PoET: Test-Integrated Nonparametric Poisson DAG Learning via Equidispersion Testing

Anonymous authors
Paper under double-blind review

Abstract

Poisson DAG models enable causal discovery for count data through a mean–variance asymmetry: conditioning on all parents yields equidispersion, whereas omitted parents induce overdispersion. Existing ranking-based methods exploit this asymmetry to construct a causal ordering by repeatedly selecting the node with the smallest overdispersion, without testing whether the selected node is statistically compatible with equidispersion. These methods also face difficulties in nonparametric conditional-mean estimation or avoid them by prespecifying nodewise link functions, which limits modeling flexibility. We propose PoET, a test-integrated framework for Poisson DAG learning that replaces score ranking with test-based layer eligibility. Its new cross-fitted tests for conditional uncorrelatedness and equidispersion use Conway–Maxwell–Poisson (CMP)-kernel smoothing to estimate conditional means nonparametrically, without prespecified nodewise link functions. To mitigate the conditioning burden of nonparametric smoothing, conditional-uncorrelatedness tests screen parent candidates to reduce conditioning dimension, while equidispersion tests guide layer assignment and parent recovery. If no candidate is eligible for layer assignment, PoET stops rather than forcing further graph decisions. We establish graph-recovery consistency under correct specification and consistent violation detection under local misspecification. Simulations show competitive recovery under heterogeneous nonlinear mechanisms and reliable detection of dispersion violations. NBA and Lending Club analyses illustrate, respectively, test-based incompatibility detection and complete layerwise learning without prescribed links.

1 Introduction

Count-valued data arise in many scientific and industrial domains, including single-cell RNA sequencing, neural spike trains, epidemiological event counts, and insurance claims (Blei et al., 2003; Love et al., 2014; Stoklosa et al., 2022). Such variables are not merely categorical. Counts are supported on the nonnegative integers, have meaningful order and scale, and often exhibit excess zeros, tail behavior, and mean–variance coupling. Causal discovery for counts should therefore neither transform counts into continuous measurements nor treat them as unordered labels, but instead exploit their distribution-specific structure (Warton, 2018; Choi et al., 2020; Choi & Ni, 2023).

Poisson directed acyclic graph (DAG) models (Park & Raskutti, 2015) provide a natural framework for count causal discovery by modeling each variable conditionally as Poisson with a mean determined by its parents. They also yield a *mean–variance asymmetry* for graph recovery: conditioning on all parents yields equidispersion, whereas omitting a relevant parent adds a positive dispersion component. Existing dispersion-based methods exploit this asymmetry for structure recovery (Park & Raskutti, 2015; 2018; Park & Park, 2019b; Zhou et al., 2022).

However, ranking-based dispersion methods suffer from score-forced selection: relative scores choose the best-ranked candidate even when none is compatible with equidispersion. Other methods such as TLDAG (Zhou et al., 2022) compare dispersion scores against thresholds rather than ranking these scores, but still lack a test-based stopping rule for equidispersion violations. A second challenge is nonparametric conditional-mean estimation. Regression-based approaches gain tractability by prescribing the functional form of the conditional mean (Park & Park, 2019b; Zhou et al., 2022;

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.

Nguyen et al., 2024), whereas exact-cell conditional-moment estimators (Park & Raskutti, 2015; Park & Park, 2019a) become unstable as conditioning configurations grow sparse.

To address these challenges, we propose PoET (Poisson DAG Learning via Equidispersion Testing), a test-integrated framework for nonparametric Poisson DAG learning. Our main contributions are as follows, with further discussion relative to the existing literature provided in Appendix A.1.

• *From score to test.* PoET replaces score rules with local equidispersion tests that determine eligibility for the next layer and guide parent recovery. If the tests do not support any next-layer assignment, it stops and labels unresolved nodes as violations rather than forcing graph decisions.

• *Flexible nonparametric learning.* PoET estimates conditional means through cross-fitted Conway–Maxwell–Poisson (CMP) kernel smoothing, accommodating potentially nonlinear parent effects without requiring knowledge of the nodewise link functions.

• *Screening for efficient learning.* Conditional-uncorrelatedness tests screen parent candidates, which can reduce the dimension of the nonparametric smoothing and improve finite-sample performance.

• *Test validity and graph consistency.* We develop cross-fitted CMP-kernel tests for equidispersion and conditional uncorrelatedness and establish their asymptotic guarantees. Building on these local tests, we prove that PoET consistently recovers the graph under correct specification and detects violations under local misspecification.

2 Preliminaries

This section introduces the notation and facts used by PoET. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ denote a count-valued random vector with distribution $P_{\boldsymbol{X}}$ on $\mathbb{N}_0^p$, where $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$.

2.1 Graph notation and topological layers

Let $V = \{1, \ldots, p\}$ and $G = (V, E)$ be a directed acyclic graph (DAG), with $k \to j$ denoting $(k, j) \in E$. For $j \in V$, let $\mathrm{Pa}_G(j)$, $\mathrm{Ch}_G(j)$, $\mathrm{An}_G(j)$, $\mathrm{De}_G(j)$, and $\mathrm{Nd}_G(j)$ denote its parents, children, ancestors, descendants, and non-descendants, respectively. The spouses and neighborhood of j are defined as $\mathrm{Sp}_G(j) := \{k \in V \setminus \{j\} : \mathrm{Ch}_G(k) \cap \mathrm{Ch}_G(j) \neq \emptyset\}$ and $\mathrm{Ne}_G(j) := \mathrm{Pa}_G(j) \cup \mathrm{Ch}_G(j) \cup \mathrm{Sp}_G(j)$, respectively. The latter is used as a local adjustment set in the conditional uncorrelatedness test employed in PoET.

Define the topological layers recursively as $\mathcal{L}_1(G) := \{j \in V : \mathrm{Pa}_G(j) = \emptyset\}$, and, for $t \geq 2$, $\mathcal{L}_t(G) := \{j \in V \setminus \mathcal{L}_{<t}(G) : \mathrm{Pa}_G(j) \subseteq \mathcal{L}_{<t}(G)\}$, $\mathcal{L}_{<t}(G) := \bigcup_{s<t} \mathcal{L}_s(G)$. The subgraph induced by the preceding layers is $G_{<t} := (V, E_{<t}(G))$, where $E_{<t}(G) := \{(a, b) \in E : a, b \in \mathcal{L}_{<t}(G)\}$.

For $S \subseteq V$, write $\boldsymbol{X}_S = (X_k)_{k \in S}$, with $\boldsymbol{X}_\emptyset$ denoting the trivial conditioning variable. We write $i \perp_{d,G} j \mid S$ for d-separation in G and assume that $\boldsymbol{X}$ is Markov with respect to G, so that $i \perp_{d,G} j \mid S$ implies $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_S$ (Spirtes et al., 2000; Lauritzen, 1996).

2.2 Poisson DAG models and their identifiability

A distribution $P_{\boldsymbol{X}}$ on $\mathbb{N}_0^p$ is a nonparametric Poisson DAG model over G if it factorizes according to G and each node-conditional mechanism is Poisson:

$$X_j \mid \boldsymbol{X}_{\mathrm{Pa}_G(j)} \sim \mathrm{Poisson}\{\mu_j(\boldsymbol{X}_{\mathrm{Pa}_G(j)})\} \tag{1}$$

for every $j \in V$, where each link function $\mu_j : \mathbb{N}_0^{|\mathrm{Pa}_G(j)|} \to [0, \infty)$ is unknown.

A conditioning set S is called *admissible* for node j if $S \subseteq \mathrm{Nd}_G(j)$. For an admissible S, define the population dispersion gap

$$\Delta_{j|S} := \mathbb{E}\left[\mathrm{Var}(X_j \mid \boldsymbol{X}_S) - \mathbb{E}(X_j \mid \boldsymbol{X}_S)\right]. \tag{2}$$

Under the Poisson DAG model in (1), the law of total variance and conditional equidispersion give $\Delta_{j|S} = \mathbb{E}\left[\mathrm{Var}\{\mu_j(\boldsymbol{X}_{\mathrm{Pa}_G(j)}) \mid \boldsymbol{X}_S\}\right] \geq 0$. Hence, if $\mathrm{Pa}_G(j) \subseteq S$, then $\Delta_{j|S} = 0$. If $\mathrm{Pa}_G(j) \not\subseteq S$ and the conditional variance above is positive with positive probability, then $\Delta_{j|S} > 0$.

We say that the link functions are non-degenerate if, for every node $j \in V$ and every admissible S for j, $\mathrm{Pa}_G(j) \not\subseteq S \Longrightarrow \Delta_{j|S} > 0$. Under non-degeneracy, omitting any parent leaves residual variation, so an admissible conditioning set restores equidispersion if and only if it contains all parents. This condition also ensures identifiability of the Poisson DAG from the observational distribution (Park & Park, 2019b).

2.3 CMP kernels for count-valued smoothing

The local tests in PoET employ nonparametric conditional mean estimation with count-valued conditioning variables. Whereas the ODS method (Park & Raskutti, 2015) relies on exact matching over observed conditioning sets, PoET uses count-aware CMP-kernel smoothing that borrows information across nearby count values (Somé & Kokonendji, 2016; Huang et al., 2022).

For $u \in \mathbb{N}_0$, mean parameter $\mu \geq 0$, and dispersion parameter $\nu \geq 0$, write

$$C(u; \mu, \nu) = \lambda(\mu, \nu)^u / \{(u!)^\nu Z(\lambda(\mu, \nu), \nu)\}, \qquad Z(\lambda, \nu) = \sum_{v=0}^{\infty} \lambda^v / (v!)^\nu, \tag{3}$$

where $\lambda(\mu, \nu)$ is chosen so that the distribution has mean μ. For a count value $z \in \mathbb{N}_0$ and bandwidth $h = 1/\nu > 0$, the one-dimensional CMP kernel is defined as $K_{z,h}(u) = C(u; z, 1/h)$. For a conditioning vector $\boldsymbol{x}_S = (x_k)_{k \in S}$, the product kernel is

$$K_{\boldsymbol{x}_S, h}(\boldsymbol{u}_S) = \prod_{k \in S} K_{x_k, h}(u_k), \tag{4}$$

with the convention that the product equals one when $S = \emptyset$.

3 PoET: Test-Integrated Poisson DAG Learning

This section implements the shift from score to test through test-based layer eligibility, screening-guided parent candidate search, and diagnostic stopping. The conditional-uncorrelatedness and equidispersion tests use cross-fitted, count-aware CMP-kernel conditional-moment estimation. Section 3.1 and Algorithm 1 summarize the learning procedure, and Section 3.2 details the local tests.

3.1 Layerwise learning with local search

For a target node j and an admissible conditioning set $S \subseteq V \setminus \{j\}$, the equidispersion hypothesis is formulated as

$$H_0^{\mathrm{eq}}(j; S) : \Delta_{j|S} = 0, \qquad H_1^{\mathrm{eq}}(j; S) : \Delta_{j|S} \neq 0, \tag{5}$$

where the population dispersion gap $\Delta_{j|S}$ is defined in Equation (2). We call node j *layer-eligible* given S at level α when $H_0^{\mathrm{eq}}(j; S)$ is not rejected. When S consists of nodes assigned to earlier layers, non-rejection is consistent with S containing all parents of j, allowing j to enter the next layer. Rejection may indicate an omitted parent or a violation of the nodewise Poisson mechanism.

For screening, we consider the following conditional-uncorrelatedness null for distinct nodes j, k and a conditioning set $C \subseteq V \setminus \{j, k\}$:

$$H_0^{\mathrm{corr}}(j, k; C) : \mathbb{E}(e_{j,C} e_{k,C} \mid \boldsymbol{X}_C) = 0 \quad \text{almost surely}, \tag{6}$$

where $e_{\ell,C} = X_\ell - \mathbb{E}(X_\ell \mid \boldsymbol{X}_C)$. In PoET, this test forms an initial screened set of earlier-layer nodes that remain residually correlated with the target node after neighborhood adjustment.

At step t, let $\widehat{\mathcal{L}}_{<t}$ denote the previously assigned nodes, $\widehat{R}_t = V \setminus \widehat{\mathcal{L}}_{<t}$ the remaining nodes, and $\widehat{G}_{<t} = (V, \widehat{E}_{<t})$ the partial graph. For $j \in \widehat{R}_t$ and $k \in \widehat{\mathcal{L}}_{<t}$, define the screening adjustment set

$$\widehat{Z}_{kj}^{(t)} := \{\mathrm{Pa}_{\widehat{G}_{<t}}(k) \cup \mathrm{Ch}_{\widehat{G}_{<t}}(k) \cup \mathrm{Sp}_{\widehat{G}_{<t}}(k)\} \setminus \{j, k\}. \tag{7}$$

The local statistics and their p-values are defined in Section 3.2.

Algorithm 1: PoET: Test-integrated Poisson DAG learning

Input: Count data $X_{1:n}$ on V, significance level α, and tolerance $\epsilon > 0$
Output: Graph $\widehat{G}$ and labels $\widehat{\ell}$

(a) Layerwise learning

Initialize $t \leftarrow 1$, $\widehat{\mathcal{L}}_{<1} \leftarrow \emptyset$, $\widehat{R}_1 \leftarrow V$, $\widehat{E}_{<1} \leftarrow \emptyset$, and $\widehat{G}_{<1} \leftarrow (V, \emptyset)$
while $\widehat{R}_t \neq \emptyset$ **do**
  Compute $\widehat{P}_j \leftarrow \textsc{LocalSearch}(j, t)$ in (b) for all $j \in \widehat{R}_t$
  $\widehat{\mathcal{L}}_t \leftarrow \{j \in \widehat{R}_t : \widehat{P}_j \neq \textsc{defer}\}$
  if $\widehat{\mathcal{L}}_t = \emptyset$ **then break**
  $\widehat{E}_{<t+1} \leftarrow \widehat{E}_{<t} \cup \{(k, j) : j \in \widehat{\mathcal{L}}_t,\ k \in \widehat{P}_j\}$
  $\widehat{G}_{<t+1} \leftarrow (V, \widehat{E}_{<t+1})$
  $\widehat{\mathcal{L}}_{<t+1} \leftarrow \widehat{\mathcal{L}}_{<t} \cup \widehat{\mathcal{L}}_t$
  $\widehat{R}_{t+1} \leftarrow \widehat{R}_t \setminus \widehat{\mathcal{L}}_t, \quad t \leftarrow t + 1$
end
Set $\widehat{\ell}_j \leftarrow \textsc{selected}$ for $j \in \widehat{\mathcal{L}}_{<t}$ and $\widehat{\ell}_j \leftarrow \textsc{violation}$ for $j \in \widehat{R}_t$
return $(\widehat{G}_{<t}, \widehat{\ell})$

(b) $\textsc{LocalSearch}(j, t)$

$\widehat{C}_j \leftarrow \{k \in \widehat{\mathcal{L}}_{<t} : p^{\mathrm{corr}}_{jk|\widehat{Z}^{(t)}_{kj}} \leq \alpha\}, \quad \widehat{S}_j \leftarrow \widehat{C}_j$
Order the nodes $k \in \widehat{\mathcal{L}}_{<t} \setminus \widehat{C}_j$ by decreasing $T^{\mathrm{corr}}_{jk|\widehat{Z}^{(t)}_{kj}}$
foreach $k \in \widehat{\mathcal{L}}_{<t} \setminus \widehat{C}_j$ **do**
  if $p^{\mathrm{eq}}_{j,\widehat{S}_j} > \alpha$ **then break**
  if $Q_{j,\widehat{S}_j \cup \{k\}} - Q_{j,\widehat{S}_j} \leq \epsilon$ **then** $\widehat{S}_j \leftarrow \widehat{S}_j \cup \{k\}$
end
if $p^{\mathrm{eq}}_{j,\widehat{S}_j} \leq \alpha$ **then return** DEFER
Order the nodes $k \in \widehat{S}_j \setminus \widehat{C}_j$ by increasing $T^{\mathrm{corr}}_{jk|\widehat{Z}^{(t)}_{kj}}$
foreach $k \in \widehat{S}_j \setminus \widehat{C}_j$ **do**
  if $p^{\mathrm{eq}}_{j,\widehat{S}_j \setminus \{k\}} > \alpha$ **then** $\widehat{S}_j \leftarrow \widehat{S}_j \setminus \{k\}$
end
return $\widehat{S}_j$

Layerwise learning and diagnostic stopping. Layerwise learning constructs the DAG by successively assigning nodes to layers and selecting their parents from previously assigned layers. At step t, Algorithm 1(a) applies LocalSearch to every remaining node to search for its parents among previously assigned nodes and determine its eligibility for the current layer. Nodes returning an estimated parent set form $\widehat{\mathcal{L}}_t$, and their estimated incoming edges are added to the graph. A node returning DEFER remains unassigned and is reconsidered at subsequent layer steps. If every remaining node returns DEFER, no further layer can be formed, and PoET stops and returns the partially learned graph. Upon termination, assigned nodes receive the label SELECTED, and unresolved nodes receive VIOLATION.

Screening-guided parent candidate search. In Algorithm 1(b), for a target node j, LocalSearch first constructs a parent candidate set $\widehat{C}_j$ from earlier-layer nodes whose conditional-uncorrelatedness null is rejected after adjustment for $\widehat{Z}^{(t)}_{kj}$. When the previous layers and their edges are correctly recovered, Lemma 3 shows that, for a true layer-t node j, nonzero population residual covariance with an earlier-layer node k implies $k \in \mathrm{Pa}_G(j)$. For targets in later layers, nonzero residual covariance need not imply a parent relationship. Initializing the conditioning set with $\widehat{C}_j$ rather than all earlier-layer nodes can reduce the dimension of conditional-mean estimation, while parents missed by screening remain available during expansion.

Equidispersion-guided expansion. Expansion searches for true parents missed by parent-candidate screening. Starting from $\widehat{S}_j = \widehat{C}_j$, Algorithm 1(b) tests $H^{\mathrm{eq}}_0(j; \widehat{S}_j)$ and returns $\widehat{C}_j$ upon non-rejection. Otherwise, it considers remaining earlier-layer nodes in decreasing order of $T^{\mathrm{corr}}_{jk|\widehat{Z}^{(t)}_{kj}}$, adding k when $Q_{j,\widehat{S}_j \cup \{k\}} - Q_{j,\widehat{S}_j} \leq \epsilon$. The null is retested after each addition, and expansion continues until non-rejection or exhaustion of the candidates. If expansion achieves non-rejection, the resulting $\widehat{S}_j$ is treated as a parent-containing set and passed to the pruning step. The ordering is a finite-sample heuristic that prioritizes more parent-like candidates, while the addition criterion is motivated by the fact that adding an admissible variable cannot increase the population dispersion gap. If rejection persists, LocalSearch returns DEFER, leaving j unassigned at layer step t.

Drop-one parent pruning. Pruning aims to remove nonparents introduced during expansion. Once a conditioning set makes j layer-eligible, the final part of Algorithm 1(b) examines $\widehat{S}_j \setminus \widehat{C}_j$ in increasing order of $T^{\mathrm{corr}}_{jk|\widehat{Z}^{(t)}_{kj}}$, prioritizing more weakly associated additions. Each candidate k is removed if $H^{\mathrm{eq}}_0(j; \widehat{S}_j \setminus \{k\})$ is not rejected, using the current set after any preceding removals. The final $\widehat{S}_j$ is returned as the estimated parent set $\widehat{P}_j$.

3.2 Local test procedures used by PoET

This subsection defines the two local tests used by PoET, both based on cross-fitted CMP-kernel residuals and truncation over count-valued conditioning variables. For node ℓ and $U \subseteq V \setminus \{\ell\}$, define $m_{\ell,U}(\boldsymbol{x}_U) := \mathbb{E}(X_\ell \mid \boldsymbol{X}_U = \boldsymbol{x}_U)$ and $e_{\ell,U} := X_\ell - m_{\ell,U}(\boldsymbol{X}_U)$. Let $\mathbb{X} = (x_{i\ell}) \in \mathbb{N}_0^{n\times p}$ be the observed data matrix and $\mathbf{x}_{i,U} = (x_{iu} : u \in U)$. With folds $I_1, \ldots, I_B$ and $b(i)$ denoting the fold containing observation i, the cross-fitted CMP-kernel estimator and residual are

$$\widehat{m}_{\ell,U}^{(-b)}(\mathbf{x}_{i,U}) = \frac{\sum_{r\notin I_b} K_{\mathbf{x}_{i,U},h}(\mathbf{x}_{r,U})x_{r\ell}}{\sum_{r\notin I_b} K_{\mathbf{x}_{i,U},h}(\mathbf{x}_{r,U})}, \qquad \widehat{e}_{i\ell,U} = x_{i\ell} - \widehat{m}_{\ell,U}^{(-b(i))}(\mathbf{x}_{i,U}). \tag{8}$$

The out-of-fold construction reduces overfitting bias in the residual-based statistics (Chernozhukov et al., 2017).

