common factor cancels and the condition orders $|a_j|$; in the Gaussian case, this is equivalent to variance ordering. Different error distributions across nodes are also allowed when $\Gamma_\beta(\varepsilon_j)$ is nondecreasing along directed edges. Therefore PASS recovers equal-error ANM and Gaussian variance-ordering conditions as special cases (Peters & Bühlmann, 2014; Xu et al., 2024), while also allowing more general error-separation orderings driven by scale, tail shape, or both.

Estimation implications. The methods estimate different quantities. HOST estimates conditional means and variances and applies a normality statistic to standardized conditional variables. SkewScore estimates the joint density score $\nabla \log p(x)$ and computes cubic moments of its coordinates; its implementation uses sliced score matching with a multilayer perceptron (MLP) score network (Lin et al., 2025). PASS avoids density-score estimation and normality testing: it fits quantile regressions and evaluates the resulting average pairwise separation. Checking the edgewise quantile-separation condition requires knowledge of the parent sets and their conditional distributions. PASS uses the same quantile-separation functional for ordering and parent recovery.

F PASS CONFIGURATION

Unless an ablation states otherwise, PASS uses $\beta = 0.3$, three folds, and $K_n = \max\{2, \lceil \log n \rceil\}$ independent $\mathrm{Uniform}(0,1)$ quantile levels, where $\log$ is the natural logarithm and n is the full sample size. One level vector is shared across all ordering and full/reduced parent fits within a run, independently of the data and fold randomization. Off-diagonal pairs use denominator $K_n(K_n - 1)$, and the implementation multiplies the score by one-half.

Predictors are rank-transformed as $\widehat{U}_{i\ell} = \mathrm{rank}(X_{i\ell})/(n+1)$ using the full sample and average ranks for ties before fold splitting. Within each training fold, continuous predictors are standardized and clipped to $[-5, 5]$. We use additive cubic splines with four equally spaced knots over the training range, constant extrapolation, and a softplus-smoothed pinball loss with an ℓ_2 penalty of weight 1.5×10^{-3} on non-intercept coefficients. Ordering responses retain their specified scale; parent responses are mean-centered and robust-scaled on the full sample before fold splitting. Fitted quantiles are clipped to $[Q_{0.01}(y_{\mathrm{tr}}) - 3s_y, Q_{0.99}(y_{\mathrm{tr}}) + 3s_y]$, where $s_y = 1.4826\,\mathrm{MAD}(y_{\mathrm{tr}})$ with sample-standard-deviation and unit-scale fallbacks, and sorted across levels. Bandwidths and clipping bounds use training-fold responses. Here y_{tr} denotes the training-fold responses, Q_τ their empirical τ-quantile, and MAD the median absolute deviation.

Parent recovery uses the standard-error-based threshold in Section 6.2, with $\alpha = 0.001$.

G Additional Experimental Results

G.1 Experimental Details

Replication and reporting. Simulation comparisons use ten matched seeds per condition (indices 0–9), except UUMC, which uses 100 per generator setting (indices 0–99). On each Danube and Sachs dataset, PASS uses ten algorithmic seeds and each comparator is run once. CausalRivers results are aggregated over subgraphs. Curves show means; tables report means and sample standard deviations where available.

Simulation settings. ERd denotes an Erdős–Rényi DAG with independent order-compatible edges, each included with probability $2d/(p-1)$. This gives dp expected edges and an expected average of d parents per node. Unless otherwise stated, the main benchmark uses ER1 graphs with $p = 50$, $n = 1000$, and linear rate 0.5; the sortability and true-order parent-selection comparisons use the same datasets. The ER3 comparison retains $p = 50$ and $n = 1000$, while the linear-rate sweeps use $p = 10$. The main noise families are Gaussian $N(0, 0.75^2)$, standard Cauchy, and uncentered Pareto(2), with $\mathbb{P}(\varepsilon > t) = t^{-2}$ for $t \geq 1$.

For linear rate λ, each non-root node independently uses a linear location function with probability λ and otherwise a tanh-MLP function following Hiremath et al. (2025). Linear functions use independently signed $\mathrm{Uniform}(0.5, 2)$ coefficients, with the summed contribution divided by $\sqrt{|\mathrm{Pa}(j)|}$. Nonlinear functions use a 10-unit tanh MLP with $\mathrm{Uniform}(-5, 5)$ weights. Hidden biases follow the same distribution in ANMs and are zero in HNMs; HNM location outputs are empirically centered. The sample-size experiments instead use additive-GP nonlinear locations at $p \in \{10, 30\}$ for ER1/ER2 and $p = 10$ for ER3, retaining linear rate 0.5. These locations sum independent, empirically centered draws from zero-mean GPs over parents, with RBF kernel $k(x, x') = \exp\{-(x - x')^2/2\}$.

ANMs use i.i.d. additive errors. The scale-calibrated HNM uses

$$X_j = f_j(X_{\mathrm{Pa}(j)}) + \sigma_j(X_{\mathrm{Pa}(j)})\varepsilon_j, \qquad \sigma_j(x) = a_j \tilde{r}_j(x),$$

where $a_j = 1.0 + 0.20 d_j$ and d_j is graph depth. The positive tanh-softplus function $\tilde{r}_j$ takes standardized parent coordinates and is normalized by its arithmetic mean on a resampled pilot set of 5,000 standardized parent observations. This calibration favors PASS's ordering condition but does not enforce the population β-moment normalization or guarantee the edgewise inequality.

Evaluation metrics. For graph recovery, we compute structural Hamming distance (SHD) as the number of false positive and false negative directed edges, counting an edge reversal as two errors. F1 is computed on directed parent-to-child adjacencies. For ordering recovery, we report A_{top}, the accuracy version of the D_{top} metric of Rolland et al. (2022). If E is the true directed edge set and $\hat{\pi}$ is the estimated ordering, then

$$A_{\mathrm{top}} = 1 - \frac{D_{\mathrm{top}}}{|E|} = \frac{1}{|E|} \sum_{u \to v \in E} \mathbf{1}\{\hat{\pi}(u) < \hat{\pi}(v)\},$$

with $A_{\mathrm{top}} = 1$ when $E = \varnothing$. Thus higher A_{top} indicates that more true directed edges are oriented consistently by the estimated topological ordering.

Official-code baselines. DirectLiNGAM uses the `lingam` package with the returned adjacency transposed to our parent-to-child convention and thresholded at 10^{-12} (Shimizu et al., 2011). TOPIC uses the official implementation for continuous data with generalized additive model (GAM) scores and extra refinement (Xu et al., 2025). FICO uses the official implementation with Fisher-information sink peeling and kernel Stein scores with $\eta_G = 10^{-3}$, using kernel conditional independence (KCI) parent pruning with level 0.05 and subsampling size $\min(n, 1000)$ in the HNM setting, and CAM pruning in the ANM setting (Chen et al., 2026). CAM uses the R package with SEMGAM scores, `selGamBoost` variable selection, `selGam` pruning, and cutoff 10^{-3} (Bühlmann et al., 2014). SCORE and NoGAM use the `dodiscover` implementations on column-standardized inputs, with `prune=True` and $\alpha = 10^{-3}$; this invokes the package's Python/`pyGAM` CAM-pruning step (Rolland et al., 2022; Montagna et al., 2023). HOST is used in both ANM and HNM settings of the main $p = 50$ benchmark, in the other HNM experiments, and in real-data benchmarks through the official code with $\eta = 10^{-4}$, learning rate 0.1, maximum 30 epochs, batch size 2048, and cutoff 10^{-3} (Duong & Nguyen, 2023).

Reimplemented baselines. SkewScore was reimplemented because no confirmed official GitHub code was available at the time of our experiments. We followed the algorithmic description and reported hyperparameter settings in the original paper as closely as possible. SkewScore follows skewness-based sink peeling; HNM runs following the paper setup estimate the density score by sliced score matching with an MLP having three hidden layers, batch size 128, learning rate 10^{-3}, and early stopping. When KCI pruning is used for SkewScore, it uses level 0.05 with subsampling size $\min(n, 1000)$ (Lin et al., 2025).

G.2 Additional Simulations

We expand the results in Table 1 and Figure 2 across graph densities, linear rates, and sample sizes, and examine an HNM setting without depth-based scale calibration.

G.2.1 Performance and Runtime of the Main Benchmarks

Tables 5 and 6 report SHD and F1 with methods as rows and ANM/HNM results in separate panels. Entries are mean (SD). Bold marks the best reported mean within each model, noise family, and metric, with ties bolded.

Table 5: SHD and F1 in the main $p = 50$, $n = 1000$ synthetic benchmark, reported as mean (SD). Lower SHD and higher F1 are better; bold marks the best reported mean in each column, with all ties bolded. The means are also shown in Table 1.

