CONDITIONAL INDEPENDENCE TESTING FROM DEPENDENT SAMPLES WITH APPLICATIONS IN CAUSAL DISCOVERY FROM TIME-SERIES DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from time-series data is a fundamental problem with many applications in several industries. Most causal discovery algorithms use conditional independences, explicitly or implicitly, between observed variables to learn the causal graph. Directly applying these algorithms to time-series data is challenging due to the dependence across samples created through the causal effects over time. The landmark PCMCI algorithm addresses this issue by a two-stage approach: While the first stage is a conservative edge pruning stage, the critical second stage tests independence between two variables by conditioning on the union of parents of both variables. This removes the dependence between samples through time, allowing us to plug in conditional independence tests designed for IID data. However, large conditioning sets in the second stage often have low statistical power with finite data since few samples are left per stratum. To address this issue, we first analyze the chi-square independence test from dependent samples. For stationary, rapidly mixing processes, we show that the test statistic still converges to a chi-square distribution when conditioning on the parents of one variable, allowing the other variable to be dependent through time. Our work, to the best of our knowledge, provides the first explicit dependence of the chi-square test on the mixing time of the underlying time series. Based on this, we develop a causal discovery algorithm, which, in its second stage, conditions on the parents of only one of the variables. The smaller conditioning set significantly improves statistical power and causal discovery performance. We show through synthetic experiments that our algorithm outperforms PCMCI and other baselines.

1 INTRODUCTION

Discovering cause-effect relations from data is a fundamental problem Pearl (2009); Spirtes et al. (2000). Understanding the qualitative causation between observed variables is necessary for making quantitative assessments of causal effect, which in turn informs decision-making. Besides decision-making, causal discovery has applications in ML ranging from root-cause analysis (RCA) to characterizing domain shifts and helping with generalization, evaluating and achieving fairness, and explainability of large models Budhathoki et al. (2022); Schölkopf et al. (2021); Kusner et al. (2017); Vig et al. (2020).

Many settings relevant for causal discovery has an abundance of *time-series* data. In microservices, for example, several metrics, such as memory usage or user-related events, are recorded with a time-stamp. Hospitals keep track of each patient's medical history, which contains the time progression of their health and diseases. In manufacturing, sensors record information related to the health of production through time. Practical causal discovery from time-series data has the potential to unlock inferences not available today across several industries Hardt et al. (2024); Harutyunyan et al. (2019); Mogensen et al. (2024); Assaad et al. (2022).

Observational causal discovery relies on signals left on the joint probability distribution by the underlying causal structure. The most common such signals are the conditional independence relations between variables. Under a general position assumption known as faithfulness, which asserts that conditional independences cannot appear by chance and must be structural, several sound and complete algorithms have been proposed over the last few decades Spirtes et al. (2000). One can classify these algorithms into constraint-based, score-based and functional assumption-based algorithms Glymour et al. (2019). Any method that does not rely on parametric assumptions relies on the conditional independence signatures in the data, either explicitly in the case of constraint-based methods, or implicitly in the case of score-based methods.

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.

In this paper, we focus on constraint-based methods from time-series data. Using Pearl's SCM framework Pearl (2009), we can model the underlying data-generating process as a collection of structural equations. Causal graph is obtained as the directed acyclic graph (DAG) where all variables that appear in the structural equation of a variable are assigned as the parents of the said variable. The goal of causal discovery is to learn this DAG from observational data. With time-series data, system is often modeled as a DAG extending through time, the so-called time-unfolded graph. An often used assumption is *structural stationarity*, which says that the parent sets of each variable are identical, just shifted in-time Peters et al. (2013); Runge et al. (2019); Gao et al. (2023).

Having time-series data is a blessing on one side: We know that the future cannot cause the past Granger (1969); Runge (2020). It is, however, also a curse in that the samples we have access to are dependent due to causation through time. This breaks the very fundamental assumption that the samples are independent, identically distributed (IID), which is made by classical constraint-based causal discovery algorithms Spirtes et al. (2000); Glymour et al. (2019); Runge (2020). It is a hard problem to getaround since, to the best of our knowledge, there is currently no conditional independence tester that can systematically handle dependent samples.

To address this problem systematically, Runge proposed PCMCI algorithm Runge et al. (2019). This two-stage algorithm first removes edges by a conservative first stage by running CI tests from dependent samples. The underlying assumption is that this phase will only detect strong independences that can be detected despite dependence among samples, and will not falsely remove edges. The algorithm, in its second stage directly addresses the dependence through time problem by conditioning on the parents of the variables being tested. Due to no unobserved confounder assumption, observed samples are independent through time, and one can plug-in a CI test for IID data in this stage. However, first stage may fail to sparsify the graph sufficiently and the conditioning set may become very large in the second stage. This means data will be stratified too much, losing statistical power to correctly execute the second stage. The culprit here is the need to convert samples dependent through time into independent samples, at the expense of using a large conditioning set.

In this paper, we address this fundamental problem. First, for stationary, rapidly mixing processes, we show that chi-square statistic converges to a chi-square distribution even when the samples of one of the variables being tested are dependent, by conditioning on the parents of the other variable. Under the null hypothesis, chi-square statistic is shown to converge to a standard chi-square distribution whereas under the alternative hypothesis, i.e., when the two variables being tested are dependent, diverges. Note that the trivial approach of waiting long enough to collect almost independent samples would result in one effective sample per mixing time. Instead, in our theory we retain alternating blocks whose length depends on the mixing time, discarding half the observations while separately accounting for exploration. We then develop a time-series causal discovery algorithm using this new CI test from dependent samples. Following PCMCI outline Runge et al. (2019), our algorithm conditions on only one of the variables in CI tests, effectively reducing the conditioning set size to its half, which significantly amplifies statistical power of a finite-length time-series.

The following are our contributions:

- For stationary, rapidly mixing processes, we prove convergence of chi-square statistic to chi-square distribution when conditioning on one parental set, even when the samples of the other variable are dependent. We also show divergence of the statistic from dependent samples under the alternative.
- We characterize sufficient run lengths for table regularity and comparison with an IID testing benchmark, as a function of mixing time of the underlying Markov chain.
- We develop a time-series causal discovery algorithm that uses the new test for practical causal discovery from time-series data.

2 Related Work

Despite its prevalence, time series data received much less attention relative to the IID setting from the causal discovery community. Among non-parametric approaches, PCMCI Runge et al. (2019) arose as the main constraint-based method, analogous to the PC algorithm. Several variants were then developed that handle multiple regimes Saggioro et al. (2020),seasonal non-stationarity Gao et al. (2023), contemporaneous edges Runge (2020), unobserved confounders Gerhardus & Runge (2020), or multiple datasets Günther et al. (2023). Specialized methods under additional assumptions arose, such as Hyvärinen et al. (2010) for linear models with non-Gaussian noise, Peters et al. (2013) that leverages independence of residuals. Similarly, Pamfil et al. (2020) applies to linear models by differentiably optimizing a score with regularization to encourage the acyclicity constraints.

We are interested in the generic, non-parametric setting, e.g., discrete variables, no functional assumptions. For any arbitrary causal graph (with no latents) it is well known that the conditional independences are the only signatures left

in the observational distribution about the causal structure. And any method needs to, either explicitly (constraint-based methods), or implicitly (score-based methods), use these signatures. This is precisely where time-series causal discovery poses unique challenges relative to the IID setting. Although independence testing is straightforward with IID samples, there is no generic independence test with dependent samples. PCMCI and variants still use off-shelf independence testers designed for IID data, but handle the errors by having two stages of pruning with different significance thresholds. First stage is assumed to obtain a superset of parents, and in the second stage, the problem is turned into the IID setting by *conditioning on the union of the parents*. However, to obtain any consistency guarantees, they assume a consistent CI test that works with dependent samples. In this paper, we make progress in this direction and analyze how chi-square statistics behave with time-series data when we allow dependence instead of conditioning on the union of both parents, which significantly reduces statistical power in practice.

3 Conditional Independence Testing from Dependent Samples

Constraint-based causal discovery determines edges using conditional independence (CI) tests. For a possible edge $X_{t-\delta} \to Y_t$, one tests

$$X_{t-\delta} \perp\!\!\!\perp Y_t \mid \mathbf{S}$$

for appropriate conditioning sets $\mathbf{S}$. With IID observations, standard tests such as Pearson's chi-square test or the G-test can detect a fixed dependence strength $\Lambda_{\min} > 0$ using n_{iid} samples, where n_{iid} depends on the desired significance level, power, and number of categories. With time-series data, however, the samples are generally dependent across t, so the usual IID calibration does not apply even when $\mathbf{S}$ is a valid d-separating set.

Let us consider two examples that illustrate why standard chi-squared independence test can not be used with time series data. The goal of these is to look at the underlying dynamics in simplified form, which could make reading or understanding the full proof, which we give later, easier.

Example 1 (false positives). Let X and Y be independent binary Markov chains, each remaining in its current state with probability $1 - \epsilon$ and flipping with probability ϵ. There is no causal path or common cause between X and Y. Each constant-state run has a geometric distribution with mean $1/\epsilon$, so over n observations each process has only $O(n\epsilon)$ transitions on average.

Suppose, for example, that each process flips only two or three times during the observed trajectory. The locations of these few transitions are random, but the resulting long runs will generally overlap unevenly: X and Y may both be 1 for a long interval, both be 0 for another, and spend relatively little time in one or both of the remaining combinations. The contingency table for (X_t, Y_{t+1}) can therefore show a strong apparent association even though the two processes are independent. In this regime, the amount of independent information is governed by the number of runs, of order $n\epsilon$, rather than by the n individual time points.

More formally, the cell counts

$$N_{ab} = \sum_{t=1}^{n} \mathbf{1}\{X_t = a, Y_{t+1} = b\}$$

are sums of positively correlated indicators: consecutive pairs tend to remain in the same cell because both processes are persistent. Their expectations under the null are unchanged, but $\mathrm{Var}(N_{ab})$ is larger by a factor of order $1/\epsilon$ than under independent sampling. The chi-square statistic therefore has heavier tails than its nominal χ^2_k distribution, so the usual critical values can produce false rejections. We focus on Pearson's chi-square test, although the same calibration issue can arise more generally for tests that assume independent observations.

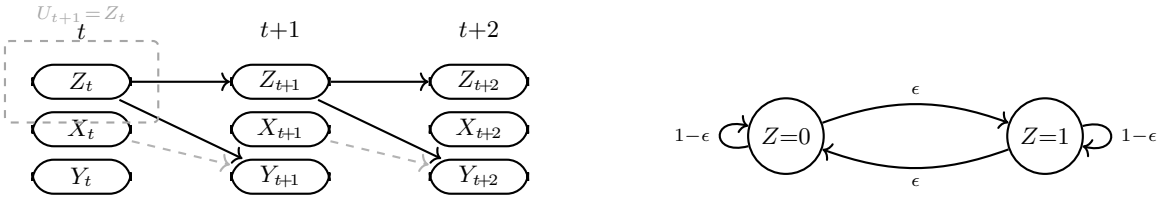
For a concrete illustration, consider large n with $n\epsilon = 2$. Using the Poisson approximation to the number of transitions, the joint process (X, Y) visits exactly three of its four possible states with probability approximately 0.38. The three visited states are likely to each occupy a nonvanishing fraction of the trajectory, in which case the nominal IID p-value tends to zero ($\Theta(n^{-1/2}e^{-Cn/2})$ for some $C > 0$), despite the independence of X and Y.

Example 2 (false negatives). Consider three observed binary processes $Z_t, X_t, Y_t \in \{0, 1\}$ with the causal structure shown in Figure 1. Let Z_t be a binary Markov chain that flips with probability ϵ, so $\tau_{\text{mix}} = \Theta(1/\epsilon)$. Let $X_t \sim \mathrm{Bernoulli}(1/2)$ independently across time, and define

$$Y_{t+1} = \begin{cases} X_t, & Z_t = 1, \\ \eta_{t+1}, & Z_t = 0, \end{cases} \qquad \eta_t \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(1/2),$$

with (η_t) independent of (X_t, Z_t). Thus X_t is a parent of Y_{t+1}, but its effect is present only in the $Z_t = 1$ regime.

If $n \lesssim 1/\epsilon$, the observed trajectory can remain almost entirely in one regime. In particular, if it remains in $Z_t = 0$, the observed Y_{t+1} values are independent fair coins regardless of X_t, and a chi-square test will fail to reject independence. Thus, for any prescribed significance and power, n can be large enough to detect the dependence from IID samples, yet the test can still fail here because the trajectory has not visited the regime in which the dependence appears.



(a) Causal graph

(b) State transitions for Z_t

Figure 1: Example 2. **(a)** The genuine edge $X_t \to Y_{t+1}$ is active only when $Z_t = 1$; testing conditions on $U_{t+1} = Z_t$ (dashed box). **(b)** The regime process Z_t is a slow-mixing two-state chain with $\tau_{\text{mix}} = \Theta(1/\epsilon)$.

These examples separate two effects of temporal dependence. The first is incorrect null calibration; the second is failure to visit informative parts of the state space. PCMCI (Runge et al., 2019), in its MCI step, conditions on estimated parents of both variables. If each parent set has size about k and each variable has at most q states, this can produce up to q^{2k} strata. Conditioning on only one parent set still gives exponentially many strata, but only up to q^k, an exponential reduction.

The difficulty is that after dropping the candidate cause's parents, the observations entering the test remain temporally dependent. Graphical separation alone therefore does not justify applying an IID chi-square test to these samples. We show that, for the discrete SCM considered here, the chi-square test is nevertheless correctly calibrated when conditioning only on the response parents. The additional structure that makes this possible is the fresh innovation generating the response at each time step. Thus the reduction in the conditioning set and the validity argument go together: the former reduces the number of strata, while the latter establishes that the remaining temporal dependence does not invalidate the test.

The remaining problem is exploration. Different strata may be visited at very different rates, and a slowly mixing process may provide few or no observations from a rare stratum. We therefore treat calibration and exploration separately: the parent-conditioning property and mixing give asymptotic test validity, while a coupling argument controls the number of observations available in each stratum, which is needed for power. Section 3.1 outlines this analysis and the resulting run-length guarantee.

Let U_t contain all parents of Y_t other than the candidate parent $X_{t-\delta}$. We assume

$$Y_t \;=\; g\big(U_t,\, X_{t-\delta},\, \eta_t\big), \qquad \eta_t \perp\!\!\!\perp \mathcal{F}_{t-1}, \tag{1}$$

where U_t and $X_{t-\delta}$ are measurable with respect to the past $\mathcal{F}_{t-1}$, and (η_t) are identically distributed. We assume a stationary, rapidly mixing process, as specified in Appendix F.4. Within this setting, $(U_t, X_{t-\delta})$ may be dependent across time and may depend on past responses.

Fix $U_t = u$ and $X_{t-\delta} = a$. Under the null that Y_t does not depend on $X_{t-\delta}$ in this stratum, the conditional distribution of Y_t is the same for every row a. Moreover, because η_t is independent of $\mathcal{F}_{t-1}$, the responses contributing to a fixed row are generated by independent innovations. The temporal dependence therefore changes how often the rows are visited, but not the conditional response distribution used in the chi-square statistic.

The test uses a separate contingency table for each value u of U_t, and sums the resulting within-stratum statistics. Our formal analysis in Appendix A uses the stratified construction. The corresponding pooled statistic decomposes into the within-stratum contributions.

For Example 1, $U_{t+1} = Y_t$, so we test $X_t \perp\!\!\!\perp Y_{t+1} \mid Y_t$. Conditional on Y_t, the next value Y_{t+1} is generated from the fresh transition innovation and does not depend on X_t. The autocorrelation in X therefore affects the row counts but not the null calibration.

For Example 2, $U_{t+1} = Z_t$. In the $Z_t = 0$ stratum, Y_{t+1} is independent of X_t; in the $Z_t = 1$ stratum, $Y_{t+1} = X_t$. Dependence in the sequence Z_t determines how many observations are obtained from each stratum, but does not introduce dependence among the response innovations conditional on (X_t, Z_t).