Conditional-uncorrelatedness test. For parent-candidate screening, PoET tests conditional uncorrelatedness between a remaining node j and a previous-layer node k after adjusting for some set C. Define the residual covariance and correlation as $\gamma_{jk|C} := \mathbb{E}(e_{j,C}e_{k,C})$ and $\rho_{jk|C} := \gamma_{jk|C}/(\gamma_{jj|C}\gamma_{kk|C})^{1/2}$, respectively. Using the CMP product kernel in Equations (3)–(4) and the cross-fitted residuals in (8), the corresponding estimators are

$$\widehat{\gamma}_{jk|C} = \sum_{i=1}^{n} \widehat{e}_{ij,C}\widehat{e}_{ik,C}\mathbb{I}_{\{\|\mathbf{x}_{i,C}\|_1\le M_n\}}/n, \qquad \widehat{\rho}_{jk|C} = \widehat{\gamma}_{jk|C}/(\widehat{\gamma}_{jj|C}\widehat{\gamma}_{kk|C})^{1/2},$$

where M_n is the truncation threshold in Assumption 1, restricting the statistic to a high-probability region and excluding rare large-count configurations where local smoothing is less stable. Let

$$\widehat{V}_{jk|C} = \sum_{i=1}^{n} \left(\widehat{e}_{ij,C}\widehat{e}_{ik,C} - \widehat{\gamma}_{jk|C}\right)^2 \mathbb{I}_{\{\|\mathbf{x}_{i,C}\|_1\le M_n\}}/n, \qquad \widehat{V}_{\rho,jk|C} = \widehat{V}_{jk|C}/(\widehat{\gamma}_{jj|C}\widehat{\gamma}_{kk|C}).$$

The test statistic and p-value are

$$T_{jk|C}^{\mathrm{corr}} = n\widehat{\rho}_{jk|C}^2/\widehat{V}_{\rho,jk|C}, \qquad p_{jk|C}^{\mathrm{corr}} = 1 - F_{\chi_1^2}\left(T_{jk|C}^{\mathrm{corr}}\right). \tag{9}$$

Local equidispersion test. For layer assignment and parent recovery, PoET tests $H_0^{\mathrm{eq}}(j; S)$ using the cross-fitted residuals in (8) with $\ell = j$ and $U = S$. The local deviation from equidispersion is estimated by

$$Q_{j,S} = \sum_{i=1}^{n} \left\{\widehat{e}_{ij,S}^2 - x_{ij}\right\} \mathbb{I}_{\{\|\mathbf{x}_{i,S}\|_1\le M_n\}}/n, \tag{10}$$

Under the Poisson DAG model, $Q_{j,S}$ estimates $\Delta_{j|S}$ for admissible S. Thus, under the local equidispersion null, $Q_{j,S}$ is centered near zero, while omitted parents induce positive values.

Let $\widehat{V}_{j,S}$ be the sample variance estimator, defined as

$$\widehat{V}_{j,S} = \sum_{i=1}^{n} \left\{\widehat{e}_{ij,S}^2 - x_{ij} - Q_{j,S}\right\}^2 \mathbb{I}_{\{\|\mathbf{x}_{i,S}\|_1\le M_n\}}/n.$$

We use the standardized statistic

$$T_{j,S}^{eq} = \sqrt{n}\, Q_{j,S}/\widehat{V}_{j,S}^{1/2}, \qquad p_{j,S}^{\mathrm{eq}} = 2\{1 - \Phi(|T_{j,S}^{eq}|)\}, \tag{11}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Small $p_{j,S}^{\mathrm{eq}}$ provides two-sided evidence against equidispersion, with positive $Q_{j,S}$ indicating overdispersion, as under omitted parents, and negative $Q_{j,S}$ indicating underdispersion-type departures.

4 Theory

This section connects the local tests in Section 3.2 to the layerwise decisions in Algorithm 1. We first establish local test guarantees, and then prove consistency for graph recovery under correct specification and violation detection under local misspecification. Throughout, p and the number of cross-fitting folds are fixed. All proofs and auxiliary results are deferred to Appendix A.2.

4.1 Local test guarantees

Assumption 1 (Tuning regime). *Let h_n and M_n be the bandwidth and truncation threshold from Sections 2.3 and 3.2. For a constant $c_h > 0$, assume that*

$$h_n = O(n^{-c_h}), \qquad M_n \to \infty, \qquad \limsup_{n\to\infty} M_n \log M_n / \log n < 1/(4p).$$

These conditions concentrate the CMP kernel around each evaluation point while preventing the truncated region from retaining too little probability mass, yielding the convergence of the cross-fitted conditional-mean estimators. In finite-sample settings, applying the threshold M_n might discard too many observations. In practice, we can use

$$\widetilde{M}_{n,C} = \max\{\widehat{q}_{\tau,n}(C), M_n\}, \tag{12}$$

where, for a fixed $\tau \in (0,1)$ and a conditioning set C, $\widehat{q}_{\tau,n}(C)$ is the empirical τ-quantile of $\|\mathbf{x}_{r,C}\|_1$ for $r = 1, \ldots, n$. This adjustment retains a non-negligible fraction of the sample when M_n is small. Its asymptotic equivalence to M_n is deferred to Appendix A.2.

Lemma 2 (Asymptotic guarantees for the local tests). *Suppose the data follow the non-degenerate Poisson DAG model in Equation* (1)*, with link functions that are Lipschitz and bounded away from zero, and Assumption 1 holds. Then the following statements hold for every fixed $\alpha \in (0,1)$.*

(i) Conditional-uncorrelatedness test. *Fix distinct nodes j, k and $C \subseteq V \setminus \{j,k\}$. Suppose that $m_{j,C}$ and $m_{k,C}$ are Lipschitz, and $H_0^{\mathrm{corr}}(j,k;C)$ holds. Then*

$$T^{\mathrm{corr}}_{jk|C} \rightsquigarrow \chi^2_1, \qquad \mathbb{P}\{p^{\mathrm{corr}}_{jk|C} \le \alpha\} \to \alpha.$$

(ii) Local equidispersion test. *Fix a node j and an admissible set S, and suppose that $m_{j,S}$ is Lipschitz. Under $H_0^{\mathrm{eq}}(j;S)$,*

$$T^{\mathrm{eq}}_{j,S} \rightsquigarrow N(0,1), \qquad \mathbb{P}\{p^{\mathrm{eq}}_{j,S} \le \alpha\} \to \alpha.$$

If S omits at least one parent of j, then, for some $c_{j,S} > 0$,

$$T^{\mathrm{eq}}_{j,S}/\sqrt{n} \xrightarrow{p} c_{j,S}, \qquad \mathbb{P}\{p^{\mathrm{eq}}_{j,S} \le \alpha\} \to 1.$$

Part (i) provides null calibration for the screening used to form initial parent candidates. Although the statistic uses averaged residual covariance whereas the null is conditional uncorrelatedness, the conditional null implies a zero averaged covariance, which is the only direction required by PoET; power against every alternative is not needed. Part (ii) distinguishes admissible sets containing all parents from those omitting at least one, supporting layer and parent recovery.

4.2 Consistency guarantees for PoET

Lemma 3 (Nonparent separation by the local neighborhood). *Fix t with $\mathcal{L}_t(G) \neq \emptyset$, $j \in \mathcal{L}_t(G)$, and $k \in \mathcal{L}_{<t}(G)$. Define the population counterpart of* (7) *by $Z^{(t)}_{kj} := Ne_{G_{<t}}(k) \setminus \{j,k\}$. Then*

$$k \notin Pa_G(j) \quad \Longrightarrow \quad k \perp_{d,G} j \mid Z^{(t)}_{kj}.$$

Thus, under correct previous-layer recovery, neighborhood-adjusted screening admits only true parents at the population level. Nodes excluded by screening remain available for inclusion during expansion, so screening does not permanently eliminate potential parents from the search.

Theorem 4 (Consistency of PoET under correct specification). *Suppose $P_{\boldsymbol{X}}$ is a non-degenerate Poisson DAG model over G, and the conditions of Lemma 2 hold for every local test used by PoET. Run PoET with fixed $\epsilon > 0$ and a common testing level satisfying $\alpha_n \to 0$ and $\log(1/\alpha_n) = o(n)$. Then*

$$\mathbb{P}\Big(\widehat{G} = G, \quad \widehat{\ell}_j = \textsc{selected} \text{ for all } j \in V\Big) \longrightarrow 1.$$

The proof mirrors Algorithm 1: screening through conditional-uncorrelatedness tests identifies true-parent subsets, equidispersion-guided expansion and pruning recover the parent sets, and test-based eligibility recovers each successive layer. The local p-values in Lemma 2 are calibrated for fixed hypotheses, whereas PoET uses them adaptively within its sequential learning procedure. Thus, Theorem 4 establishes asymptotic graph-recovery consistency under correct specification, rather than finite-sample simultaneous or post-selection error control over the local decisions.

Theorem 5 (Consistency of violation detection under local model misspecification)**.** *Suppose that $P_{\boldsymbol{X}}$ factorizes according to a DAG G, and let $\mathcal{W} \subseteq V$ be a nonempty set of misspecified nodes. Let Equation* (1) *hold with non-degenerate link functions at every $j \in V \setminus \mathcal{W}$, and the population dispersion gap $\Delta_{k|S}$ be nonzero for every $k \in \mathcal{W}$ and every admissible S, with $\mathbb{E}(X_k^4) < \infty$ and $\mathrm{Var}(e_{k,S}^2 - X_k) > 0$. Assume Assumption 1. Suppose further that the conditional mean functions are Lipschitz and bounded away from zero, and that $m_{j,S}$ is Lipschitz for every node j and admissible S. Run PoET under the same ϵ and α_n regime as in Theorem 4, and define $\mathcal{V}_{\mathcal{W}} := \bigcup_{k\in\mathcal{W}}\big(\{k\} \cup De_G(k)\big)$. Then*

$$\mathbb{P}\left(\widehat{\ell}_j = \textsc{selected}\ \forall j \in V \setminus \mathcal{V}_{\mathcal{W}}, \quad \widehat{\ell}_j = \textsc{violation}\ \forall j \in \mathcal{V}_{\mathcal{W}}\right) \to 1.$$

This result formalizes diagnostic stopping. The label VIOLATION identifies the unresolved region affected by local misspecification, including both the misspecified nodes in W and their descendants, which together form $\mathcal{V}_W$. Since misspecified mechanisms need not inherit the moment bounds automatically guaranteed by correctly specified Poisson DAG models, we impose moment conditions to establish the local-test asymptotics used in the proof. The result covers multiple local violations, including binomial, negative-binomial, and latent-variable models that satisfy the stated assumptions.

5 Numerical Experiments

We evaluate PoET under misspecified and correctly specified Poisson DAG models, focusing on violation detection and graph recovery performance, respectively.

5.1 Simulation settings

For the main experiments, we consider $p = 30$, $n \in \{1000, 3000, 5000, 7000, 10000\}$, and 100 replications. We use hub graphs, Barabási–Albert (BA), and Erdős–Rényi (ER) with graph density $g \in \{2, 4\}$ for each graph family, corresponding to sparse and dense settings, respectively. Each Poisson mechanism independently uses one of four link functions: Box–Cox, sine-sum, parity, or periodic jump. Both local tests use $\alpha_n = 2\{1 - \Phi(n^{1/4}/3)\}$, 5-fold cross-fitting, and CMP-kernel bandwidth $h_n = 0.5n^{-1/(2+d)}$, where d is the conditioning-set size. We use the adaptive truncation threshold $\widetilde{M}_{n,C} = \max\{\widehat{q}_{\tau,n}(C), M_n\}$ in (12), with $\tau = 0.1$ and $M_n = \frac{1}{8(p+1)}\frac{\log n}{\log\log n}$, and set the expansion tolerance to $\epsilon = 10^{-6}$. Graph generation, link specifications, additional simulation results, tuning-parameter sensitivity, and diagnostic analyses are provided in Appendix A.3.

5.2 Detection under model violations

Table 1: Average violation-detection F1 scores (standard errors) of PoET under misspecified models with $g \in \{2, 4\}$. Results including $n \in \{3000, 7000\}$ are provided in Appendix A.3.4.

Graph density	n	Binomial			Negative Binomial			Latent Poisson		
		Hub	BA	ER	Hub	BA	ER	Hub	BA	ER
$g = 2$	1000	0.88 (0.016)	0.77 (0.017)	0.78 (0.012)	0.36 (0.032)	0.23 (0.028)	0.24 (0.020)	0.38 (0.032)	0.40 (0.032)	0.43 (0.027)
	5000	0.95 (0.009)	0.85 (0.014)	0.84 (0.012)	0.90 (0.022)	0.81 (0.020)	0.80 (0.018)	0.68 (0.030)	0.66 (0.027)	0.76 (0.024)
	10000	0.96 (0.007)	0.88 (0.013)	0.89 (0.010)	0.94 (0.018)	0.88 (0.017)	0.91 (0.011)	0.75 (0.030)	0.82 (0.019)	0.87 (0.016)
$g = 4$	1000	0.85 (0.016)	0.77 (0.017)	0.77 (0.011)	0.42 (0.034)	0.44 (0.038)	0.39 (0.017)	0.50 (0.027)	0.58 (0.030)	0.65 (0.019)
	5000	0.93 (0.008)	0.87 (0.013)	0.88 (0.009)	0.88 (0.019)	0.83 (0.023)	0.86 (0.014)	0.77 (0.024)	0.82 (0.020)	0.86 (0.015)
	10000	0.94 (0.007)	0.90 (0.012)	0.91 (0.008)	0.95 (0.012)	0.92 (0.016)	0.91 (0.011)	0.87 (0.016)	0.89 (0.014)	0.91 (0.011)

We first evaluate whether PoET detects violations of the conditional Poisson assumption under three mechanisms: binomial, negative binomial, and latent variables. For the first two, one node is

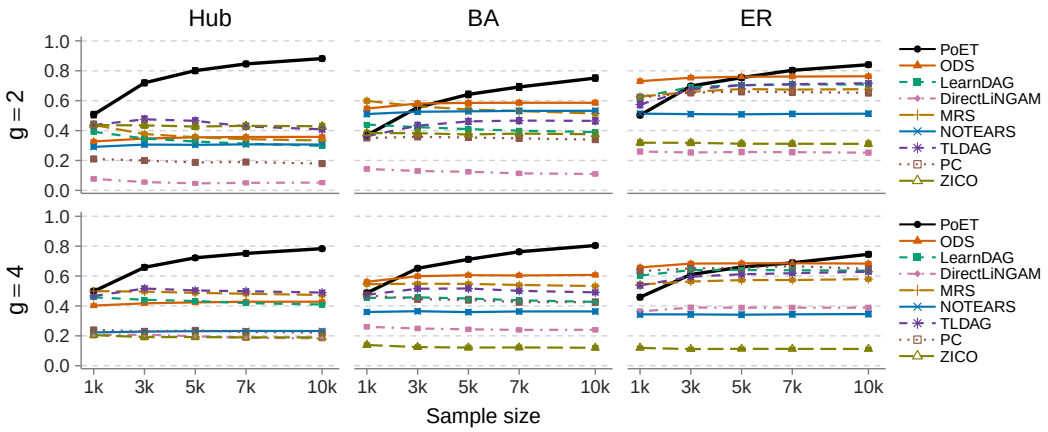


Figure 1: Average graph-recovery F1 scores with one-standard-error bars under correctly specified Poisson DAG models.

selected uniformly at random and assigned the corresponding non-Poisson mechanism, while each remaining node independently follows either the same mechanism or one of the four conditional Poisson mechanisms in Section 5.1, with equal probability. For latent-variable violations, we sample one to three latent variables uniformly, generate a Poisson DAG on the augmented node set with each latent variable having at least two observed children, and remove the latent variables before applying PoET. Details are provided in Appendix A.3.3. We report average *violation-detection* F1 scores, treating nodes labeled VIOLATION as predicted positives and $\mathcal{V}_{\mathcal{W}}$ as the true positives set. Precision is the fraction of predicted positives belonging to the target set, and recall is the fraction of the target set detected. We compute $\mathrm{F1} = 2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall})$. Table 1 shows that violation-detection F1 generally improves with sample size.

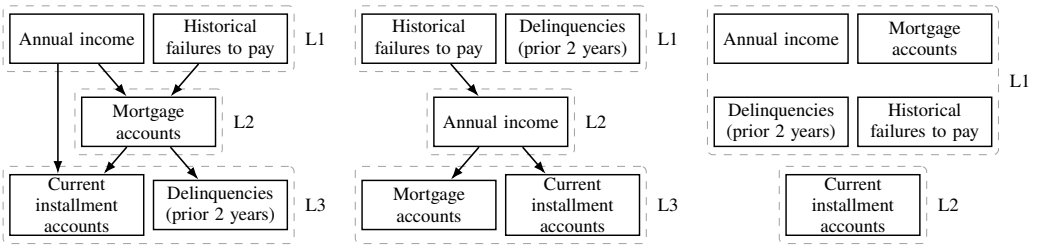
5.3 Graph recovery under correct specification

We next evaluate PoET under correctly specified Poisson DAG models, where each node follows (1) and its link function μ_j is sampled uniformly from the four specifications in Section 5.1. PoET is compared with ODS (Park & Raskutti, 2015), MRS (Park & Park, 2019b), TLDAG (Zhou et al., 2022), LearnDAG (Nguyen et al., 2024), ZICO (Sato et al., 2026), PC (Spirtes et al., 2000), DirectLiNGAM (Shimizu et al., 2011), and NOTEARS (Zheng et al., 2018). The first five are designed for count data, whereas the remaining methods are general-purpose DAG learners. For PC, we use Gaussian conditional-independence tests on log-transformed counts and extend the resulting CPDAG to a DAG via the Dor–Tarsi procedure following Nguyen et al. (2024). For NOTEARS, we replace the least-squares loss with a Poisson loss to accommodate the count-valued observations. Other settings follow the original studies. We compare average *graph-recovery* F1 scores across all methods, with precision and recall computed over directed edges. When PoET terminates early with VIOLATION labels, its score is computed from the returned partial graph.

Figure 1 likewise shows improving graph-recovery F1 scores across graph families. Performance of PoET is particularly strong on hub graphs, as their shallower topological layering requires fewer sequential tests. While it can underperform some competitors at small n due to the finite-sample cost of nonparametric estimation (Stone, 1982; Lavergne & Patilea, 2008) and the additional burden of checking the model assumptions, it outperforms most competitors at larger n. Its graph-recovery F1 scores generally increase with sample size, consistent with the theoretical guarantee in Theorem 4.

6 Real Data Analysis

We use two datasets to illustrate complementary behaviors of PoET. The NBA data examine whether the marginal equidispersion implication required to initiate layer learning is supported, whereas the Lending Club data illustrate benefits of PoET under unspecified nodewise conditional link functions. We use the same tuning rules as in the numerical experiments, using the conventional significance level $\alpha_n = 0.05$.



(a) PoET (b) ODS (c) TLDAG

Figure 2: Estimated Lending Club DAGs: (a) PoET, (b) ODS, and (c) TLDAG.

NBA player dataset: We first analyze the 2009–2010 NBA player statistics dataset in the R package `SportsAnalytics`, which contains $n = 441$ players and has been used in previous studies of identifiable DAG learning (Park & Raskutti, 2018; Zhou et al., 2022). Following standard preprocessing, we retain $p = 18$ integer-valued variables; details are provided in Appendix A.4.

PoET rejects marginal equidispersion for all 18 variables, with p-values ranging from 2.62×10^{-123} to 1.32×10^{-4}, hence no node is assigned to the first layer and the procedure stops without forcing subsequent graph decisions. In particular, all source candidates identified by ODS and TLDAG also reject this necessary marginal implication of the Poisson DAG model. This example highlights a key distinction between score- and test-based learning: a score can always select the most source-like node, even when no node is sufficiently source-like, whereas PoET can reject the equidispersion null for every source candidate and stop.

Lending Club dataset: We additionally analyze the Lending Club dataset available in the R package `openintro`. After preprocessing, we retain $n = 455$ observations and $p = 5$ variables. Detailed preprocessing and variable definitions are provided in Appendix A.4.

In contrast to the NBA data, PoET completes the layerwise procedure without issuing a violation label. Figure 2 compares its graph with those estimated by ODS and TLDAG. PoET yields several directions with natural interpretations: annual income points to mortgage and current installment accounts, historical failures to pay points to mortgage accounts, and mortgage accounts point to current installment accounts and recent delinquencies. These relationships are consistent with differences in borrowers' credit access and repayment histories. By comparison, ODS includes less interpretable directions, such as historical failures to pay pointing to annual income, while TLDAG does not identify any directed edge. Although the true DAG is unknown, this example shows that PoET can produce a test-based layerwise structure without prespecifying the functional forms of the nodewise conditional link functions.

7 Conclusion and Future Work

We introduced a test-integrated framework for nonparametric Poisson DAG discovery from count-valued data. Rather than using the equidispersion–overdispersion separation only as a score or ordering criterion, we embed theoretically justified local equidispersion tests into topological-layer learning, parent recovery, and diagnostic stopping. Cross-fitted CMP-kernel smoothing estimates conditional means without prescribing nodewise link functions. Conditional-uncorrelatedness tests initialize the parent search, while equidispersion tests guide expansion and pruning of the candidate sets. Testing thus guides graph learning while checking the equidispersion implications of the underlying Poisson DAG model. When no remaining node is layer-eligible, PoET returns the partially learned graph, with SELECTED and VIOLATION labels distinguishing assigned and unresolved nodes.

Several directions remain open. The same test-integrated perspective could be extended beyond Poisson DAGs to broader count mechanisms, including QVF, negative-binomial, zero-inflated, or conditional CMP-type models. Sharper graph-level error control for adaptive test-based DAG learning is another important direction. Finally, extensions to dependent count observations such as INGARCH-type time series would further broaden the scope of inferential count causal discovery.