	Gaussian		Pareto(2)		Cauchy	
Method	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑
ANM						
PASS	**32.4 (6.6)**	0.682 (0.052)	**24.0 (9.9)**	**0.753 (0.094)**	**23.1 (6.0)**	**0.759 (0.061)**
FICO	37.0 (9.6)	**0.704 (0.068)**	69.3 (15.1)	0.557 (0.069)	90.4 (10.1)	0.388 (0.068)
TOPIC	62.3 (10.7)	0.483 (0.061)	103.9 (12.8)	0.184 (0.048)	108.4 (11.8)	0.055 (0.033)
CAM	50.6 (15.5)	0.578 (0.098)	105.3 (12.0)	0.269 (0.051)	128.4 (14.0)	0.100 (0.028)
SCORE	208.9 (36.9)	0.186 (0.041)	607.2 (39.5)	0.104 (0.020)	678.2 (44.5)	0.071 (0.017)
NoGAM	217.8 (36.1)	0.181 (0.040)	613.4 (33.7)	0.106 (0.015)	683.9 (42.2)	0.072 (0.016)
DirectLiNGAM	104.3 (24.9)	0.351 (0.081)	55.1 (13.8)	0.587 (0.087)	48.2 (8.9)	0.525 (0.093)
HOST	134.0 (37.4)	0.338 (0.073)	164.0 (20.0)	0.113 (0.023)	150.1 (53.5)	0.208 (0.045)
SkewScore	108.0 (17.4)	0.295 (0.060)	114.0 (19.9)	0.359 (0.061)	139.6 (21.6)	0.199 (0.047)
HNM						
PASS	**24.2 (5.0)**	**0.751 (0.055)**	**18.6 (3.0)**	**0.807 (0.026)**	**19.8 (6.0)**	**0.790 (0.057)**
FICO	418.7 (64.2)	0.186 (0.040)	180.1 (14.7)	0.315 (0.018)	67.4 (11.8)	0.424 (0.072)
TOPIC	77.9 (9.1)	0.351 (0.058)	104.6 (7.6)	0.158 (0.028)	104.6 (14.0)	0.053 (0.037)
CAM	67.8 (14.6)	0.440 (0.108)	116.8 (7.9)	0.218 (0.048)	128.1 (12.6)	0.082 (0.029)
SCORE	204.7 (40.6)	0.151 (0.040)	601.4 (42.6)	0.102 (0.023)	683.4 (31.0)	0.063 (0.011)
NoGAM	202.1 (43.2)	0.150 (0.059)	598.1 (35.5)	0.110 (0.024)	684.4 (28.8)	0.065 (0.010)
DirectLiNGAM	85.7 (17.4)	0.411 (0.077)	68.3 (12.0)	0.523 (0.059)	49.6 (9.9)	0.496 (0.071)
HOST	115.3 (42.1)	0.402 (0.112)	152.5 (18.0)	0.123 (0.055)	116.0 (18.4)	0.243 (0.047)
SkewScore	92.3 (17.9)	0.329 (0.056)	120.7 (17.0)	0.313 (0.055)	145.4 (25.9)	0.151 (0.052)

Method	Pooled runtime (s)
PASS	1541.1 (1566.5)
FICO	1990.3 (1169.3)
TOPIC	100.6 (26.7)
CAM	93.0 (24.7)
SCORE	1030.6 (64.4)
NoGAM	1229.3 (73.3)
DirectLiNGAM	11.3 (1.8)
HOST	1363.4 (486.4)
SkewScore	886.0 (152.7)

Table 6: SHD and F1 in the $p = 10$, $n = 1000$, ER1 linear-rate benchmark at linear rate 0.5, reported as mean (SD). Lower SHD and higher F1 are better; bold marks the best reported mean in each column, with all ties bolded. HOST is evaluated only under HNM. The entries correspond to the linear-rate 0.5 slice of Figures 3 and 4.

	Gaussian		Pareto(2)		Cauchy	
Method	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑
ANM						
PASS	4.2 (5.8)	0.832 (0.180)	**1.6 (1.5)**	**0.926 (0.059)**	**2.8 (2.7)**	**0.856 (0.122)**
FICO	**3.3 (5.1)**	**0.893 (0.156)**	4.3 (3.1)	0.829 (0.084)	8.4 (5.0)	0.600 (0.176)
TOPIC	10.6 (6.3)	0.503 (0.249)	19.8 (8.3)	0.160 (0.060)	19.5 (8.5)	0.062 (0.075)
CAM	12.0 (9.7)	0.520 (0.260)	18.9 (7.7)	0.229 (0.132)	20.8 (7.8)	0.114 (0.094)
SCORE	17.6 (14.8)	0.395 (0.269)	14.4 (8.1)	0.496 (0.126)	18.7 (13.3)	0.291 (0.174)
NoGAM	16.9 (12.2)	0.375 (0.196)	13.8 (9.3)	0.526 (0.133)	18.2 (13.1)	0.304 (0.179)
DirectLiNGAM	15.3 (9.7)	0.436 (0.152)	4.5 (2.0)	0.770 (0.112)	5.3 (2.9)	0.691 (0.164)
SkewScore	17.6 (15.7)	0.410 (0.267)	17.3 (10.1)	0.364 (0.178)	18.0 (9.6)	0.287 (0.147)
HNM						
PASS	**3.2 (3.9)**	**0.873 (0.127)**	**2.2 (2.1)**	**0.898 (0.084)**	**3.3 (2.7)**	**0.844 (0.096)**
FICO	9.7 (6.6)	0.657 (0.150)	4.2 (2.6)	0.828 (0.072)	6.8 (4.5)	0.670 (0.164)
TOPIC	13.4 (8.0)	0.420 (0.195)	19.5 (7.4)	0.160 (0.159)	16.8 (6.4)	0.063 (0.089)
CAM	12.5 (6.7)	0.477 (0.206)	18.9 (9.4)	0.247 (0.110)	18.1 (6.1)	0.140 (0.097)
SCORE	16.7 (9.1)	0.352 (0.142)	14.2 (8.2)	0.496 (0.123)	15.6 (8.4)	0.333 (0.094)
NoGAM	16.7 (9.0)	0.352 (0.116)	14.3 (8.7)	0.511 (0.162)	15.7 (8.4)	0.325 (0.086)
DirectLiNGAM	15.2 (5.3)	0.335 (0.188)	7.5 (3.6)	0.653 (0.145)	6.8 (3.9)	0.583 (0.163)
HOST	11.9 (11.2)	0.587 (0.250)	21.1 (10.7)	0.191 (0.101)	19.9 (8.6)	0.226 (0.132)
SkewScore	15.7 (9.9)	0.441 (0.175)	14.7 (7.5)	0.472 (0.068)	13.8 (6.5)	0.417 (0.119)

Method	Pooled runtime (s)
PASS	15.0 (2.4)
FICO	59.0 (50.4)
TOPIC	6.3 (3.3)
CAM	14.2 (6.5)
SCORE	9.0 (4.2)
NoGAM	15.8 (7.4)
DirectLiNGAM	0.2 (0.1)
HOST	88.1 (32.3)
SkewScore	213.9 (81.4)

Runtime interpretation. Runtime is reported as mean (SD), pooled over the displayed noise families and model classes; HOST at $p = 10$ uses only HNM. The $p = 10$ PASS times were remeasured on one CPU thread using the same data, seeds, and settings, reproducing the reported graphs; baseline times come from the original runs. The $p = 50$ runs also used one CPU thread per method call. Timings reflect the respective shared-system execution conditions.

G.2.2 ER3 GRAPH-DENSITY COMPARISON AT $p = 50$

To assess graph-density sensitivity at the main benchmark dimension, we change the ER coefficient from one to three while retaining $p = 50$, $n = 1000$, linear rate 0.5, and the same ANM/HNM mechanisms and Gaussian, Pareto(2), and Cauchy noise families. We use the methods in Table 1, including HOST in both model classes, with their unchanged configurations. Dataset seed indices and PASS fold and quantile seeds match the ER1 benchmark; the ER3 observations are generated afresh and shared across methods. Graph recovery is evaluated by SHD and F1, without a runtime comparison. This protocol is separate from the $p = 10$ additive-GP ER3 sample-size experiment in Appendix G.2.6.

Table 7 reports the comparison. PASS has the lowest mean SHD in all six settings, whereas CAM leads F1 for Gaussian ANMs and DirectLiNGAM for the other settings. PASS selects substantially fewer edges than the true graphs, so its SHD advantage does not imply superior recall or F1.

Table 7: SHD and F1 at $p = 50$, $n = 1000$, ER3, reported as mean (SD). Lower SHD and higher F1 are better; bold marks the best reported mean in each column within each model class, with all ties bolded. Configurations match the ER1 comparison in Table 5.