Theorem 3.1 (Chi-square calibration from dependent samples)**.** *Let $Y_t = g(U_t, X_{t-\delta}, \eta_t)$ with identically distributed $\eta_t \perp\!\!\!\perp \mathcal{F}_{t-1}$, where U_t and $X_{t-\delta}$ are $\mathcal{F}_{t-1}$-measurable. Assume the process is stationary and rapidly mixing (Assump-*

tion F.8), with fixed alphabets of sizes $r, c \geq 2$. Fix a stratum u with positive row and column probabilities and a fixed parental conditioning set. Use a deterministic set of observations whose size tends to infinity. Then:

1. ***Null calibration.*** *If Y_t does not depend on $X_{t-\delta}$ within stratum u, then $\chi^2_n(u) \Rightarrow \chi^2_{(r-1)(c-1)}$.*

2. ***Power.*** *If the conditional distribution of Y_t given $X_{t-\delta}$ varies across rows within stratum u, then $\chi^2_n(u) \xrightarrow{p} \infty$.*

Feedback from past Y into future (U, X) is allowed.

3.1 Proof Sketch

The full proof is given in Appendix A. Under the null, each response Y_t is generated from a fresh innovation $\eta_t \perp\!\!\!\perp \mathcal{F}_{t-1}$. The centered responses therefore have zero conditional mean given the past, even though the sequence of row visits may be temporally dependent. Mixing makes the row frequencies converge to their stationary probabilities. Replacing the random row counts by their stationary expectations contributes a vanishing error, and a martingale central limit theorem then gives the usual chi-square limit. Under a fixed alternative, the row-specific response probabilities differ, and the statistic diverges as observations accumulate in the informative rows.

Calibration alone is not enough for power. A slowly mixing process may provide too few observations from a rare or informative row-stratum combination. We control this shortfall using the blocking argument in Appendix A. The trajectory is divided into blocks of length $m \asymp \tau_{\mathrm{mix}} \log n$, alternating blocks are retained, and Berbee's coupling replaces the retained blocks by independent copies with vanishing total-variation error. A concentration bound can then be applied to the resulting row counts. Theorem F.1 gives the corresponding comparison with the original process.

This argument separates the two contributions to the required run length: n_{iid} is an IID benchmark for the number of observations in a stratum, while mixing contributes an additional cost needed to ensure that the relevant cells are visited sufficiently often.

3.2 Combined Sample Complexity

Combining the IID sample-size benchmark with the row-count bound gives the sufficient scaling

$$n \;=\; \underbrace{\frac{c_0 \cdot n_{\mathrm{iid}}}{\pi_u}}_{\substack{\text{signal}\\ \text{accumulation}}} \;+\; \underbrace{O\!\left(\frac{\tau_{\mathrm{mix}} \log^2 n}{\pi_{\min}}\right)}_{\substack{\text{exploration}\\ \text{regularity}}}, \tag{2}$$

where π_u is the stationary probability of the tested stratum, $\pi_{\min}$ is that of the least frequent row-stratum combination, and c_0 is the blocking constant. The hidden constant depends on the fixed alphabet sizes and positive conditional probabilities. The first term accumulates n_{iid} samples in a stratum visited at rate π_u. The second controls the additional trajectory length needed for adequate exploration; see Theorem F.10. This is a count guarantee relative to an IID benchmark, rather than a finite-sample power guarantee.

Interpretation. Statistical testing and exploration impose separate requirements. Even for IID observations, a stratum occurring with probability π_u requires n_{iid}/π_u total samples in expectation to obtain n_{iid} observations in that stratum. Separately, a signal confined to a stratum cannot be detected before that stratum is visited; in slowly switching models this can take order τ_{mix} steps. These observations motivate the two contributions but do not establish a minimax lower bound.

This differs from subsampling the trajectory at intervals of order τ_{mix}. Such subsampling retains only about n/τ_{mix} observations. Accumulating the same in-stratum sample-size benchmark then requires, in expectation, order

$$\frac{n_{\mathrm{iid}} \tau_{\mathrm{mix}}}{\pi_u} \quad \text{time steps.}$$

In our bound the benchmark and exploration costs are additive rather than multiplicative, up to the additional factors in equation 2.

4 Time-series Causal Discovery Algorithm

We apply our test to constraint-based time series causal discovery problems, where independence tests are used to determine the presence or absence of causal relationships among variables across different times. We propose **PC-OP**

(**PC** with **O**ne **P**arental set), a constraint-based causal discovery algorithm for time series that extends the PCMCI framework (Runge et al., 2019) with a modified conditioning strategy in the momentary conditional independence (MCI) testing step. PC-OP proceeds in two main stages, with an optional intermediate refinement step.

We make several standard assumptions, namely Causal Markov condition, stationarity, and standard faithfulness assumptions. The testing theorem concerns a fixed parental conditioning set; calibration after estimating that set from the same data is not established here. In addition, we also assume the max Markov order be known, and assume no intra-slice connections or contemporaneous edges. The last two assumptions are standard in many time series causal discovery work (Runge et al., 2019).

Algorithm 1 PC with One-Parent-Set (PC-OP) Algorithm for Time Series

Input: Dataset D, significance level α, Markov Order k
Output: Estimated causal graph G
Step 1: Find parents' supersets of each variable
Initialize G a complete graph as $G_{i,j} = 1, i \neq j$ and $G_{i,j} = 0, i = j$, with nodes $\{V_i\}_{i=1}^{k}$.
Step 1 of PC algorithm (one-sided test for independence), with threshold α.
Step 2a (Optional): verify independence needs a different threshold (higher).
$\alpha' \leftarrow \frac{\alpha}{\text{the number of missing edges}}$
for each pair of variables (X, Y) not adjacent in G **do**
    P $\leftarrow$ Parent set of Y in G, excluding X
    test $X \perp\!\!\!\perp Y \mid P$ is indeed true
end for
Step 2: verify dependence with two-sided tests
for each pair of variables (X, Y) adjacent in G **do**
    P $\leftarrow$ Parent set of Y in G, excluding X
    test $X \not\perp\!\!\!\perp Y \mid P$ is indeed true with α'
end for
Return G

We introduce our proposed algorithm, PC with One-Parent-Set (PC-OP), shown in Algorithm 1. It has two major steps, with an optional check for which we call PC-OP+.

Stage 1: Condition selection via the PC algorithm. In the first stage, PC-OP applies the iterative PC-stable algorithm to identify, for each variable X_j, a superset of its lagged parents $\widetilde{\mathcal{P}}(X_j^t)$. Specifically, we initialize the graph G as a complete graph. Then we use conditional independence tests to eliminate each edge, conditioning on subsets of all other variables in order of increasing sizes, and remove edges between any pairs if conditional independence holds true. We stop the procedure early if the conditioned set size is equal to the parental set size of each variable. This step is the same as the typical PC algorithm. This step is identical to the condition-selection phase of PCMCI and serves to reduce the dimensionality of subsequent conditional independence tests by removing clearly irrelevant conditioning variables. The significance threshold for this stage is set to $\alpha_{\text{PC}} = 0.2$, following the recommendation of treating it as a regularization parameter rather than a strict significance level. For completeness, we include the full algorithm in Algorithm 2 of Appendix I

In Step 2, we verify each edge is present (or missing in the optional Step 2a) by conditioning on all parents of one of the variables. To confirm an edge's absence or presence, we perform an independence test conditioning on parents of Y excluding X (i.e., testing the edge $X \to Y$).

Stage 2a (optional): independency confirmation. When the independency confirmation is desired (denoted by the "+" suffix in variant names), PC-OP performs an additional refinement step after the initial skeleton discovery. For each variable pair $(X_i^{t-\tau}, X_j^t)$ that was *not* identified as a parent-child relationship in Stage 1, a confirmatory conditional independence test is conducted. A Bonferroni-style threshold is applied across the confirmatory tests: the significance threshold for each individual test is set to $\alpha_{\text{edge}} = 1/M$, where M is the total number of non-parent edges being tested. If the test rejects the null hypothesis of conditional independence at the corrected threshold, the corresponding edge is added back to the estimated parent set.

Stage 2b: Oriented Parent Set (OP). The key innovation of PC-OP lies in its modification of the MCI testing step. In our method, each candidate edge $X_i^{t-\tau} \to Y_j^t$ is tested for conditional independence given the estimated parents of

Table 1: Empirical Type-I error under the null (nominal $\alpha = 0.05$) at two cardinalities. Cells are `rate(se)` over $R = 2000$ replicates, with the parenthesized digit the Monte-Carlo standard error in units of the last decimal place.

C	Test	Cond.	n=100	n=200	n=1000	n=2000	n=5000
3	Chi-squared	Y-parents	0.054(5)	0.055(5)	0.054(5)	0.051(5)	0.046(5)
	Chi-squared	both parents	0.062(5)	0.052(5)	0.043(5)	0.049(5)	0.052(5)
4	Chi-squared	Y-parents	0.096(7)	0.100(7)	0.057(5)	0.052(5)	0.056(5)
	Chi-squared	both parents	0.105(7)	0.099(7)	0.071(6)	0.052(5)	0.051(5)

Y_j, excluding the candidate:

$$X_i^{t-\tau} \perp\!\!\!\perp Y_j^t \mid \widetilde{\mathcal{P}}(Y_j^t) \setminus \{X_i^{t-\tau}\}. \tag{3}$$

5 Experiments

We evaluate the proposed PC-OP with 3 sets of experiments: the CI test in isolation (Section 5.1), end-to-end recovery on synthetic categorical series (Section 5.2), and the real-world CausalRivers benchmark (Section 5.3). For a more complete results and discussion, please refer to Appendix J.

Setup. Synthetic data are generated from random dynamic DAGs over d categorical variables with maximum lag L, in-degree k_{inter}, and cardinality C; conditional probability tables are drawn from a symmetric Dirichlet prior to obtain samples. The reference condition is $d = 10$, $L = 2$, $S = 1000$, $C = 2$, $k_{\text{inter}} = 2$, averaged over five seeds. All constraint-based methods share $\alpha_{\text{PC}} = 0.2$, $\alpha_{\text{MCI}} = 0.05$ and the same χ^2 test, so PCMCI and PC-OP differ *only* in the Step-2 conditioning set. "PC-OP+" denotes PC-OP with the Step-2a dependency-verification. All experiments are run on a single machine with 3.6 GHz CPU and 64 GB memory.

Baselines. We compare PC-OP against a diverse set of established time series causal discovery algorithms, spanning constraint-based, score-based, and neural approaches. The baselines include PCMCI (Runge et al., 2019), Multivariate Granger causality (Granger-MV) (Granger, 1969), DYNOTEARS (Pamfil et al., 2020), NTS-NOTEARS (Sun et al., 2021), and TCDF (Nauta et al., 2019). Granger, TCDF, NTS-NOTEARS, and DYNOTEARS do not rely on an external conditional independence test.

Evaluation Protocols. We evaluate the quality of recovered causal graphs using two standard metrics: 1) Structural Hamming Distance (SHD), which is the total number of edge additions, deletions, and reversals required to transform the learned graph into the ground-truth graph, and 2) F1 score, which is the harmonic mean of precision and recall, providing a single balanced measure of edge recovery quality. For each experimental condition, all four metrics are computed independently for each of the five random seeds. We report the mean and standard error across replications. Table 6 in Appendix summarizes the key hyperparameters used across all methods. We first compare different methods on synthetic datasets and perform ablation studies on various approaches.

5.1 Validating the CI Test in Isolation

We first verify Theorem 3.1 on a single test, using a three-node dynamic DAG in which $X \perp Y \mid Z$ holds by d-separation under the null hypothesis. The alternative adds $X^{t-1} \to Y^t$ with effect size β mixing the CPT between an X-independent ($\beta = 0$) and an X-dependent distributions. We run $R = 2000$ replicates at $\alpha = 0.05$ for $C \in \{3, 4\}$, comparing the proposed χ^2 tests under both conditioning strategies.

The chi-squared test is shown to be calibrated with sufficient samples (as seen from Table 1): at $C = 3$ it stays within Monte-Carlo error of nominal, and at $C = 4$ a mild small-sample inflation (0.100 at $n = 200$) has vanished by $n = 1000$ (0.057). Conditioning on Y-parents set shows lower or similar Type-1 errors than conditioning on both sets under the null when the sample sizes are limited. We also include power of these results in Tables 2 and 3, reporting power as a function of sample size (fixed $\beta = 0.3$) and of effect size (fixed $n = 1000$), respectively. Power increases monotonically toward 1 in both n and β for every test and conditioning strategy. Crucially, at $\beta = 0$ the measured power equals the Type-I error (≈ 0.05 for chi-squared), confirming that the alternative reduces to the null as designed. The two conditioning strategies yield near-identical calibration and power when $C = 3$ and Y-parents is slightly favored when $C = 4$, especially at lower n. This indicates the validity of one-parent-set CI test. Full details and results are included in the Appendix J.6.

Table 2: Power under the alternative as a function of sample size n, at fixed effect size $\beta = 0.3$. Cells are `rate(se)` over $R = 2000$ replicates. Comparing across the two blocks isolates the cardinality effect at fixed n.

C	Test	Cond.	n=200	n=500	n=1000	n=2000	n=5000
3	Chi-squared	Y-parents	0.128(7)	0.265(10)	0.508(11)	0.817(9)	0.990(2)
	Chi-squared	both parents	0.131(8)	0.254(10)	0.508(11)	0.832(8)	0.990(2)
4	Chi-squared	Y-parents	0.137(8)	0.196(9)	0.284(10)	0.537(11)	0.948(5)
	Chi-squared	both parents	0.132(8)	0.164(8)	0.272(10)	0.565(11)	0.951(5)

Table 3: Power under the alternative as a function of effect size β, at fixed sample size $n = 1000$. Cells are `rate(se)` over $R = 2000$ replicates.

C	Test	Cond.	β=0.0	β=0.1	β=0.2	β=0.3	β=0.5
3	Chi-squared	Y-parents	0.057(5)	0.096(7)	0.239(10)	0.531(11)	0.915(6)
	Chi-squared	both parents	0.051(5)	0.084(6)	0.248(10)	0.536(11)	0.906(7)
4	Chi-squared	Y-parents	0.063(5)	0.087(6)	0.141(8)	0.275(10)	0.699(10)
	Chi-squared	both parents	0.061(6)	0.082(6)	0.138(8)	0.279(10)	0.711(10)

5.2 Synthetic Dataset

We generate synthetic categorical time series from a structural time series model composed of two causal graph components and a set of conditional probability tables (CPTs). Due to page limit, the data-generating process can be found in appendix J.1. We systematically vary the following data-generating parameters across experimental conditions to study their effect on causal discovery performance: 1) **Number of variables** (d): We consider $d \in \{5, 10, 20, 30\}$ to assess scalability with respect to the dimensionality of the system. 2) **Sample size** (S): We test $S \in \{10, 100, 500, 1000\}$, corresponding to total time series lengths $T = S \times d$, to evaluate performance under limited and abundant data regimes. 3) **Markov order** (L): We vary $L \in \{1, 2, 3\}$ to examine how the depth of temporal dependencies affects recovery accuracy. 4) **Inter-slice in-degree** (k_{inter}): We set $k_{\text{inter}} \in \{1, 2, 3\}$ to control the density of lagged causal connections.

Comparison with Baselines. We show the the performance of our method across different data-generating parameters. Figure 2a shows F1 scores of PC-OP with all the baseline methods together with different numbers of variables, and Figure 2b shows methods with different numbers of samples. Overall, our approaches clearly perform the best in F1 scores. SHD results are included in the appendix.