AI USE STATEMENT

In this work, we used generative AI tools to get feedback on experiments. We did not use generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, implement methods, clean or reformat datasets, or interpret results. Assistance with translation, and supporting qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the research paper to improve readability. LLM-generated code and AI-assisted written content were reviewed and verified by two authors for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT.

For reproducibility, we provide a comprehensive description of the experimental configuration, hyperparameter settings, and implementation procedures in the appendix. Upon publication, we will make the source code and processed datasets publicly available.

REFERENCES

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

A Colin Cameron and Pravin K Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364, 1990.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

Junsouk Choi and Yang Ni. Model-based causal discovery for zero-inflated count data. *Journal of Machine Learning Research*, 24(200):1–32, 2023.

Junsouk Choi, Robert Chapkin, and Yang Ni. Bayesian causal structural learning with zero-inflated Poisson Bayesian networks. *Advances in neural information processing systems*, 33:5887–5897, 2020.

Charmaine Dean and Jerald Franklin Lawless. Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472, 1989.

Charmaine B Dean. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

Alan Huang, Lucas Sippel, and Thomas Fung. Consistent second-order discrete kernel smoothing using dispersed conway–maxwell–poisson kernels. *Computational Statistics*, 37(2):551–563, 2022.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Pascal Lavergne and Valentin Patilea. Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics*, 143(1):103–122, 2008.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

Simos G Meintanis, Joseph Ngatchou-Wandji, Leonard Santana, and Marius Smuts. Goodness-of-fit tests for discrete response models with covariates. *Statistical Papers*, 66(3):54, 2025.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pp. 745–752, 2009.

Thi Kim Hue Nguyen, Monica Chiogna, and Davide Risso. Unguided structure learning of dags for count data, 2024. URL https://arxiv.org/abs/2406.04994.

Gunwoong Park and Hyewon Park. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 158–166. PMLR, 2019a.

Gunwoong Park and Sion Park. High-dimensional poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20(95):1–41, 2019b.

Gunwoong Park and Garvesh Raskutti. Learning large-scale Poisson DAG models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, volume 28, pp. 631–639, 2015.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (QVF) DAG models via overdispersion scoring (ODS). *Journal of Machine Learning Research*, 18(224):1–44, 2018.

Noriaki Sato, Marco Scutari, Shuichi Kawano, Rui Yamaguchi, and Seiya Imoto. DAG learning from zero-inflated count data using continuous optimization. In *Fifth Conference on Causal Learning and Reasoning*, 2026.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020. doi: 10.1214/19-AOS1857.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, 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-JMLR*, 12(Apr):1225–1248, 2011.

Sobom M Somé and Célestin C Kokonendji. Effects of associated kernels in nonparametric multiple regressions. *Journal of Statistical Theory and Practice*, 10(2):456–471, 2016.

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

Jakub Stoklosa, Rachel V Blakey, and Francis KC Hui. An overview of modern applications of negative binomial modelling in ecology and biodiversity. *Diversity*, 14(5):320, 2022.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Y Samuel Wang, Mladen Kolar, and Mathias Drton. Confidence sets for causal orderings. *Journal of the American Statistical Association*, 121(553):690–703, 2026.

David I Warton. Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74 (1):362–368, 2018.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pp. 5561–5570. PMLR, 2018.

Ruixuan Zhao, Xin He, and Junhui Wang. Learning linear non-Gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research*, 23(269):1–34, 2022.

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, 2018.

Wei Zhou, Xin He, Wei Zhong, and Junhui Wang. Efficient learning of quadratic variance function directed acyclic graphs via topological layers. *Journal of Computational and Graphical Statistics*, 31(4):1269–1279, 2022.

Wei Zhou, Xueqian Kang, Wei Zhong, and Junhui Wang. Efficient learning of DAG structures in heavy-tailed data. *Statistica Sinica*, 37(3), 2027. doi: 10.5705/ss.202024.0199. Forthcoming.

A Appendix

A.1 Comparison with Related Work

Table 2 compares PoET with related work along the four contributions summarized in Section 1, highlighting the methodological differences and theoretical guarantees. Beyond the comparisons, we discuss connections to existing work.

Equidispersion and discrete goodness-of-fit testing. Testing equidispersion is classical for Poisson regressions with prespecified conditional means (Dean & Lawless, 1989; Cameron & Trivedi, 1990; Dean, 1992). Broader goodness-of-fit procedures assess specified discrete distributions through Stein characterizations (Yang et al., 2018) or conditional count-response models through probability-generating functions (Meintanis et al., 2025). PoET targets the equidispersion implication used for Poisson-DAG identification and incorporates its tests into layer assignment, parent recovery, and diagnostic stopping, while leaving nodewise link functions unspecified.

Residual-based testing and cross-fitting. The generalised covariance measure constructs tests from products of regression residuals after nonparametric conditional-mean estimation (Shah & Peters, 2020). Cross-fitting separates nuisance estimation from score evaluation and is a central ingredient of double/debiased machine learning (Chernozhukov et al., 2017). PoET shares these methodological ideas and develops their use in count-data DAG learning through CMP-kernel smoothing. Its conditional-uncorrelatedness test screens parent candidates, while equidispersion tests guide layer assignment, parent recovery, and stopping.

Testing in causal discovery. Mooij et al. (2009) use residual-independence tests to select successive sink nodes in additive-noise models and stop when no candidate passes the test. PoET shares this principle of test-based selection and stopping, but uses conditional equidispersion rather than residual independence to identify successive layers and recover parents in nonparametric Poisson DAGs. Goodness-of-fit tests have also been used to construct confidence sets of causal orderings under additive-error models (Wang et al., 2026).

A.2 Proofs of the theoretical results

We follow the data-matrix notation of Section 3.2. Thus, $X_{i\ell}$ denotes the random variable corresponding to the observed value $x_{i\ell}$, and $\mathbf{X}_{i,U} := (X_{iu} : u \in U)$. The row vectors $(X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, are independent and identically distributed with the same law as $\boldsymbol{X}$. The number of folds B is fixed, and $|I_b| \asymp n$ and $|I_b^c| \asymp n$ uniformly in b.

The proofs have two common ingredients. First, the cross-fitted CMP–NW estimator must approximate each conditional mean uniformly on the region where the test statistic is evaluated. Second, after replacing the fitted residuals by the corresponding population residuals, the remaining leading term is an average of i.i.d. random variables and can be handled by a law of large numbers or a central limit theorem. We therefore prove the uniform regression result first.

A.2.1 A moment bound for the Poisson DAG

Lemma 6 (Exponential moment of a finite Poisson DAG)**.** *Suppose that G is finite and acyclic and that every nodewise link function is globally Lipschitz. Then there exists $\eta > 0$ such that*

$$\mathbb{E}exp\{\eta \|\boldsymbol{X}\|_1\} < \infty. \tag{13}$$

Consequently, every polynomial moment of every coordinate of $\boldsymbol{X}$ is finite.

Proof. Let $v_1, \ldots, v_p$ be a topological ordering of G. Global Lipschitz continuity gives constants $a_r, L_r < \infty$ such that

$$\mu_{v_r}(\boldsymbol{x}_{\mathrm{Pa}_G(v_r)}) \le a_r + L_r \sum_{s<r} x_{v_s}, \qquad r = 1, \ldots, p. \tag{14}$$

For the first node, whose parent set is empty,

$$\mathbb{E}\exp(tX_{v_1}) = \exp\{\mu_{v_1}(e^t - 1)\} < \infty$$

Table 2: Positioning of PoET. The four rows correspond directly to the four contributions summarized in Section 1.

PoET contribution	Related line of work	What is missing	What PoET adds
From score to test	ODS and QVF-DAG exploit the equidispersion–overdispersion separation as an ordering or structure-learning criterion (Park & Raskutti, 2015; Zhou et al., 2022).	The mean–variance asymmetry is used primarily as a numerical score, without directly testing whether the local equidispersion implication used for identification is supported by the data.	Converts the equidispersion implication into local p-values for layer eligibility, conditioning-set expansion, and parent pruning, and permits termination when no next layer is supported.
Flexible nonparametric learning	Regression-based Poisson DAG methods often use prespecified conditional-mean structures, while ODS estimates conditional moments through exact-cell matching. Discrete associated-kernel methods provide nonparametric smoothing on count support (Park & Raskutti, 2015; Park & Park, 2019b; Somé & Kokonendji, 2016; Huang et al., 2022).	Prespecified links restrict functional flexibility, whereas exact-cell estimators can become unstable when conditioning configurations are sparsely observed. Existing count-kernel methods are not integrated into Poisson DAG learning.	Leaves the nodewise conditional link functions unspecified and estimates the conditional moments required for graph learning through cross-fitted CMP-kernel smoothing.
Screening for efficient learning	Layer-based DAG methods restrict parent candidates to previously recovered layers, while nonparametric local testing requires smoothing over the selected conditioning variables (Zhou et al., 2022; Zhao et al., 2022; Zhou et al., 2027; Stone, 1982; Lavergne & Patilea, 2008).	The set of earlier-layer candidates may remain large, increasing the number of local tests, the computational burden of kernel estimation, and the dimension of the nonparametric smoothing problem.	Uses a CMP-kernel conditional uncorrelatedness test to construct target-specific parent candidates, which empirically reduces the size of conditioning sets.
Test validity and graph consistency	Poisson and QVF-DAG methods establish identifiability and consistent graph recovery under their respective model assumptions (Park & Raskutti, 2015; 2018; Zhou et al., 2022).	Graph-recovery guarantees under correct specification do not by themselves establish calibration of local tests or detection of model violations.	Establishes asymptotic properties for both local tests, and consistency of PoET for graph recovery under correct specification and violation detection under local misspecification.

for every $t > 0$. Suppose inductively that, for some $\eta_{r-1} > 0$,

$$\mathbb{E}\exp\left\{\eta_{r-1}\sum_{s<r} X_{v_s}\right\} < \infty.$$

Choose $\eta_r > 0$ sufficiently small that

$$\eta_r + L_r(e^{\eta_r} - 1) \le \eta_{r-1}. \tag{15}$$

This is possible because the left-hand side converges to zero as $\eta_r \downarrow 0$. Using the conditional moment-generating function of a Poisson variable and (14),

$$\begin{aligned}
&\mathbb{E}\exp\left\{\eta_r \sum_{s\le r} X_{v_s}\right\} \\
&\quad = \mathbb{E}\left[\exp\left\{\eta_r \sum_{s<r} X_{v_s}\right\} \exp\left\{\mu_{v_r}(\boldsymbol{X}_{\mathrm{Pa}_{G(v_r)}})(e^{\eta_r}-1)\right\}\right] \\
&\quad \le \exp\{a_r(e^{\eta_r}-1)\}\mathbb{E}\exp\left[\{\eta_r + L_r(e^{\eta_r}-1)\}\sum_{s<r} X_{v_s}\right] \\
&\quad < \infty
\end{aligned}$$

by (15). Induction up to $r = p$ proves (13). The moment conclusion follows from the elementary bound $x^q \le C_{q,\eta} e^{\eta x}$ for $x \ge 0$. □

A.2.2 Uniform convergence of the cross-fitted CMP–NW estimator

Lemma 7 (Uniform convergence of the cross-fitted CMP–NW estimator)**.** *Fix a node ℓ and a conditioning set $U \subseteq V \setminus \{\ell\}$. Put $d = |U|$ and define*

$$\mathcal{S}_n := \{\boldsymbol{z} \in \mathbb{N}_0^d : \|\boldsymbol{z}\|_1 \le M_n\}, \quad f_U(\boldsymbol{z}) := \mathbb{P}(\boldsymbol{X}_U = \boldsymbol{z}), \quad \delta_n(U) := \min_{\boldsymbol{z}\in\mathcal{S}_n} f_U(\boldsymbol{z}).$$

For $U = \emptyset$, use the convention $\mathcal{S}_n = \{\emptyset\}$ and $\delta_n(U) = 1$. Let $\widehat{m}_{\ell,U}^{(-b)}$ be the out-of-fold estimator in Equation (8)*, evaluated at a generic $\boldsymbol{z} \in \mathcal{S}_n$ by the same formula. Suppose the data are generated by the Poisson DAG model in Equation* (1)*, the nodewise link functions are globally Lipschitz and bounded away from zero, $m_{\ell,U}$ is Lipschitz, and Assumption 1 holds. Then*

$$\max_{1\le b\le B}\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\widehat{m}_{\ell,U}^{(-b)}(\boldsymbol{z}) - m_{\ell,U}(\boldsymbol{z})\right| = O_p\left[\frac{exp\{-1/[2h_n(M_n+1)]\} + \sqrt{\log(2|\mathcal{S}_n|)/n}}{\delta_n(U)}\right] = o_p(n^{-1/4}).$$

Moreover, with probability tending to one, every denominator in Equation (8) *is positive uniformly over $\boldsymbol{z} \in \mathcal{S}_n$ and $b = 1, \dots, B$.*

Proof. The claim is immediate for $U = \emptyset$. Hence assume $d \ge 1$. For fold b, let $N_b := |I_b^c|$ and write

$$\widehat{m}_{\ell,U}^{(-b)}(\boldsymbol{z}) = \frac{\widehat{r}_{\ell,U}^{(-b)}(\boldsymbol{z})}{\widehat{f}_U^{(-b)}(\boldsymbol{z})},$$

where

$$\widehat{r}_{\ell,U}^{(-b)}(\boldsymbol{z}) := \frac{1}{N_b}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})X_{i\ell}, \quad \widehat{f}_U^{(-b)}(\boldsymbol{z}) := \frac{1}{N_b}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U}).$$

The expectation of the denominator is

$$\mathbb{E}\{\widehat{f}_U^{(-b)}(\boldsymbol{z})\} = \sum_{\boldsymbol{u}\in\mathbb{N}_0^d} K_{\boldsymbol{z},h_n}(\boldsymbol{u}) f_U(\boldsymbol{u}).$$

We first bound the cell probabilities, then the denominator, and finally the numerator.

Step 1: lower bound for the conditioning-cell probabilities. Fix $\boldsymbol{z} \in \mathcal{S}_n$. Marginalizing the joint Poisson-DAG probability over the coordinates outside U gives

$$f_U(\boldsymbol{z}) = \sum_{\boldsymbol{x}_{V\setminus U}\in\mathbb{N}_0^{p-d}} \prod_{v=1}^{p} \frac{\mu_v(\boldsymbol{x}_{\mathrm{Pa}_G(v)})^{x_v}\exp\{-\mu_v(\boldsymbol{x}_{\mathrm{Pa}_G(v)})\}}{x_v!}, \tag{16}$$

where the coordinates indexed by U are fixed at $\boldsymbol{z}$. All summands are nonnegative. Retain the summand in which every coordinate outside U is zero. In this configuration, $0 \le x_v \le M_n$ for every v.

Let $\omega_{\min} > 0$ be a common lower bound for the nodewise conditional means. By global Lipschitz continuity, there are constants $C_0, C_g < \infty$ such that, on the selected configuration,

$$\omega_{\min} \le \mu_v(\boldsymbol{x}_{\mathrm{Pa}_{G(v)}}) \le C_0 + C_g M_n, \qquad v = 1, \dots, p.$$

Consequently,

$$\mu_v(\boldsymbol{x}_{\mathrm{Pa}_{G(v)}})^{x_v} \ge \min\{1, \omega_{\min}^{M_n}\}, \qquad e^{-\mu_v(\boldsymbol{x}_{\mathrm{Pa}_{G(v)}})} \ge e^{-C_0 - C_g M_n},$$

and, because $x_v \le M_n$,

$$x_v! \le (M_n + 1)^{M_n + 1}.$$

Substitution into the selected summand of (16) yields

$$f_U(\boldsymbol{z}) \ge \prod_{v=1}^{p} \frac{\min\{1, \omega_{\min}^{M_n}\} \exp(-C_0 - C_g M_n)}{(M_n + 1)^{M_n + 1}}.$$

Thus, using

$$-\log \min\{1, \omega_{\min}^{M_n}\} \le M_n |\log \omega_{\min}|,$$

we obtain

$$\begin{aligned} -\log f_U(\boldsymbol{z}) &\le p(M_n + 1)\log(M_n + 1) + pM_n |\log \omega_{\min}| \\ &\quad + pC_g M_n + pC_0 \\ &= pM_n \log M_n + O(M_n), \end{aligned}$$

uniformly over $\boldsymbol{z} \in \mathcal{S}_n$.

We now use Assumption 1. Since

$$\limsup_{n \to \infty} \frac{M_n \log M_n}{\log n} < \frac{1}{4p},$$

we can choose a constant $c_\delta > p$ and $\beta < 1/4$ such that

$$c_\delta \limsup_{n \to \infty} \frac{M_n \log M_n}{\log n} < \beta < \frac{1}{4}. \tag{17}$$

Then, for all sufficiently large n,

$$-\log f_U(\boldsymbol{z}) \le c_\delta M_n \log M_n$$

uniformly over $\boldsymbol{z} \in \mathcal{S}_n$. Hence

$$\delta_n(U) \ge \exp\{-c_\delta M_n \log M_n\}. \tag{18}$$

Equation (17) also implies that, for all sufficiently large n,

$$c_\delta M_n \log M_n \le \beta \log n.$$

Combining this inequality with (18) gives

$$\delta_n(U) \ge \exp(-\beta \log n) = n^{-\beta}. \tag{19}$$

We will also need the number of lattice points in the truncation region. By the stars-and-bars formula,

$$\begin{aligned} |\mathcal{S}_n| &= \#\left\{ \boldsymbol{z} \in \mathbb{N}_0^d : \sum_{r=1}^{d} z_r \le M_n \right\} \\ &= \binom{\lfloor M_n \rfloor + d}{d} = O(M_n^d). \end{aligned}$$

Assumption 1 implies $M_n \log M_n = O(\log n)$, and hence $M_n = O(\log n)$. Therefore, since d is fixed,

$$|\mathcal{S}_n| = O\{(\log n)^d\}, \qquad \log(2|\mathcal{S}_n|) = O(\log \log n), \qquad \frac{|\mathcal{S}_n|}{n} \to 0. \tag{20}$$

The logarithmic bound in (20) is used in the Hoeffding and Bernstein union bounds below, while $|\mathcal{S}_n|/n \to 0$ is used in the uniform control of the numerator.

Step 2: uniform control of the denominator. For fixed $\boldsymbol{z}$, the summands in $\widehat{f}_U^{(-b)}(\boldsymbol{z})$ lie in $[0,1]$. Hoeffding's inequality and a union bound give

$$\mathbb{P}\left(\sup_{\boldsymbol{z}\in\mathcal{S}_n}|\widehat{f}_U^{(-b)}(\boldsymbol{z})-\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z})|>t\right) \leq 2|\mathcal{S}_n|\exp(-2N_bt^2).$$

Since $N_b \asymp n$,

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n}|\widehat{f}_U^{(-b)}(\boldsymbol{z})-\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z})| = O_p\left\{\sqrt{\frac{\log(2|\mathcal{S}_n|)}{n}}\right\}. \tag{21}$$

It remains to control $\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z})-f_U(\boldsymbol{z})$, which is the smoothing bias of the kernel estimator of the cell probability.

We first quantify how rapidly the one-dimensional CMP kernel decreases away from its center. For $z \geq 1$, write

$$\lambda_n(z) := \lambda(z, 1/h_n)$$

and define

$$R_{+,n}(z) := \frac{K_{z,h_n}(z+1)}{K_{z,h_n}(z)} = \frac{\lambda_n(z)}{(z+1)^{1/h_n}},$$

and

$$R_{-,n}(z) := \frac{K_{z,h_n}(z-1)}{K_{z,h_n}(z)} = \frac{z^{1/h_n}}{\lambda_n(z)}.$$

Their product is

$$R_{+,n}(z)R_{-,n}(z) = \left(\frac{z}{z+1}\right)^{1/h_n}.$$

For $1 \leq z \leq M_n$,

$$\begin{aligned} R_{+,n}(z)R_{-,n}(z) &= \left(1-\frac{1}{z+1}\right)^{1/h_n} \\ &\leq \exp\left\{-\frac{1}{h_n(z+1)}\right\} \\ &\leq \exp\left\{-\frac{1}{h_n(M_n+1)}\right\}. \end{aligned} \tag{22}$$

The first inequality uses $1-x \leq e^{-x}$, and the second follows from $z+1 \leq M_n+1$.

The mean constraint of the CMP kernel,

$$\sum_{x=0}^{\infty}(x-z)K_{z,h_n}(x) = 0,$$

divided by $K_{z,h_n}(z)$, gives

$$\sum_{q=1}^{z} q\,\frac{\{z(z-1)\cdots(z-q+1)\}^{1/h_n}}{\lambda_n(z)^q} = \sum_{q=1}^{\infty} q\,\frac{\lambda_n(z)^q}{\{(z+1)(z+2)\cdots(z+q)\}^{1/h_n}}. \tag{23}$$

We first show that both neighboring ratios are uniformly smaller than one for all sufficiently large n. Suppose, to the contrary, that this fails for $R_{+,n}$. Then there exist a sequence $n_m \to \infty$ and integers $z_m \in \{1,\ldots,\lfloor M_{n_m}\rfloor\}$ such that

$$R_{+,n_m}(z_m) \geq 1 \qquad \text{for every } m.$$

By (22),

$$R_{-,n_m}(z_m) \leq R_{+,n_m}(z_m)R_{-,n_m}(z_m) \longrightarrow 0.$$

The right-hand side of (23) is at least its $q = 1$ term, namely $R_{+,n_m}(z_m) \geq 1$. On the other hand, its left-hand side is bounded by

$$\sum_{q=1}^{\infty} qR_{-,n_m}(z_m)^q = \frac{R_{-,n_m}(z_m)}{\{1 - R_{-,n_m}(z_m)\}^2} \longrightarrow 0,$$

which contradicts the equality in (23). Hence

$$\sup_{1 \leq z \leq M_n} R_{+,n}(z) < 1$$

for all sufficiently large n. By symmetry, $\sup_{1 \leq z \leq M_n} R_{-,n}(z) < 1$ for all sufficiently large n.