	Gaussian		Pareto(2)		Cauchy	
Method	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑
ANM						
PASS	**152.0 (17.4)**	0.351 (0.059)	**121.4 (20.1)**	0.465 (0.085)	**97.2 (16.2)**	0.557 (0.074)
FICO	219.3 (33.8)	0.436 (0.054)	165.4 (38.4)	0.567 (0.076)	184.3 (29.2)	0.446 (0.059)
TOPIC	177.1 (20.7)	0.377 (0.045)	231.3 (13.0)	0.173 (0.034)	236.7 (23.7)	0.076 (0.034)
CAM	163.1 (22.4)	**0.472 (0.069)**	207.7 (26.4)	0.352 (0.063)	246.1 (19.3)	0.153 (0.041)
SCORE	658.2 (80.2)	0.142 (0.027)	790.1 (42.7)	0.182 (0.014)	852.9 (25.0)	0.133 (0.009)
NoGAM	660.3 (85.1)	0.140 (0.031)	778.6 (52.4)	0.195 (0.023)	846.1 (38.1)	0.137 (0.013)
DirectLiNGAM	355.0 (56.0)	0.317 (0.046)	130.5 (31.1)	**0.645 (0.075)**	97.9 (18.8)	**0.621 (0.073)**
HOST	462.6 (124.8)	0.195 (0.046)	469.6 (92.3)	0.109 (0.042)	367.4 (82.8)	0.212 (0.056)
SkewScore	374.9 (30.7)	0.207 (0.035)	336.0 (28.7)	0.285 (0.037)	385.7 (24.5)	0.166 (0.028)
HNM						
PASS	**125.8 (14.6)**	0.434 (0.059)	**104.5 (19.6)**	0.529 (0.079)	**107.1 (13.8)**	0.482 (0.073)
FICO	429.3 (51.0)	0.309 (0.038)	262.5 (27.7)	0.399 (0.035)	143.6 (17.2)	0.433 (0.059)
TOPIC	187.2 (19.2)	0.291 (0.047)	223.0 (21.4)	0.134 (0.038)	225.3 (23.4)	0.053 (0.038)
CAM	164.6 (24.3)	0.442 (0.070)	220.4 (30.3)	0.285 (0.067)	245.9 (17.3)	0.110 (0.016)
SCORE	540.8 (96.3)	0.180 (0.038)	720.7 (23.9)	0.199 (0.018)	814.8 (72.1)	0.126 (0.017)
NoGAM	554.9 (73.6)	0.174 (0.029)	714.1 (24.9)	0.205 (0.022)	813.8 (64.8)	0.125 (0.016)
DirectLiNGAM	196.3 (44.3)	**0.495 (0.089)**	168.2 (36.2)	**0.546 (0.074)**	108.4 (13.7)	**0.540 (0.084)**
HOST	422.4 (161.9)	0.214 (0.066)	365.8 (53.4)	0.151 (0.039)	303.3 (57.2)	0.238 (0.066)
SkewScore	268.1 (40.8)	0.280 (0.057)	255.5 (29.4)	0.344 (0.042)	339.4 (56.2)	0.160 (0.031)

G.2.3 Full ANM/HNM Linear Rate Sweeps

This section compares the linear-rate sweep with the main $p = 50, n = 1000$, ER1, linear-rate 0.5 synthetic benchmark reported in Table 1 and Table 5. We use matched $p = 10$ ANM/HNM sweeps with the nodewise linear/tanh-MLP generators described in Appendix G.1. The ANM sweep compares PASS, FICO, TOPIC, CAM, SCORE, NoGAM, DirectLiNGAM, and SkewScore; the HNM sweep compares all nine methods. Figures 3 and 4 report the full SHD and F1 trajectories across linear rates.

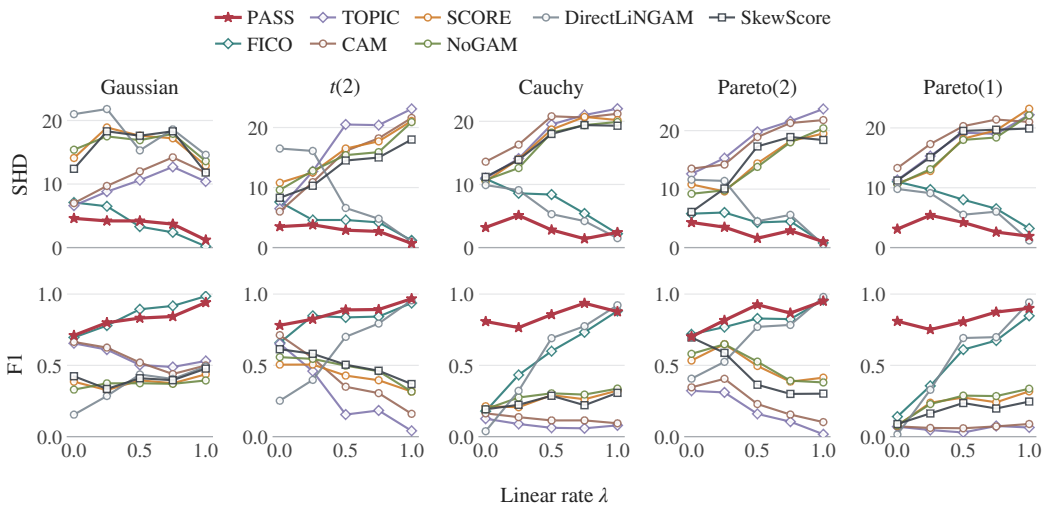


Figure 3: ANM linear rate sweep across Gaussian, $t(2)$, Cauchy, Pareto(2), and Pareto(1) noise. The setting uses $p = 10$, $n = 1000$, and an ER1 graph. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

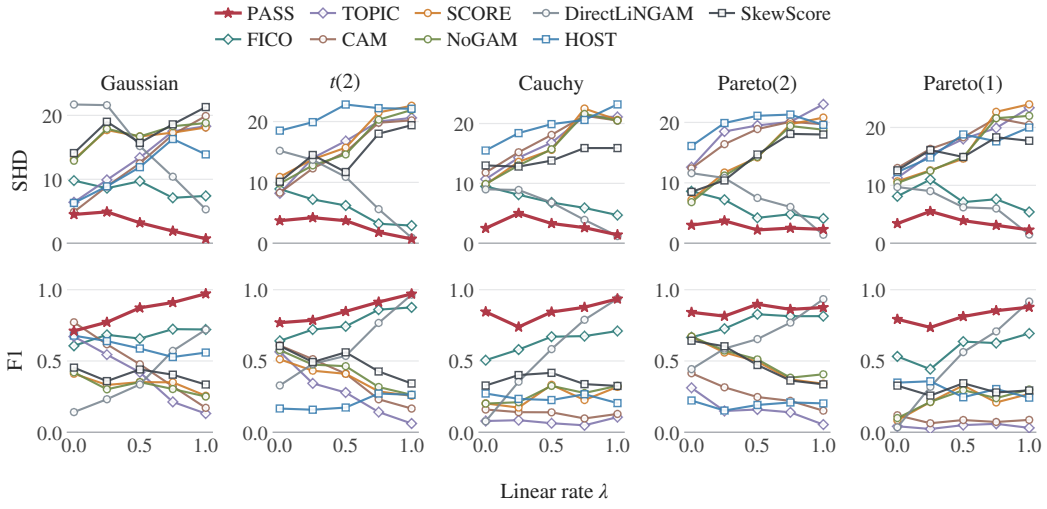


Figure 4: HNM linear rate sweep across Gaussian, $t(2)$, Cauchy, Pareto(2), and Pareto(1) noise. The setting uses $p = 10$, $n = 1000$, an ER1 graph, and the scale-calibrated HNM setting. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

G.2.4 ER1 Sample-Size Trends

This section complements the main benchmark by checking whether graph recovery accuracy improves as n increases. We use additive-GP mechanisms, an ER1 graph, linear rate 0.5, and the sample-size grid $n \in \{100, 250, 500, 1000, 2000, 3000\}$, with Gaussian and Cauchy noise. The displayed n-axis is logarithmic. ANM runs use i.i.d. additive errors, and HNM runs use the scale-calibrated HNM setting described in Appendix G.1. The nodewise scale calibration is designed to approximate the edgewise quantile-separation condition in Corollary 5.1. Figures 5 and 6 report SHD and F1 separately for ANM and HNM. All methods use matched dataset seeds at each displayed sample size. For DirectLiNGAM under ER1 ANMs at $p \in \{10, 30\}$ and $n \geq 250$, the inputs are the same cached arrays used by PASS, with the DirectLiNGAM implementation and edge threshold specified in Appendix G.1.