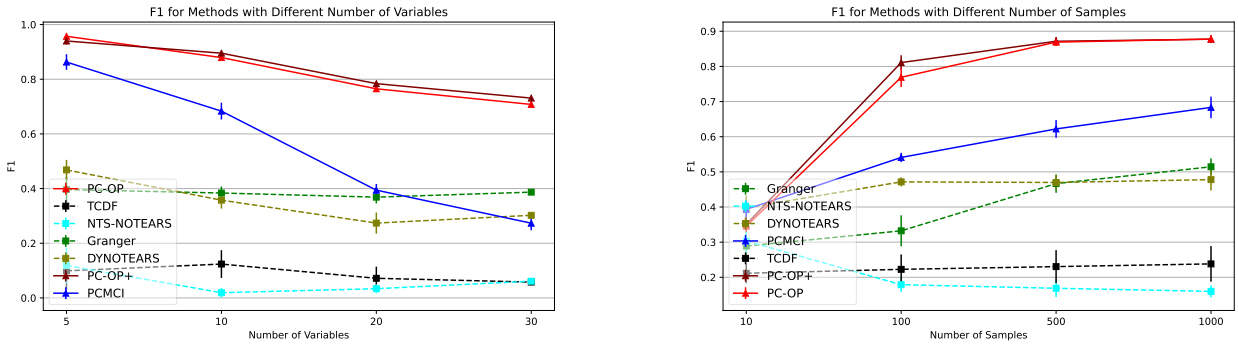
We also study if the same pattern holds along other factors that enlarges the conditioning set (Appendix J.7): 1) going from $L = 1$ to $L = 3$ reduces the performance of PC-OP by 0.054 F1 but PCMCI 0.529, 2) raising the in-degree from 2 to 3 reduces PC-OP nothing but PCMCI by 0.242, and 3) raising C from 2 to 3 reduces PC-OP 0.028 versus 0.396 for PCMCI. A threshold sweep over the full 4×4 grid of $(\alpha_{\text{PC}}, \alpha_{\text{MCI}})$ confirms the advantage is not a tuning artifact and additionally shows PC-OP to be far more stable. At $d = 30$ the SHD spread across the grid is 164 for PC-OP versus 1326 for PCMCI (Appendix J.9). Additional experiments on different methods with different number of time lags and different number of graph in-degree can also be found in Appendix J.

Oracle ablation. The comparisons above could in principle confound conditioning-set *size* with Step 1 screening differences. To isolate the mechanism we supply both methods with the *true* parent sets (Table 4). Recall is then saturated at 1.000 for both, so neither misses edges; yet PC-OP conditions on half as many variables ($|Z| = 3.9$ vs. 7.8 at $d = 20$, $L = 2$) and its false-positive rate is correspondingly lower (0.053 vs. 0.089), giving F1 of 0.808 versus 0.715. With parent sets exact and recall at ceiling, this result mostly confirms that the gain comes from the conditioning set itself.

In appendix, we include an ablation study on some variants of our proposed algorithm (J.10) and also study whether our method can work with other tests such as G-test.

5.3 Real Datasets: CausalRivers

To complement the above synthetic experiments, we test our algorithm in a real benchmark dataset for time series causal



(a) Different Numbers of Variables

(b) Different Numbers of Samples

Figure 2: Comparison of Different Methods on Synthetic Datasets

Table 4: Oracle-parent ablation ($L = 2$).Both methods receive the *true* parent sets, removing stage-1 screening as a confound. Recall is saturated for both, so the F1 gap is driven entirely by the false-positive rate, which is lower for the smaller conditioning set in all four conditions.

Config	Method	Recall	FPR	F1	$\lvert Z \rvert$
d5	Oracle-PCMCI	1.000 ± 0.000	0.113 ± 0.037	0.924 ± 0.023	7.19
	Oracle-PC-OP	1.000 ± 0.000	0.053 ± 0.013	0.962 ± 0.010	3.60
d10	Oracle-PCMCI	1.000 ± 0.000	0.090 ± 0.011	0.849 ± 0.017	7.60
	Oracle-PC-OP	1.000 ± 0.000	0.050 ± 0.003	0.909 ± 0.005	3.80
d20	Oracle-PCMCI	1.000 ± 0.000	0.089 ± 0.004	0.715 ± 0.009	7.80
	Oracle-PC-OP	1.000 ± 0.000	0.053 ± 0.006	0.808 ± 0.019	3.90
d30	Oracle-PCMCI	1.000 ± 0.000	0.072 ± 0.003	0.666 ± 0.009	7.87
	Oracle-PC-OP	1.000 ± 0.000	0.049 ± 0.002	0.746 ± 0.008	3.93

discovery with graph structures available. CausalRivers (Stein et al., 2025) features an extensive dataset on river discharge that covers the complete eastern German territory (666 measurement stations) and the state of Bavaria (494 measurement stations). It spans the years 2019 to 2023 with a 15-minute temporal resolution. These average daily discharges of rivers in the upper Danube basin are made available by the Bavarian Environmental Agency. We use the benchmark datasets with default setting (mainly 6 hour resolution) and include PCMCI as comparison since only PCMCI algorithm is competitive in the discrete variable setting. Since the data is continuous, we discretized the data first (using two bins), and, for completeness, we also include a version of these algorithms using partial correlation as independence tests. We report the average AUROC performance of different methods across 11 types of graphs. As one can seen from Table 5, both PC-OP and PC-OP+ outperform PCMCI with both independence test settings. Full detailed results per type is included in Appendix J.11.

Method	Ind. Test	Discretized	AUROC Mean
PC-OP	chi-square	True	**0.670**
PC-OP+	chi-square	True	0.663
PCMCI	chi-square	True	0.651
PC-OP	parcorr	False	**0.686**
PC-OP+	parcorr	False	0.682
PCMCI	parcorr	False	0.672

Table 5: Average AUROC of Different Methods across 11 sub-types in CausalRivers benchmark.

6 Conclusions

Under the stated mixing and parent-conditioning assumptions, we show that even with dependence left in one of the variables, chi-square statistic converges to the chi square distribution under the null hypothesis. Similarly, it diverges under the alternative hypothesis. This allowed us to develop a causal discovery algorithm, with one-parent-set chi-square tests, that has been shown to have higher statistical power than the baseline PCMCI algorithm. Empirical results validate the performance of the new tests and the superior performance of our algorithms. Future work could study and fully justify the usage of other types of independence tests with dependence samples.

USE OF AI TOOLS

In this work, we used generative AI tools for assisting in the writing of proofs. The key idea that dependence in one of the variables should not affect independence testing under the stated parent-conditioning and mixing assumptions was developed by the authors without AI assistance. The initial proof layout and core ideas, including the shortfall argument, belong to the authors. We used AI to supply standard machinery from probability theory, such as martingale central limit theorems and concentration inequalities, and to assemble the write-up, and we verified the derivations manually after several rounds of revisions. We have not used generative AI tools for: generating synthetic data sets, designing or providing feedback on experimental methodology or experiments, implementing the empirical methods, supporting qualitative and thematic data analysis.

Additionally, we used generative AI's assistance in drafting parts of the research paper and editing to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Charles K. Assaad, Emilie Devijver, and Éric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022. doi: 10.1613/jair.1.13428. URL https://www.jair.org/index.php/jair/article/view/13428.

Henry C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*. Number 112 in Mathematical Centre Tracts. Mathematisch Centrum, Amsterdam, 1979.

Yvonne MM Bishop, Stephen E Fienberg, and Paul W Holland. *Discrete multivariate analysis theory and practice*. Springer, 1975.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2357–2369. PMLR, 2022. URL https://proceedings.mlr.press/v162/budhathoki22a.html.

P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 31(2):393–427, 1995.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series. *Advances in Neural Information Processing Systems*, 36:46624–46657, 2023.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in neural information processing systems*, 33:12615–12625, 2020.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00524/full.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2023.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

Michaela Hardt, William Roy Orchard, Patrick Blöbaum, Elke Kirschbaum, and Shiva Kasiviswanathan. The PetShop dataset—finding causes of performance issues across microservices. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 957–978. PMLR, 2024. URL https://proceedings.mlr.press/v236/hardt24a.html.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://www.nature.com/articles/s41597-019-0103-9.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Søren Wengel Mogensen, Karin Rathsman, and Per Nilsson. Causal discovery in a complex industrial system: A time series benchmark. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1218–1236. PMLR, 2024. URL https://proceedings.mlr.press/v236/mogensen24a.html.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161. URL https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pp. 1388–1397. Pmlr, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11), 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2000. ISBN 9780262194402. URL https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers–scaling up benchmarking of causal discovery for real-world time-series. *arXiv preprint arXiv:2503.17452*, 2025.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

V. A. Volkonskiĭ and Y. A. Rozanov. Some limit theorems for random functions, I. *Theory of Probability and Its Applications*, 4(2):178–197, 1959.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

A PROOFS

We study conditional independence testing for time-indexed categorical data in which the parent/design process may be dependent over time under the stationary mixing assumption below and may exhibit feedback from past outcomes. Fix a lag $\delta \geq 1$ and a discrete label U_t encoding the values of all screened parents of Y_t other than the candidate parent $X_{t-\delta}$; by construction U_t depends only on variables realized strictly before time t. Within a fixed stratum $u \in \mathcal{U}$, we analyze Pearson's chi-square test applied to the contingency table of $(X_{t-\delta}, Y_t)$ among times with $U_t = u$. The key technical point is to avoid conditioning on the entire realized parent path (which can fail under feedback); instead we derive the classical central χ^2 null calibration using a martingale CLT in Pearson tangent coordinates. We also give an edge-present divergence result with an explicit signal rate, deduce consistency of the resulting stratumwise conditional-independence test under fixed alternatives, and discuss sufficient run-length conditions for table regularity via blocking in dependent designs.

After coming up with the main idea and the proof outline, some technical details from probability theory are put together by an LLM, which was verified by the authors.

B SETUP AND STATISTIC

Fix finite alphabets $[r] = \{1, \ldots, r\}$ and $[c] = \{1, \ldots, c\}$ with $r, c \geq 2$, and a fixed lag $\delta \geq 1$. For each $t \geq 1$ we observe

$$X_{t-\delta} \in [r], \qquad U_t \in \mathcal{U}, \qquad Y_t \in [c],$$

where $\mathcal{U}$ is finite.

Standing mixing and support assumptions. Throughout, the process $(X_{t-\delta}, U_t, Y_t)$ is strictly stationary and rapidly mixing in the sense of Assumption F.8 below (geometric β-mixing). The model, alphabets, lag, and screening set are fixed. For every tested stratum u, assume $\pi_{au} := \mathbb{P}(U_t = u, X_{t-\delta} = a) > 0$ for every row a and $\gamma_{ub} := \mathbb{P}(U_t = u, Y_t = b) > 0$ for every column b; write $\pi_u := \sum_a \pi_{au}$.

Boundary convention. Because $X_{t-\delta}$ appears in the table construction, all arguments are implicitly restricted to times $t \geq \delta + 1$. Equivalently, one may absorb the finitely many boundary times $t \leq \delta$ into the deterministic used-index set $\mathcal{S}$ and never select them. This has no asymptotic effect but avoids repeatedly carrying initial-condition notation in the sequel.

Screening interpretation. U_t is a discrete encoding of the values of *all true parents of* Y_t *except* $X_{t-\delta}$, using only variables realized strictly before time t (no contemporaneous conditioning with Y_t). We test within stratum values $u \in \mathcal{U}$.

Used indices. Let $\mathcal{S} \subseteq \{1, \ldots, n\}$ be a deterministic set of used indices (e.g. the indices inside retained blocks from the blocking scheme described in Section F). The set $\mathcal{S}$ is fixed before any data are observed; it does not depend on the realized process. This determinism guarantees that $\mathbb{1}\{t \in \mathcal{S}\}$ is trivially $\mathcal{F}_{t-1}$-measurable for every t, which is needed for the martingale construction in Section D. For asymptotics write this set as $\mathcal{S}_n$, with deterministic $s_n := |\mathcal{S}_n| \to \infty$, suppressing the subscript on $\mathcal{S}_n$ below. When $\mathcal{S}_n$ changes with n, every martingale below is understood as a triangular array: for each terminal n, keep $\mathcal{S}_n$ fixed while forming partial sums over $t \leq n$.

Fix $u \in \mathcal{U}$ and define

$$I(u) := \{t \in \mathcal{S} : U_t = u\}, \qquad N_u := |I(u)|.$$

Note that $\mathcal{S}$ and $I(u)$ play different roles: $\mathcal{S}$ is a deterministic filter removing time points for mixing/blocking reasons, while $I(u)$ is a random subset of $\mathcal{S}$ selecting those time points where the stratum label equals u. The contingency table is built from indices that survive *both* filters.

Define contingency table counts:

$$N_{ab} := \sum_{t \in I(u)} \mathbb{1}\{X_{t-\delta} = a,\ Y_t = b\}, \qquad N_{a\cdot} := \sum_{b=1}^{c} N_{ab}, \qquad N_{\cdot b} := \sum_{a=1}^{r} N_{ab}, \qquad \widehat{q}_b := \frac{N_{\cdot b}}{N_u}.$$

Pearson's statistic is

$$\chi^2_n(u) = \sum_{a=1}^{r} \sum_{b=1}^{c} \frac{(N_{ab} - N_{a\cdot}\widehat{q}_b)^2}{N_{a\cdot}\widehat{q}_b}. \tag{4}$$

Assumption B.1 (Regularity for asymptotics)**.** Let $m_n := \min_{1\le a\le r} N_{a\cdot}$. Assume

$$m_n \xrightarrow{p} \infty, \qquad \min_{1\le a\le r} \frac{N_{a\cdot}}{N_u} \ge \varepsilon_{\mathrm{row}} > 0 \text{ with probability } 1-o(1), \qquad \min_{1\le b\le c} \widehat{q}_b \ge c_0 > 0 \text{ with probability } 1-o(1),$$

for some constants $\varepsilon_{\mathrm{row}}, c_0 > 0$. (Necessarily $\varepsilon_{\mathrm{row}} \le 1/r$ and $c_0 \le 1/c$, since the row fractions $N_{a\cdot}/N_u$ and the column marginals $\widehat{q}_b$ each sum to 1.)

Count limits under mixing. Let $V_t := (X_{t-\delta}, U_t, Y_t)$ and let h be any bounded function. Geometric mixing implies $\sum_{\ell\ge 1} |\operatorname{Cov}(h(V_0), h(V_\ell))| < \infty$. There are at most s_n pairs of retained indices with any given positive time difference, so stationarity gives

$$\operatorname{Var}\left(\frac{1}{s_n}\sum_{t\in\mathcal{S}_n} h(V_t)\right) \le \frac{1}{s_n}\left(\operatorname{Var}(h(V_0)) + 2\sum_{\ell\ge 1} |\operatorname{Cov}(h(V_0), h(V_\ell))|\right).$$

Applying Chebyshev's inequality to the row, stratum, and column indicators yields

$$\frac{N_{a\cdot}}{s_n} = \pi_{au} + O_p(s_n^{-1/2}), \qquad \frac{N_u}{s_n} = \pi_u + O_p(s_n^{-1/2}), \qquad \frac{N_{\cdot b}}{s_n} = \gamma_{ub} + O_p(s_n^{-1/2}). \tag{5}$$

Positive support therefore gives Assumption B.1, with row fractions converging to π_{au}/π_u and column fractions to γ_{ub}/π_u. The same argument applies jointly to finitely many strata. Statistics may be defined arbitrarily when a denominator is zero; this event has probability tending to zero.

C SCM ASSUMPTIONS AND HYPOTHESES

We work with a filtration $(\mathcal{F}_t)_{t\ge 0}$ representing information up to time t, and we take

$$\mathcal{F}_t := \sigma\big(\{X_s, U_s, Y_s : 1 \le s \le t\}\big)$$

(possibly enlarged to include any additional observed lagged variables used to form U_t; however, this is a modeling choice, not an automatic monotonicity statement. If one works with an enlarged filtration, then Assumption C.3 below must be understood relative to that enlarged filtration, since the requirement $\eta_t \perp \mathcal{F}_{t-1}$ becomes strictly stronger when more past information is included.)

C.1 LAG-ONLY SCREENING LABEL AND FRESH INNOVATIONS

Assumption C.1 (Lag-only screening label)**.** The stratum label U_t is a measurable function of variables realized strictly before time t. Equivalently, U_t is $\mathcal{F}_{t-1}$-measurable for each t. We also assume $\delta \ge 1$ (so $X_{t-\delta} \in \mathcal{F}_{t-1}$).

Lemma C.2 (Predictability (derived, not assumed))**.** *Under Assumption C.1, for each t the pair $(U_t, X_{t-\delta})$ is $\mathcal{F}_{t-1}$-measurable. Consequently, for any fixed $u \in \mathcal{U}$ and $a \in [r]$,*

$$\mathbb{1}\{U_t = u,\ X_{t-\delta} = a\} \in \mathcal{F}_{t-1}.$$

Proof. By Assumption C.1, $U_t \in \mathcal{F}_{t-1}$. Since $\delta \ge 1$, $t-\delta \le t-1$, so $X_{t-\delta} \in \mathcal{F}_{t-1}$. Hence the pair is $\mathcal{F}_{t-1}$-measurable. □

Assumption C.3 (Time-homogeneous fresh innovations for Y)**.** There exist i.i.d. innovations $(\eta_t)_{t\ge 1}$ and a measurable map g such that for all t,

$$Y_t = g(U_t, X_{t-\delta}, \eta_t), \qquad \eta_t \ \perp\ \mathcal{F}_{t-1}.$$

Assumption C.3 implies that, given $\mathcal{F}_{t-1}$, the only randomness in Y_t comes from a fresh draw η_t. Beyond the standing mixing assumption, no restriction is imposed on the temporal dependence of $(U_t, X_{t-\delta})$; in particular, feedback from past Y into future (U, X) is allowed.