In fact, both suprema converge to zero. For example, if $\sup_{1 \leq z \leq M_n} R_{+,n}(z)$ did not converge to zero, then for some $\epsilon > 0$ there would exist $n_m \to \infty$ and $z_m \leq M_{n_m}$ such that

$$R_{+,n_m}(z_m) \geq \epsilon.$$

Equation (22) would then imply

$$R_{-,n_m}(z_m) \leq \frac{R_{+,n_m}(z_m)R_{-,n_m}(z_m)}{\epsilon} \longrightarrow 0.$$

But the right-hand side of (23) would be at least ϵ, while its left-hand side would again converge to zero, a contradiction. Thus

$$\sup_{1 \leq z \leq M_n} R_{+,n}(z) \to 0.$$

The same argument gives

$$\sup_{1 \leq z \leq M_n} R_{-,n}(z) \to 0.$$

Consequently, both ratios are at most $1/2$ for all sufficiently large n, uniformly over $1 \leq z \leq M_n$. From (23),

$$R_{+,n}(z) \leq \frac{R_{-,n}(z)}{\{1 - R_{-,n}(z)\}^2} \leq 4R_{-,n}(z),$$

and symmetrically,

$$R_{-,n}(z) \leq 4R_{+,n}(z).$$

Hence,

$$\begin{aligned} \max\{R_{+,n}(z), R_{-,n}(z)\} &\leq 2\sqrt{R_{+,n}(z)R_{-,n}(z)} \\ &\leq 2\text{exp}\left\{-\frac{1}{2h_n(M_n+1)}\right\}. \end{aligned}$$

For $z = 0$, the mean-zero CMP distribution is degenerate at zero, so the upper-neighbor ratio is zero.

For $\boldsymbol{z} \in \mathbb{N}_0^d$, define the set of one-step lattice neighbors by

$$\mathcal{N}_1(\boldsymbol{z}) := \left\{\boldsymbol{u} \in \mathbb{N}_0^d : \|\boldsymbol{u} - \boldsymbol{z}\|_1 = 1\right\},$$

and define

$$R_{\max,n}(\boldsymbol{z}) := \max_{\boldsymbol{u} \in \mathcal{N}_1(\boldsymbol{z})} \frac{K_{\boldsymbol{z},h_n}(\boldsymbol{u})}{K_{\boldsymbol{z},h_n}(\boldsymbol{z})}.$$

For the product kernel, the factors corresponding to unchanged coordinates cancel in this ratio. Hence $R_{\max,n}(\boldsymbol{z})$ is the largest among $R_{+,n}(z_r)$ and, when $z_r \geq 1$, $R_{-,n}(z_r)$, $r = 1, \ldots, d$. Therefore,

$$\overline{R}_n := \sup_{\boldsymbol{z} \in \mathcal{S}_n} R_{\max,n}(\boldsymbol{z}) \leq 2\text{exp}\left\{-\frac{1}{2h_n(M_n+1)}\right\} \longrightarrow 0.$$

For any $\boldsymbol{u} \neq \boldsymbol{z}$, multiplying the one-step ratios along each coordinate gives

$$\frac{K_{\boldsymbol{z},h_n}(\boldsymbol{u})}{K_{\boldsymbol{z},h_n}(\boldsymbol{z})} \leq R_{\max,n}(\boldsymbol{z})^{\|\boldsymbol{u}-\boldsymbol{z}\|_1}.$$

Since $K_{\boldsymbol{z},h_n}(\boldsymbol{z}) \le 1$,

$$K_{\boldsymbol{z},h_n}(\boldsymbol{u}) \le R_{\max,n}(\boldsymbol{z})^{\|\boldsymbol{u}-\boldsymbol{z}\|_1}. \tag{24}$$

We now return to the smoothing bias. Since $\sum_{\boldsymbol{u}} K_{\boldsymbol{z},h_n}(\boldsymbol{u}) = 1$,

$$\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z}) - f_U(\boldsymbol{z}) = \sum_{\boldsymbol{u}\in\mathbb{N}_0^d} K_{\boldsymbol{z},h_n}(\boldsymbol{u})\{f_U(\boldsymbol{u}) - f_U(\boldsymbol{z})\}.$$

The term $\boldsymbol{u} = \boldsymbol{z}$ is zero. For each integer $q \ge 1$, the number of lattice points satisfying $\|\boldsymbol{u}-\boldsymbol{z}\|_1 = q$ is at most $(2d)^q$: such a point can be reached by q unit moves, and at each move there are at most $2d$ coordinate directions. Thus, using $|f_U(\boldsymbol{u}) - f_U(\boldsymbol{z})| \le 1$,

$$\begin{aligned}\left|\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z}) - f_U(\boldsymbol{z})\right| &\le \sum_{q=1}^{\infty}(2d)^q R_{\max,n}(\boldsymbol{z})^q \\ &= \frac{2dR_{\max,n}(\boldsymbol{z})}{1-2dR_{\max,n}(\boldsymbol{z})}.\end{aligned}$$

Because $\overline{R}_n \to 0$, for all sufficiently large n, $2d\overline{R}_n \le 1/2$. Hence

$$\begin{aligned}\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z}) - f_U(\boldsymbol{z})\right| &\le \frac{2d\overline{R}_n}{1-2d\overline{R}_n} \\ &\le 4d\overline{R}_n \\ &\le 8d\exp\left\{-\frac{1}{2h_n(M_n+1)}\right\}.\end{aligned}$$

Therefore,

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\mathbb{E}\widehat{f}_U^{(-b)}(\boldsymbol{z}) - f_U(\boldsymbol{z})\right| = O\left(\exp\left\{-\frac{1}{2h_n(M_n+1)}\right\}\right). \tag{25}$$

Combining (21) and (25), and using (19), gives

$$\begin{aligned}&\frac{\sqrt{\log(2|\mathcal{S}_n|)/n} + \exp\{-1/[2h_n(M_n+1)]\}}{\delta_n(U)} \\ &\le n^{\beta-1/2}\sqrt{\log(2|\mathcal{S}_n|)} + n^{\beta}\exp\{-1/[2h_n(M_n+1)]\} = o(1).\end{aligned}$$

For the second term, we used $1/\{h_n(M_n+1)\}/\log n \to \infty$. It follows that

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n}|\widehat{f}_U^{(-b)}(\boldsymbol{z}) - f_U(\boldsymbol{z})| = o_p\{\delta_n(U)\}.$$

Hence, with probability tending to one,

$$\inf_{\boldsymbol{z}\in\mathcal{S}_n}\widehat{f}_U^{(-b)}(\boldsymbol{z}) \ge \frac{\delta_n(U)}{2}. \tag{26}$$

This holds simultaneously over the fixed number of folds.

Step 3: uniform control of the numerator. Subtracting $m_{\ell,U}(\boldsymbol{z})$ from the Nadaraya–Watson ratio gives

$$\begin{aligned}&\widehat{m}_{\ell,U}^{(-b)}(\boldsymbol{z}) - m_{\ell,U}(\boldsymbol{z}) \\ &\quad= \frac{N_b^{-1}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})\{X_{i\ell} - m_{\ell,U}(\boldsymbol{z})\}}{\widehat{f}_U^{(-b)}(\boldsymbol{z})}.\end{aligned}$$

With $e_{\ell,U,i} := X_{i\ell} - m_{\ell,U}(\mathbf{X}_{i,U})$, the numerator is $A_b(\boldsymbol{z}) + E_b(\boldsymbol{z})$, where

$$\begin{aligned}A_b(\boldsymbol{z}) &:= \frac{1}{N_b}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})\{m_{\ell,U}(\mathbf{X}_{i,U}) - m_{\ell,U}(\boldsymbol{z})\}, \\ E_b(\boldsymbol{z}) &:= \frac{1}{N_b}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})e_{\ell,U,i}.\end{aligned}$$

Let L_m be a Lipschitz constant for $m_{\ell,U}$. Since the observations in I_b^c are i.i.d.,

$$\mathbb{E}A_b(\boldsymbol{z}) = \sum_{\boldsymbol{u}\in\mathbb{N}_0^d} K_{\boldsymbol{z},h_n}(\boldsymbol{u})\{m_{\ell,U}(\boldsymbol{u}) - m_{\ell,U}(\boldsymbol{z})\}f_U(\boldsymbol{u}).$$

The term $\boldsymbol{u} = \boldsymbol{z}$ is zero. For $\|\boldsymbol{u} - \boldsymbol{z}\|_1 = q$, Lipschitz continuity gives

$$|m_{\ell,U}(\boldsymbol{u}) - m_{\ell,U}(\boldsymbol{z})| \le L_m q,$$

and (24) gives $K_{\boldsymbol{z},h_n}(\boldsymbol{u}) \le R_{\max,n}(\boldsymbol{z})^q$. Hence

$$|\mathbb{E}A_b(\boldsymbol{z})| \le L_m \sum_{q=1}^{\infty} q(2d)^q R_{\max,n}(\boldsymbol{z})^q.$$

For all sufficiently large n, $2dR_{\max,n}(\boldsymbol{z}) \le 1/2$, and therefore

$$\begin{aligned}\sum_{q=1}^{\infty} q(2d)^q R_{\max,n}(\boldsymbol{z})^q &= \frac{2dR_{\max,n}(\boldsymbol{z})}{\{1 - 2dR_{\max,n}(\boldsymbol{z})\}^2} \\ &\le 8dR_{\max,n}(\boldsymbol{z}).\end{aligned}$$

Thus

$$|\mathbb{E}A_b(\boldsymbol{z})| \le 8dL_m R_{\max,n}(\boldsymbol{z}). \tag{27}$$

Similarly,

$$\begin{aligned}\mathrm{Var}\{A_b(\boldsymbol{z})\} &= \frac{1}{N_b}\mathrm{Var}\left[K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})\{m_{\ell,U}(\mathbf{X}_{i,U}) - m_{\ell,U}(\boldsymbol{z})\}\right] \\ &\le \frac{1}{N_b}\mathbb{E}\left[K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})^2\{m_{\ell,U}(\mathbf{X}_{i,U}) - m_{\ell,U}(\boldsymbol{z})\}^2\right] \\ &\le \frac{L_m^2}{N_b}\sum_{q=1}^{\infty} q^2(2d)^q R_{\max,n}(\boldsymbol{z})^{2q}.\end{aligned}$$

Writing $y = 2dR_{\max,n}(\boldsymbol{z})^2$ and using

$$\sum_{q=1}^{\infty} q^2 y^q = \frac{y(1+y)}{(1-y)^3},$$

we obtain, for all sufficiently large n such that $y \le 1/2$,

$$\mathrm{Var}\{A_b(\boldsymbol{z})\} \le \frac{24dL_m^2}{N_b} R_{\max,n}(\boldsymbol{z})^2.$$

Consequently, for every $t > 0$, Chebyshev's inequality and a union bound over $\mathcal{S}_n$ give

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{z}\in\mathcal{S}_n} |A_b(\boldsymbol{z}) - \mathbb{E}A_b(\boldsymbol{z})| > t\overline{R}_n\sqrt{\frac{|\mathcal{S}_n|}{N_b}}\right] \\ &\le \sum_{\boldsymbol{z}\in\mathcal{S}_n} \frac{24dL_m^2 R_{\max,n}(\boldsymbol{z})^2/N_b}{t^2\overline{R}_n^2|\mathcal{S}_n|/N_b} \\ &\le \frac{24dL_m^2}{t^2}.\end{aligned}$$

Therefore,

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n} |A_b(\boldsymbol{z}) - \mathbb{E}A_b(\boldsymbol{z})| = O_p\left(\overline{R}_n\sqrt{\frac{|\mathcal{S}_n|}{n}}\right).$$

By (20), $|\mathcal{S}_n|/n \to 0$, so this centered fluctuation is $o_p(\overline{R}_n)$. Combining this with (27) yields

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n} |A_b(\boldsymbol{z})| = O_p(\overline{R}_n) = O_p\left(\exp\left\{-\frac{1}{2h_n(M_n+1)}\right\}\right).$$

It remains to control $E_b(\boldsymbol{z})$. Fix $c_2 \in (0,1)$ and $\gamma > 1/(1-c_2)$, and let

$$B_n = c_b(\log n)^\gamma, \qquad c_b > 1.$$

For each observation i, define

$$e^{\text{center}}_{\ell,U,i} := e_{\ell,U,i}\mathbb{1}_{\{|e_{\ell,U,i}|\le B_n\}} - \mathbb{E}\left[e_{\ell,U,i}\mathbb{1}_{\{|e_{\ell,U,i}|\le B_n\}} \mid \mathbf{X}_{i,U}\right],$$
$$e^{\text{tail}}_{\ell,U,i} := e_{\ell,U,i}\mathbb{1}_{\{|e_{\ell,U,i}|> B_n\}} - \mathbb{E}\left[e_{\ell,U,i}\mathbb{1}_{\{|e_{\ell,U,i}|> B_n\}} \mid \mathbf{X}_{i,U}\right].$$

Since $\mathbb{E}(e_{\ell,U,i} \mid \mathbf{X}_{i,U}) = 0$,

$$e^{\text{center}}_{\ell,U,i} + e^{\text{tail}}_{\ell,U,i} = e_{\ell,U,i},$$

and both terms have conditional mean zero given $\mathbf{X}_{i,U}$.

Tail part. Fix $M > 3$ and let $\tau_n = c_z \log n$ for some $c_z > 0$.

Define

$$F_n := \left\{ \max_{1\le i\le n} \max_{v\in U\cup \text{Pa}_G(\ell)} X_{iv} \le \tau_n \right\}.$$

This single event controls both the parent variables entering the Poisson rate and all coordinates of the conditioning vector. In particular, on F_n,

$$\|\mathbf{X}_{i,U}\|_1 \le d\tau_n.$$

Let $\eta > 0$ be as in Lemma 6. By Markov's inequality and a union bound,

$$\begin{aligned}\mathbb{P}(F_n^c) &\le n \sum_{v\in U\cup \text{Pa}_G(\ell)} \mathbb{P}(X_v > \tau_n) \\ &\le n\,|U\cup \text{Pa}_G(\ell)|\mathbb{E}e^{\eta\|\boldsymbol{X}\|_1} e^{-\eta\tau_n} \\ &= O(n^{1-\eta c_z}).\end{aligned}$$

Choosing $c_z > (M+1)/\eta$ gives

$$\mathbb{P}(F_n^c) = o(n^{-M}).$$

On F_n, Lipschitz continuity of $m_{\ell,U}$ and μ_ℓ gives

$$m_{\ell,U}(\mathbf{X}_{i,U}) \le m_{\ell,U}(\mathbf{0}) + L_m\|\mathbf{X}_{i,U}\|_1 \le C_m \log n,$$
$$\mu_\ell(\mathbf{X}_{i,\text{Pa}_G(\ell)}) \le \mu_\ell(\mathbf{0}) + L_\mu \sum_{v\in \text{Pa}_G(\ell)} X_{iv} \le C_\mu \log n,$$

for finite constants C_m, C_μ independent of n and i.

Because $B_n/\log n \to \infty$ and $X_{i\ell} \ge 0$,

$$e_{\ell,U,i} = X_{i\ell} - m_{\ell,U}(\mathbf{X}_{i,U}) \ge -C_m \log n > -B_n$$

on F_n for all sufficiently large n. Hence only the upper residual tail remains on F_n. Moreover,

$$\{e_{\ell,U,i} > B_n\} \cap F_n \subseteq \{X_{i\ell} > B_n\} \cap \left\{ \max_{v\in \text{Pa}_G(\ell)} X_{iv} \le \tau_n \right\}.$$

Therefore,

$$\begin{aligned}\mathbb{P}(|e_{\ell,U,i}| > B_n) &\le \mathbb{P}(F_n^c) \\ &\quad + \mathbb{E}\left[\mathbb{P}\left(X_{i\ell} > B_n \mid \mathbf{X}_{i,\text{Pa}_G(\ell)}\right)\mathbb{1}_{\{\max_{v\in \text{Pa}_{G(\ell)}} X_{iv}\le\tau_n\}}\right].\end{aligned}$$

Conditional on $\mathbf{X}_{i,\text{Pa}_G(\ell)}$, the variable $X_{i\ell}$ is Poisson with mean $\mu_\ell(\mathbf{X}_{i,\text{Pa}_G(\ell)})$.

For $W \sim \text{Poisson}(\lambda)$ and $x > \lambda$, Chernoff's inequality gives, for any $c_2 \in (0,1)$,

$$\begin{aligned}\mathbb{P}(W \ge x) &\le \exp\{-x\log(x/\lambda) + x - \lambda\} \\ &\le \exp\left\{C_{c_2}\lambda^{1/(1-c_2)} - c_2 x\log x\right\},\end{aligned}$$

where $C_{c_2} < \infty$ depends only on c_2. We apply this bound conditionally with $\lambda = \mu_\ell(\mathbf{X}_{i,\mathrm{Pa}_{G(\ell)}})$. Since $0 < \mu_\ell(\mathbf{X}_{i,\mathrm{Pa}_{G(\ell)}}) \le C_\mu \log n$ on F_n, it is enough to bound the Poisson tail uniformly over all $0 < \lambda \le C_\mu \log n$. Taking $x = B_n = c_b(\log n)^\gamma$ therefore gives

$$\begin{aligned}
&\log\left[n^M \sup_{0<\lambda\le C_\mu \log n} \mathbb{P}\{\mathrm{Poisson}(\lambda) > B_n\}\right] \\
&\quad \le M\log n + C_{c_2} C_\mu^{1/(1-c_2)} (\log n)^{1/(1-c_2)} - c_2 c_b (\log n)^\gamma \{\log c_b + \gamma \log\log n\} \longrightarrow -\infty,
\end{aligned}$$

because $\gamma > 1/(1-c_2)$. Hence

$$\sup_{0<\lambda\le C_\mu \log n} \mathbb{P}\{\mathrm{Poisson}(\lambda) > B_n\} = o(n^{-M}),$$

and consequently

$$\mathbb{P}(|e_{\ell,U,i}| > B_n) = o(n^{-M}).$$

Since $\mathbb{E}(e_{\ell,U,i}^2) < \infty$, Cauchy–Schwarz yields

$$\begin{aligned}
\mathbb{E}\left[|e_{\ell,U,i}| \mathbb{1}_{\{|e_{\ell,U,i}|>B_n\}}\right] &\le \{\mathbb{E}(e_{\ell,U,i}^2)\}^{1/2} \mathbb{P}(|e_{\ell,U,i}| > B_n)^{1/2} \\
&= o(n^{-M/2}).
\end{aligned}$$

By the definition of the conditionally centered tail term and conditional Jensen's inequality,

$$\begin{aligned}
\mathbb{E}|e_{\ell,U,i}^{\mathrm{tail}}| &\le \mathbb{E}\left[|e_{\ell,U,i}| \mathbb{1}_{\{|e_{\ell,U,i}|>B_n\}}\right] \\
&\quad + \mathbb{E}\left|\mathbb{E}\left[e_{\ell,U,i} \mathbb{1}_{\{|e_{\ell,U,i}|>B_n\}} \mid \mathbf{X}_{i,U}\right]\right| \\
&\le 2\mathbb{E}\left[|e_{\ell,U,i}| \mathbb{1}_{\{|e_{\ell,U,i}|>B_n\}}\right] \\
&= o(n^{-M/2}).
\end{aligned}$$

Let $a_n = n^{-M/2+1}$. Since $0 \le K_{\boldsymbol{z},h_n} \le 1$, Markov's inequality followed by a union bound gives

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{z}\in\mathcal{S}_n} \left|\frac{1}{N_b} \sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U}) e_{\ell,U,i}^{\mathrm{tail}}\right| > a_n\right] \\
&\quad \le \frac{|\mathcal{S}_n|}{a_n} \mathbb{E}|e_{\ell,U,i}^{\mathrm{tail}}| \\
&\quad = |\mathcal{S}_n|\, o(n^{-1}) = o(1),
\end{aligned}$$

where the last equality uses $|\mathcal{S}_n| = O\{(\log n)^d\}$ from (20). Therefore,

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n} \left|\frac{1}{N_b} \sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U}) e_{\ell,U,i}^{\mathrm{tail}}\right| = o_p(n^{-M/2+1}).$$

Centered part. For fixed $\boldsymbol{z}$, define

$$Y_{i,n}(\boldsymbol{z}) := K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U}) e_{\ell,U,i}^{\mathrm{center}}, \qquad i \notin I_b.$$

Because the kernel weight is a function of $\mathbf{X}_{i,U}$ and $\mathbb{E}(e_{\ell,U,i}^{\mathrm{center}} \mid \mathbf{X}_{i,U}) = 0$,

$$\begin{aligned}
\mathbb{E}Y_{i,n}(\boldsymbol{z}) &= \mathbb{E}\left[K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U}) \mathbb{E}\left(e_{\ell,U,i}^{\mathrm{center}} \mid \mathbf{X}_{i,U}\right)\right] \\
&= 0.
\end{aligned}$$

Moreover,

$$|Y_{i,n}(\boldsymbol{z})| \le 2B_n,$$

because $K_{\boldsymbol{z},h_n} \le 1$ and $|e_{\ell,U,i}^{\mathrm{center}}| \le 2B_n$.