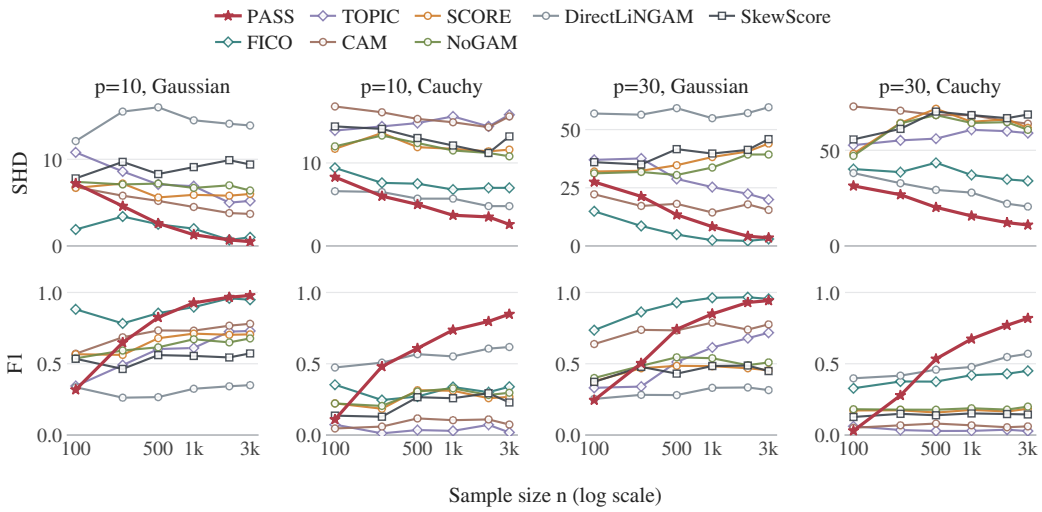


Figure 5: ANM sample-size trends under additive-GP mechanisms with i.i.d. additive errors. Columns report $p = 10$ Gaussian, $p = 10$ Cauchy, $p = 30$ Gaussian, and $p = 30$ Cauchy settings; rows report SHD and F1. The compared methods are PASS, FICO, TOPIC, CAM, SCORE, NoGAM, DirectLiNGAM, and SkewScore, and the n-axis is logarithmic. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

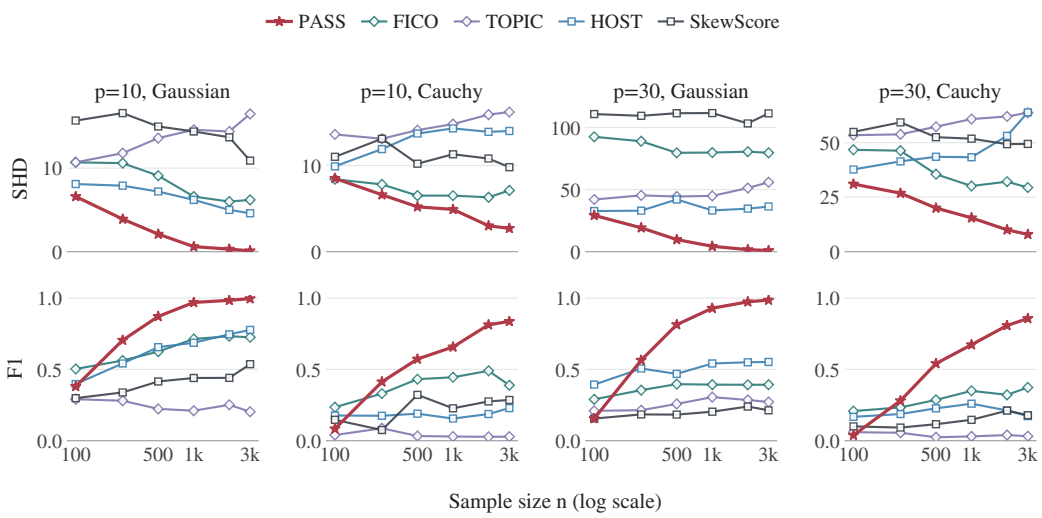


Figure 6: HNM sample-size trends under additive-GP location functions in the scale-calibrated HNM setting. Columns report $p = 10$ Gaussian, $p = 10$ Cauchy, $p = 30$ Gaussian, and $p = 30$ Cauchy settings; rows report SHD and F1. The compared methods are PASS, FICO, TOPIC, HOST, and SkewScore, and the n-axis is logarithmic. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

G.2.5 ER2 Sample-Size Trends

We repeat the ANM and HNM sample-size checks under denser ER2 graphs. ANM runs use i.i.d. additive errors, and HNM runs use the scale-calibrated HNM setting described in Appendix G.1, with additive-GP location functions. This isolates graph-density effects while keeping the additive-GP mechanism and evaluation metrics fixed. Figures 7 and 8 report the denser-graph sample-size trends on a logarithmic n-axis.

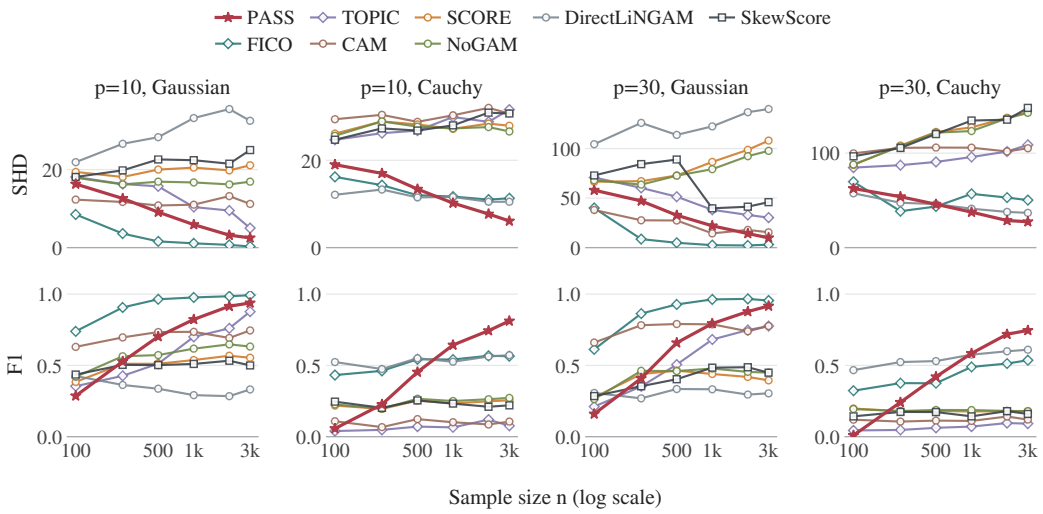


Figure 7: ANM sample-size trends under denser ER2 graphs with i.i.d. additive errors. The plots report SHD and F1 across Gaussian and Cauchy noise for $p = 10$ and $p = 30$. The compared methods are PASS, FICO, TOPIC, CAM, SCORE, NoGAM, DirectLiNGAM, and SkewScore, and the n-axis is logarithmic. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

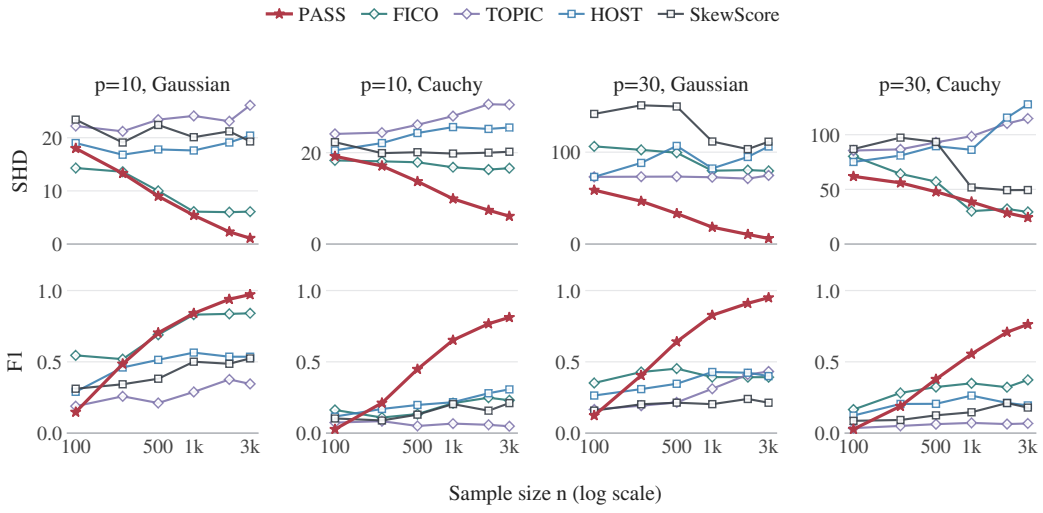


Figure 8: HNM sample-size trends under denser ER2 graphs with the scale-calibrated HNM setting. The plots report SHD and F1 across Gaussian and Cauchy noise for $p = 10$ and $p = 30$. The compared methods are PASS, FICO, TOPIC, HOST, and SkewScore, and the n-axis is logarithmic. Curves report means; lower SHD and higher F1 are better. SHD ranges vary across panels.