C.2 Structural null and structural alternative (within a fixed stratum)

Fix a stratum value $u \in \mathcal{U}$. We test

$$H_0: \quad Y_t \perp X_{t-\delta} \mid (U_t = u), \qquad H_1: \quad Y_t \not\perp X_{t-\delta} \mid (U_t = u).$$

Assumption C.4 (Structural null in stratum u (no edge))**.** There exists a measurable map g_0 such that for every $a \in [r]$,

$$g(u, a, \eta) = g_0(u, \eta) \quad \text{for } \eta \sim \eta_1 \text{ almost surely.}$$

Equivalently, whenever $U_t = u$ we have $Y_t = g_0(u, \eta_t)$ almost surely.

Lemma C.5 (Null pmf as a derived object)**.** *Assume Assumptions C.1, C.3, and C.4. Define*

$$q_b := \mathbb{P}\big(g_0(u, \eta_1) = b\big), \qquad b \in [c].$$

Then $q = (q_1, \ldots, q_c)$ is a pmf, and for every t and $b \in [c]$,

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = q_b \quad \textit{on the event } \{U_t = u\}.$$

Equivalently, for all $a \in [r]$,

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}, U_t = u, X_{t-\delta} = a) = q_b.$$

Remark C.6 (On the "on the event" language). Recall that $\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1})$ is a random variable (a function of ω), defined as a conditional expectation via the Radon–Nikodym theorem (see, e.g., (Durrett, 2019, Ch. 5), (Williams, 1991, Ch. 9)). The statement that this random variable equals q_b *on the event* $\{U_t = u\}$ means: for those ω where $U_t(\omega) = u$, we have $\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1})(\omega) = q_b$. Off the event $\{U_t = u\}$, the conditional probability may be anything—we make no claim there.

This is *not* the same as saying "conditioned on $U_t = u$," which would modify the conditioning sigma-field to $\sigma(\mathcal{F}_{t-1}, U_t)$. However, since U_t is $\mathcal{F}_{t-1}$-measurable (Assumption C.1), it is already "known" given $\mathcal{F}_{t-1}$, so the two formulations are equivalent here. We prefer the "on the event" formulation because it keeps the conditioning sigma-field as $\mathcal{F}_{t-1}$, which is what the martingale CLT requires.

Proof of Lemma C.5. Fix t and work on the event $\{U_t = u\}$. By Lemma C.2, U_t and $X_{t-\delta}$ are $\mathcal{F}_{t-1}$-measurable. By Assumption C.3, $\eta_t \perp \mathcal{F}_{t-1}$. By Assumption C.4, on $\{U_t = u\}$ we have $Y_t = g_0(u, \eta_t)$ and the law of η_t equals that of η_1. Hence

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = \mathbb{P}(g_0(u, \eta_t) = b \mid \mathcal{F}_{t-1}) = \mathbb{P}(g_0(u, \eta_1) = b) = q_b,$$

where the second equality uses $\eta_t \perp \mathcal{F}_{t-1}$ together with the fact that u is the realized value of the $\mathcal{F}_{t-1}$-measurable random variable U_t on this event (see Remark C.6). Conditioning further on $X_{t-\delta} = a$ does not change the conclusion because $X_{t-\delta}$ is already $\mathcal{F}_{t-1}$-measurable. □

Assumption C.7 (Structural alternative in stratum u (row-dependent response))**.** For each $a \in [r]$ define

$$p_{ab} := \mathbb{P}\big(g(u, a, \eta_1) = b\big), \qquad b \in [c],$$

and assume each p_a is a pmf with entries in $[0, 1]$. For each column b, assume $\sum_a p_{ab} > 0$, as follows from the standing positive column-support assumption. In addition, *the alternative is genuinely row-dependent*: there exist $a \neq a'$ such that

$$p_a \neq p_{a'}.$$

Equivalently, the family $\{p_a : a \in [r]\}$ is not constant across rows.

Lemma C.8 (Alternative conditional pmf as a derived object)**.** *Assume Assumptions C.1, C.3, and C.7. Then for every t and $b \in [c]$,*

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = p_{X_{t-\delta}, b} \quad \textit{on the event } \{U_t = u\},$$

and in particular

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}, U_t = u, X_{t-\delta} = a) = p_{ab}.$$

Proof. On $\{U_t = u\}$, we have $Y_t = g(u, X_{t-\delta}, \eta_t)$ by Assumption C.3. By Lemma C.2, $(u, X_{t-\delta})$ is $\mathcal{F}_{t-1}$-measurable, and $\eta_t \perp \mathcal{F}_{t-1}$. Therefore

$$\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = \mathbb{P}\big(g(u, X_{t-\delta}, \eta_t) = b \mid \mathcal{F}_{t-1}\big) = \sum_{a=1}^{r} \mathbb{1}\{X_{t-\delta} = a\}\, \mathbb{P}\big(g(u, a, \eta_1) = b\big) = p_{X_{t-\delta}, b}.$$

□

C.3 THE INDICATOR-REPAIR LEMMA

The following one-line lemma is the workhorse used in every martingale construction below (Lemmas D.4, D.12, E.4, and E.8). It packages the recurring step "the centering identity holds only on the event $\{U_t = u\}$, and the indicator W vanishes off this event, so the product is a global martingale difference."

Lemma C.9 (Indicator-repair)**.** *Let W be a bounded $\mathcal{F}_{t-1}$-measurable random variable, let E be an event, and assume $\{W \neq 0\} \subseteq E$. Let Z be an integrable random variable for which the centering*

$$\mathbb{E}[Z \mid \mathcal{F}_{t-1}](\omega) = c(\omega) \quad \textit{for almost every } \omega \in E$$

holds, where c is bounded and $\mathcal{F}_{t-1}$-measurable. Then

$$\mathbb{E}\big[W\,(Z-c)\,\big|\,\mathcal{F}_{t-1}\big] = 0 \quad \textit{a.s.}$$

Proof. By the pull-out property ((Williams, 1991, Theorem 9.7(e))), $\mathbb{E}[W(Z-c) \mid \mathcal{F}_{t-1}] = W \cdot \mathbb{E}[Z-c \mid \mathcal{F}_{t-1}]$. On E the second factor vanishes by hypothesis; off E the first factor vanishes since $W = 0$. Hence the product is zero a.s. □

D NULL CALIBRATION: CENTRAL CHI-SQUARE VIA MARTINGALE GEOMETRY

Theorem D.1 (Central χ^2 calibration under the structural null)**.** *In the standing stationary mixing setting, assume Assumptions C.1, C.3, C.4, and B.1. Assume moreover that the null pmf induced by Lemma C.5 satisfies $q_b > 0$ for all b. Then*

$$\chi^2_n(u) \Rightarrow \chi^2_{(r-1)(c-1)}.$$

The proof proceeds in four steps:

1. Introduce Pearson tangent coordinates and establish a martingale CLT under mixing.
2. Derive an exact projection identity relating a denominator-fixed Pearson statistic to a quadratic form in the studentized martingale.
3. Bound the error from replacing random row counts by their stationary expectations, and obtain the χ^2 limit.
4. Show that replacing the true denominators q by estimated denominators $\widehat{q}$ is asymptotically negligible.

D.1 STEP 1: MARTINGALE CLT IN PEARSON TANGENT COORDINATES

Let $q = (q_1, \ldots, q_c)$ be as in Lemma C.5.

One-hot encoding. Let $e_b \in \mathbb{R}^c$ denote the b-th standard basis vector and define

$$E_t := e_{Y_t} \in \mathbb{R}^c.$$

E_t is a random vector: if $Y_t(\omega) = b$, then $E_t(\omega)$ has a 1 in position b and 0 elsewhere. (Equivalently, $(E_t)_b = \mathbb{1}\{Y_t = b\}$.) The centered observation $E_t - q$ satisfies $\mathbf{1}^\top(E_t - q) = 0$, since $\sum_b (E_t)_b = 1$ and $\sum_b q_b = 1$. Hence $E_t - q$ lies in the hyperplane $\{v \in \mathbb{R}^c : \mathbf{1}^\top v = 0\}$, the tangent space of the simplex.

Pearson weighting. Define $D_q := \mathrm{diag}(q_1, \ldots, q_c)$ and $s := \sqrt{q} \in \mathbb{R}^c$ (i.e., $s_b = \sqrt{q_b}$). Multiplying by $D_q^{-1/2} = \mathrm{diag}(q_1^{-1/2}, \ldots, q_c^{-1/2})$ rescales each coordinate by $1/\sqrt{q_b}$, amplifying rare categories and compressing common ones. This is the weighting that appears in Pearson's chi-square statistic.

After this rescaling, the constraint changes direction: $E_t - q$ was orthogonal to $\mathbf{1}$, while $D_q^{-1/2}(E_t - q)$ is orthogonal to $s = \sqrt{q}$. To see this, compute directly:

$$s^\top D_q^{-1/2}(E_t - q) = \sum_{b=1}^{c} \sqrt{q_b} \cdot \frac{(E_t)_b - q_b}{\sqrt{q_b}} = \sum_{b=1}^{c} \big((E_t)_b - q_b\big) = 0.$$

More generally, applying a linear map A to a vector orthogonal to w produces a vector orthogonal to $A^{-\top}w$; here $A = D_q^{-1/2}$ is symmetric, so $A^{-\top} = D_q^{1/2}$, and $D_q^{1/2}\mathbf{1} = s$.

Projection to full-rank coordinates. Let $H \in \mathbb{R}^{(c-1)\times c}$ be any matrix with orthonormal rows spanning $s^\perp$ (so $Hs = 0$ and $HH^\top = I_{c-1}$). Since $D_q^{-1/2}(E_t - q)$ lies in $s^\perp \subset \mathbb{R}^c$, the matrix H acts as an isometry from this $(c-1)$-dimensional subspace to $\mathbb{R}^{c-1}$, preserving all norms and inner products. No information is lost because $D_q^{-1/2}(E_t - q)$ has zero component along s. The specific choice of H does not matter: different choices correspond to rotations within $\mathbb{R}^{c-1}$, and the Pearson statistic (being a sum of squared norms) is rotation-invariant.

Remark D.2 (Why project at all?). Without H, we would work in $\mathbb{R}^c$ with the rank-deficient covariance $I - ss^\top$, which is not invertible. The matrix H drops the zero-eigenvalue direction so that the resulting covariance is the full-rank matrix I_{c-1}, enabling self-normalization in the martingale CLT.

Whitened tangent increment. Define

$$\xi_t := HD_q^{-1/2}(E_t - q) \in \mathbb{R}^{c-1}.$$

This maps the centered observation through three stages:

$$\underbrace{E_t - q}_{\text{centered in } \mathbb{R}^c} \xrightarrow{D_q^{-1/2}} \underbrace{\text{Pearson-weighted}}_{\text{in } s^\perp \subset \mathbb{R}^c} \xrightarrow{H} \underbrace{\xi_t}_{\text{in } \mathbb{R}^{c-1}}.$$

Design indicator. For each row $a \in [r]$, define

$$W_t(a,u) := \mathbb{1}\{t \in \mathcal{S},\ U_t = u,\ X_{t-\delta} = a\}.$$

Since $\mathcal{S}$ is deterministic and $(U_t, X_{t-\delta})$ is $\mathcal{F}_{t-1}$-measurable by Lemma C.2, we have $W_t(a,u) \in \mathcal{F}_{t-1}$.

Remark D.3 (Notational convention). We use W_t for the indicator to avoid confusion with the contrast matrix $H \in \mathbb{R}^{(c-1)\times c}$.

Row-wise martingale sums. Define

$$M_{a,n} := \sum_{t=1}^{n} W_t(a,u)\,\xi_t \ \in\ \mathbb{R}^{c-1}, \qquad M_n := (M_{1,n}^\top, \ldots, M_{r,n}^\top)^\top \in \mathbb{R}^{r(c-1)}.$$

Lemma D.4 (Martingale property of M_n)**.** *Under Assumptions C.1, C.3, and C.4, M_n is a martingale with respect to $(\mathcal{F}_n)$.*

Proof. We must show $\mathbb{E}[W_t(a,u)\,\xi_t \mid \mathcal{F}_{t-1}] = 0$ for each t and a. The vector $\xi_t = HD_q^{-1/2}(E_t - q)$ has, on $\{U_t = u\}$, conditional mean

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = HD_q^{-1/2}\big(\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] - q\big) = 0$$

by Lemma C.5. The indicator $W_t(a,u) = \mathbb{1}\{t \in \mathcal{S}, U_t = u, X_{t-\delta} = a\}$ is $\mathcal{F}_{t-1}$-measurable (Lemma C.2) and supported on $\{U_t = u\}$. Applying the indicator-repair lemma (Lemma C.9) componentwise to ξ_t with $W = W_t(a,u)$, $E = \{U_t = u\}$, and $c = 0$ gives $\mathbb{E}[W_t(a,u)\,\xi_t \mid \mathcal{F}_{t-1}] = 0$ a.s., as required. □

Lemma D.5 (Conditional covariance of ξ_t)**.** *On the event $\{U_t = u\}$, under the null,*

$$\mathbb{E}[\xi_t \xi_t^\top \mid \mathcal{F}_{t-1}] = I_{c-1}.$$

Proof. Since $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ on $\{U_t = u\}$ (Lemma D.4), the conditional second moment equals the conditional covariance on this event. We now compute this conditional covariance by restricting to $\omega \in \{U_t = u\}$ throughout.

Conditional covariance of E_t. The (b,b')-entry of $\mathbb{E}[E_t E_t^\top \mid \mathcal{F}_{t-1}]$ is $\mathbb{E}[(E_t)_b (E_t)_{b'} \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{1}\{Y_t = b\}\mathbb{1}\{Y_t = b'\} \mid \mathcal{F}_{t-1}]$. For $b \neq b'$, Y_t cannot equal both b and b', so this is 0. For $b = b'$, restricting to the event $\{U_t = u\}$ and applying Lemma C.5, this equals $\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = q_b$. Hence, on $\{U_t = u\}$,

$$\mathbb{E}[E_t E_t^\top \mid \mathcal{F}_{t-1}] = D_q, \qquad \mathrm{Cov}(E_t \mid \mathcal{F}_{t-1}) = D_q - qq^\top.$$

This is the standard multinomial covariance for a single draw with probability vector q.

Transform through $HD_q^{-1/2}$. Using $\mathrm{Cov}(AX \mid \mathcal{G}) = A\,\mathrm{Cov}(X \mid \mathcal{G})\,A^\top$ for any deterministic matrix A:

$$\mathbb{E}[\xi_t \xi_t^\top \mid \mathcal{F}_{t-1}] = HD_q^{-1/2}(D_q - qq^\top)D_q^{-1/2}H^\top.$$

Expanding: $D_q^{-1/2} D_q D_q^{-1/2} = I$ (each diagonal entry is $q_b^{-1/2} \cdot q_b \cdot q_b^{-1/2} = 1$) and $D_q^{-1/2} q = s$ (each entry is $q_b^{-1/2} \cdot q_b = \sqrt{q_b} = s_b$). Therefore $D_q^{-1/2}(D_q - qq^\top)D_q^{-1/2} = I - ss^\top$. Finally, $H(I - ss^\top)H^\top = HH^\top - Hss^\top H^\top = I_{c-1} - 0 = I_{c-1}$, since $Hs = 0$ (the rows of H are orthogonal to s) and $HH^\top = I_{c-1}$ (the rows are orthonormal). □

Predictable quadratic variation. Define the increment $\Delta M_t := M_t - M_{t-1}$ (i.e., the change from time $t-1$ to time t). Explicitly,

$$\Delta M_t = \big(W_t(1,u)\,\xi_t^\top,\ \ldots,\ W_t(r,u)\,\xi_t^\top\big)^\top \in \mathbb{R}^{r(c-1)}.$$

Since $M_n \in \mathbb{R}^{r(c-1)}$ is organized as r sub-vectors of dimension $c-1$, the matrix $\mathbb{E}[\Delta M_t \Delta M_t^\top \mid \mathcal{F}_{t-1}]$ has a natural $r \times r$ block structure, with each block of size $(c-1)\times(c-1)$. The (a,a')-block arises from the outer product of $W_t(a,u)\,\xi_t$ and $W_t(a',u)\,\xi_t$: a scalar indicator times a $(c-1)$-dimensional vector. Concretely, the (a,a')-block of $\mathbb{E}[\Delta M_t \Delta M_t^\top \mid \mathcal{F}_{t-1}]$ is

$$\mathbb{E}\big[W_t(a,u)\,\xi_t\xi_t^\top\,W_t(a',u) \mid \mathcal{F}_{t-1}\big] = W_t(a,u)\,W_t(a',u)\,\mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_{t-1}],$$

where we pulled out both $\mathcal{F}_{t-1}$-measurable indicators. At any time t, $X_{t-\delta}$ takes exactly one value, so for $a \neq a'$ the indicators $W_t(a,u)$ and $W_t(a',u)$ cannot both be 1; their product is 0. Hence all off-diagonal blocks vanish. For the diagonal block ($a = a'$), $W_t(a,u)^2 = W_t(a,u)$ (it is 0 or 1), and on the event where $W_t(a,u) = 1$ we have $U_t = u$, so Lemma D.5 gives $\mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_{t-1}] = I_{c-1}$. Thus each diagonal block contributes $W_t(a,u)\cdot I_{c-1}$.