For the second moment,

$$\mathbb{E}Y_{i,n}(\boldsymbol{z})^2 = \sum_{\boldsymbol{u}\in\mathbb{N}_0^d} K_{\boldsymbol{z},h_n}(\boldsymbol{u})^2 \mathbb{E}\left[(e^{\text{center}}_{\ell,U,i})^2 \mid \mathbf{X}_{i,U} = \boldsymbol{u}\right] f_U(\boldsymbol{u}).$$

Conditional centering implies

$$\begin{aligned}\mathbb{E}\left[(e^{\text{center}}_{\ell,U,i})^2 \mid \mathbf{X}_{i,U}\right] &= \text{Var}\left(e_{\ell,U,i}\mathbb{1}_{\{|e_{\ell,U,i}|\le B_n\}} \mid \mathbf{X}_{i,U}\right) \\ &\le \mathbb{E}(e^2_{\ell,U,i} \mid \mathbf{X}_{i,U}).\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}(e^2_{\ell,U,i} \mid \mathbf{X}_{i,U} = \boldsymbol{u}) &= \text{Var}(X_{i\ell} \mid \mathbf{X}_{i,U} = \boldsymbol{u}) \\ &\le \mathbb{E}(X^2_{i\ell} \mid \mathbf{X}_{i,U} = \boldsymbol{u}).\end{aligned}$$

Multiplying by $f_U(\boldsymbol{u})$ gives

$$\begin{aligned}&\mathbb{E}(e^2_{\ell,U,i} \mid \mathbf{X}_{i,U} = \boldsymbol{u}) f_U(\boldsymbol{u}) \\ &\quad\le \mathbb{E}(X^2_{i\ell} \mid \mathbf{X}_{i,U} = \boldsymbol{u})\mathbb{P}(\mathbf{X}_{i,U} = \boldsymbol{u}) \\ &\quad= \mathbb{E}\left[X^2_{i\ell}\mathbb{1}_{\{\mathbf{X}_{i,U}=\boldsymbol{u}\}}\right].\end{aligned}$$

Therefore, since $K_{\boldsymbol{z},h_n}(\boldsymbol{u})^2 \le 1$,

$$\begin{aligned}\mathbb{E}Y_{i,n}(\boldsymbol{z})^2 &\le \sum_{\boldsymbol{u}\in\mathbb{N}_0^d} \mathbb{E}\left[X^2_{i\ell}\mathbb{1}_{\{\mathbf{X}_{i,U}=\boldsymbol{u}\}}\right] \\ &= \mathbb{E}(X^2_{i\ell}) =: V_0 < \infty.\end{aligned}$$

Bernstein's inequality gives, for every $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N_b}\sum_{i\notin I_b} Y_{i,n}(\boldsymbol{z})\right| > t\right) \le 2\text{exp}\left\{-\frac{N_b t^2}{2V_0 + (4/3)B_n t}\right\}.$$

Set

$$t_n = C\sqrt{\frac{\log(2|\mathcal{S}_n|)}{n}}.$$

By (20), $\log(2|\mathcal{S}_n|) = O(\log\log n)$, and hence

$$B_n t_n = O\left[(\log n)^\gamma \sqrt{\frac{\log\log n}{n}}\right] \longrightarrow 0.$$

Since $N_b \asymp n$, there exists a constant $c_I > 0$ such that $N_b \ge c_I n$ for all sufficiently large n. Hence, for such n,

$$\frac{N_b t_n^2}{2V_0 + (4/3)B_n t_n} \ge c\, C^2 \log(2|\mathcal{S}_n|)$$

for a constant $c > 0$ independent of n, $\boldsymbol{z}$, and b. Choosing C sufficiently large and applying a union bound over $\mathcal{S}_n$ gives

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\frac{1}{N_b}\sum_{i\notin I_b} Y_{i,n}(\boldsymbol{z})\right| > t_n\right] \\ &\quad\le 2|\mathcal{S}_n|\text{exp}\{-cC^2\log(2|\mathcal{S}_n|)\} \longrightarrow 0.\end{aligned}$$

Thus,

$$\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\frac{1}{N_b}\sum_{i\notin I_b} K_{\boldsymbol{z},h_n}(\mathbf{X}_{i,U})e^{\text{center}}_{\ell,U,i}\right| = O_p\left\{\sqrt{\frac{\log(2|\mathcal{S}_n|)}{n}}\right\}.$$

Finally, divide the numerator bound by (26). By (19),

$$\delta_n(U)^{-1} \le n^{\beta}$$

for all sufficiently large n. Hence, using $\log(2|\mathcal{S}_n|) = O(\log\log n)$,

$$\begin{aligned}\frac{\sqrt{\log(2|\mathcal{S}_n|)/n}}{\delta_n(U)} &\le n^{\beta-1/2}\sqrt{\log(2|\mathcal{S}_n|)} \\ &\le Cn^{\beta-1/2}\sqrt{\log\log n} \\ &= n^{-1/4}\left\{Cn^{\beta-1/4}\sqrt{\log\log n}\right\} \\ &= o(n^{-1/4}),\end{aligned}$$

because $\beta < 1/4$.

For the kernel-bias term, Assumption 1 implies $M_n = O(\log n)$, while $h_n = O(n^{-c_h})$. Therefore,

$$\frac{1}{h_n(M_n+1)\log n} \longrightarrow \infty.$$

Consequently, for every fixed $a > 0$,

$$\exp\left\{-\frac{1}{2h_n(M_n+1)}\right\} = o(n^{-a}).$$

Taking any $a > \beta + 1/4$ gives

$$\begin{aligned}\frac{\exp\{-1/[2h_n(M_n+1)]\}}{\delta_n(U)} &\le n^{\beta}\exp\left\{-\frac{1}{2h_n(M_n+1)}\right\} \\ &= o(n^{-1/4}).\end{aligned}$$

Therefore,

$$\max_{1\le b\le B}\sup_{\boldsymbol{z}\in\mathcal{S}_n}\left|\widehat{m}_{\ell,U}^{(-b)}(\boldsymbol{z}) - m_{\ell,U}(\boldsymbol{z})\right| = o_p(n^{-1/4}).$$

□

Adaptive truncation. Fix a conditioning set U and $\tau \in (0,1)$. Because $\|\boldsymbol{X}_U\|_1$ is finite almost surely, there is a finite integer q_U such that $\mathbb{P}(\|\boldsymbol{X}_U\|_1 \le q_U) > \tau$. The empirical proportion of observations satisfying $\|\mathbf{X}_{i,U}\|_1 \le q_U$ converges to this probability. Hence the empirical τ-quantile obeys $\widehat{q}_{\tau,n}(U) \le q_U$ with probability tending to one. At the same time, $M_n \to \infty$, so eventually $M_n > q_U$. Therefore

$$\widetilde{M}_{n,U} = \max\{\widehat{q}_{\tau,n}(U), M_n\} = M_n$$

with probability tending to one. Since p is fixed, only finitely many conditioning sets can be used by the algorithm, and the same conclusion holds simultaneously for all local tests. Thus replacing M_n by the adaptive threshold does not change any of the asymptotic results below.

A.2.3 Proof of the conditional-uncorrelatedness test

Proof of Part (i) of Lemma 2. Fix distinct nodes j, k and an adjustment set Z. For observation i, write

$$e_{j,i} := X_{ij} - m_{j,Z}(\mathbf{X}_{i,Z}), \qquad e_{k,i} := X_{ik} - m_{k,Z}(\mathbf{X}_{i,Z}).$$

For $i \in I_b$, define the two conditional-mean estimation errors

$$D_{j,i}^{(b)} := \widehat{m}_{j,Z}^{(-b)}(\mathbf{X}_{i,Z}) - m_{j,Z}(\mathbf{X}_{i,Z}), \qquad D_{k,i}^{(b)} := \widehat{m}_{k,Z}^{(-b)}(\mathbf{X}_{i,Z}) - m_{k,Z}(\mathbf{X}_{i,Z}).$$

The corresponding fitted residuals satisfy

$$\widehat{e}_{ij,Z} = e_{j,i} - D_{j,i}^{(b)}, \qquad \widehat{e}_{ik,Z} = e_{k,i} - D_{k,i}^{(b)}, \qquad i \in I_b.$$

Step 1: decomposition of the covariance estimator. Let $\gamma_{jk|Z} := \mathbb{E}(e_{j,Z}e_{k,Z})$. Expanding $(e_{j,i} - D_{j,i}^{(b)})(e_{k,i} - D_{k,i}^{(b)})$ and adding and subtracting the mean of the truncated population residual product gives

$$\sqrt{n}\{\widehat{\gamma}_{jk|Z} - \gamma_{jk|Z}\} = L_{1,n} + L_{2,n} + L_{3,n} + L_{4,n} + L_{5,n}, \tag{28}$$

where

$$\begin{aligned}
L_{1,n} &:= -\sqrt{n}\,\mathbb{E}\big[e_{j,Z}e_{k,Z}\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 > M_n\}}\big], \\
L_{2,n} &:= -\frac{1}{\sqrt{n}}\sum_{b=1}^{B}\sum_{i\in I_b} e_{j,i}D_{k,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}, \\
L_{3,n} &:= -\frac{1}{\sqrt{n}}\sum_{b=1}^{B}\sum_{i\in I_b} e_{k,i}D_{j,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}, \\
L_{4,n} &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[e_{j,i}e_{k,i}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} - \mathbb{E}\big\{e_{j,Z}e_{k,Z}\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 \le M_n\}}\big\}\right], \\
L_{5,n} &:= \frac{1}{\sqrt{n}}\sum_{b=1}^{B}\sum_{i\in I_b} D_{j,i}^{(b)}D_{k,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}.
\end{aligned}$$

The five terms have distinct roles. $L_{1,n}$ is the difference between the untruncated population covariance and the truncated covariance estimated by the statistic. $L_{2,n}$ and $L_{3,n}$ are the first-order effects of estimating the two conditional means. $L_{4,n}$ is the sampling fluctuation of the population residual product. Finally, $L_{5,n}$ is the product of the two conditional-mean estimation errors.

Step 2: the truncation term $L_{1,n}$. Under $H_0^{\mathrm{corr}}(j,k;Z)$, $\mathbb{E}(e_{j,Z}e_{k,Z} \mid \boldsymbol{X}_Z) = 0$ almost surely. For every n, the tower property gives

$$\mathbb{E}\big[e_{j,Z}e_{k,Z}\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 > M_n\}}\big] = \mathbb{E}\big[\mathbb{E}(e_{j,Z}e_{k,Z} \mid \boldsymbol{X}_Z)\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 > M_n\}}\big] = 0.$$

Hence $L_{1,n} = 0$.

Step 3: the linear terms $L_{2,n}$ and $L_{3,n}$. By symmetry it is enough to analyze $L_{2,n}$. Let $\mathcal{F}_b$ denote the sigma-field generated by the observations in the training set I_b^c. Conditional on $\mathcal{F}_b$, the fitted function $\widehat{m}_{k,Z}^{(-b)}$ is fixed. The evaluation observations indexed by I_b remain independent of the training data and independent of one another.

For $i \in I_b$, the factor $D_{k,i}^{(b)}$ is a fixed function of $\mathbf{X}_{i,Z}$ once $\mathcal{F}_b$ is given. Since $\mathbb{E}(e_{j,i} \mid \mathbf{X}_{i,Z}) = 0$, and, by cross-fitting, we have $\mathbb{E}(e_{j,i} \mid \mathbf{X}_{i,Z}, \mathcal{F}_b) = \mathbb{E}(e_{j,i} \mid \mathbf{X}_{i,Z}) = 0$. Hence,

$$\begin{aligned}
&\mathbb{E}\Big[e_{j,i}D_{k,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \mid \mathcal{F}_b\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}(e_{j,i} \mid \mathbf{X}_{i,Z})D_{k,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \mid \mathcal{F}_b\Big] = 0.
\end{aligned}$$

Thus each summand is conditionally centered. Because different evaluation observations are conditionally independent, all cross products with two distinct indices vanish in the conditional second moment of the fold sum. Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\left\{\frac{1}{\sqrt{n}}\sum_{i\in I_b} e_{j,i}D_{k,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}\right\}^2 \,\middle|\, \mathcal{F}_b\right] \\
&\quad = \frac{1}{n}\sum_{i\in I_b}\mathbb{E}\left[e_{j,i}^2(D_{k,i}^{(b)})^2\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \,\middle|\, \mathcal{F}_b\right].
\end{aligned}$$

On the truncation region, $|D_{k,i}^{(b)}|$ is bounded by the uniform error of the fold-b regression estimator. Therefore the last display is at most

$$\frac{|I_b|}{n}\left\{\sup_{\boldsymbol{z}\in\mathcal{S}_n}|\widehat{m}_{k,Z}^{(-b)}(\boldsymbol{z}) - m_{k,Z}(\boldsymbol{z})|\right\}^2 \mathbb{E}(e_{j,Z}^2) = o_p(1),$$

by Lemma 7. Conditional Chebyshev's inequality therefore gives $L_{2,n} = o_p(1)$, and interchanging the roles of j and k gives $L_{3,n} = o_p(1)$.

Step 4: the population residual-product term $L_{4,n}$. Define

$$W_{i,n} := e_{j,i}e_{k,i}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} - \mathbb{E}\big[e_{j,Z}e_{k,Z}\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 \le M_n\}}\big].$$

Then $L_{4,n} = n^{-1/2}\sum_{i=1}^n W_{i,n}$. For each n, the variables $W_{1,n}, \dots, W_{n,n}$ are i.i.d. and centered. The finite acyclic graph and the Lipschitz link functions imply $\mathbb{E}(e_{j,Z}^4 e_{k,Z}^4) < \infty$ and $\sup_n \mathbb{E}(|W_{i,n}|^4) < \infty$. The dominated convergence theorem gives

$$\mathrm{Var}\big(e_{j,Z}e_{k,Z}\mathbb{1}_{\{\|\boldsymbol{X}_Z\|_1 \le M_n\}}\big) \longrightarrow \mathrm{Var}(e_{j,Z}e_{k,Z}) = V_{jk|Z} > 0.$$

The Lindeberg–Feller central limit theorem therefore yields

$$L_{4,n} \rightsquigarrow N(0, V_{jk|Z}).$$

Step 5: the quadratic term $L_{5,n}$. The absolute value of $L_{5,n}$ is bounded by the number of summands times the product of the two largest regression errors on the truncation region:

$$|L_{5,n}| \le \sqrt{n}\left\{\max_b \sup_{\boldsymbol{z}\in\mathcal{S}_n} |\widehat{m}_{j,Z}^{(-b)}(\boldsymbol{z}) - m_{j,Z}(\boldsymbol{z})|\right\} \times \left\{\max_b \sup_{\boldsymbol{z}\in\mathcal{S}_n} |\widehat{m}_{k,Z}^{(-b)}(\boldsymbol{z}) - m_{k,Z}(\boldsymbol{z})|\right\}.$$

Each supremum is $o_p(n^{-1/4})$ by Lemma 7. Their product is therefore $o_p(n^{-1/2})$, and multiplication by $\sqrt{n}$ gives $L_{5,n} = o_p(1)$.

Combining Steps 2–5 with (28) leaves only the limit of $L_{4,n}$:

$$\sqrt{n}\,\widehat{\gamma}_{jk|Z} \rightsquigarrow N(0, V_{jk|Z}). \tag{29}$$

Here we used $\gamma_{jk|Z} = 0$ under the conditional-moment null.

Step 6: consistency of the quantities used for studentization. The test statistic divides the squared estimated covariance by an estimated variance. We now show that these normalizing quantities converge to their population counterparts.

First consider the residual variances. For $\ell \in \{j, k\}$, the identity $\widehat{e}_{i\ell,Z} = e_{\ell,i} - D_{\ell,i}^{(b)}$ gives

$$\begin{aligned}\widehat{\gamma}_{\ell\ell|Z} = \frac{1}{n}\sum_{i=1}^n e_{\ell,i}^2\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} &- \frac{2}{n}\sum_{b=1}^B\sum_{i\in I_b} e_{\ell,i}D_{\ell,i}^{(b)}\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \\ &+ \frac{1}{n}\sum_{b=1}^B\sum_{i\in I_b}(D_{\ell,i}^{(b)})^2\mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}.\end{aligned}$$

The first term converges to $\mathbb{E}(e_{\ell,Z}^2)$ by the weak law of large numbers and dominated convergence. The absolute value of the second term is at most

$$2\left\{\max_b \sup_{\boldsymbol{z}\in\mathcal{S}_n} |\widehat{m}_{\ell,Z}^{(-b)}(\boldsymbol{z}) - m_{\ell,Z}(\boldsymbol{z})|\right\}\frac{1}{n}\sum_{i=1}^n |e_{\ell,i}|,$$

which is $o_p(1)$ because the supremum is $o_p(1)$ and the sample average is $O_p(1)$. The third term is bounded by the square of the same supremum and is also $o_p(1)$. Thus

$$\widehat{\gamma}_{jj|Z} \xrightarrow{p} \gamma_{jj|Z}, \qquad \widehat{\gamma}_{kk|Z} \xrightarrow{p} \gamma_{kk|Z}. \tag{30}$$

Both limits are strictly positive by assumption.

Next consider the variance of the residual product. Let

$$g_i := e_{j,i}e_{k,i}, \qquad \widehat{g}_i := \widehat{e}_{ij,Z}\widehat{e}_{ik,Z}.$$

For $i \in I_b$, direct expansion gives

$$\widehat{g}_i - g_i = -e_{j,i} D^{(b)}_{k,i} - e_{k,i} D^{(b)}_{j,i} + D^{(b)}_{j,i} D^{(b)}_{k,i}.$$

On the truncation region, the two D terms are uniformly $o_p(1)$. Using $|a^2 - b^2| = |a - b|\,|a + b|$, the preceding expansion, Cauchy–Schwarz, and Lemma 6 give

$$\frac{1}{n}\sum_{i=1}^{n} |\widehat{g}_i^2 - g_i^2| \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} = o_p(1). \tag{31}$$

Expanding the sample variance gives

$$\begin{aligned}\widehat{V}_{jk|Z} = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i^2 \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \\ - 2\widehat{\gamma}_{jk|Z} \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \\ + \widehat{\gamma}^2_{jk|Z} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}}.\end{aligned}$$

The middle sample average is exactly $\widehat{\gamma}_{jk|Z}$, and the last sample average converges to one. Therefore

$$\widehat{V}_{jk|Z} = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i^2 \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} - \widehat{\gamma}^2_{jk|Z} + o_p(1).$$

By (31), the first term can be replaced by the corresponding average of g_i^2. The weak law and dominated convergence give

$$\frac{1}{n}\sum_{i=1}^{n} g_i^2 \mathbb{1}_{\{\|\mathbf{X}_{i,Z}\|_1 \le M_n\}} \xrightarrow{p} \mathbb{E}(g_i^2).$$

Under the null, $\mathbb{E}(g_i) = 0$ and $\widehat{\gamma}_{jk|Z} = o_p(1)$ by (29). Hence

$$\widehat{V}_{jk|Z} \xrightarrow{p} \mathbb{E}(g_i^2) = \mathrm{Var}(g_i) = V_{jk|Z}.$$

Finally, on the event that the residual-variance estimators are positive, substituting the definitions of $\widehat{\rho}_{jk|Z}$ and $\widehat{V}_{\rho,jk|Z}$ shows that their common denominator cancels:

$$\begin{aligned}T^{\mathrm{corr}}_{jk|Z} &= n \frac{\widehat{\gamma}^2_{jk|Z}}{\widehat{\gamma}_{jj|Z}\widehat{\gamma}_{kk|Z}} \left\{ \frac{\widehat{V}_{jk|Z}}{\widehat{\gamma}_{jj|Z}\widehat{\gamma}_{kk|Z}} \right\}^{-1} \\ &= \frac{n\widehat{\gamma}^2_{jk|Z}}{\widehat{V}_{jk|Z}}.\end{aligned}$$

Equation (30) implies that this event has probability tending to one. Slutsky's theorem and (29) now give

$$T^{\mathrm{corr}}_{jk|Z} \rightsquigarrow \chi^2_1.$$

Because the χ^2_1 distribution is continuous, the associated p-value satisfies $\mathbb{P}\{p^{\mathrm{corr}}_{jk|Z} \le \alpha\} \to \alpha$ for every fixed $\alpha \in (0, 1)$. The adaptive truncation threshold gives the same limit by the preceding adaptive-truncation argument. □

A.2.4 Proof of the local equidispersion test

Proof of Part (ii) of Lemma 2. Fix a node j and an admissible conditioning set S. For observation i, define

$$e_i := X_{ij} - m_{j,S}(\mathbf{X}_{i,S}), \qquad \psi_i := e_i^2 - X_{ij}.$$

For $i \in I_b$, define the conditional-mean estimation error

$$D^{(b)}_i := \widehat{m}^{(-b)}_{j,S}(\mathbf{X}_{i,S}) - m_{j,S}(\mathbf{X}_{i,S}).$$

Since the fitted residual is $\widehat{e}_{ij,S} = e_i - D^{(b)}_i$, its square is $\widehat{e}^2_{ij,S} = e_i^2 - 2e_i D^{(b)}_i + (D^{(b)}_i)^2$.