G.2.6 ER3 Sample-Size Trends

The ER3 experiments use $p = 10$, linear rate 0.5, and Gaussian/Cauchy noise under ANM/HNM settings. We use additive-GP location mechanisms and scale-calibrated HNM noise with depth scale $a_j = 1 + 0.20d_j$. For each seed, the samples at $n \in \{100, 250, 500, 1000, 2000, 3000\}$ are nested prefixes of one generated dataset. PASS uses $\beta = 0.3$ and $K_n = \max\{2, \lceil \log n \rceil\}$ independent $\mathrm{Uniform}(0, 1)$ quantile levels, giving $K_n \in \{5, 6, 7, 7, 8, 9\}$ at these sample sizes. One sampled level vector is shared across all ordering and parent fits within each run. The estimator uses three folds and cubic additive splines with four knots, an ℓ_2 coefficient-penalty weight of 1.5×10^{-3}, and the standard-error-based parent threshold with $\alpha = 10^{-3}$.

All methods receive the same generated data within each condition and replication. The comparison includes PASS, FICO, TOPIC, CAM, SCORE, NoGAM, DirectLiNGAM, and SkewScore under ANM, and PASS, FICO, TOPIC, HOST, and SkewScore under HNM, matching the ER1 method sets. The lower row of Figure 2 reports SHD, and Figure 9 reports F1.

At $n = 3000$, PASS has the lowest mean SHD and highest mean F1 among the compared methods for Gaussian HNM and both Cauchy settings. FICO performs better for Gaussian ANM (SHD 2.8 and F1 0.955, versus PASS's 5.4 and 0.896). At small sample sizes, PASS can select overly sparse graphs: for Cauchy ANM at $n = 100$, it selects 0.4 edges on average against 28.1 true edges, with mean F1 0.026. Thus the ER3 results show improvement with sample size, rather than uniform superiority across sample sizes and mechanisms.

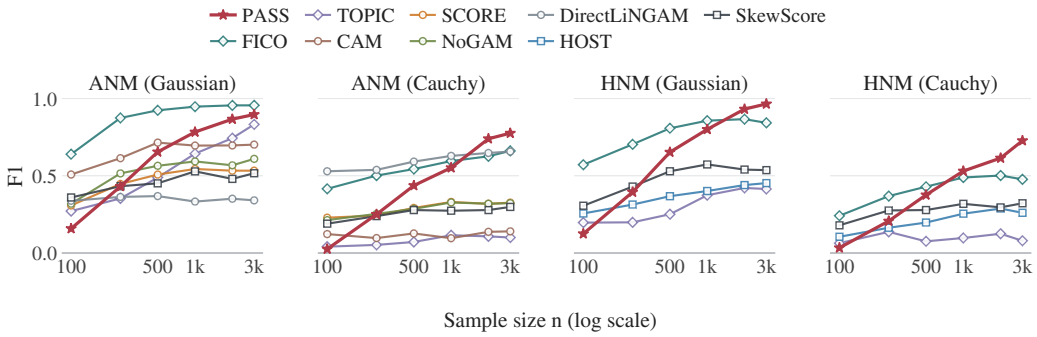


Figure 9: ER3 sample-size trends at $p = 10$, reported as mean F1. Columns show ANM and scale-calibrated HNM settings with Gaussian and Cauchy noise. The n-axis is logarithmic; higher F1 is better.

G.2.7 HNM Trends with RBF-GP Location and Sigmoid Scale Functions

We evaluate HNM sample-size trends under the RBF-GP location family, sigmoid scale family, and error distributions used in the SkewScore HNM experiments (Lin et al., 2025). For each node j, the setting uses the location–scale form

$$X_j = f_j(X_{\mathrm{Pa}(j)}) + h_j(X_{\mathrm{Pa}(j)})\varepsilon_j.$$

For root nodes, $f_j = 0$ and $h_j = 1$. For non-root nodes, the location vector $\{f_j(X^{(i)}_{\mathrm{Pa}(j)})\}_{i=1}^n$ is sampled from a zero-mean RBF Gaussian process with kernel

$$K_j(i, i') = \exp\{-\|X^{(i)}_{\mathrm{Pa}(j)} - X^{(i')}_{\mathrm{Pa}(j)}\|_2^2/(2\ell^2)\}, \qquad \ell = 1.$$

The raw scale profile is

$$\tilde{h}_j(x) = \{1 + \exp[-(w_j^\top x + b_j)]\}^{-1},$$

where $w_j \sim N(0, I_{|\mathrm{Pa}(j)|}/|\mathrm{Pa}(j)|)$ and $b_j \sim N(0, 0.25^2)$, independently across nodes. The scale used in the simulations is the empirical-mean normalized profile

$$h_j(x) = \tilde{h}_j(x)/\bar{h}_j, \qquad \bar{h}_j = n^{-1}\sum_{i=1}^n \tilde{h}_j(X^{(i)}_{\mathrm{Pa}(j)}).$$

This setting uses the same broad location family, sigmoid scale-profile family, and error distributions as the SkewScore HNM experiments; unlike the scale-calibrated HNM setting, it does not standardize parent inputs or impose the depth-ordered scale calibration described in Appendix G.1. In the Gaussian setting $\varepsilon_j \sim N(0, 0.75^2)$. In the Student-$t$ mixture setting, each replicate draws a degree of freedom from $\{2, 3, 4, 5\}$ and uses the corresponding standard Student-t error family. The sample-size grid is $n \in \{100, 250, 500, 1000, 2000, 3000\}$ with $p \in \{10, 30\}$, ER1 graphs, no additional linear component, and no interaction term; the displayed n-axis is logarithmic.

This setting remains an HNM, but the normalized sigmoid scale function alone does not enforce the sufficient PASS ordering condition. Since $\sigma_j(x) = h_j(x)$ in this design, the HNM ordering condition would require $\mathbb{E}[h_u(X_{\mathrm{Pa}(u)})^\beta] \le \mathbb{E}[h_v(X_{\mathrm{Pa}(v)})^\beta]$ for every edge $u \to v$. The normalization fixes the average scale level, but not the β-scale moments along graph depth, and the sigmoid weights are sampled nodewise rather than assigned through the depth-ordered scale calibration. We therefore treat the experiment as a robustness check for the location–scale setting of Lin et al. (2025), not as a separate sufficient-condition construction. Figure 10 reports SHD and F1 across both noise settings and both dimensions.

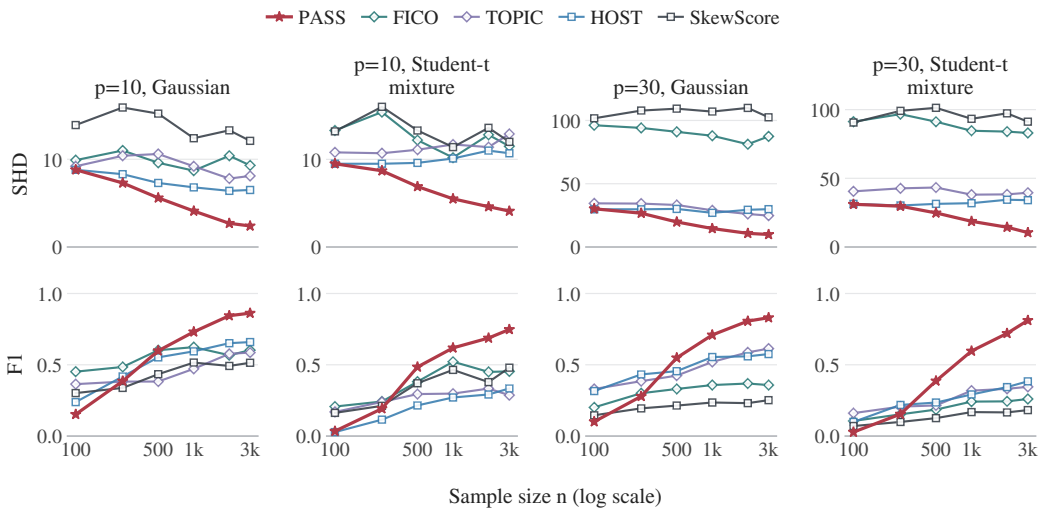


Figure 10: HNM sample-size trends under RBF-GP location functions and sigmoid scale functions matched to the SkewScore HNM setup (Lin et al., 2025). Columns show $p = 10$ Gaussian, $p = 10$ Student-t mixture, $p = 30$ Gaussian, and $p = 30$ Student-t mixture; rows report SHD and F1. Lower SHD and higher F1 are better. The compared methods are PASS, FICO, TOPIC, HOST, and SkewScore, and the n-axis is logarithmic. Curves report means.

G.3 Diagnostics and Sensitivity

G.3.1 UUMC Settings and Ordering Behavior

Purpose and common setting. This benchmark compares how PASS, variance, and R^2 ordering respond to different coefficient and noise-scale constructions within linear Gaussian SCMs. We examine whether PASS recovers the causal order in settings where the two sorting methods provide weak or reversed ordering information. Ordering accuracy evaluates ordering independently of parent selection, whereas SHD also reflects parent selection. No PASS identifying condition is imposed when generating the benchmark.