Summing over $t = 1,\ldots,n$, the (a,a)-block of the total predictable quadratic variation becomes $\big(\sum_{t=1}^n W_t(a,u)\big)\cdot I_{c-1} = N_{a\cdot}\cdot I_{c-1}$ (recall that all contingency-table counts are within the fixed stratum u). This gives:

$$\langle M\rangle_n := \sum_{t=1}^n \mathbb{E}[\Delta M_t\,\Delta M_t^\top \mid \mathcal{F}_{t-1}] = \mathrm{diag}(N_{1\cdot},\ldots,N_{r\cdot}) \otimes I_{c-1}. \tag{6}$$

The Kronecker structure reflects two facts: across rows of the contingency table (i.e., across values of $a \in [r]$), the variance is proportional to the number of observations $N_{a\cdot}$ in that row; within each contingency-table row, the $(c-1)\times(c-1)$ covariance block is identity because of the Pearson whitening.

Studentization. The martingale M_n grows in magnitude with n; to obtain a distributional limit we normalize by the predictable quadratic variation. Define the *studentized* vector

$$U_n := \langle M\rangle_n^{-1/2} M_n = (U_{1,n}^\top,\ldots,U_{r,n}^\top)^\top, \qquad U_{a,n} = \frac{M_{a,n}}{\sqrt{N_{a\cdot}}}. \tag{7}$$

This is the martingale analogue of a t-statistic: each row's accumulated residual is divided by the square root of the number of observations in that row. Crucially, $N_{a\cdot}$ is itself random (it depends on which rows the process visited), so this is *self-normalization*—division by a data-dependent scale.

Martingale convergence. The increments ΔM_t are uniformly bounded, since ξ_t is bounded and at most one row indicator is nonzero at each time. They vanish off $\mathcal{S}_n$. By equation 5,

$$\frac{\langle M\rangle_n}{s_n} = \mathrm{diag}\left(\frac{N_{1\cdot}}{s_n},\ldots,\frac{N_{r\cdot}}{s_n}\right)\otimes I_{c-1} \xrightarrow{p} V_u := \mathrm{diag}(\pi_{1u},\ldots,\pi_{ru})\otimes I_{c-1}.$$

The limit is deterministic and positive definite. Lemma G.1 therefore gives $M_n/\sqrt{s_n} \Rightarrow \mathcal{N}(0,V_u)$ and

$$U_n \Rightarrow Z \sim \mathcal{N}(0, I_{r(c-1)}). \tag{8}$$

Remark D.6 (Role of mixing). Mixing supplies the stationary row-count limits. The martingale property still allows past responses to affect future row assignments; independence of the responses from the entire realized row-assignment sequence is not assumed.

D.2 Step 2: exact projection identity for the denominator-fixed statistic

Define the denominator-fixed Pearson statistic (using the true null pmf q in the denominator instead of the estimated $\widehat{q}$):

$$\chi^2_{n,q}(u) := \sum_{a=1}^r \sum_{b=1}^c \frac{(N_{ab} - N_{a\cdot}\widehat{q}_b)^2}{N_{a\cdot}q_b}.$$

Define the row weight vector (the subscript n indicates dependence on sample size, not dimension):

$$w_n := \big(\sqrt{N_{1\cdot}/N_u},\ldots,\sqrt{N_{r\cdot}/N_u}\big)^\top \in \mathbb{R}^r, \qquad \|w_n\|_2 = 1, \tag{9}$$

and the projection matrices

$$P_n := (w_n w_n^\top)\otimes I_{c-1}, \qquad A_n := I_{r(c-1)} - P_n. \tag{10}$$

Then A_n is a symmetric idempotent (orthogonal projection) matrix of rank $(r-1)(c-1)$ for every n. Note that w_n, P_n, and A_n are all random (they depend on the row counts which change with n) and measurable with respect to the row counts.

Lemma D.7 (Exact projection identity)**.** *With U_n defined in equation 7,*

$$\chi^2_{n,q}(u) = U_n^\top A_n U_n.$$

Proof. This is a purely algebraic identity (no probabilistic inputs).

Setup. Let $N_{a\bullet} := (N_{a1}, \dots, N_{ac})^\top \in \mathbb{R}^c$ be the vector of cell counts in row a, and let $\delta := \widehat{q} - q$. Then

$$N_{ab} - N_{a\cdot}\widehat{q}_b = (N_{ab} - N_{a\cdot}q_b) - N_{a\cdot}\delta_b.$$

Expand the statistic. Squaring, dividing by $N_{a\cdot}q_b$, and summing:

$$\chi^2_{n,q}(u) = \underbrace{\sum_{a=1}^{r}\sum_{b=1}^{c}\frac{(N_{ab} - N_{a\cdot}q_b)^2}{N_{a\cdot}q_b}}_{\text{Term 1}} \; \underbrace{- \, 2\sum_{a=1}^{r}\sum_{b=1}^{c}\frac{(N_{ab} - N_{a\cdot}q_b)\, N_{a\cdot}\delta_b}{N_{a\cdot}q_b}}_{\text{Term 2}}$$
$$+ \underbrace{\sum_{a=1}^{r}\sum_{b=1}^{c}\frac{N_{a\cdot}^2\delta_b^2}{N_{a\cdot}q_b}}_{\text{Term 3}}.$$

Define

$$R_{a,n} := \frac{1}{\sqrt{N_{a\cdot}}} D_q^{-1/2}(N_{a\bullet} - N_{a\cdot}q) \in \mathbb{R}^c.$$

Then **Term 1** $= \sum_a \|R_{a,n}\|_2^2$ by definition (the b-th component of $R_{a,n}$ is $(N_{ab} - N_{a\cdot}q_b)/(\sqrt{N_{a\cdot}}\sqrt{q_b})$).

For **Term 3**: $\sum_a N_{a\cdot}\sum_b \delta_b^2/q_b = N_u \|D_q^{-1/2}\delta\|_2^2$, using $\sum_a N_{a\cdot} = N_u$.

For **Term 2**: cancel $N_{a\cdot}$ and swap the order of summation:

$$\text{Term 2} = 2\sum_{b=1}^{c}\frac{\delta_b}{q_b}\sum_{a=1}^{r}(N_{ab} - N_{a\cdot}q_b) = 2\sum_{b=1}^{c}\frac{\delta_b}{q_b}(N_{\cdot b} - q_b N_u) = 2\sum_{b=1}^{c}\frac{\delta_b}{q_b}N_u\delta_b = 2N_u\|D_q^{-1/2}\delta\|_2^2,$$

where we used $\sum_a N_{ab} = N_{\cdot b}$ and $N_{\cdot b} - q_b N_u = N_u(\widehat{q}_b - q_b) = N_u\delta_b$.

Combining: $\chi^2_{n,q}(u) = \sum_a \|R_{a,n}\|_2^2 - 2N_u\|D_q^{-1/2}\delta\|_2^2 + N_u\|D_q^{-1/2}\delta\|_2^2 = \sum_a \|R_{a,n}\|_2^2 - N_u\|D_q^{-1/2}\delta\|_2^2$.

Using $\sum_a N_{a\bullet} = N_u\widehat{q}$ and $\sum_a N_{a\cdot} = N_u$, we obtain

$$\delta = \widehat{q} - q = \frac{1}{N_u}\sum_{a=1}^{r}(N_{a\bullet} - N_{a\cdot}q).$$

Rewrite as a quadratic form. Since $\delta = \frac{1}{N_u}\sum_a(N_{a\bullet} - N_{a\cdot}q)$, we have $D_q^{-1/2}\delta = \frac{1}{\sqrt{N_u}}\sum_{a=1}^{r} w_{n,a}\, R_{a,n}$ (using $w_{n,a} = \sqrt{N_{a\cdot}/N_u}$), so

$$N_u\|D_q^{-1/2}\delta\|_2^2 = \Big\|\sum_a w_{n,a}\, R_{a,n}\Big\|_2^2.$$

Now observe that each $R_{a,n}$ lies in $s^\perp$ (because $s^\top R_{a,n} = (D_q^{1/2}\mathbf{1})^\top D_q^{-1/2}(N_{a\bullet} - N_{a\cdot}q)/\sqrt{N_{a\cdot}} = \mathbf{1}^\top(N_{a\bullet} - N_{a\cdot}q)/\sqrt{N_{a\cdot}} = 0$).

By the definitions of $U_{a,n}$, $M_{a,n}$, and ξ_t,

$$U_{a,n} = \frac{1}{\sqrt{N_{a\cdot}}}\sum_{t=1}^{n} W_t(a,u)\,\xi_t = \frac{1}{\sqrt{N_{a\cdot}}} H D_q^{-1/2}\sum_{t=1}^{n} W_t(a,u)(E_t - q) = HR_{a,n},$$

since $\sum_t W_t(a,u)E_t = N_{a\bullet}$ and $\sum_t W_t(a,u) = N_{a\cdot}$.

Since H is an isometry on $s^\perp$ and $U_{a,n} = HR_{a,n}$:

$$\|R_{a,n}\|_2 = \|U_{a,n}\|_2$$

Since each $R_{a,n} \in s^{\perp}$, their weighted sum $\sum_a w_{n,a} R_{a,n}$ also lies in $s^{\perp}$. By linearity of H,

$$\sum_a w_{n,a} U_{a,n} = \sum_a w_{n,a} H R_{a,n} = H\Big(\sum_a w_{n,a} R_{a,n}\Big).$$

Because H is an isometry on $s^{\perp}$, it preserves the norm of this sum:

$$\Big\| \sum_a w_{n,a} R_{a,n} \Big\|_2 = \Big\| \sum_a w_{n,a} U_{a,n} \Big\|_2.$$

Lemma D.8 (Block projection identity)**.** *Let* $w = (w_1, \dots, w_r)^\top \in \mathbb{R}^r$ *satisfy* $\|w\|_2 = 1$*, and let*

$$x = (x_1^\top, \dots, x_r^\top)^\top \in \mathbb{R}^{rd}, \qquad x_a \in \mathbb{R}^d.$$

Then

$$x^\top \big((ww^\top) \otimes I_d\big) x = \Big\| \sum_{a=1}^r w_a x_a \Big\|_2^2.$$

Consequently,

$$x^\top \big(I_{rd} - (ww^\top) \otimes I_d\big) x = \sum_{a=1}^r \|x_a\|_2^2 - \Big\| \sum_{a=1}^r w_a x_a \Big\|_2^2.$$

Proof. Set

$$\bar{x} := \sum_{a=1}^r w_a x_a \in \mathbb{R}^d.$$

By the block action of the Kronecker product,

$$\big((ww^\top) \otimes I_d\big) x = \big((w_1 \bar{x})^\top, \dots, (w_r \bar{x})^\top\big)^\top.$$

Therefore

$$x^\top \big((ww^\top) \otimes I_d\big) x = x^\top \begin{pmatrix} w_1 \bar{x} \\ \vdots \\ w_r \bar{x} \end{pmatrix} = \sum_{a=1}^r w_a \, x_a^\top \bar{x} = \bar{x}^\top \bar{x} = \|\bar{x}\|_2^2 = \Big\| \sum_{a=1}^r w_a x_a \Big\|_2^2.$$

This proves the first identity.

For the second, note that

$$x^\top x = \sum_{a=1}^r \|x_a\|_2^2,$$

so

$$x^\top \big(I_{rd} - (ww^\top) \otimes I_d\big) x = x^\top x - x^\top \big((ww^\top) \otimes I_d\big) x = \sum_{a=1}^r \|x_a\|_2^2 - \Big\| \sum_{a=1}^r w_a x_a \Big\|_2^2.$$

□

Applying Lemma D.8 with $d = c - 1$, $x_a = U_{a,n}$, and $w = w_n$, we obtain

$$\chi^2_{n,q}(u) = \sum_a \|U_{a,n}\|_2^2 - \Big\| \sum_a w_{n,a} U_{a,n} \Big\|_2^2 = U_n^\top \big(I_{r(c-1)} - (w_n w_n^\top) \otimes I_{c-1}\big) U_n = U_n^\top A_n U_n.$$

□

Remark D.9 (Interpretation of the projection identity). The identity decomposes the Pearson statistic into total variation minus pooled variation. The term $\sum_a \|U_{a,n}\|_2^2$ is the total rowwise Pearson fluctuation, while $\big\| \sum_a w_{n,a} U_{a,n} \big\|_2^2$ is the fluctuation explained by a single pooled column-shift shared across rows. Thus the Pearson independence statistic keeps only the component of the rowwise fluctuations that cannot be explained by a common pooled shift.

Equivalently, the matrix A_n removes the subspace $\{\, w_n \otimes v : v \in \mathbb{R}^{c-1} \,\}$, which consists of stacked block-vectors whose row pattern is proportional to w_n, and retains the orthogonal complement, i.e. the between-row variation.

D.3 STEP 3: CONCLUDE $\chi^2_{n,q}(u) \Rightarrow \chi^2_{(r-1)(c-1)}$

The expansion in Step 2 gives the exact identity

$$\chi^2_{n,q}(u) = \sum_a \frac{\|M_{a,n}\|_2^2}{N_{a\cdot}} - \frac{\|\sum_a M_{a,n}\|_2^2}{N_u}.$$

Keep the same row sums and replace only their random denominators by stationary expectations:

$$Q_n(u) := \sum_a \frac{\|M_{a,n}\|_2^2}{s_n \pi_{au}} - \frac{\|\sum_a M_{a,n}\|_2^2}{s_n \pi_u}.$$

Let $\Delta_n := \max_a |N_{a\cdot}/s_n - \pi_{au}|$. On the event $N_{a\cdot} \geq s_n \pi_{au}/2$ for every a, subtracting the two expressions and using $\|\sum_a M_{a,n}\|_2^2 \leq r \sum_a \|M_{a,n}\|_2^2$ gives

$$|\chi^2_{n,q}(u) - Q_n(u)| \leq C\Delta_n \frac{\sum_a \|M_{a,n}\|_2^2}{s_n} = O_p(s_n^{-1/2}). \tag{11}$$

Here C depends only on the fixed positive row probabilities and r. The event has probability tending to one by equation 5; also $\Delta_n = O_p(s_n^{-1/2})$ and $M_n/\sqrt{s_n} = O_p(1)$ by Step 1.

Set $w := (\sqrt{\pi_{au}/\pi_u})_{a=1}^r$ and $A := (I_r - ww^\top) \otimes I_{c-1}$. Then A is a deterministic orthogonal projection of rank $k = (r-1)(c-1)$, and

$$Q_n(u) = \left(V_u^{-1/2} \frac{M_n}{\sqrt{s_n}}\right)^\top A \left(V_u^{-1/2} \frac{M_n}{\sqrt{s_n}}\right) \Rightarrow Z^\top A Z \sim \chi^2_k$$

by Step 1 and Lemma G.4. Together with equation 11, this proves

$$\chi^2_{n,q}(u) \Rightarrow \chi^2_{(r-1)(c-1)}. \tag{12}$$

Remark D.10 (Expectation of the comparison statistic). Martingale orthogonality and stationarity give $\mathbb{E}\|M_{a,n}\|_2^2 = (c-1)s_n\pi_{au}$ and $\mathbb{E}\|\sum_a M_{a,n}\|_2^2 = (c-1)s_n\pi_u$. Consequently $\mathbb{E}Q_n(u) = (r-1)(c-1)$ exactly, even with feedback. This does not assert an exact finite-sample expectation for the statistic with random denominators.