As in the preceding proof, the last two terms vanish because of cross-fitting and uniform consistency, leaving the population residual as the leading term.

Step 1: decomposition of the test statistic. Substituting the preceding squared-residual expansion into $Q_{j,S}$ gives

$$\sqrt{n}\, Q_{j,S} = U_{1,n} + U_{2,n} + U_{3,n},$$

where

$$U_{1,n} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}},$$

$$U_{2,n} := -\frac{2}{\sqrt{n}} \sum_{b=1}^{B} \sum_{i \in I_b} e_i D_i^{(b)} \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}},$$

$$U_{3,n} := \frac{1}{\sqrt{n}} \sum_{b=1}^{B} \sum_{i \in I_b} (D_i^{(b)})^2 \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}}.$$

The same conditional-centering calculation used for $L_{2,n}$ applies to $U_{2,n}$. Conditional on the training data outside fold b, $D_i^{(b)}$ is a fixed function of $\mathbf{X}_{i,S}$, while $\mathbb{E}(e_i \mid \mathbf{X}_{i,S}) = 0$. Therefore the fold sum has conditional mean zero, and its conditional variance is bounded by

$$\frac{4|I_b|}{n} \left\{ \sup_{\boldsymbol{z} \in \mathcal{S}_n} |\widehat{m}_{j,S}^{(-b)}(\boldsymbol{z}) - m_{j,S}(\boldsymbol{z})| \right\}^2 \mathbb{E}(e_i^2).$$

Lemma 7 makes this bound $o_p(1)$, so $U_{2,n} = o_p(1)$. For the quadratic term,

$$|U_{3,n}| \le \sqrt{n} \left\{ \max_b \sup_{\boldsymbol{z} \in \mathcal{S}_n} |\widehat{m}_{j,S}^{(-b)}(\boldsymbol{z}) - m_{j,S}(\boldsymbol{z})| \right\}^2 = o_p(1).$$

Thus the asymptotic behavior of $\sqrt{n} Q_{j,S}$ is determined by $U_{1,n}$.

Step 2: null distribution of the population score. The Poisson dispersion identity gives, for every value of the conditioning variables,

$$\begin{aligned} \mathbb{E}(\psi_i \mid \mathbf{X}_{i,S}) &= \mathrm{Var}(X_{ij} \mid \mathbf{X}_{i,S}) - \mathbb{E}(X_{ij} \mid \mathbf{X}_{i,S}) \\ &= \mathrm{Var}\{\mu_j(\boldsymbol{X}_{i,\mathrm{Pa}_{G(j)}}) \mid \mathbf{X}_{i,S}\} \ge 0. \end{aligned} \tag{32}$$

Under $H_0^{\mathrm{eq}}(j; S)$, the expectation of the last nonnegative random variable is zero. A nonnegative random variable with expectation zero must itself be zero almost surely. Hence

$$\mathbb{E}(\psi_i \mid \mathbf{X}_{i,S}) = 0 \quad \text{almost surely under } H_0^{\mathrm{eq}}(j; S).$$

Because the truncation indicator is measurable with respect to $\mathbf{X}_{i,S}$, the tower property then gives

$$\mathbb{E}\big[\psi_i \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}}\big] = 0$$

for every n. Thus the summands in $U_{1,n}$ are exactly centered.

Lemma 6 supplies the finite moments needed for the central limit theorem. Dominated convergence gives

$$\mathrm{Var}\big(\psi_i \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}}\big) \longrightarrow \mathrm{Var}(\psi_i) =: V_{j,S}.$$

The finite fourth moment supplied by Lemma 6 verifies the Lyapunov condition for the triangular array created by the truncation indicator. The Lindeberg–Feller central limit theorem yields

$$U_{1,n} \rightsquigarrow N(0, V_{j,S}).$$

Together with $U_{2,n}, U_{3,n} = o_p(1)$, this proves

$$\sqrt{n}\, Q_{j,S} \rightsquigarrow N(0, V_{j,S}). \tag{33}$$

Step 3: consistency of the variance estimator. Let $\widehat{\psi}_i := \widehat{e}_{ij,S}^2 - X_{ij}$. Since $\widehat{e}_{ij,S} = e_i - D_i^{(b)}$,

$$\widehat{\psi}_i - \psi_i = -2e_i D_i^{(b)} + (D_i^{(b)})^2.$$

On the truncation region this gives

$$|\widehat{\psi}_i - \psi_i| \le 2|e_i|\,|D_i^{(b)}| + |D_i^{(b)}|^2.$$

Using $|a^2 - b^2| = |a-b|\,|a+b|$, the uniform rate in Lemma 7, the weak law of large numbers, and the finite fourth moment of ψ_i, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} |\widehat{\psi}_i^2 - \psi_i^2| \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} = o_p(1). \tag{34}$$

The weak law and dominated convergence also give

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^2 \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} \xrightarrow{p} \mathbb{E}(\psi_i^2),$$
$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} \xrightarrow{p} 1.$$

Expanding the definition of the sample variance and using $Q_{j,S} = n^{-1}\sum_i \widehat{\psi}_i \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}}$ yields

$$\widehat{V}_{j,S} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i^2 \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} - Q_{j,S}^2 + o_p(1).$$

Under the null, $Q_{j,S} = o_p(1)$ by (33). Equation (34) then gives

$$\widehat{V}_{j,S} \xrightarrow{p} \mathbb{E}(\psi_i^2) = \mathrm{Var}(\psi_i) = V_{j,S}.$$

Slutsky's theorem proves $T_{j,S} \rightsquigarrow N(0,1)$. Since the standard normal distribution is continuous, $\mathbb{P}\{p_{j,S}^{\mathrm{eq}} \le \alpha\} \to \alpha$ for every fixed $\alpha \in (0,1)$.

Step 4: omitted-parent alternatives. Suppose S omits at least one parent of j. By non-degeneracy and the Poisson dispersion identity,

$$\Delta_{j|S} = \mathbb{E}\Big[\mathrm{Var}\{\mu_j(\boldsymbol{X}_{\mathrm{Pa}_G(j)}) \mid \boldsymbol{X}_S\}\Big] > 0.$$

Equation (32) shows that $\mathbb{E}(\psi_i) = \Delta_{j|S}$. The weak law and dominated convergence therefore give

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} \xrightarrow{p} \Delta_{j|S}. \tag{35}$$

Let

$$R_n := \max_b \sup_{\boldsymbol{z} \in \mathcal{S}_n} |\widehat{m}_{j,S}^{(-b)}(\boldsymbol{z}) - m_{j,S}(\boldsymbol{z})|.$$

Lemma 7 gives $R_n = o_p(1)$, and hence

$$\left|\frac{2}{n}\sum_{b=1}^{B}\sum_{i\in I_b} e_i D_i^{(b)} \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}}\right| \le 2R_n \frac{1}{n}\sum_{i=1}^{n}|e_i| = o_p(1),$$
$$\frac{1}{n}\sum_{b=1}^{B}\sum_{i\in I_b} (D_i^{(b)})^2 \mathbb{1}_{\{\|\mathbf{X}_{i,S}\|_1 \le M_n\}} \le R_n^2 = o_p(1).$$

Combining these bounds with (35) gives $Q_{j,S} \xrightarrow{p} \Delta_{j|S}$.

The same expansion used in Step 3, now without imposing the null, shows

$$\widehat{V}_{j,S} \xrightarrow{p} \mathrm{Var}(\psi_i) =: V_{j,S,1} > 0.$$

Therefore

$$\frac{T_{j,S}}{\sqrt{n}} = \frac{Q_{j,S}}{\widehat{V}_{j,S}^{1/2}} \xrightarrow{p} \frac{\Delta_{j|S}}{V_{j,S,1}^{1/2}} =: c_{j,S} > 0.$$

Thus $T_{j,S}$ grows at order $\sqrt{n}$ under an omitted-parent alternative, whereas the fixed normal critical value remains bounded. It follows that $\mathbb{P}\{p_{j,S}^{\mathrm{eq}} \le \alpha\} \to 1$.

The adaptive truncation threshold yields the same null and alternative limits by the adaptive-truncation argument above. Combining the null and alternative arguments, for every fixed admissible conditioning set A considered in the theorem,

$$Q_{j|A} \xrightarrow{p} \Delta_{j|A}.$$

□

The sequential expansion step of Algorithm 1 adds a candidate node only when its inclusion increases the local equidispersion statistic by at most the tolerance ϵ. The following lemma verifies that a missing true parent satisfies this expansion criterion with probability tending to one.

Lemma 8 (Expansion eligibility under a positive tolerance)**.** *Fix a node j, an admissible set S, and a node $k \notin S$ such that $S \cup \{k\}$ is also admissible. Then*

$$\Delta_{j|S\cup\{k\}} \le \Delta_{j|S}.$$

Suppose that the conditions of Part (ii) of Lemma 2 hold with conditioning sets S and $S \cup \{k\}$. Then, for every fixed $\epsilon > 0$,

$$\mathbb{P}\left\{Q_{j,S\cup\{k\}} - Q_{j,S} \le \epsilon\right\} \to 1.$$

Proof. By the conditional variance decomposition,

$$\begin{aligned}\mathbb{E}\{\mathrm{Var}(X_j \mid \boldsymbol{X}_S)\} = \mathbb{E}\{\mathrm{Var}(X_j \mid \boldsymbol{X}_{S\cup\{k\}})\} \\ + \mathbb{E}\left[\mathrm{Var}\left\{\mathbb{E}(X_j \mid \boldsymbol{X}_{S\cup\{k\}}) \mid \boldsymbol{X}_S\right\}\right].\end{aligned}$$

Since

$$\mathbb{E}\{\mathbb{E}(X_j \mid \boldsymbol{X}_S)\} = \mathbb{E}\{\mathbb{E}(X_j \mid \boldsymbol{X}_{S\cup\{k\}})\} = \mathbb{E}(X_j),$$

subtracting $\mathbb{E}(X_j)$ from both sides gives

$$\Delta_{j|S} - \Delta_{j|S\cup\{k\}} = \mathbb{E}\left[\mathrm{Var}\left\{\mathbb{E}(X_j \mid \boldsymbol{X}_{S\cup\{k\}}) \mid \boldsymbol{X}_S\right\}\right] \ge 0.$$

Hence

$$\Delta_{j|S\cup\{k\}} \le \Delta_{j|S}.$$

To prove the second claim, fix $A \in \{S, S \cup \{k\}\}$ and consider two cases. If $\mathrm{Pa}_G(j) \subseteq A$, then $\Delta_{j|A} = 0$, and the null part of the proof of Part (ii) of Lemma 2 gives

$$\sqrt{n}\, Q_{j,A} = O_p(1).$$

Therefore,

$$Q_{j,A} - \Delta_{j|A} = o_p(1).$$

If $\mathrm{Pa}_G(j) \not\subseteq A$, the alternative part of the same proof gives

$$Q_{j,A} \xrightarrow{p} \Delta_{j|A}.$$

Thus, in either case,

$$Q_{j,A} - \Delta_{j|A} = o_p(1).$$

Consequently,

$$Q_{j,S\cup\{k\}} - Q_{j,S} = \Delta_{j|S\cup\{k\}} - \Delta_{j|S} + o_p(1) \le o_p(1).$$

It follows that, for every fixed $\epsilon > 0$,

$$\mathbb{P}\left\{Q_{j,S\cup\{k\}} - Q_{j,S} \le \epsilon\right\} \to 1.$$

□

A.2.5 Proofs of the structural results

Proof of Lemma 3. Fix $j \in \mathcal{L}_t(G)$ and $k \in \mathcal{L}_{<t}(G)$ with $k \notin \mathrm{Pa}_G(j)$, and put

$$Z = Z_{kj}^{(t)} = \mathrm{Ne}_{G_{<t}}(k) \setminus \{j, k\}.$$

We must show that every path from k to j is blocked after conditioning on Z. There are two ways in which a path is blocked. A noncollider on the path blocks the path if it is included in Z. A collider blocks the path if neither the collider nor any of its descendants is included in Z. The set Z contains the parents, earlier-layer children, and earlier-layer spouses of k in the already recovered graph. We show that an arbitrary path is blocked by such Z.

Let $\pi = (v_0, v_1, \ldots, v_m)$ be an arbitrary path from $v_0 = k$ to $v_m = j$.

Case 1: the first edge points into k. Suppose $v_1 \to k$. Then v_1 is a parent of k. Because k lies in an earlier layer, its parent also lies in an earlier layer, so $v_1 \in \mathrm{Pa}_{G_{<t}}(k) \subseteq Z$. Hence v_1 is a noncollider on π. Since this noncollider is conditioned on, the path is blocked at v_1.

Case 2: the first edge points from k to another earlier-layer node. Suppose $k \to v_1$ and $v_1 \in \mathcal{L}_{<t}(G)$. Then v_1 is an earlier-layer child of k, so $v_1 \in \mathrm{Ch}_{G_{<t}}(k) \subseteq Z$. If v_1 is a noncollider on the path, conditioning on v_1 blocks the path immediately.

The only remaining possibility is that v_1 is a collider. Since the first edge is $k \to v_1$, the next edge on the path must also point into v_1, say $v_2 \to v_1$. The node v_2 is then another parent of the child v_1 and is therefore a spouse of k. Moreover, a parent of an earlier-layer node must itself lie in an earlier layer. Hence $v_2 \in \mathrm{Sp}_{G_{<t}}(k) \subseteq Z$ and v_2 is a noncollider on the path. Conditioning on v_2 therefore blocks the path.

Case 3: the first edge points from k to a node in layer t or a later layer. Suppose $k \to v_1$ and $v_1 \notin \mathcal{L}_{<t}(G)$. If the next edge is $v_2 \to v_1$, then v_1 is a collider. The conditioning set Z contains only earlier-layer nodes. Since a descendant of v_1 cannot lie in an earlier layer than v_1, neither v_1 nor any descendant of v_1 belongs to Z. The collider therefore remains closed and blocks the path.

It remains to consider paths that continue in the direction of the arrows. Follow the maximal initial directed segment

$$k = v_0 \to v_1 \to \cdots \to v_r.$$

This directed segment cannot end at j. If $r = 1$ and $v_1 = j$, then the edge $k \to j$ is present, contradicting $k \notin \mathrm{Pa}_G(j)$. If $r \geq 2$ and $v_r = j$, then v_1 is a strict ancestor of j. Every strict ancestor of a layer-t node lies in a layer earlier than t, contradicting $v_1 \notin \mathcal{L}_{<t}(G)$. Therefore the path must eventually reverse direction: its next edge is $v_{r+1} \to v_r$. The two adjacent edges on the path then point into v_r, so v_r is a collider. The node v_r and all of its descendants lie in layer t or later and hence are absent from Z. This collider blocks the path.

The three cases cover every possible first step of a path from k to j. Thus every such path is blocked by Z, proving $k \perp_{d,G} j \mid Z$. By the DAG Markov property, $X_k \perp\!\!\!\perp X_j \mid \boldsymbol{X}_Z$. Conditional independence implies conditional uncorrelatedness of the centered variables:

$$\begin{aligned}\mathbb{E}(e_{j,Z} e_{k,Z} \mid \boldsymbol{X}_Z) &= \mathbb{E}\big[\{X_j - \mathbb{E}(X_j \mid \boldsymbol{X}_Z)\}\{X_k - \mathbb{E}(X_k \mid \boldsymbol{X}_Z)\} \;\big|\; \boldsymbol{X}_Z\big] \\ &= 0 \quad \text{almost surely.}\end{aligned}$$

Therefore every earlier-layer nonparent satisfies the screening null. □

Proof of Theorem 4. We first define an event on which all local testing and expansion decisions required for graph recovery are correct. We show that this event has probability tending to one. We then suppose that this event occurs and prove, by induction over the true topological layers, that PoET recovers every layer and every parent set exactly.

Step 1: simultaneous correctness of the local decisions. Consider first a fixed local null hypothesis. Under either an equidispersion null or a screening null, Lemma 2 implies that the corresponding statistic has a nondegenerate limiting distribution and is therefore $O_p(1)$. Since $\alpha_n \to 0$, the corresponding critical value diverges. Hence the probability of falsely rejecting any fixed null tends to zero.

Next consider a fixed equidispersion test whose admissible conditioning set C omits at least one parent of j. By non-degeneracy, $\Delta_{j|C} > 0$, and Part (ii) of Lemma 2 gives

$$\frac{T_{j,C}}{\sqrt{n}} \xrightarrow{p} c_{j,C} > 0.$$

Let

$$z_n = \Phi^{-1}(1 - \alpha_n/2)$$

be the two-sided normal critical value. The standard Gaussian tail bound gives

$$z_n^2 \leq 2\log(2/\alpha_n).$$

Since $\log(1/\alpha_n) = o(n)$,

$$\frac{z_n}{\sqrt{n}} \to 0.$$

Therefore the test rejects every fixed omitted-parent alternative with probability tending to one.

Finally, Lemma 8 implies that, for every fixed $\epsilon > 0$, adding any missing true parent to an admissible conditioning set satisfies the expansion criterion

$$Q_{j,C\cup\{k\}} - Q_{j,C} \leq \epsilon$$

with probability tending to one.

Because p is fixed, only finitely many node pairs and conditioning sets can arise in Algorithm 1. Define $\mathcal{E}_n$ as the event on which all of the following properties hold simultaneously:

1. Every equidispersion test whose admissible conditioning set contains all parents of its target fails to reject.
2. Every equidispersion test whose admissible conditioning set omits at least one parent rejects.
3. Every conditional-uncorrelatedness-based screening test corresponding to an earlier-layer nonparent, as characterized by Lemma 3, fails to reject.
4. For every true layer node j, every admissible current conditioning set C considered in the sequential expansion step, and every missing true parent $k \in \mathrm{Pa}_G(j) \setminus C$,
$$Q_{j,C\cup\{k\}} - Q_{j,C} \leq \epsilon.$$

Each individual failure probability tends to zero, and only finitely many such decisions are possible. A union bound therefore yields

$$\mathbb{P}(\mathcal{E}_n) \longrightarrow 1. \tag{36}$$

Step 2: recovery of the true layers on $\mathcal{E}_n$. For the remainder of the proof, suppose that $\mathcal{E}_n$ occurs. Assume inductively that, before layer step t, Algorithm 1 has recovered exactly the nodes in $\mathcal{L}_{<t}(G)$ and every edge whose child lies in these layers. For $t = 1$, the induction hypothesis is vacuous.

First consider a true layer-t node $j \in \mathcal{L}_t(G)$. By definition of the topological layers,

$$\mathrm{Pa}_G(j) \subseteq \mathcal{L}_{<t}(G).$$

Stage 1: screening-guided parent candidate search. For any $k \in \mathcal{L}_{<t}(G) \setminus \mathrm{Pa}_G(j)$, the induction hypothesis implies that the learned partial graph used in Algorithm 1 coincides with the true partial graph on the previous layers. Lemma 3 therefore implies that the corresponding conditional-uncorrelatedness null is true. By property 3 of $\mathcal{E}_n$, this null is not rejected. Consequently, no earlier-layer nonparent enters the initial screened set, and

$$\widehat{C}_j \subseteq \mathrm{Pa}_G(j). \tag{37}$$

A true parent may fail to enter $\widehat{C}_j$ at this screening step, but such a parent remains available for the subsequent sequential expansion.

Stage 2: equidispersion-guided expansion. If $\widehat{C}_j$ already contains all parents of j, then (37) implies $\widehat{C}_j = \mathrm{Pa}_G(j)$.

By property 1 of $\mathcal{E}_n$, the equidispersion test with $\widehat{S}_j = \widehat{C}_j$ fails to reject, so Algorithm 1 assigns j to $\widehat{\mathcal{L}}_t$.

Now suppose that $\widehat{S}_j$ omits at least one parent. By property 2, the initial equidispersion test rejects. Algorithm 1 then enters the sequential expansion step. At any point in this expansion before all parents have been included, the current set $\widehat{S}_j$ is contained in $\mathcal{L}_{<t}(G)$ and still omits at least one parent. Every missing parent

$$k \in \mathrm{Pa}_G(j) \setminus \widehat{S}_j$$

also belongs to $\mathcal{L}_{<t}(G) \setminus \widehat{S}_j$ and is therefore among the variables considered for addition. By property 4 of $\mathcal{E}_n$, when such a parent k is considered, it satisfies

$$Q_{j,\widehat{S}_j \cup \{k\}} - Q_{j,\widehat{S}_j} \leq \epsilon,$$

so Algorithm 1 adds it to $\widehat{S}_j$.

Moreover, by property 2, as long as at least one parent remains missing, the equidispersion test after an accepted addition continues to reject. Hence the sequential expansion cannot terminate before all parents have entered $\widehat{S}_j$. Since there are finitely many earlier-layer variables and every true parent of j belongs to $\mathcal{L}_{<t}(G)$, the expansion must eventually produce a set satisfying

$$\mathrm{Pa}_G(j) \subseteq \widehat{S}_j.$$

At that point, property 1 implies that the equidispersion test fails to reject, and Algorithm 1 assigns j to $\widehat{\mathcal{L}}_t$.