We use the six linear Gaussian generator settings of Herman et al. (2025): 50–50, DaO, IPA, UUMC, iSCM, and unit-variance noise. Data are generated with the official UUMCdata package (revision `aec581da`, with `daosim` 0.0.5), using Erdős–Rényi initialization with edge probability 0.5, $p = 20$, and $n = 3000$. Graph, parameter, and observation seed streams are shared across settings, with replications as specified in Appendix G.1. VarSort and R^2Sort abbreviate Var-SortnRegress and

R^2-SortnRegress. PASS uses $\beta = 0.3$, three folds, and $K_n = \lceil \log(3000) \rceil = 9$ independent $\mathrm{Uniform}(0, 1)$ quantile levels.

Here, 50–50 balances signal and noise variances at non-root nodes. DaO (DAG adaptation of the onion method) constructs DAG-compatible correlation matrices. IPA (independent parents assumption) normalizes coefficients and noise scales using variances calculated as if the parents were independent. UUMC (unitless, unrestricted, Markov-consistent) samples coefficient vectors within a unit ball and uses parental covariances to normalize coefficients and noise scales. iSCM (internally standardized SCM) standardizes each generated variable before generating its descendants. Unit-variance noise fixes every Gaussian error variance to one (Herman et al., 2025).

For DaO, the pinned UUMCdata version uses the innovation variances returned by `daosim` as Gaussian standard deviations. Our results follow this implementation and do not assume unit marginal variances for this setting.

Observed ordering behavior. The sorting-baseline columns in Table 3 report the fraction of true edges consistent with each estimated order. With unit-variance noise, VarSort and R^2Sort achieve ordering accuracies of 0.934 and 0.870. Under 50–50, VarSort is near chance (0.485), while R^2Sort places most true edges in reverse order (0.127). Under DaO, VarSort also places most edges in reverse order (0.148), whereas R^2Sort is slightly above chance (0.566); under IPA, both methods are near chance. UUMC and iSCM have near-chance variance ordering and below-chance R^2 ordering. For iSCM, variance-based rankings reflect numerical tie-breaking after sample-variance normalization. These values summarize ordering accuracy over the sampled datasets.

Input reconstruction. The PASS inputs are regenerated using the specified generator version, transformations, and component seeds, and checked against the saved reference graphs and diagnostics; the original observation arrays were not archived.

Table 8: Full UUMC generator-family results. Entries are mean (SD). Precision and recall refer to directed edges. Bold marks the best reported mean within each generator setting and metric, with all ties bolded; SDs are not used to determine bolding.

Generator setting	Method	$A_{\mathrm{top}} \uparrow$	SHD↓	F1↑	Precision↑	Recall↑
50–50	PASS	**0.911 (0.054)**	**49.9 (10.8)**	**0.667 (0.081)**	**0.897 (0.071)**	**0.538 (0.093)**
	VarSort	0.485 (0.084)	155.7 (19.2)	0.358 (0.072)	0.294 (0.064)	0.458 (0.083)
	R^2Sort	0.127 (0.049)	229.4 (13.0)	0.087 (0.034)	0.069 (0.026)	0.116 (0.046)
DaO	PASS	0.365 (0.065)	140.9 (14.1)	0.187 (0.043)	0.209 (0.050)	0.171 (0.041)
	VarSort	0.148 (0.043)	229.7 (14.1)	0.095 (0.029)	0.076 (0.024)	0.128 (0.039)
	R^2Sort	**0.566 (0.079)**	**138.2 (20.6)**	**0.417 (0.066)**	**0.351 (0.062)**	**0.517 (0.072)**
IPA	PASS	**0.497 (0.089)**	**117.4 (16.6)**	**0.347 (0.073)**	**0.369 (0.083)**	0.330 (0.072)
	VarSort	0.434 (0.099)	176.6 (22.2)	0.309 (0.078)	0.246 (0.066)	**0.417 (0.098)**
	R^2Sort	0.436 (0.088)	180.6 (19.0)	0.303 (0.068)	0.239 (0.057)	0.414 (0.087)
UUMC	PASS	**0.604 (0.098)**	**113.2 (13.8)**	0.295 (0.059)	**0.364 (0.082)**	0.252 (0.054)
	VarSort	0.505 (0.081)	155.3 (20.3)	**0.356 (0.064)**	0.295 (0.062)	**0.452 (0.069)**
	R^2Sort	0.423 (0.095)	170.3 (20.2)	0.302 (0.072)	0.248 (0.063)	0.390 (0.089)
iSCM	PASS	**0.591 (0.087)**	**110.3 (16.7)**	**0.391 (0.083)**	**0.411 (0.089)**	0.376 (0.085)
	VarSort	0.509 (0.080)	165.8 (16.8)	0.354 (0.061)	0.281 (0.051)	**0.482 (0.079)**
	R^2Sort	0.451 (0.097)	176.0 (20.3)	0.318 (0.075)	0.252 (0.063)	0.433 (0.094)
Unit-variance noise	PASS	**0.977 (0.021)**	56.3 (14.3)	0.724 (0.057)	0.686 (0.081)	0.773 (0.057)
	VarSort	0.934 (0.036)	**44.5 (22.5)**	**0.799 (0.092)**	**0.715 (0.119)**	**0.914 (0.048)**
	R^2Sort	0.870 (0.053)	82.5 (23.8)	0.660 (0.087)	0.549 (0.098)	0.834 (0.060)

G.3.2 Sortability on the Main $p = 50$ Benchmark

We examine variance and R^2 sortability separately from the algorithm comparison in Table 1. The diagnostics and sorting baselines use the same datasets as the main benchmark across its six

ANM/HNM–noise conditions at $p = 50$ and $n = 1000$. No new PASS fits are used in this comparison. Variance and R^2 sortability are computed from each realized sample and reference graph using the official implementation (Reisach et al., 2021; 2023). They measure agreement along directed paths and differ from the edge-based ordering accuracy A_{top}.

Table 9: Sample sortability of the main $p = 50$ datasets, reported as mean (SD). These are finite-sample diagnostics; in particular, the Cauchy results do not assert the existence of population variances.

Setting	Variance sortability	R^2 sortability
ANM, Gaussian	0.868 (0.037)	0.585 (0.068)
ANM, Pareto(2)	0.817 (0.080)	0.543 (0.086)
ANM, Cauchy	0.652 (0.066)	0.491 (0.082)
HNM, Gaussian	0.896 (0.062)	0.562 (0.084)
HNM, Pareto(2)	0.907 (0.052)	0.504 (0.125)
HNM, Cauchy	0.730 (0.065)	0.421 (0.111)

Table 10: PASS and official sorting baselines on the same $p = 50$ datasets. Entries are mean (SD); higher A_{top} and F1 and lower SHD are better. Bold marks the best reported mean within each setting and metric, with all ties bolded. VarSort and R^2Sort denote Var-SortnRegress and R^2-SortnRegress.

Setting	Method	$A_{\text{top}} \uparrow$	SHD$\downarrow$	F1$\uparrow$
ANM, Gaussian	PASS	**0.899 (0.029)**	**32.4 (6.6)**	**0.682 (0.052)**
ANM, Gaussian	VarSort	0.846 (0.040)	65.1 (20.0)	0.570 (0.087)
ANM, Gaussian	R^2Sort	0.535 (0.063)	132.9 (28.7)	0.288 (0.060)
ANM, Pareto(2)	PASS	**0.926 (0.030)**	**24.0 (9.9)**	**0.753 (0.094)**
ANM, Pareto(2)	VarSort	0.791 (0.054)	69.6 (18.0)	0.524 (0.076)
ANM, Pareto(2)	R^2Sort	0.544 (0.079)	115.3 (30.3)	0.314 (0.079)
ANM, Cauchy	PASS	**0.896 (0.041)**	**23.1 (6.0)**	**0.759 (0.061)**
ANM, Cauchy	VarSort	0.650 (0.079)	69.0 (14.4)	0.386 (0.102)
ANM, Cauchy	R^2Sort	0.503 (0.076)	83.5 (17.9)	0.290 (0.075)
HNM, Gaussian	PASS	**0.946 (0.038)**	**24.2 (5.0)**	**0.751 (0.055)**
HNM, Gaussian	VarSort	0.859 (0.076)	64.2 (32.6)	0.591 (0.152)
HNM, Gaussian	R^2Sort	0.510 (0.054)	142.5 (22.8)	0.257 (0.041)
HNM, Pareto(2)	PASS	**0.972 (0.023)**	**18.6 (3.0)**	**0.807 (0.026)**
HNM, Pareto(2)	VarSort	0.881 (0.033)	65.6 (16.5)	0.555 (0.079)
HNM, Pareto(2)	R^2Sort	0.494 (0.084)	121.1 (24.6)	0.274 (0.070)
HNM, Cauchy	PASS	**0.927 (0.027)**	**19.8 (6.0)**	**0.790 (0.057)**
HNM, Cauchy	VarSort	0.697 (0.056)	68.3 (13.9)	0.373 (0.064)
HNM, Cauchy	R^2Sort	0.428 (0.070)	90.0 (13.8)	0.233 (0.039)

PASS has lower mean SHD and higher mean F1 than either sorting baseline in all six conditions. Variance sortability nevertheless exceeds 0.5 throughout this benchmark, and variance ordering is informative in several settings. These comparisons therefore do not establish that PASS is unaffected by the generator's scale structure. The UUMC results and variable-scaling experiment in Section 7.2 examine this dependence across other generator settings and coordinate scales.