Remark D.11 (Scope of the error bound). The bound in equation 11 controls replacement of the row counts by their stationary expectations. It is not a rate of distributional convergence to the chi-square law.

D.4 STEP 4: REPLACE DENOMINATORS q BY $\widehat{q}$ (NEGLIGIBLE)

Lemma D.12 (Column-marginal rate under the null)**.** *Under Assumptions C.1, C.3, and C.4, for each $b \in [c]$,*

$$\widehat{q}_b - q_b = O_p(N_u^{-1/2}).$$

Proof. Fix b and define the scalar process

$$K_{n,b} := \sum_{t=1}^n \mathbb{1}\{t \in \mathcal{S},\ U_t = u\}\,(\mathbb{1}\{Y_t = b\} - q_b) = N_{\cdot b} - q_b N_u.$$

Let

$$\Delta K_{t,b} := K_{t,b} - K_{t-1,b} = \mathbb{1}\{t \in \mathcal{S},\ U_t = u\}\,(\mathbb{1}\{Y_t = b\} - q_b).$$

By Lemma C.2, the weight $\mathbb{1}\{t \in \mathcal{S}, U_t = u\}$ is $\mathcal{F}_{t-1}$-measurable. By Lemma C.5, on $\{U_t = u\}$ we have $\mathbb{E}[\mathbb{1}\{Y_t = b\} \mid \mathcal{F}_{t-1}] = q_b$. Applying the indicator-repair lemma (Lemma C.9) with $W = \mathbb{1}\{t \in \mathcal{S}, U_t = u\}$, $E = \{U_t = u\}$, $Z = \mathbb{1}\{Y_t = b\}$, and $c = q_b$ gives $\mathbb{E}[\Delta K_{t,b} \mid \mathcal{F}_{t-1}] = 0$. Hence $(K_{n,b})$ is a martingale with increments bounded by 1 in absolute value.

Its predictable quadratic variation is

$$\langle K^{(b)} \rangle_n := \sum_{t=1}^n \mathbb{E}\big[(\Delta K_{t,b})^2 \mid \mathcal{F}_{t-1}\big].$$

Moreover,

$$\langle K^{(b)}\rangle_n = \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\, U_t = u\}\, \mathbb{E}\Big[\big(\mathbb{1}\{Y_t = b\} - q_b\big)^2 \,\Big|\, \mathcal{F}_{t-1}\Big] \le \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\, U_t = u\} = N_u,$$

because on $\{U_t = u\}$ the conditional variance is the Bernoulli variance $q_b(1-q_b) \le 1$.

By equation 5, $N_u/s_n \to \pi_u > 0$ in probability. Also $0 \le V_n \le s_n$ for the choice below, so $\mathbb{E}[V_n] \le s_n$. Thus Lemma G.3 applies with $V_n := N_u$ and yields

$$K_{n,b} = O_p(\sqrt{N_u}).$$

Dividing by N_u gives

$$\widehat{q}_b - q_b = \frac{K_{n,b}}{N_u} = O_p(N_u^{-1/2}).$$

□

Lemma D.13 (Denominator replacement under the null)**.** *Under the assumptions of Theorem D.1,*

$$\chi^2_n(u) - \chi^2_{n,q}(u) = o_p(1).$$

Proof. Write

$$\chi^2_n(u) - \chi^2_{n,q}(u) = \sum_{a=1}^{r}\sum_{b=1}^{c} \frac{(N_{ab} - N_{a\cdot}\widehat{q}_b)^2}{N_{a\cdot} q_b} \left(\frac{q_b}{\widehat{q}_b} - 1\right).$$

On the event $\min_b \widehat{q}_b \ge c_0$, the elementwise ratio satisfies

$$\left\| \frac{q}{\widehat{q}} - \mathbf{1} \right\|_\infty = \max_b \frac{|q_b - \widehat{q}_b|}{\widehat{q}_b} \le \frac{1}{c_0} \|\widehat{q} - q\|_\infty,$$

where $q/\widehat{q}$ denotes the vector $(q_1/\widehat{q}_1, \dots, q_c/\widehat{q}_c)$ and $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^c$.

Also,

$$\sum_{a,b} \frac{(N_{ab} - N_{a\cdot}\widehat{q}_b)^2}{N_{a\cdot} q_b} = \chi^2_{n,q}(u).$$

Therefore,

$$|\chi^2_n(u) - \chi^2_{n,q}(u)| \le \frac{1}{c_0} \|\widehat{q} - q\|_\infty\, \chi^2_{n,q}(u).$$

By Lemma D.12,

$$\|\widehat{q} - q\|_\infty = O_p(N_u^{-1/2}).$$

Next, by equation 12, $\chi^2_{n,q}(u) \Rightarrow \chi^2_{(r-1)(c-1)}$. Since $\chi^2_{(r-1)(c-1)}$ is a proper probability distribution on $[0, \infty)$, for every $\varepsilon > 0$ there exists $M_\varepsilon < \infty$ such that

$$\mathbb{P}\big(\chi^2_{(r-1)(c-1)} > M_\varepsilon\big) < \varepsilon.$$

Hence

$$\mathbb{P}\big(\chi^2_{n,q}(u) > M_\varepsilon\big) \to \mathbb{P}\big(\chi^2_{(r-1)(c-1)} > M_\varepsilon\big) < \varepsilon,$$

so $(\chi^2_{n,q}(u))_n$ is bounded in probability, i.e.

$$\chi^2_{n,q}(u) = O_p(1).$$

Since $N_u \to \infty$ in probability (because $N_u = \sum_a N_{a\cdot} \ge r\, m_n$ and $m_n \to \infty$ in probability), we have $N_u^{-1/2} = o_p(1)$. Therefore

$$\|\widehat{q} - q\|_\infty\, \chi^2_{n,q}(u) = O_p(N_u^{-1/2})\, O_p(1) = o_p(1),$$

and the claim follows. □

Conclusion. By equation 12 and Lemma D.13,

$$\chi^2_n(u) \Rightarrow \chi^2_{(r-1)(c-1)}.$$

This proves Theorem D.1. □

E EDGE-PRESENT CASE: DIVERGENCE AND CONSISTENCY

The null analysis in Section D gives the reference distribution needed for calibration. For a valid conditional-independence test under fixed alternatives, the remaining task is not to identify a second reference law, but to show that the statistic diverges whenever the structural alternative is truly row-dependent. This section establishes exactly that point.

Remark E.1 (Tools from Section D used under the alternative). Several results from the null analysis carry over to the present section because they do not require the structural null (Assumption C.4):

- Lemma C.2 (predictability of $(U_t, X_{t-\delta})$): requires only Assumption C.1.
- Lemma C.8 (conditional distribution of Y_t under the alternative): requires Assumptions C.1, C.3, and C.7. This is the alternative-hypothesis counterpart of Lemma C.5: it provides the centering $\mathbb{P}(Y_t = b \mid \mathcal{F}_{t-1}) = p_{X_{t-\delta},b}$ on $\{U_t = u\}$.
- The indicator-repair lemma (Lemma C.9): the structural reasoning is assumption-agnostic—it only needs (a) $\mathcal{F}_{t-1}$-measurability of the design indicator (Lemma C.2) and (b) a centering property on $\{U_t = u\}$ (Lemma C.5 under the null, Lemma C.8 under the alternative). It therefore applies in both settings.

Throughout this section, we invoke these tools without further comment; the reader should verify that only the assumptions listed above are used.

E.1 ORACLE COLUMN MARGINAL AND DIVERGENCE RATE

Assume Assumption C.7 and let p_{ab} be the induced row pmfs, i.e., for each contingency-table row $a \in [r]$, the vector $(p_{a1}, \dots, p_{ac})$ is the conditional distribution of Y_t given $X_{t-\delta} = a$ within the fixed stratum u. Define the random row fractions

$$w_{a,n} := \frac{N_{a\cdot}}{N_u}, \qquad a \in [r],$$

and the *oracle column marginal*

$$\bar{q}_{b,n} := \sum_{a=1}^{r} w_{a,n} p_{ab}, \qquad \bar{q}_n := (\bar{q}_{1,n}, \dots, \bar{q}_{c,n}).$$

Thus $\bar{q}_n$ is the column marginal predicted by the true conditional response distributions $(p_{a1}, \dots, p_{ac})$ for each contingency-table row $a \in [r]$, weighted by the observed row frequencies. It is random only through the design counts $N_{a\cdot}$.

The natural signal size is the weighted Pearson separation of the row distributions from their mixture:

$$\Lambda_n^{\mathrm{orc}} := \sum_{a=1}^{r} w_{a,n} \sum_{b=1}^{c} \frac{(p_{ab} - \bar{q}_{b,n})^2}{\bar{q}_{b,n}}, \qquad \lambda_n^{\mathrm{orc}} := N_u \, \Lambda_n^{\mathrm{orc}}. \tag{13}$$

The quantity Λ_n^{orc} is the per-observation signal, while λ_n^{orc} is the cumulative signal over the N_u observations in the stratum.

Remark E.2 (Interpretation of λ_n^{orc}). If all rows share the same response law, then $p_{ab} = \bar{q}_{b,n}$ for all a, b and $\Lambda_n^{\mathrm{orc}} = \lambda_n^{\mathrm{orc}} = 0$. Under the structural alternative, the rows have different response distributions, so the row-wise deviations $p_a - \bar{q}_n$ cannot cancel after averaging over rows with positive weight. The role of Lemma E.3 below is to make this "no cancellation" statement quantitative.

Lemma E.3 (Positive signal margin). *Assume Assumptions C.7 and B.1. Define, for w in the simplex $\{w \in \mathbb{R}^r : w_a \geq 0,\ \sum_a w_a = 1\}$,*

$$\Lambda(w) := \sum_{a=1}^{r} w_a \sum_{b=1}^{c} \frac{(p_{ab} - \bar{q}_b(w))^2}{\bar{q}_b(w)}, \qquad \bar{q}_b(w) := \sum_{a=1}^{r} w_a p_{ab}.$$

Terms with $\bar{q}_b(w) = 0$ are defined to be zero; then $p_{ab} = 0$ for every row with $w_a > 0$ in that column. Then:

(i) *For every w with $w_a > 0$ for all a, one has $\Lambda(w) > 0$.*

(ii) *There exists $\Lambda_{\min} > 0$ such that*

$$\mathbb{P}\big(\Lambda_n^{\mathrm{orc}} \geq \Lambda_{\min}\big) \geq 1 - o(1).$$

Proof. For part (i), note first that $\Lambda(w) \geq 0$. If $\Lambda(w) = 0$, then every summand in the definition of $\Lambda(w)$ must vanish, so $p_{ab} = \bar{q}_b(w)$ for every a with $w_a > 0$ and every b. When all $w_a > 0$, this implies $p_a = \bar{q}(w)$ for all a, hence all row distributions are identical. That contradicts Assumption C.7, which requires the family $\{p_a\}_{a=1}^r$ to be non-constant. Therefore $\Lambda(w) > 0$ whenever w lies in the interior of the simplex.

For part (ii), the map $w \mapsto \Lambda(w)$ is continuous on the compact set

$$\mathcal{W} := \Big\{ w : \min_a w_a \geq \varepsilon_{\mathrm{row}},\ \sum_a w_a = 1 \Big\},$$

where $\varepsilon_{\mathrm{row}} > 0$ is the constant from Assumption B.1. (The set $\mathcal{W}$ is nonempty precisely when $\varepsilon_{\mathrm{row}} \leq 1/r$, which always holds since the row fractions $N_{a\cdot}/N_u$ sum to one.) On $\mathcal{W}$, each denominator satisfies $\bar{q}_b(w) \geq \varepsilon_{\mathrm{row}} \sum_a p_{ab} > 0$. Thus the denominators are uniformly bounded away from zero, and Λ is continuous on $\mathcal{W}$. By Assumption B.1, $\min_a N_{a\cdot}/N_u \geq \varepsilon_{\mathrm{row}}$ with probability $1 - o(1)$, which is exactly the condition for $w_n \in \mathcal{W}$. Hence

$$\mathbb{P}\big(w_n \in \mathcal{W}\big) = 1 - o(1), \qquad w_n := (N_{1\cdot}/N_u, \ldots, N_{r\cdot}/N_u).$$

By part (i), $\Lambda(w) > 0$ for every $w \in \mathcal{W}$, so continuity and compactness give

$$\Lambda_{\min} := \min_{w \in \mathcal{W}} \Lambda(w) > 0.$$

Since $\Lambda_n^{\mathrm{orc}} = \Lambda(w_n)$, the claimed high-probability lower bound follows. Explicitly: on the event $\{w_n \in \mathcal{W}\}$, we have $\Lambda_n^{\mathrm{orc}} = \Lambda(w_n) \geq \Lambda_{\min}$, and this event has probability $1 - o(1)$. □

Recall from Section 1 that $\widehat{q}_b = N_{\cdot b}/N_u$ is the empirical column marginal in stratum u.

Lemma E.4 (Column-marginal rate under the alternative)**.** *Assume Assumptions C.1, C.3, and C.7. Then for each $b \in [c]$,*

$$\widehat{q}_b - \bar{q}_{b,n} = O_p(N_u^{-1/2}).$$

Remark E.5 (Why the oracle marginal is the right centering). Under the null there is a single fixed column law q shared by all rows. Under the alternative there is no such common law: the conditional response distribution in row a is p_a. The quantity $\bar{q}_{b,n}$ is therefore the right pooled benchmark, because it averages the true row-specific column probabilities using the row frequencies actually realized in the sample. Lemma E.4 says that the empirical pooled column marginal $\widehat{q}_b$ fluctuates around this oracle benchmark at the usual $N_u^{-1/2}$ rate.

Proof. Fix $b \in [c]$ and define

$$K_{n,b}^{\mathrm{alt}} := \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u\} \big(\mathbb{1}\{Y_t = b\} - p_{X_{t-\delta},b}\big).$$

By Lemma C.2, both $\mathbb{1}\{t \in \mathcal{S},\ U_t = u\}$ and $p_{X_{t-\delta},b}$ are $\mathcal{F}_{t-1}$-measurable. By Lemma C.8, on the event $\{U_t = u\}$ we have $\mathbb{E}[\mathbb{1}\{Y_t = b\} \mid \mathcal{F}_{t-1}] = p_{X_{t-\delta},b}$. Applying the indicator-repair lemma (Lemma C.9) with $W = \mathbb{1}\{t \in \mathcal{S}, U_t = u\}$, $E = \{U_t = u\}$, $Z = \mathbb{1}\{Y_t = b\}$, and $c = p_{X_{t-\delta},b}$ gives $\mathbb{E}[\Delta K_t^{\mathrm{alt}} \mid \mathcal{F}_{t-1}] = 0$. Therefore $(K_{n,b}^{\mathrm{alt}})_{n \geq 1}$ is a martingale with increments bounded by 1 in absolute value.