Stage 3: drop-one parent pruning. The candidate set $\widehat{S}_j$ obtained in Stage 2 may contain variables that were introduced during sequential expansion. Algorithm 1 applies the drop-one equidispersion test only to $k \in \widehat{S}_j \setminus \widehat{C}_j$. If such a node k is a true parent, then $\widehat{S}_j \setminus \{k\}$ omits at least one parent. By property 2 of $\mathcal{E}_n$, the corresponding equidispersion null is rejected, so Algorithm 1 retains k. If k is not a parent, removing it leaves a conditioning set that still contains all parents. By property 1, the corresponding equidispersion null is not rejected, so Algorithm 1 removes k. Together with (37), which guarantees that the initial screened set contains no nonparent, Stage 3 therefore yields $\widehat{S}_j = \mathrm{Pa}_G(j)$. Thus every node in the true layer $\mathcal{L}_t(G)$ is assigned at step t, and all of its incoming edges are recovered exactly.

It remains to show that no node from a later layer is assigned at step t. Consider $j \in V \setminus \mathcal{L}_{<t+1}(G)$. By the definition of the topological layers, $\mathrm{Pa}_G(j) \not\subseteq \mathcal{L}_{<t}(G)$, so at least one parent of j lies outside the previously recovered layers. Every conditioning set considered for j at layer step t is a subset of $\mathcal{L}_{<t}(G)$ and therefore omits this parent. By property 2 of $\mathcal{E}_n$, every corresponding equidispersion test rejects. Hence Algorithm 1 cannot assign j to $\widehat{\mathcal{L}}_t$.

We have therefore shown that, whenever $\mathcal{E}_n$ occurs, Algorithm 1 assigns exactly the nodes in $\mathcal{L}_t(G)$ at layer step t and recovers their parent sets exactly. This completes the induction over t.

Since every true topological layer is nonempty until all nodes have been assigned, Algorithm 1 never invokes the VIOLATION stopping rule on $\mathcal{E}_n$. Consequently,

$$\mathcal{E}_n \subseteq \left\{ \widehat{G} = G, \quad \widehat{\ell}_j = \textsc{selected} \text{ for all } j \in V \right\}.$$

Combining this inclusion with (36) proves the theorem.

□

Proof of Theorem 5. We first establish an event on which all local testing and expansion decisions needed for violation detection are correct. We show that this event has probability tending to one, and then prove the claimed labels by induction over the estimated layers.

Put $\mathcal{A} := V \setminus \mathcal{V}_{\mathcal{W}}$. This set is ancestral in G, and the marginal distribution of $\boldsymbol{X}_{\mathcal{A}}$ is a Poisson DAG model over the induced graph on $\mathcal{A}$. We only need local guarantees for admissible conditioning sets contained in $\mathcal{A}$, as verified by the induction below.

Step 1: simultaneous correctness of the local decisions. First, every coordinate has a finite fourth moment. For nodes in $\mathcal{W}$, this is assumed. For the remaining nodes, it follows by induction along a topological ordering from the Lipschitz bound on the link functions and the identity

$$\mathbb{E}\{X_j^4 \mid \boldsymbol{X}_{\mathrm{Pa}_G(j)}\} = \mu_j^4 + 6\mu_j^3 + 7\mu_j^2 + \mu_j, \qquad j \notin \mathcal{W},$$

where $\mu_j = \mu_j(\boldsymbol{X}_{\mathrm{Pa}_G(j)})$.

Fix a node j and an admissible set $S \subseteq \mathcal{A}$. Use the notation $\mathcal{S}_n$, $\delta_n(S)$, N_b, $A_b(\boldsymbol{z})$, and $E_b(\boldsymbol{z})$ from the proof of Lemma 7, with $(\ell, U) = (j, S)$. Since $\boldsymbol{X}_{\mathcal{A}}$ follows a Poisson DAG model, the cell-probability and denominator bounds in that proof remain valid. In particular, for some $\beta < 1/4$,

$$\delta_n(S) \geq n^{-\beta}, \qquad |\mathcal{S}_n| = O\{(\log n)^{|S|}\}.$$

The bound for $A_b(\boldsymbol{z})$ also remains valid because $m_{j,S}$ is Lipschitz. For $E_b(\boldsymbol{z})$, conditional centering and $0 \leq K_{\boldsymbol{z},h_n} \leq 1$ give

$$\mathrm{Var}\{E_b(\boldsymbol{z})\} \leq \frac{\mathbb{E}(e_{j,S}^2)}{N_b}.$$

Chebyshev's inequality and a union bound therefore yield

$$\max_b \sup_{\boldsymbol{z} \in \mathcal{S}_n} |E_b(\boldsymbol{z})| = O_p\left(\sqrt{\frac{|\mathcal{S}_n|}{n}}\right).$$

Combining these bounds as in Lemma 7 gives

$$\begin{aligned}
&\max_b \sup_{\boldsymbol{z} \in \mathcal{S}_n} \left|\widehat{m}_{j,S}^{(-b)}(\boldsymbol{z}) - m_{j,S}(\boldsymbol{z})\right| \\
&\qquad = O_p\left[\frac{\exp\{-1/[2h_n(M_n+1)]\} + \sqrt{|\mathcal{S}_n|/n}}{\delta_n(S)}\right] = o_p(n^{-1/4}).
\end{aligned}$$

The case $S = \emptyset$ follows directly from the sample-mean rate.

As in the proof of Part (ii) of Lemma 2, let $e_i := X_{ij} - m_{j,S}(\mathbf{X}_{i,S})$ and $\psi_i := e_i^2 - X_{ij}$. The fourth-moment bound and conditional Jensen's inequality imply $\mathbb{E}(\psi_i^2) < \infty$. The residual expansions and laws of large numbers used in Steps 3–4 of that proof require only these moment bounds and uniform consistency. They therefore give

$$Q_{j,S} \xrightarrow{p} \Delta_{j|S}, \qquad \widehat{V}_{j,S} \xrightarrow{p} \mathrm{Var}(\psi_i).$$

The limiting variance is positive by assumption for $j \in \mathcal{W}$ and by conditional Poisson full support for $j \notin \mathcal{W}$. Consequently, whenever $\Delta_{j|S} \neq 0$,

$$\frac{|T_{j,S}|}{\sqrt{n}} \xrightarrow{p} \frac{|\Delta_{j|S}|}{\mathrm{Var}(\psi_i)^{1/2}} =: c_{j,S} > 0.$$

For $j \notin \mathcal{W}$, the conditional Poisson mechanism and non-degeneracy give

$$\Delta_{j|S} = 0 \quad \Longleftrightarrow \quad \mathrm{Pa}_G(j) \subseteq S.$$

Such a parent-containing set $S \subseteq \mathcal{A}$ is possible only when $j \in \mathcal{A}$, since every node in $\mathcal{V}_{\mathcal{W}} \setminus \mathcal{W}$ has a parent in $\mathcal{V}_{\mathcal{W}}$. For these null hypotheses, Part (ii) of Lemma 2 applies directly to the Poisson marginal $\boldsymbol{X}_{\mathcal{A}}$ and gives $T_{j,S} = O_p(1)$.

Let $z_n = \Phi^{-1}(1 - \alpha_n/2)$. As in the proof of Theorem 4,

$$z_n \to \infty, \qquad \frac{z_n}{\sqrt{n}} \to 0.$$

Thus every parent-containing test at a correctly specified node fails to reject, whereas every test with a nonzero dispersion gap rejects, with probability tending to one. The adaptive truncation threshold gives the same conclusions by the argument in Lemma 7.

Finally, for admissible sets $S, S \cup \{k\} \subseteq \mathcal{A}$, the variance decomposition in Lemma 8 and the convergence of Q give

$$Q_{j,S\cup\{k\}} - Q_{j,S} = \Delta_{j|S\cup\{k\}} - \Delta_{j|S} + o_p(1) \leq o_p(1).$$

Hence every missing parent of a node in $\mathcal{A}$ satisfies the expansion criterion with probability tending to one.

Define $\mathcal{E}'_n$ as the event on which the following properties hold simultaneously for admissible conditioning sets contained in $\mathcal{A}$:

1. Every equidispersion test for a node in $\mathcal{W}$ rejects.
2. For every node in $V \setminus \mathcal{W}$, a parent-containing test fails to reject, whereas a parent-omitting test rejects.
3. For every node in $\mathcal{A}$, every missing true parent satisfies the expansion criterion.

Since p is fixed, only finitely many such decisions are possible. A union bound therefore yields

$$\mathbb{P}(\mathcal{E}'_n) \longrightarrow 1. \tag{38}$$

Step 2: layer assignment on $\mathcal{E}'_n$. For the remainder of the proof, suppose that $\mathcal{E}'_n$ occurs. Assume inductively that, before layer step t,

$$\widehat{\mathcal{L}}_{<t} \subseteq \mathcal{A}, \qquad \widehat{\mathcal{L}}_{<t} \text{ is ancestral in } G.$$

For $t = 1$, this holds because $\widehat{\mathcal{L}}_{<1} = \emptyset$. Every conditioning set considered for a remaining node is a subset of $\widehat{\mathcal{L}}_{<t}$ and hence admissible. Indeed, if it contained a descendant of the target, ancestry would imply that the target had already been selected.

First consider $j \in \mathcal{V}_{\mathcal{W}}$. If $j \in \mathcal{W}$, every equidispersion test for j rejects by property 1 of $\mathcal{E}'_n$. Otherwise, a directed path from a node in $\mathcal{W}$ to j contains a parent of j in $\mathcal{V}_{\mathcal{W}}$. This parent is absent from $\widehat{\mathcal{L}}_{<t}$, so every available conditioning set omits a parent of j. By property 2, every corresponding test rejects. Thus no node in $\mathcal{V}_{\mathcal{W}}$ is selected at step t. Moreover, any selected node must have all its parents in $\widehat{\mathcal{L}}_{<t}$ by property 2. The enlarged selected set therefore remains ancestral and contained in $\mathcal{A}$.

It remains to show that the algorithm makes progress whenever $\mathcal{A} \setminus \widehat{\mathcal{L}}_{<t} \neq \emptyset$. Choose a source j of the subgraph induced by these remaining nodes. Since $\mathcal{A}$ is ancestral,

$$\mathrm{Pa}_G(j) \subseteq \widehat{\mathcal{L}}_{<t}.$$

If the initial set $\widehat{C}_j$ contains all parents, its equidispersion test fails to reject by property 2. Otherwise, sequential expansion considers every missing parent, and property 3 ensures that each is added when considered. Property 2 prevents acceptance while a parent remains missing and ensures acceptance once all parents have been included. Thus j is assigned to $\widehat{\mathcal{L}}_t$. This argument only uses $\widehat{C}_j \subseteq \widehat{\mathcal{L}}_{<t}$ and an ordering of the remaining candidates; any fixed convention for undefined screening statistics and ties is sufficient.

Consequently, Algorithm 1 selects at least one node at each step until all nodes in $\mathcal{A}$ have been selected, while no node in $\mathcal{V}_{\mathcal{W}}$ is selected. The remaining set is then exactly $\mathcal{V}_{\mathcal{W}}$. Every test for every remaining node rejects, so the next estimated layer is empty and the algorithm assigns VIOLATION to all remaining nodes. Therefore,

$$\mathcal{E}'_n \subseteq \left\{ \widehat{\ell}_j = \textsc{selected}\ \forall j \in V \setminus \mathcal{V}_{\mathcal{W}}, \quad \widehat{\ell}_j = \textsc{violation}\ \forall j \in \mathcal{V}_{\mathcal{W}} \right\}.$$

Combining this inclusion with (38) proves the theorem. □

A.3 Numerical experiment details

This section provides details of graph generation and data-generating mechanisms for the experiments in Section 5. Supplementary evaluations cover additional graph-recovery metrics and sample sizes, larger graphs, linear predictors, false violation detection rates under correct Poisson DAG models, tuning-parameter sensitivity, and an ablation of screening and expansion.

Table 3: Poisson-compatible conditional-mean specifications.

Specification	Conditional mean $\mu_j(\boldsymbol{X}_{\mathrm{Pa}_{G(j)}})$
Box–Cox	$\left\lvert 1 + \sum_{k \in \mathrm{Pa}_{G(j)}} \beta_{jk} \frac{(1+X_k)^{0.7} - 1}{0.7} \right\rvert$
Sine-sum	$\left\lvert 1 + \sum_{k \in \mathrm{Pa}_{G(j)}} \beta_{jk} X_k + \sin\left\{ \frac{\pi}{2} \sum_{k \in \mathrm{Pa}_{G(j)}} X_k \right\} \right\rvert$
Parity	$\left\lvert 1 + \sum_{k \in \mathrm{Pa}_{G(j)}} \beta_{jk} \mathbb{1}\{X_k \text{ is odd}\} \right\rvert$
Periodic jump	$\left\lvert 1 + \sum_{k \in \mathrm{Pa}_{G(j)}} \beta_{jk} \psi(X_k) \right\rvert$

A.3.1 Graph generation

All graphs follow the causal ordering $1, \ldots, p$, with edges directed from earlier to later nodes. Unless otherwise stated, we set $p = 30$. For each graph family, $g = 2$ and $g = 4$ specify the sparse and dense settings, respectively, through the constructions below.

Hub graphs. Node 1 is a hub, and $g-1$ additional hubs are sampled uniformly without replacement from $\{2, \ldots, p-1\}$. Each hub is a parent of every later node. This gives two-hub and four-hub graphs for $g = 2$ and $g = 4$, respectively.

Barabási–Albert graphs. Let $m = g/2$. We initialize a complete graph on $m+1$ nodes, oriented according to the causal ordering. Each subsequent node selects m earlier nodes as parents without replacement, with probabilities proportional to their current total degrees. This gives $mp - m(m+1)/2$ edges.

Erdős–Rényi graphs. Each edge $k \to j$, $k < j$, is included independently with probability $g/(p-1)$, giving expected average total degree g.

A.3.2 Poisson-compatible link functions

For the correctly specified nonlinear models, $X_j \mid \boldsymbol{X}_{\mathrm{Pa}_{G(j)}} \sim \mathrm{Poisson}\{\mu_j(\boldsymbol{X}_{\mathrm{Pa}_{G(j)}})\}$, and each node independently selects one of the four specifications in Table 3 with equal probability. We independently draw coefficient magnitudes $|\beta_{jk}| \sim \mathrm{Unif}(0.5, 1.5)$, assigning each sign with probability $1/2$.

For the periodic-jump specification, let $r(x) = x - 10\lfloor x/10 \rfloor$ and define

$$\psi(x) = \begin{cases} r(x), & 0 \le r(x) < 5, \\ 15 - r(x), & 5 \le r(x) < 10. \end{cases}$$

This transformation has period 10 and jumps at $r(x) = 5$ and the period boundary.

A.3.3 Violation mechanisms

For each local misspecification experiment, one node is selected uniformly and assigned the specified non-Poisson mechanism. Each remaining node independently uses the same mechanism with probability 0.2, and otherwise selects one of the four Poisson specifications uniformly. Let $\mathcal{W}$ denote all misspecified nodes. In these experiments, Poisson nodes have an intercept of 1 and signed coefficients with magnitudes drawn from $\mathrm{Unif}(0.5, 1.5)$. For $j \in \mathcal{W}$, we use signed coefficients with magnitudes

drawn from $\mathrm{Unif}(0.15, 0.45)$ and set

$$\eta_j = \left| \alpha_j + \sum_{k \in \mathrm{Pa}_G(j)} \beta_{jk} \log(1 + X_k) \right|.$$

Binomial. We generate $X_j \mid \boldsymbol{X}_{\mathrm{Pa}_G(j)} \sim \mathrm{Binomial}(20, \pi_j)$, where $\pi_j = \exp(\eta_j)/\{1 + \exp(\eta_j)\}$. We draw $u_j \sim \mathrm{Unif}(0.20, 0.45)$ and set $\alpha_j = \log\{u_j/(1 - u_j)\}$. The conditional variance is $20\pi_j(1 - \pi_j)$, below the conditional mean $20\pi_j$.

Negative binomial. We generate $X_j \mid \boldsymbol{X}_{\mathrm{Pa}_G(j)} \sim \mathrm{NB}(2, \mu_j)$, where $\mu_j = \exp(\eta_j)$ and $\alpha_j = 0$. The conditional variance is $\mu_j + \mu_j^2/2$.

Latent variables. We sample $L \sim \mathrm{Unif}\{1, 2, 3\}$ and generate a graph on $p + L$ nodes. We select L latent nodes from positions having at least two later observed nodes. For each latent node, the number of added observed children is sampled uniformly between two and the number of later observed nodes, and that many children are selected uniformly without replacement. Existing edges are retained. All nodes follow Poisson mechanisms with intercept of 1 and signed coefficient magnitudes drawn from $\mathrm{Unif}(0.5, 1.5)$. The four equally likely links are Box-Cox, sine-sum, parity, and periodic jump. We then hide the latent variables and apply PoET to the p observed variables.

A.3.4 Violation detection at additional sample sizes

Table 4 extends the violation-detection results in the main text by including $n = 3000$ and $n = 7000$, with all other experimental settings unchanged.

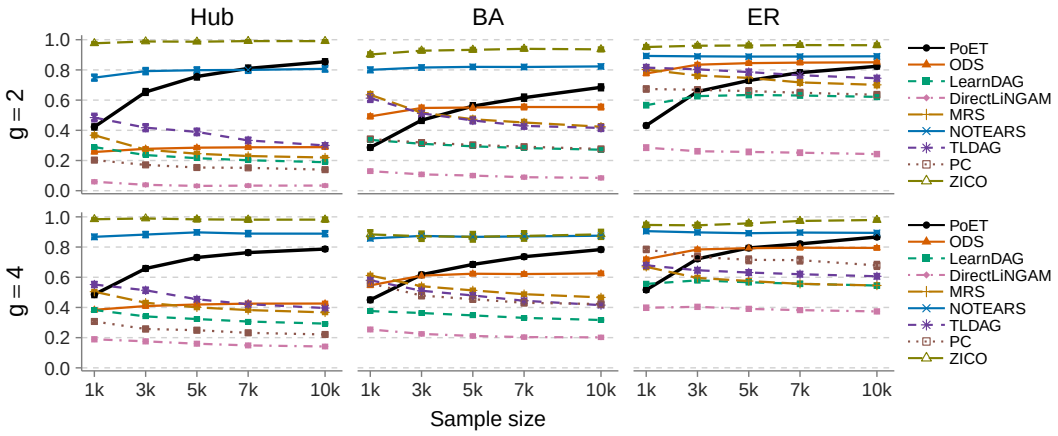
Table 4: Average violation-detection F1 scores (standard errors) of PoET under misspecified models.

Graph density	n	Binomial			Negative Binomial			Latent Poisson		
		Hub	BA	ER	Hub	BA	ER	Hub	BA	ER
$g = 2$	1000	0.88 (0.016)	0.77 (0.017)	0.78 (0.012)	0.36 (0.032)	0.23 (0.028)	0.24 (0.020)	0.38 (0.032)	0.40 (0.032)	0.43 (0.027)
	3000	0.93 (0.012)	0.82 (0.015)	0.83 (0.013)	0.81 (0.024)	0.65 (0.027)	0.64 (0.022)	0.58 (0.032)	0.56 (0.031)	0.64 (0.025)
	5000	0.95 (0.009)	0.85 (0.014)	0.84 (0.012)	0.90 (0.022)	0.81 (0.020)	0.80 (0.018)	0.68 (0.030)	0.66 (0.027)	0.76 (0.024)
	7000	0.96 (0.008)	0.88 (0.013)	0.87 (0.011)	0.93 (0.019)	0.88 (0.016)	0.85 (0.016)	0.73 (0.029)	0.72 (0.025)	0.83 (0.017)
	10000	0.96 (0.007)	0.88 (0.013)	0.89 (0.010)	0.94 (0.018)	0.88 (0.017)	0.91 (0.011)	0.75 (0.030)	0.82 (0.019)	0.87 (0.016)
$g = 4$	1000	0.85 (0.016)	0.77 (0.017)	0.77 (0.011)	0.42 (0.034)	0.44 (0.038)	0.39 (0.017)	0.50 (0.027)	0.58 (0.030)	0.65 (0.019)
	3000	0.90 (0.010)	0.85 (0.013)	0.85 (0.010)	0.79 (0.023)	0.77 (0.027)	0.75 (0.020)	0.75 (0.024)	0.77 (0.022)	0.79 (0.017)
	5000	0.93 (0.008)	0.87 (0.013)	0.88 (0.009)	0.88 (0.019)	0.83 (0.023)	0.86 (0.014)	0.77 (0.024)	0.82 (0.020)	0.86 (0.015)
	7000	0.94 (0.007)	0.89 (0.012)	0.90 (0.008)	0.92 (0.016)	0.88 (0.019)	0.89 (0.011)	0.85 (0.020)	0.85 (0.018)	0.88 (0.013)
	10000	0.94 (0.007)	0.90 (0.012)	0.91 (0.008)	0.95 (0.012)	0.92 (0.016)	0.91 (0.011)	0.87 (0.016)	0.89 (0.014)	0.91 (0.011)

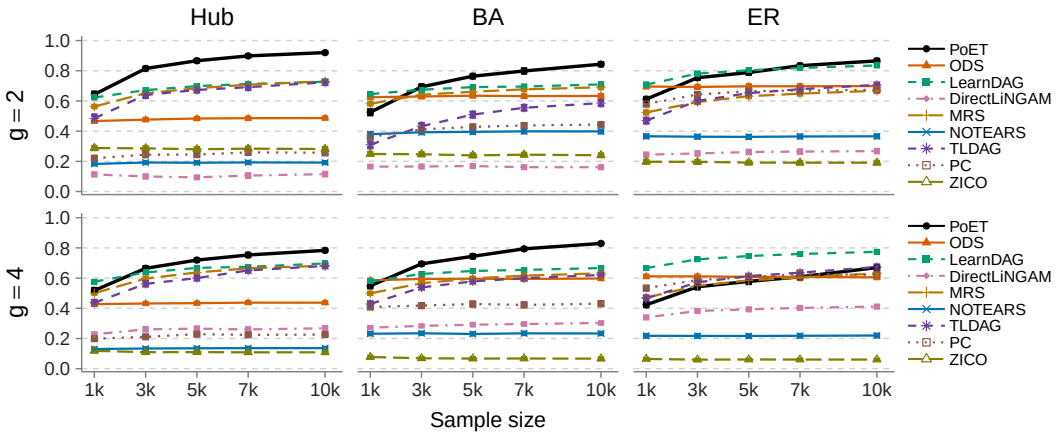
A.3.5 Additional graph-recovery metrics

We use the same experimental settings as in the main text and report precision, recall, and structural Hamming distance (SHD) to complement the F1 scores. Precision and recall distinguish false positive edges from missed edges, allowing the conservativeness of the estimated graphs to be assessed more directly. SHD summarizes structural discrepancies by counting edge additions, deletions, and reversals, each with unit cost.