G.3.3 Parent Selection with the True Order

We isolate parent selection using the same 60 datasets as the main benchmark: $p = 50$, $n = 1000$, ER1 graphs, ANM/HNM mechanisms, and three noise families. Every method receives the same observations and true topological order; no ordering estimator is run. We compare the parent-selection stage of PASS with the official CausalDisco BIC/Lasso regression, a local additive-spline

implementation with backward BIC selection, and the official CaPS CAM/GAM pruning procedure. PASS retains its main configuration and quantile seeds. The spline implementation uses rank-transformed predictors and cubic splines with four knots; CAM/GAM starts from the complete DAG compatible with the supplied order and uses pruning cutoff 10^{-3}. Each method retains its original data transformations.

Table 11 shows that PASS has the lowest mean SHD and highest mean F1 under Cauchy noise in both model classes. CAM/GAM performs best under Gaussian noise, while spline BIC performs best under Pareto(2) noise. Thus the parent-selection advantage depends on the noise distribution. This comparison separates that behavior from errors in estimating the order; it does not establish finite-sample error control for the standard-error-based threshold.

Table 11: Parent selection with a supplied true order on the main $p = 50$, $n = 1000$, ER1 datasets. Entries are mean (SD). All methods use the same true order. Bold marks the best mean within each row; lower SHD and higher F1 are better.

Setting	Metric	PASS	BIC-Lasso	Spline BIC	CAM/GAM
ANM, Gaussian	SHD↓	14.7 (6.7)	22.0 (8.8)	8.9 (4.5)	**7.6 (2.4)**
	F1↑	0.856 (0.056)	0.827 (0.056)	0.922 (0.036)	**0.932 (0.020)**
ANM, Pareto(2)	SHD↓	14.0 (4.3)	38.6 (8.3)	**11.7 (2.7)**	44.0 (8.5)
	F1↑	0.856 (0.043)	0.707 (0.052)	**0.890 (0.027)**	0.695 (0.050)
ANM, Cauchy	SHD↓	**12.9 (5.1)**	46.4 (9.3)	35.7 (3.8)	79.6 (8.5)
	F1↑	**0.869 (0.053)**	0.539 (0.096)	0.551 (0.042)	0.445 (0.050)
HNM, Gaussian	SHD↓	15.6 (5.2)	21.8 (9.5)	8.9 (8.6)	**6.9 (3.2)**
	F1↑	0.837 (0.060)	0.825 (0.067)	0.920 (0.075)	**0.937 (0.028)**
HNM, Pareto(2)	SHD↓	14.1 (4.3)	49.6 (14.5)	**13.9 (4.9)**	54.7 (7.7)
	F1↑	0.855 (0.046)	0.653 (0.072)	**0.868 (0.049)**	0.640 (0.041)
HNM, Cauchy	SHD↓	**15.1 (5.3)**	47.9 (8.8)	40.5 (7.6)	84.7 (11.8)
	F1↑	**0.842 (0.053)**	0.513 (0.066)	0.448 (0.095)	0.395 (0.051)

G.3.4 Scale-Ordering Perturbations

We perturb the nodewise scale targets in HNMs with additive-GP location functions, $p = 10$, $n = 5000$, ER1 graphs, and linear rate 0.5. We compare PASS with oracle-order PASS, which uses the true topological ordering in the same parent-selection step, and the HNM baseline methods under Gaussian, Pareto(2), and Cauchy noise.

Each replicate starts from depth-ordered targets $\rho_j^{(0)} = \rho_0(1 + 0.2d_j)$, where $\rho_0 = 1/\sqrt{\pi}$ is the common baseline target and d_j is graph depth. For a requested $\eta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, we sample Gaussian perturbation vectors $z^{(b)}$, indexed by b, and search $s \in [0, 3]$ for targets $\rho_j^{(b)}(s) = \rho_j^{(0)} \exp(sz_j^{(b)})$ whose fraction of reversed true edges is closest to η. These targets calibrate the node scales, while the tanh-softplus profiles are normalized by the arithmetic pilot means defined in Appendix G.1. Consequently, η describes reversals of the scale targets; it is not a measured violation fraction for the β-moment condition in Corollary 5.1.

Figure 11 reports SHD and A_{top} along this perturbation sweep. PASS ordering accuracy declines as the target reversal fraction increases, while oracle-order PASS remains more stable. This diagnostic separates the contribution of ordering errors from parent-selection errors given a correct order.

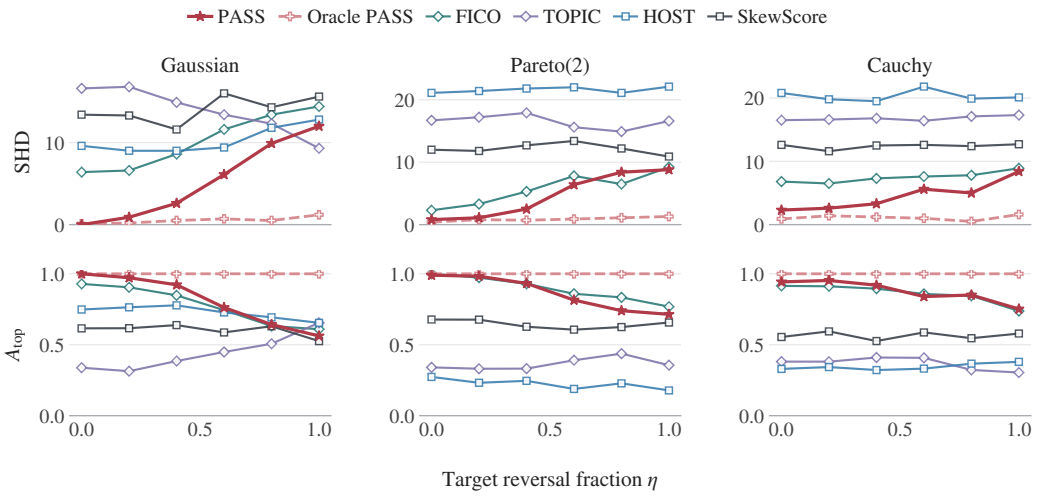


Figure 11: Target scale-ordering perturbations under HNMs with $p = 10$, $n = 5000$, an ER1 graph, and linear rate 0.5. Columns correspond to Gaussian, Pareto(2), and Cauchy noise; rows report mean SHD and mean A_{top}, the fraction of true directed edges consistent with the estimated ordering. Lower SHD and higher A_{top} are better. The horizontal axis is the target reversal fraction η.

G.3.5 Beta-Exponent Selection

The exponent β determines the moment requirement and the scale of the estimated scores. We use additive-GP ANMs with $p = 10$, an ER1 graph, linear rate 0.5, and $n \in \{100, 250, 500, 1000, 2000, 3000\}$. The compared exponents are $\beta \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, evaluated under Gaussian and $t(2)$ noise. The main experiments use $\beta = 0.3$ without tuning it separately for each noise family. Figure 13 reports the results.

Figure 12 additionally examines Student-t_ν noise and Gaussian small-sample behavior. Smaller β relaxes the moment requirement but reduces the contrast between large and small quantile separations, which can reduce finite-sample discrimination.

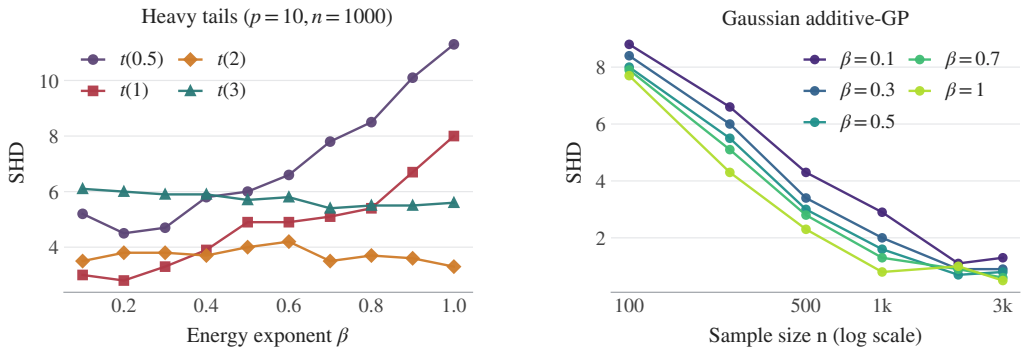


Figure 12: Effect of the separation exponent. Left: Student-t_ν noise and the moment condition $2\beta < \nu$. Right: Gaussian additive-GP ANMs and the finite-sample tradeoff in choosing β. Lower SHD is better. Curves report means.