Let $\Delta K_t^{\mathrm{alt}}$ denote the increment at time t. Its predictable quadratic variation satisfies

$$\begin{aligned}
\langle K_b^{\mathrm{alt}} \rangle_n &:= \sum_{t=1}^{n} \mathbb{E}\big[(\Delta K_t^{\mathrm{alt}})^2 \mid \mathcal{F}_{t-1}\big] \\
&= \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u\}\, \mathbb{E}\Big[\big(\mathbb{1}\{Y_t = b\} - p_{X_{t-\delta},b}\big)^2 \,\Big|\, \mathcal{F}_{t-1}\Big] \\
&\leq \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u\} = N_u,
\end{aligned}$$

where the second equality pulls out the $\mathcal{F}_{t-1}$-measurable indicator (Lemma C.2), and the inequality uses the fact that on $\{U_t = u\}$, Lemma C.8 gives $\mathbb{E}[\mathbb{1}\{Y_t = b\} \mid \mathcal{F}_{t-1}] = p_{X_{t-\delta},b}$, so the conditional second moment is the Bernoulli variance $p_{X_{t-\delta},b}(1 - p_{X_{t-\delta},b}) \leq 1$. By equation 5, $N_u/s_n \to \pi_u > 0$ in probability. Also $0 \leq V_n \leq s_n$ for the choice below, so $\mathbb{E}[V_n] \leq s_n$. Thus Lemma G.3 applies with $V_n := N_u$ and yields

$$K_{n,b}^{\mathrm{alt}} = O_p(\sqrt{N_u}).$$

It remains to connect $K^{\mathrm{alt}}_{n,b}$ to $\widehat{q}_b - \bar{q}_{b,n}$. Using $N_{\cdot b} = \sum_{t\in I(u)} \mathbb{1}\{Y_t = b\}$ and $\sum_a N_{a\cdot} p_{ab} = \sum_a \big(\sum_{t\in I(u)} \mathbb{1}\{X_{t-\delta} = a\}\big) p_{ab} = \sum_{t\in I(u)} \sum_a \mathbb{1}\{X_{t-\delta} = a\}\, p_{ab} = \sum_{t\in I(u)} p_{X_{t-\delta},b}$, where the last step uses the fact that $X_{t-\delta}$ takes exactly one value at each t, we obtain

$$N_{\cdot b} - N_u \bar{q}_{b,n} = \sum_{t\in I(u)} \big(\mathbb{1}\{Y_t = b\} - p_{X_{t-\delta},b}\big) = K^{\mathrm{alt}}_{n,b}.$$

Dividing by N_u yields

$$\widehat{q}_b - \bar{q}_{b,n} = \frac{K^{\mathrm{alt}}_{n,b}}{N_u} = O_p(N_u^{-1/2}).$$

□

Oracle Pearson coordinates. For each n, let

$$D_{\bar{q}_n} := \mathrm{diag}(\bar{q}_{1,n}, \ldots, \bar{q}_{c,n}), \qquad \bar{s}_n := \sqrt{\bar{q}_n},$$

and let $G_n \in \mathbb{R}^{(c-1)\times c}$ have orthonormal rows spanning $\bar{s}_n^{\perp}$. Define

$$R^{\mathrm{orc}}_{a,n} := \frac{1}{\sqrt{N_{a\cdot}}} D^{-1/2}_{\bar{q}_n} \big(N_{a\bullet} - N_{a\cdot}\bar{q}_n\big) \in \mathbb{R}^c, \qquad U^{\mathrm{orc}}_{a,n} := G_n R^{\mathrm{orc}}_{a,n} \in \mathbb{R}^{c-1},$$

and stack the row blocks as

$$U^{\mathrm{orc}}_n := \big((U^{\mathrm{orc}}_{1,n})^\top, \ldots, (U^{\mathrm{orc}}_{r,n})^\top\big)^\top \in \mathbb{R}^{r(c-1)}.$$

In words, $R^{\mathrm{orc}}_{a,n}$ measures how much row a's empirical column distribution deviates from the overall pooled benchmark $\bar{q}_n$, in Pearson-standardized units normalized by the row sample size $\sqrt{N_{a\cdot}}$. Its squared norm $\|R^{\mathrm{orc}}_{a,n}\|_2^2$ is row a's contribution to the chi-square statistic: under the null all rows look alike and these are small (pure noise), while under the structural alternative at least some rows deviate systematically.

Remark E.6 (Projection at aggregate vs. observation level). In Section D, the projection H was applied at the observation level (defining $\xi_t \in \mathbb{R}^{c-1}$) because the martingale CLT required whitened individual increments with identity conditional covariance. Here no CLT is needed—the goal is purely algebraic: decomposing χ^2 into signal and noise. Accordingly, the projection G_n is applied directly to the aggregated row vectors $R^{\mathrm{orc}}_{a,n}$, paralleling the algebraic identity in Lemma D.7 rather than the martingale construction.

Also define

$$w_n := \big(\sqrt{N_{1\cdot}/N_u}, \ldots, \sqrt{N_{r\cdot}/N_u}\big)^\top, \qquad A_n := I_{r(c-1)} - (w_n w_n^\top) \otimes I_{c-1}.$$

Lemma E.7 (Exact oracle projection identity)**.** *Define the oracle-denominator Pearson statistic*

$$\chi^2_{n,\bar{q}}(u) := \sum_{a=1}^{r}\sum_{b=1}^{c} \frac{(N_{ab} - N_{a\cdot}\widehat{q}_b)^2}{N_{a\cdot}\bar{q}_{b,n}}.$$

Then

$$\chi^2_{n,\bar{q}}(u) = (U^{\mathrm{orc}}_n)^\top A_n U^{\mathrm{orc}}_n.$$

Proof. The proof is the same algebra as Lemma D.7, carried out with the random reference vector $\bar{q}_n$ in place of the fixed null vector q. Because the argument is pointwise in the realized table, no probabilistic input is needed.

Set $\delta^{\mathrm{orc}}_n := \widehat{q} - \bar{q}_n$. Since $N_{ab} - N_{a\cdot}\widehat{q}_b = (N_{ab} - N_{a\cdot}\bar{q}_{b,n}) - N_{a\cdot}\delta^{\mathrm{orc}}_{n,b}$, squaring, dividing by $N_{a\cdot}\bar{q}_{b,n}$, and summing over a, b produces three terms—exactly as in the proof of Lemma D.7 with $\bar{q}_n$ replacing q. The cross term evaluates to $-2N_u\|D^{-1/2}_{\bar{q}_n}\delta^{\mathrm{orc}}_n\|_2^2$ (using $\sum_a (N_{ab} - N_{a\cdot}\bar{q}_{b,n}) = N_u\delta^{\mathrm{orc}}_{n,b}$), and the squared-shift term is $+N_u\|D^{-1/2}_{\bar{q}_n}\delta^{\mathrm{orc}}_n\|_2^2$. Combining gives

$$\chi^2_{n,\bar{q}}(u) = \sum_{a=1}^{r} \big\|R^{\mathrm{orc}}_{a,n}\big\|_2^2 - N_u\big\|D^{-1/2}_{\bar{q}_n}\delta^{\mathrm{orc}}_n\big\|_2^2.$$

Next,

$$D^{-1/2}_{\bar{q}_n}\delta^{\mathrm{orc}}_n = \frac{1}{\sqrt{N_u}}\sum_{a=1}^{r} w_{n,a} R^{\mathrm{orc}}_{a,n},$$

and each $R^{\mathrm{orc}}_{a,n}$ lies in $\bar{s}_n^{\perp}$ because

$$\bar{s}_n^{\top} R^{\mathrm{orc}}_{a,n} = \frac{1}{\sqrt{N_{a\cdot}}} \sum_{b=1}^{c} \big(N_{ab} - N_{a\cdot}\bar{q}_{b,n}\big) = 0.$$

Since G_n is an isometry on $\bar{s}_n^{\perp}$,

$$\big\|R^{\mathrm{orc}}_{a,n}\big\|_2 = \big\|U^{\mathrm{orc}}_{a,n}\big\|_2, \qquad \Big\| \sum_a w_{n,a} R^{\mathrm{orc}}_{a,n} \Big\|_2 = \Big\| \sum_a w_{n,a} U^{\mathrm{orc}}_{a,n} \Big\|_2.$$

Substituting these identities into the previous display yields

$$\chi^2_{n,\bar{q}}(u) = \sum_{a=1}^{r} \big\|U^{\mathrm{orc}}_{a,n}\big\|_2^2 - \Big\| \sum_{a=1}^{r} w_{n,a} U^{\mathrm{orc}}_{a,n} \Big\|_2^2 = (U^{\mathrm{orc}}_n)^{\top} A_n U^{\mathrm{orc}}_n.$$

□

Lemma E.8 (Row-wise fluctuation bound in oracle coordinates)**.** *Assume Assumptions C.1, C.3, C.7, and B.1. For each $a \in [r]$, define the centered row vector*

$$B_{a,n} := \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u,\ X_{t-\delta} = a\}\,(e_{Y_t} - p_a) \in \mathbb{R}^c.$$

Then $\|B_{a,n}\|_2 = O_p(\sqrt{N_u})$. Consequently, on the event $\{\min_b \bar{q}_{b,n} \geq \bar{c}_0,\ \min_a N_{a\cdot}/N_u \geq \varepsilon_{\mathrm{row}}\}$,

$$Z_{a,n} := \frac{1}{\sqrt{N_{a\cdot}}} G_n D_{\bar{q}_n}^{-1/2} B_{a,n} \quad \textit{satisfies} \quad \|Z_{a,n}\|_2 = O_p(1).$$

Hence the stacked vector $Z_n := (Z_{1,n}^{\top}, \ldots, Z_{r,n}^{\top})^{\top}$ satisfies $\|Z_n\|_2 = O_p(1)$.

Proof. Fix $a \in [r]$. For each coordinate $b \in [c]$,

$$B_{a,n,b} = \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u,\ X_{t-\delta} = a\}\,\big(\mathbb{1}\{Y_t = b\} - p_{ab}\big).$$

By Lemma C.2 (which requires only Assumption C.1), the indicator is $\mathcal{F}_{t-1}$-measurable. By Lemma C.8 (which requires Assumptions C.1, C.3, and C.7), $\mathbb{E}[\mathbb{1}\{Y_t = b\} \mid \mathcal{F}_{t-1}] = p_{ab}$ on $\{U_t = u, X_{t-\delta} = a\}$. Applying Lemma C.9 with $W = \mathbb{1}\{t \in \mathcal{S}, U_t = u, X_{t-\delta} = a\}$, $E = \{U_t = u, X_{t-\delta} = a\}$, $Z = \mathbb{1}\{Y_t = b\}$, and $c = p_{ab}$ gives $\mathbb{E}[\Delta B_{a,n,b} \mid \mathcal{F}_{t-1}] = 0$ globally, so $(B_{a,n,b})_{n\geq 1}$ is a martingale.

Its predictable quadratic variation satisfies

$$\langle B_{a,\cdot,b}\rangle_n = \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u,\ X_{t-\delta} = a\}\,\mathbb{E}\Big[\big(\mathbb{1}\{Y_t = b\} - p_{ab}\big)^2 \,\Big|\, \mathcal{F}_{t-1}\Big] \leq \sum_{t=1}^{n} \mathbb{1}\{t \in \mathcal{S},\ U_t = u,\ X_{t-\delta} = a\} = N_{a\cdot},$$

because on $\{U_t = u, X_{t-\delta} = a\}$ the conditional variance is the Bernoulli variance $p_{ab}(1 - p_{ab}) \leq 1$. By equation 5, $N_{a\cdot}/s_n \to \pi_{au} > 0$ in probability. Also $0 \leq V_n \leq s_n$ for the choice below, so $\mathbb{E}[V_n] \leq s_n$. Thus Lemma G.3 applies with $V_n := N_{a\cdot}$ and yields

$$B_{a,n,b} = O_p(\sqrt{N_{a\cdot}}).$$

Because c is fixed, this implies

$$\|B_{a,n}\|_2 = O_p(\sqrt{N_{a\cdot}}).$$

On the event stated in the lemma,

$$\|D_{\bar{q}_n}^{-1/2}\|_{\mathrm{op}} \leq \bar{c}_0^{-1/2}, \qquad \|G_n\|_{\mathrm{op}} = 1.$$

(The first bound holds because $D_{\bar{q}_n}^{-1/2}$ is diagonal with entries $\bar{q}_{b,n}^{-1/2} \leq \bar{c}_0^{-1/2}$; the second because G_n has orthonormal rows and is therefore a partial isometry.) Hence

$$\|Z_{a,n}\|_2 = \frac{1}{\sqrt{N_{a\cdot}}} \|G_n D_{\bar{q}_n}^{-1/2} B_{a,n}\|_2 \leq \frac{1}{\sqrt{N_{a\cdot}}} \|G_n\|_{\mathrm{op}} \|D_{\bar{q}_n}^{-1/2}\|_{\mathrm{op}} \|B_{a,n}\|_2 \leq \bar{c}_0^{-1/2} \frac{\|B_{a,n}\|_2}{\sqrt{N_{a\cdot}}} = O_p(1).$$

Because r is fixed, stacking over $a = 1, \ldots, r$ gives $\|Z_n\|_2 = O_p(1)$. □

Theorem E.9 (Edge-present divergence)**.** *In the standing stationary mixing setting, assume Assumptions C.1, C.3, C.7, and B.1. Assume also that there exists $\bar{c}_0 > 0$ such that*

$$\mathbb{P}\big(\min_{1\le b\le c} \bar{q}_{b,n} \ge \bar{c}_0\big) = 1 - o(1)$$

(for example, it is enough that $p_{ab} \ge \bar{c}_0$ for all a, b). Then

$$\chi^2_n(u) = \lambda^{\mathrm{orc}}_n + O_p\big(\sqrt{\lambda^{\mathrm{orc}}_n}\big),$$

since $\lambda^{\mathrm{orc}}_n \to \infty$ in probability under the positivity assumption, so any $O_p(1)$ remainder is dominated by $O_p(\sqrt{\lambda^{\mathrm{orc}}_n})$. In particular,

$$\frac{\chi^2_n(u)}{\lambda^{\mathrm{orc}}_n} \xrightarrow{p} 1, \qquad \chi^2_n(u) \xrightarrow{p} \infty.$$

Proof. Write

$$N_{a\bullet} - N_{a\cdot}\bar{q}_n = N_{a\cdot}(p_a - \bar{q}_n) + B_{a,n},$$

where $B_{a,n}$ is the centered row fluctuation from Lemma E.8. Applying the oracle Pearson coordinates gives the decomposition

$$U^{\mathrm{orc}}_{a,n} = \nu_{a,n} + Z_{a,n},$$

where

$$\nu_{a,n} := \sqrt{N_{a\cdot}}\, G_n D^{-1/2}_{\bar{q}_n}(p_a - \bar{q}_n), \qquad Z_{a,n} := \frac{1}{\sqrt{N_{a\cdot}}} G_n D^{-1/2}_{\bar{q}_n} B_{a,n}.$$

Stacking over rows yields

$$U^{\mathrm{orc}}_n = \nu_n + Z_n, \qquad \|Z_n\|_2 = O_p(1)$$

by Lemma E.8.