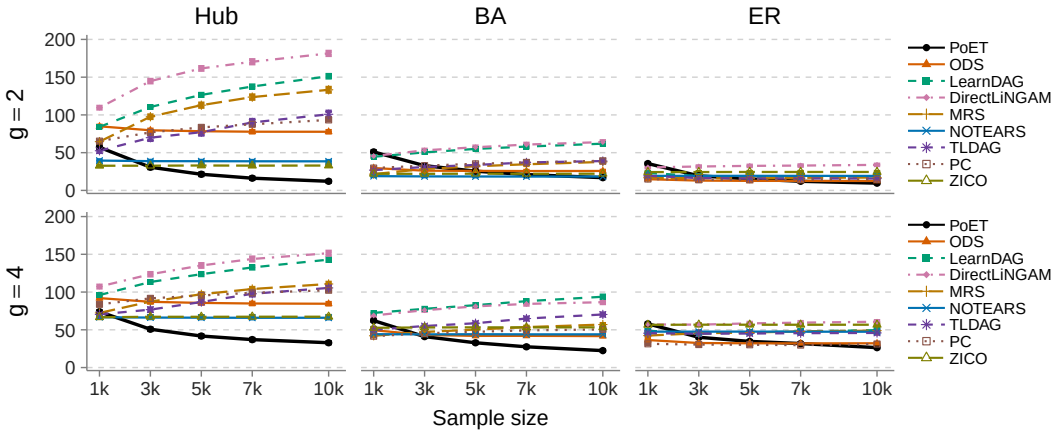
Figure 3 shows that both precision and recall of PoET generally increase with sample size across all considered graph types and density settings. Some competing methods exhibit precision close to one but low recall, consistent with very sparse, nearly empty graph estimates that miss a large fraction of the true edges. For PoET, improvement in both precision and recall indicates that larger samples allow more true edges to be recovered while reducing the proportion of false positives among the estimated edges. These improvements are accompanied by generally decreasing SHD, with particularly clear reductions for hub and BA graphs.



(a) Precision



(b) Recall



(c) Structural Hamming distance (SHD)

Figure 3: Average precision, recall, and SHD with one-standard-error bars under correctly specified Poisson DAG models.

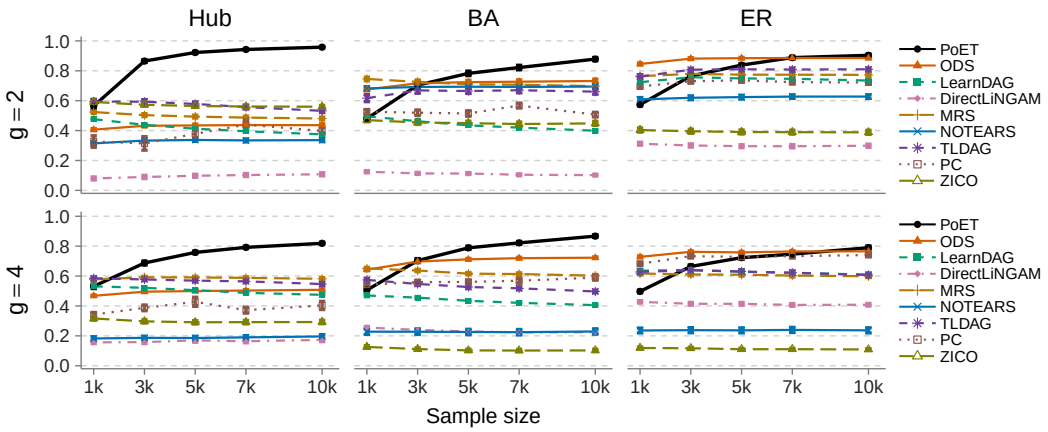


Figure 4: Average graph-recovery F1 scores with one-standard-error bars under correctly specified Poisson DAG models with linear links.

A.3.6 Graph recovery in higher dimensions

We repeat the graph-recovery comparison on Hub graphs with $g = 2$, increasing the number of nodes to $p = 100$ while keeping the remaining data-generating and tuning settings unchanged. Table 5 reports average graph-recovery F1 scores and standard errors across sample sizes. LearnDAG performs comparatively well at smaller sample sizes, while PoET improves substantially as the sample size increases. PoET achieves the highest average F1 at $n \in \{5000, 7000, 10000\}$, reaching 0.608 at $n = 10000$.

Table 5: Average graph-recovery F1 scores (standard errors). Bold indicates the highest F1 at each sample size.

n	PoET	ODS	LearnDAG	DirectLiNGAM	MRS	NOTEARS	TLDAG	PC	ZICO
1000	0.223 (0.011)	0.138 (0.002)	**0.405** (0.005)	0.033 (0.004)	0.250 (0.004)	0.006 (0.001)	0.243 (0.012)	0.215 (0.021)	0.348 (0.020)
3000	0.396 (0.017)	0.157 (0.002)	**0.471** (0.008)	0.028 (0.003)	0.183 (0.004)	0.005 (0.001)	0.273 (0.018)	0.214 (0.022)	0.341 (0.020)
5000	**0.508** (0.019)	0.162 (0.002)	0.454 (0.008)	0.029 (0.003)	0.161 (0.008)	0.006 (0.001)	0.292 (0.021)	0.213 (0.020)	0.339 (0.020)
7000	**0.564** (0.017)	0.164 (0.002)	0.433 (0.008)	0.025 (0.002)	0.147 (0.007)	0.006 (0.001)	0.295 (0.022)	0.214 (0.024)	0.337 (0.020)
10000	**0.608** (0.015)	0.165 (0.002)	0.404 (0.007)	0.025 (0.002)	0.137 (0.007)	0.005 (0.001)	0.261 (0.018)	0.211 (0.027)	0.338 (0.020)

A.3.7 Graph recovery under linear link functions

To complement the nonlinear mechanisms considered in the main text, we additionally evaluate all methods using linear predictors with the same mechanism as in Appendix A.3.2. Specifically, at every node

$$\mu_j(\boldsymbol{X}_{\mathrm{Pa}_G(j)}) = \left| 1 + \sum_{k \in \mathrm{Pa}_G(j)} \beta_{jk} X_k \right|.$$

The coefficient distribution and all other experimental settings remain unchanged.

Figure 4 shows that the graph-recovery F1 scores of PoET increase with sample size across all graph families and both density settings. PoET exhibits clearer advantages on BA and hub graphs as the sample size increases, although ODS remains competitive on ER graphs, particularly at smaller sample sizes.

A.3.8 False violation detection rates under correct models

Table 6 reports the node-level false-violation rates of PoET under correctly specified Poisson DAG models on hub graphs with $p = 30$ and $g = 2$. Each entry is the average fraction of nodes labeled VIOLATION across replications. The rates are close to zero at larger sample sizes, consistent with the expected behavior under correct model specification.

Table 6: Average node-level false violation detection rates (Monte Carlo standard errors) under correctly specified Poisson DAG models.

n	1000	3000	5000	7000	10000
False violation detection rate	0.025 (0.014)	0.001 (0.001)	0.000 (0.000)	0.001 (0.000)	0.000 (0.000)

A.3.9 Sensitivity to tuning parameters

We assess one parameter at a time using Hub graphs with $g = 2$ and $n = 1000$, under a correctly specified Poisson DAG model and binomial model. Each configuration uses 30 replications, with matched data and cross-fitting seeds across configurations. The baseline significance levels are $\alpha_n^{\text{corr}} = \alpha_n^{\text{eq}} = 2\{1 - \Phi(n^{1/4}/3)\}$. We assess tuning-parameter sensitivity by varying one parameter at a time: multiplying either significance level by 0.8 or 1.2, setting $\tau \in \{0.5, 0.9\}$, varying the CMP bandwidth multiplier $c \in \{0.4, 0.6\}$ in $h_n = cn^{-1/(2+d)}$, or setting $\epsilon \in \{0, 10^{-4}\}$. All other parameters are held at their baseline values. Cross-fitting remains fixed at 5 folds. Table 7 reports average graph-recovery F1 under correct specification and average violation-detection F1 under binomial misspecification, with standard errors. These comparisons assess sensitivity of multiple tuning parameters involved in PoET.

Table 7: Tuning-parameter sensitivity of PoET. One parameter is varied at a time. Entries are means (standard errors).

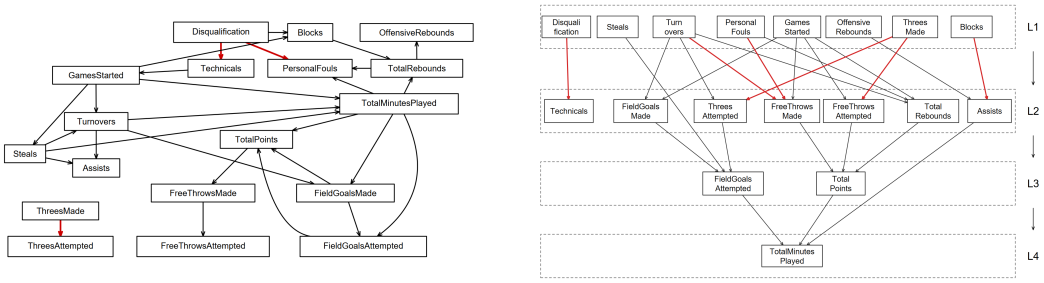
Setting	Correct model Graph-recovery F1	Binomial Violation-detection F1
Baseline	0.516 (0.025)	0.850 (0.034)
$0.8\alpha_{\text{eq}}$	0.503 (0.025)	0.862 (0.033)
$1.2\alpha_{\text{eq}}$	0.529 (0.025)	0.861 (0.033)
$0.8\alpha_{\text{corr}}$	0.518 (0.025)	0.855 (0.032)
$1.2\alpha_{\text{corr}}$	0.503 (0.024)	0.849 (0.032)
$\tau = 0.5$	0.661 (0.029)	0.861 (0.028)
$\tau = 0.9$	0.679 (0.030)	0.825 (0.036)
$c = 0.4$	0.508 (0.025)	0.865 (0.030)
$c = 0.6$	0.511 (0.025)	0.848 (0.036)
$\epsilon = 0$	0.516 (0.025)	0.858 (0.034)
$\epsilon = 10^{-4}$	0.516 (0.025)	0.858 (0.034)

A.3.10 Ablation of screening and expansion

With the data-generating and tuning settings unchanged on Hub graphs with $g = 2$, we compare the full procedure with a variant that conditions on all previously selected nodes, bypassing parent-candidate screening and sequential expansion while retaining pruning and the remaining settings. Table 8 reports the same two F1 metrics with 30 replications. The full procedure improves graph recovery, increasing F1 from 0.395 to 0.516, while also improving violation-detection F1 from 0.610 to 0.850. These results support the role of screening and expansion in avoiding unnecessarily large conditioning sets, which is particularly important for nonparametric conditional-moment estimation and equidispersion tests.

Table 8: Ablation of screening and expansion. Entries are means (standard errors).

Setting	Correct model Graph-recovery F1	Binomial Violation-detection F1
PoET (screening + expansion)	0.516 (0.025)	0.850 (0.034)
All previous nodes	0.395 (0.036)	0.610 (0.044)



(a) ODS estimate reported by Park and Raskutti.

(b) TLDAG estimate reported by Zhou et al.

Figure 5: Previously reported DAG estimates for the 2009–2010 NBA player-statistics data under the conditional Poisson specification.

A.4 Details of the real data analysis

NBA player statistics. The original dataset contains 25 variables. Following Park & Raskutti (2018), we exclude the league indicator, the non-numerical variables Name, Team, and Position, and GamesPlayed. We further remove Ejections and FlagrantFouls, both of which are identically zero in this dataset. The resulting dataset consists of $(p = 18)$ integer-valued variables: TotalMinutesPlayed, FieldGoalsMade, FieldGoalsAttempted, ThreesMade, ThreesAttempted, FreeThrowsMade, FreeThrowsAttempted, OffensiveRebounds, TotalRebounds, Assists, Steals, Turnovers, Blocks, PersonalFouls, Disqualifications, TotalPoints, Technicals, and GamesStarted.

The proposed procedure fails to identify a first layer. Under the Poisson DAG model, a node without parents should satisfy marginal equidispersion. However, the equidispersion hypothesis is rejected for all variables, so none of them can be assigned to the first layer. Consequently, the subsequent layer and edge recovery steps cannot be performed. Several variables identified as root nodes in previous studies also violate marginal equidispersion. As shown in Figure 5, some edges originating from these root nodes have counterintuitive directions, which are highlighted in red. These results suggest that violations of the Poisson DAG assumptions may contribute to such edge directions.

A score can identify the most root-like node even when no node is sufficiently root-like; a test can conclude that no root layer is supported.

Lending Club variables. We analyze the Lending Club dataset in the R package `openintro`. To reduce observed heterogeneity, we restrict the sample to grade-A debt-consolidation loans submitted as individual applications by borrowers with mortgage homeownership, yielding $n = 455$ observations. The analysis uses $p = 5$ variables: annual income, the number of mortgage accounts, the number of current installment accounts, delinquencies in the previous two years, and historical failures to pay. Annual income captures economic resources, while the two account counts represent distinct forms of credit holdings. Delinquencies in the previous two years and historical failures to pay capture complementary aspects of adverse credit history and are included as potential common influences on both account types. This selection narrows the scope of the analysis to borrowers' credit-account holdings and their relationships with financial resources and repayment history. Annual income ranges from \$15,000 to \$1,050,000 and is replaced by the corresponding bin index from 0 to 19 using 20 equal-width intervals of \$51,750.

Sensitivity analysis for the Lending Club data. We examine sensitivity to income binning and the sample restrictions, keeping the original five variables, $\alpha_n = 0.05$, and all other tuning parameters fixed. In the baseline sample, we relabel the original 20 income intervals from 0–19 to 1–20. We also consider 15 and 25 equal-width intervals over the same observed income range, labeled 0–14 and 0–24. For the sample comparison, we allow all observed categories of application type, homeownership, loan purpose, or grade separately, while retaining the other three restrictions. We also remove all four restrictions together. The baseline income-bin cutoffs are held fixed in these sample comparisons, with values outside the baseline range assigned to the nearest endpoint bin.

Table 9: Sensitivity of PoET for the Lending Club data to income binning and sample restrictions at $p = 5$ and $\alpha_n = 0.05$.

Change	Specification	n	Layers	Nodes	Edges	Result
Baseline	20 bins, coded 0–19	455	3	5/5	5	Complete
Income binning	Same 20 bins, coded 1–20	455	3	5/5	5	Complete
	15 bins, coded 0–14	455	3	4/5	3	Incomplete
	25 bins, coded 0–24	455	3	5/5	5	Complete
Sample restrictions	Allow all application types	566	1	2/5	0	Incomplete
	Allow all homeownership categories	928	1	1/5	0	Incomplete
	Allow all loan purposes	1068	1	1/5	0	Incomplete
	Allow all grades (A–G)	2007	1	1/5	0	Incomplete
	Remove all four restrictions	10000	1	1/5	0	Incomplete

Relabeling the same 20 income intervals from 0–19 to 1–20 leaves all three layers and five edges unchanged. Under the 15-bin setting, the variable recording current installment accounts remains unassigned, while the other four variables retain their baseline layers and three edges. The 25-bin setting assigns all five variables and preserves four of the five baseline edges; the variable recording delinquencies in the previous two years enters an earlier layer, with annual income replacing mortgage accounts as its parent. These comparisons indicate that key relationships are relatively robust to income binning, although layer assignments and some edges depend on the chosen setting. Binning is used to include annual income as an important potential confounder, since borrowers' financial resources may affect access to both mortgage and installment credit. Although the resulting bin index is not a naturally observed count, this approximation allows income to be considered in the count-valued analysis rather than omitted.

Relaxing any one sample restriction leaves only one or two variables assigned to the first layer, and no subsequent layer is identified. Only the historical-failures-to-pay variable enters the first layer, except when all application types are allowed, in which case annual income also enters. Pooling these categories introduces greater heterogeneity in borrower characteristics. Homeownership and application type may reflect differences in housing and household financial circumstances, whereas grade and loan purpose may serve as proxies for borrower credit characteristics. These underlying characteristics may act as unmeasured common causes of the five variables. The incomplete DAG recovery in the broader samples is therefore consistent with possible unmeasured confounding introduced by pooling these categories.

Table 10: Variables in the Lending Club dataset. Variable names and descriptions follow the OpenIntro documentation. Variable types are analysis-specific classifications based on measurement meaning rather than storage type. Monetary amounts and ratios are treated as continuous, and term as categorical with contractual levels of 36 and 60 months. Employment length and recency variables are integer-valued durations rather than event counts and need not satisfy a conditional Poisson model.

Variable type	Variable	Description
Count-valued (event/account)	delinq_2y	Number of delinquencies on credit accounts during the preceding two years.
	inquiries_last_12m	Number of inquiries into the applicant's credit during the preceding 12 months.
	total_credit_lines	Total number of credit lines in the applicant's credit history.
	open_credit_lines	Number of credit lines that are currently open.
	num_collections_last_12m	Number of non-medical collections during the preceding 12 months.
	num_historical_failed_to_pay	Number of derogatory public records, approximately measuring historical failures to pay.
	current_accounts_delinq	Number of accounts on which the applicant is currently delinquent.
	current_installment_accounts	Number of installment accounts, such as fixed-payment automobile loans.
	accounts_opened_24m	Number of new credit lines opened during the preceding 24 months.
	num_satisfactory_accounts	Number of satisfactory accounts.
	num_accounts_120d_past_due	Number of current accounts that are 120 days past due.
	num_accounts_30d_past_due	Number of current accounts that are 30 days past due.
	num_active_debit_accounts	Number of bank cards currently active.
	num_total_cc_accounts	Total number of credit-card accounts in the applicant's credit history.
	num_open_cc_accounts	Number of credit-card accounts that are currently open.
	num_cc_carrying_balance	Number of credit cards carrying an outstanding balance.
	num_mort_accounts	Number of mortgage accounts.
	public_record_bankrupt	Number of bankruptcies appearing in the applicant's public records.
Integer-valued (duration/recency)	emp_length	Number of years in the applicant's job, rounded down and top-coded at 10.
	months_since_last_delinq	Number of months since the applicant's most recent delinquency.
	months_since_90d_late	Number of months since the applicant was most recently 90 days late on a payment.
	months_since_last_credit_inquiry	Number of months since the applicant's most recent credit inquiry.
Categorical	emp_title	Applicant's job title.
	state	Two-letter state code.
	homeownership	Ownership status of the applicant's residence.
	verified_income	Verification status of the applicant's individual income.
	verification_income_joint	Verification status of joint income.
	loan_purpose	Purpose reported for the loan.
	application_type	Indicates whether the loan application is individual or joint.
	term	Duration of the loan in months.
	grade	Loan grade.
	sub_grade	Loan subgrade.
	issue_month	Calendar month in which the loan was issued.
	loan_status	Status recorded for the loan.
	initial_listing_status	Status recorded when the loan was initially listed.
	disbursement_method	How the loan proceeds were disbursed.
Continuous	annual_income	Applicant's annual income.
	debt_to_income	Applicant's debt-to-income ratio.
	annual_income_joint	Combined annual income of both applicants for a joint application.
	debt_to_income_joint	Combined debt-to-income ratio of both applicants for a joint application.
	total_credit_limit	Total available credit limit, excluding mortgage credit.
	total_credit_utilized	Total utilized credit balance, excluding mortgage balances.
	total_collection_amount_ever	Total monetary amount the applicant has had in collections.
	total_debit_limit	Sum of the credit limits across all bank cards.
	account_never_delinq_percent	Percentage of the applicant's credit lines that have never been delinquent.
	loan_amount	Amount of the loan received by the applicant.
	interest_rate	Interest rate assigned to the loan.
	installment	Required monthly installment payment.
	balance	Current outstanding balance on the loan.
	paid_total	Total amount paid on the loan to date.
	paid_principal	Original loan amount minus the current loan balance.
	paid_interest	Amount of interest paid to date.
	paid_late_fees	Amount of late fees paid to date.
Other	earliest_credit_line	Calendar year in which the applicant opened the earliest recorded credit line.
	tax_liens	Numeric variable; no detailed definition is provided in the OpenIntro documentation.