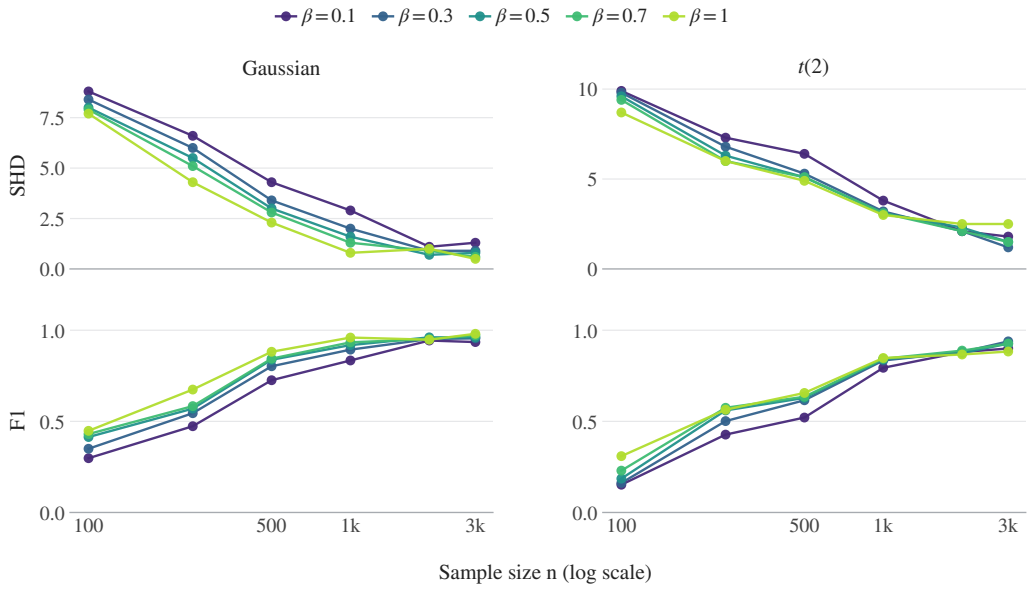


Figure 13: Sensitivity to the separation exponent under additive-GP ANMs. Columns correspond to Gaussian and $t(2)$ noise; rows report SHD and F1 on a logarithmic n-axis. The main experiments fix $\beta = 0.3$ as the default. Curves report means.

G.3.6 Additive and Pairwise Spline Bases

To assess the additive approximation used in the experiments, we repeat the ER1 additive-GP sample-size study in Figure 2 with PASS-pairwise, which augments the additive B-spline basis with pairwise tensor-product terms. Both methods use the same datasets, quantile levels, and remaining settings from Appendix F. Adding pairwise terms does not improve the displayed mean SHD or F1 (Figure 14).

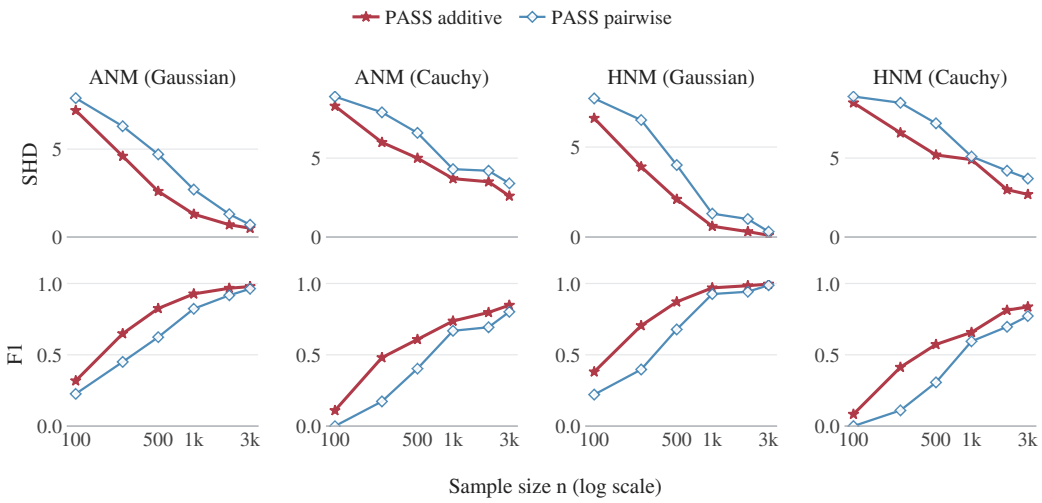


Figure 14: PASS and PASS-pairwise on the ER1 datasets from Figure 2. Curves report mean SHD and F1.

G.4 Real-Data Benchmarks

All three benchmarks use the PASS configuration in Appendix F.

G.4.1 Sachs and the 12-Station Danube Network

Sachs uses the 853×11 observational array from the CaPS data release, with its supplied reference DAG, without an additional logarithmic transform, standardization, or filtering. The main Danube benchmark uses the $4{,}600 \times 12$ daily-discharge matrix of Gnecco et al. (2021), comprising summer observations from 1960–2009. We apply no additional coordinate transformation. These sample sizes give $K_n = 7$ for Sachs and $K_n = 9$ for Danube.

PASS varies its algorithmic seed and random quantile levels on each fixed dataset. Its mean and sample SD in Table 4 measure algorithmic variability on the same observations. Comparators are run once, so no SD is estimated for them. Results are specific to the stated preprocessing and coordinate scales.

G.4.2 CausalRivers

The reference benchmark is the original CausalRivers Random-5 collection of $7{,}521$ official connected five-node subgraphs from RiversEastGermany (Stein et al., 2025). Raw 15-minute discharge series are aggregated to 6-hour resolution without interpolation, and rows with missing station values are removed. This preprocessing excludes 35 subgraphs with insufficient observations: 26 retain only one complete observation and nine retain none. The remaining $7{,}486$ subgraphs (median $n = 5{,}711$) form the common evaluation set in Table 4. PASS uses one execution per eligible subgraph and the default sample-size-dependent quantile count.

CausalRivers SDs in Table 4 describe variation across subgraphs, rather than repeated algorithmic seeds; comparator results reuse the existing benchmark evaluations. Since these subgraphs overlap and the river observations are temporally dependent, these summaries do not provide independent-replicate standard errors or establish a statistically significant ranking. The reported F1 evaluates each returned graph at its fixed decision rule; it is distinct from the threshold-maximized F1 used in the extended evaluation of Stein et al. (2025).

G.4.3 Official CausalRivers Protocol

We additionally evaluate random-quantile PASS on all $7{,}521$ Random-5 subgraphs using the official preprocessing with six-hour aggregation, interpolation, boundary trimming, and no normalization (Stein et al., 2025). PASS uses the default $\beta = 0.3$, three folds, and $K_n = 9$ random quantile levels, with one prespecified execution per graph.

Table 12 compares PASS with the released Random-5 benchmark results. F1-max is obtained by maximizing F1 over thresholds separately for each labeled graph and then averaging over graphs; it is distinct from the fixed-output F1 in Table 4. Comparator entries retain the officially selected configurations, which vary in time resolution and normalization. VAR and RP+B were also reproduced locally on the six-hour interpolated inputs, recovering their released F1-max values. The other comparator entries are reported reference results.

VAR uses linear vector-autoregressive Granger causality, and CDMI uses neural forecast residuals under knockoff interventions. CP (Causal Pretraining) predicts graphs with Transformer or GRU networks pretrained on synthetic data. RP (Reverse Physical) orients links from lower to higher mean discharge; CC uses peak lagged cross-correlation, and RPCC combines these rules. The suffixes +N, +B, and +C retain at most one successor per source: respectively the next larger, largest, or most strongly cross-correlated river (Stein et al., 2025).

PASS has F1-max 0.604, below VAR, VarLiNGAM, and CDMI but above the remaining reported comparators. The ranking depends on the metric: on the same interpolated inputs, fixed-output F1 is 0.493 for PASS and 0.495 for RP+B. These are descriptive comparisons over overlapping subgraphs.

Table 12: Official CausalRivers Random-5 comparison. Entries are mean per-graph F1-max, except that the null model follows the released pooled-label convention. Higher is better; bold marks the highest value. PASS uses random quantile levels, and all comparator entries are released benchmark results from Stein et al. (2025).

Method	F1-max↑
VAR	**0.649**
VarLiNGAM	0.633
CDMI	0.628
PASS	0.604
DynoTears	0.537
CP (Transformer)	0.530
RP+B	0.529
PCMCI	0.522
RP	0.522
RPCC+C	0.522
CC+C	0.519
RPCC	0.508
CC	0.493
RPCC+B	0.493
CP (GRU)	0.490
RP+N	0.490
RPCC+N	0.490
Null model	0.335