By Lemma E.7,

$$\chi^2_{n,\bar{q}}(u) = (U^{\mathrm{orc}}_n)^\top A_n U^{\mathrm{orc}}_n.$$

We next identify the deterministic signal term. Because $\bar{q}_n = \sum_a w_{a,n} p_a$,

$$\sum_{a=1}^{r} w_{n,a}\nu_{a,n} = \sqrt{N_u}\, G_n D^{-1/2}_{\bar{q}_n} \sum_{a=1}^{r} w_{a,n}(p_a - \bar{q}_n) = 0.$$

Hence $P_n\nu_n = 0$ and therefore $A_n\nu_n = \nu_n$. Moreover, each vector $D^{-1/2}_{\bar{q}_n}(p_a - \bar{q}_n)$ lies in $\bar{s}_n^{\perp}$: since $\bar{s}_n^\top D^{-1/2}_{\bar{q}_n} = \mathbf{1}^\top$, we have $\bar{s}_n^\top D^{-1/2}_{\bar{q}_n}(p_a - \bar{q}_n) = \mathbf{1}^\top(p_a - \bar{q}_n) = 0$ because both p_a and $\bar{q}_n$ are pmfs. Hence G_n acts isometrically on this vector. Consequently,

$$\begin{aligned}
\nu_n^\top A_n \nu_n &= \|\nu_n\|_2^2 \\
&= \sum_{a=1}^{r} N_{a\cdot} \left\| D^{-1/2}_{\bar{q}_n}(p_a - \bar{q}_n)\right\|_2^2 \\
&= \sum_{a=1}^{r} N_{a\cdot} \sum_{b=1}^{c} \frac{(p_{ab} - \bar{q}_{b,n})^2}{\bar{q}_{b,n}} \\
&= N_u \sum_{a=1}^{r} w_{a,n} \sum_{b=1}^{c} \frac{(p_{ab} - \bar{q}_{b,n})^2}{\bar{q}_{b,n}} \\
&= \lambda^{\mathrm{orc}}_n.
\end{aligned}$$

Therefore

$$\begin{aligned}
\chi^2_{n,\bar{q}}(u) &= (\nu_n + Z_n)^\top A_n(\nu_n + Z_n) \\
&= \lambda^{\mathrm{orc}}_n + 2\nu_n^\top A_n Z_n + Z_n^\top A_n Z_n.
\end{aligned}$$

Since A_n is an orthogonal projection, $\|A_n\|_{\mathrm{op}} \le 1$. Using $A_n\nu_n = \nu_n$ and $\|Z_n\|_2 = O_p(1)$,

$$|\nu_n^\top A_n Z_n| \le \|\nu_n\|_2 \, \|Z_n\|_2 = \sqrt{\lambda^{\mathrm{orc}}_n}\, O_p(1) = O_p\big(\sqrt{\lambda^{\mathrm{orc}}_n}\big),$$

and

$$0 \le Z_n^\top A_n Z_n \le \|Z_n\|_2^2 = O_p(1).$$

Thus

$$\chi^2_{n,\bar{q}}(u) = \lambda_n^{\mathrm{orc}} + O_p\big(\sqrt{\lambda_n^{\mathrm{orc}}}\big) + O_p(1). \tag{14}$$

We now replace the oracle denominator $\bar{q}_{b,n}$ by the empirical denominator $\widehat{q}_b$. On the event

$$E_n := \Big\{ \min_b \widehat{q}_b \ge c_0,\ \min_b \bar{q}_{b,n} \ge \bar{c}_0 \Big\},$$

Assumption B.1 and the standing positivity assumption imply $\mathbb{P}(E_n) = 1 - o(1)$. On E_n,

$$\left\| \frac{\bar{q}_n}{\widehat{q}} - \mathbf{1} \right\|_\infty \le \frac{1}{\min(c_0, \bar{c}_0)} \, \|\widehat{q} - \bar{q}_n\|_\infty = O_p(N_u^{-1/2})$$

by Lemma E.4. Therefore

$$|\chi^2_n(u) - \chi^2_{n,\bar{q}}(u)| \le O_p(N_u^{-1/2})\, \chi^2_{n,\bar{q}}(u).$$

From equation 14 and the positivity assumptions, Λ_n^{orc} is bounded above in probability and bounded below away from 0 with high probability: Lemma E.3 gives $\Lambda_n^{\mathrm{orc}} \ge \Lambda_{\min} > 0$ w.h.p., while on $\{\min_b \bar{q}_{b,n} \ge \bar{c}_0\}$,

$$\Lambda_n^{\mathrm{orc}} \le \sum_{a=1}^{r} w_{a,n} \sum_{b=1}^{c} \frac{1}{\bar{c}_0} = \frac{c}{\bar{c}_0}.$$

Hence $\lambda_n^{\mathrm{orc}} = N_u \Lambda_n^{\mathrm{orc}}$ is of the same order in probability as N_u. It follows that

$$O_p(N_u^{-1/2})\, \chi^2_{n,\bar{q}}(u) = O_p\big(\sqrt{\lambda_n^{\mathrm{orc}}}\big) + O_p(1).$$

Combining this with equation 14 yields

$$\chi^2_n(u) = \lambda_n^{\mathrm{orc}} + O_p\big(\sqrt{\lambda_n^{\mathrm{orc}}}\big) + O_p(1).$$

Finally, Lemma E.3 and Assumption B.1 imply $\lambda_n^{\mathrm{orc}} \ge \Lambda_{\min} N_u \to \infty$ in probability, because $N_u \ge r m_n$ and $m_n \to \infty$. Hence $1 = o_p(\sqrt{\lambda_n^{\mathrm{orc}}})$, so the $O_p(1)$ remainder is absorbed into $O_p(\sqrt{\lambda_n^{\mathrm{orc}}})$, giving the form stated in the theorem. The relative error is $o_p(1)$, so

$$\frac{\chi^2_n(u)}{\lambda_n^{\mathrm{orc}}} \xrightarrow{p} 1, \qquad \chi^2_n(u) \xrightarrow{p} \infty.$$

□

Corollary E.10 (Consistency of the stratumwise test)**.** *Let $k := (r-1)(c-1)$ and let $c_{k,\alpha}$ denote the $(1-\alpha)$-quantile of χ^2_k. Consider the test*

$$\phi_n(u) := \mathbb{1}\{\chi^2_n(u) > c_{k,\alpha}\}.$$

Under Assumptions C.1, C.3, C.4, and B.1,

$$\mathbb{P}_{H_0}\big(\phi_n(u) = 1\big) \to \alpha.$$

Under Assumptions C.1, C.3, C.7, B.1, and the positivity assumption of Theorem E.9,

$$\mathbb{P}_{H_1}\big(\phi_n(u) = 1\big) \to 1.$$

Proof. The null statement is exactly Theorem D.1. Under the alternative, Theorem E.9 gives $\chi^2_n(u) \to \infty$ in probability, whereas the critical value $c_{k,\alpha}$ is fixed. Therefore $\mathbb{P}_{H_1}(\chi^2_n(u) > c_{k,\alpha}) \to 1$. □

Assumption E.11 (Joint regularity across strata)**.** For every stratum $u \in \mathcal{U}$, let

$$m_n(u) := \min_{1 \le a \le r} N_{a\cdot}(u), \qquad N_u := \sum_{a=1}^{r} N_{a\cdot}(u), \qquad \widehat{q}_b(u) := \frac{N_{\cdot b}(u)}{N_u}.$$

Assume that the regularity conditions of Assumption B.1 hold simultaneously across all strata, i.e.

$$\min_{u \in \mathcal{U}} m_n(u) \xrightarrow{p} \infty,$$

and for some constants $\varepsilon_{\mathrm{row}}, c_0 > 0$,

$$\mathbb{P}\left(\min_{u \in \mathcal{U}} \min_{1 \le a \le r} \frac{N_{a\cdot}(u)}{N_u} \ge \varepsilon_{\mathrm{row}} \right) \to 1, \qquad \mathbb{P}\left(\min_{u \in \mathcal{U}} \min_{1 \le b \le c} \widehat{q}_b(u) \ge c_0 \right) \to 1.$$

Proposition F.7 gives one explicit sufficient condition under which Assumption E.11 holds.

Proposition E.12 (Combined conditional-independence test across strata)**.** *Suppose that $\mathcal{U}$ is finite and that Assumptions C.1 and C.3 hold. Assume Assumption E.11. For each stratum $u \in \mathcal{U}$, let $q^{(u)}$ denote the null pmf from Lemma C.5, and assume $q_b^{(u)} > 0$ for all $u \in \mathcal{U}$ and $b \in [c]$. Define the combined statistic*

$$\chi^2_{\text{total}} := \sum_{u \in \mathcal{U}} \chi^2_n(u).$$

Under the full null $H_0 \colon Y_t \perp X_{t-\delta} \mid U_t$ (assuming Assumption C.4 holds for every $u \in \mathcal{U}$),

$$\chi^2_{\text{total}} \;\Rightarrow\; \chi^2_{|\mathcal{U}|(r-1)(c-1)}.$$

Under the alternative that Assumption C.7 holds in at least one stratum, together with the positivity condition of Theorem E.9 in that stratum,

$$\chi^2_{\text{total}} \;\xrightarrow{p}\; \infty.$$

Proof. For each $u \in \mathcal{U}$, fix a matrix $H_u \in \mathbb{R}^{(c-1)\times c}$ whose rows form an orthonormal basis of $\sqrt{q^{(u)}}^{\perp}$, and define the whitened increment

$$\xi_t^{(u)} := H_u D_{q^{(u)}}^{-1/2}\big(e_{Y_t} - q^{(u)}\big) \in \mathbb{R}^{c-1}.$$

Also define

$$W_t(a,u) := \mathbb{1}\{t \in \mathcal{S},\, U_t = u,\, X_{t-\delta} = a\}, \qquad M_{a,n}(u) := \sum_{t=1}^{n} W_t(a,u)\,\xi_t^{(u)}.$$

Stacking over (u,a) gives a martingale

$$M_n^{\text{tot}} := \big(M_{a,n}(u)^\top\big)_{u\in\mathcal{U},\, a\in[r]} \in \mathbb{R}^{|\mathcal{U}|\,r(c-1)}.$$

Exactly as in Lemma D.4, applying Lemma C.9 blockwise shows that M_n^{tot} is a martingale. Moreover, its predictable quadratic variation is block diagonal: for fixed u, the within-stratum (a,a')-blocks are the same as in equation 6, and for $u \neq u'$ the cross-stratum blocks vanish because U_t can take only one value at time t, so $W_t(a,u)W_t(a',u') = 0$ identically. Therefore

$$\langle M^{\text{tot}}\rangle_n = \operatorname{diag}\big(\operatorname{diag}(N_{1\cdot}(u),\dots,N_{r\cdot}(u)) \otimes I_{c-1}\big)_{u\in\mathcal{U}},$$

where $N_{a\cdot}(u)$ denotes the row count in stratum u.

Let

$$U_n^{\text{tot}} := \langle M^{\text{tot}}\rangle_n^{-1/2} M_n^{\text{tot}}.$$

By equation 5, $\langle M^{\text{tot}}\rangle_n / s_n$ converges in probability to the deterministic positive definite block matrix $\operatorname{diag}(\pi_{au})_{u,a} \otimes I_{c-1}$. The increments are uniformly bounded and vanish off $\mathcal{S}_n$. Lemma G.1 gives

$$U_n^{\text{tot}} \Rightarrow Z^{\text{tot}} \sim \mathcal{N}(0, I_{|\mathcal{U}|r(c-1)}).$$

For each $u \in \mathcal{U}$, let $A_n(u)$ be the projection matrix from equation 10 formed with the row counts in that stratum, and let

$$A_n^{\text{tot}} := \operatorname{diag}\big(A_n(u)\big)_{u\in\mathcal{U}}.$$

By Lemma D.7,

$$\chi^2_{n,q^{(u)}}(u) = U_n(u)^\top A_n(u) U_n(u),$$

so after stacking,

$$\sum_{u\in\mathcal{U}} \chi^2_{n,q^{(u)}}(u) = (U_n^{\text{tot}})^\top A_n^{\text{tot}} U_n^{\text{tot}}.$$

Each $A_n(u)$ is a symmetric idempotent matrix of rank $(r-1)(c-1)$, hence A_n^{tot} is symmetric, idempotent, and has rank $|\mathcal{U}|(r-1)(c-1)$.

By the row-fraction limits, $A_n^{\text{tot}} \to A^{\text{tot}}$ in probability, where A^{tot} is a deterministic orthogonal projection of rank $|\mathcal{U}|(r-1)(c-1)$. Slutsky's theorem yields

$$(U_n^{\text{tot}})^\top A_n^{\text{tot}} U_n^{\text{tot}} \Rightarrow (Z^{\text{tot}})^\top A^{\text{tot}} Z^{\text{tot}} \sim \chi^2_{|\mathcal{U}|(r-1)(c-1)}.$$

Thus

$$\sum_{u\in\mathcal{U}} \chi^2_{n,q^{(u)}}(u) \Rightarrow \chi^2_{|\mathcal{U}|(r-1)(c-1)}.$$

Finally, Lemma D.13 applies in each stratum, and since $\mathcal{U}$ is finite,

$$\chi^2_{\text{total}} - \sum_{u\in\mathcal{U}} \chi^2_{n,q^{(u)}}(u) = \sum_{u\in\mathcal{U}} \big(\chi^2_n(u) - \chi^2_{n,q^{(u)}}(u)\big) = o_p(1).$$

This proves the null statement.

For the alternative statement, if Assumption C.7 and the positivity condition of Theorem E.9 hold in at least one stratum $u_\star$, then Theorem E.9 gives

$$\chi^2_n(u_\star) \xrightarrow{p} \infty.$$

All other summands are nonnegative, so

$$\chi^2_{\text{total}} = \chi^2_n(u_\star) + \sum_{u\neq u_\star} \chi^2_n(u) \geq \chi^2_n(u_\star) \xrightarrow{p} \infty.$$

□

Remark E.13 (Why the analysis stops at divergence). For the fixed-alternative conditional-independence test studied here, null calibration plus divergence is enough: Theorem D.1 supplies the asymptotic level, and Theorem E.9 supplies consistency. A noncentral-χ^2 description under local alternatives would be useful for weak-signal power approximations, but it is not needed for the validity of the test itself.

F Blocking and run-length (optional sufficient conditions)

The core results in Section D and Theorem E.9 do *not* use a shortfall bound. Their count limits follow from the standing mixing and support assumptions. The present section gives quantitative bounds for verifying table regularity at a given run length.

F.1 Explicit lower bound on the signal margin

For each $a \in [r]$, let $p_a = (p_{a1}, \ldots, p_{ac})$ be the row pmf from Assumption C.7. For any weight vector $w = (w_1, \ldots, w_r)$ in the probability simplex, define the mixture column marginal

$$\bar{q}_b(w) := \sum_{a=1}^{r} w_a\, p_{ab}, \qquad b \in [c],$$

and the weighted Pearson divergence of the rows from their mixture:

$$\Lambda(w) := \sum_{a=1}^{r} w_a \sum_{b=1}^{c} \frac{(p_{ab} - \bar{q}_b(w))^2}{\bar{q}_b(w)}.$$

Under the structural alternative (Assumption C.7), $\Lambda(w) > 0$ for every w in the interior of the simplex (Lemma E.3(i)).

The signal margin relevant to the run-length bound is obtained by minimizing Λ over the set of row-weight vectors that the blocking argument guarantees with high probability. Define

$$\Lambda_{\min} \;:=\; \min\left\{ \Lambda(w) :\; w \in \mathbb{R}^r,\; \sum_{a=1}^{r} w_a = 1,\; \min_{1\leq a\leq r} w_a \geq \frac{\pi_{\min}}{2} \right\}, \tag{15}$$

where

$$\pi_u := \mathbb{P}(U_t = u), \qquad \pi_{\min} := \min_{1\leq a\leq r} \mathbb{P}(U_t = u,\, X_{t-\delta} = a).$$

The constraint set in equation 15 is compact (closed and bounded) and every point lies in the interior of the simplex (since $\pi_{\min}/2 > 0$), so by Lemma E.3(i) the function Λ is continuous and strictly positive on it. By the extreme-value theorem the minimum is attained, and Lemma E.3(i) gives $\Lambda(w) > 0$ at every point of the set; hence $\Lambda_{\min} > 0$.

This quantity depends only on the model primitives:

- the row-conditional pmfs $\{p_{ab}\}_{a\in[r],\, b\in[c]}$ (from the structural alternative), and
- the stationary probabilities $\pi_{au} = \mathbb{P}(U_t = u, X_{t-\delta} = a)$ (from the joint process),

and is free of any regularity assumption on the observed table. The factor $1/2$ in the lower bound $\pi_{\min}/2$ arises from the shortfall bound (Theorem F.1 with $\theta = 1/2$), which guarantees $N_{a\cdot} \geq \frac{1}{2} mK\pi_{au}$ with high probability under the keep-one/skip-one blocking scheme; combined with the deterministic bound $N_u \leq |\mathcal{S}| = mK$, this yields $N_{a\cdot}/N_u \geq \pi_{au}/2 \geq \pi_{\min}/2$.

F.2 KEEP-ONE/SKIP-ONE BLOCKING AND SHORTFALL

Throughout this subsection we work abstractly: let $(\zeta_t)_{t\geq 1}$ be a strictly stationary $\{0,1\}$-valued process with mean $\pi := \mathbb{E}[\zeta_1]$. The mapping to our application—in which ζ_t is the indicator of a row, column, or row–stratum event derived from the joint process (X_t, U_t, Y_t)—is deferred to Appendix H, where the proof of Theorem F.1 is carried out in the same abstract setting.

The β-*mixing coefficient* at lag ℓ is

$$\beta(\ell) \;:=\; \sup_{n\geq 1}\; \mathbb{E}\left[\sup_{B\,\in\,\sigma(\zeta_{n+\ell},\zeta_{n+\ell+1},\dots)} \big|\mathbb{P}(B \mid \sigma(\zeta_1,\dots,\zeta_n)) - \mathbb{P}(B)\big|\right],$$

equivalently Volkonskiĭ & Rozanov (1959) the total variation between the joint law of past and future and the product of their marginals. β-mixing is also called *absolute regularity*; this is the formulation used in Berbee's coupling theorem (Appendix H). When $\zeta_t = f(Z_t)$ for a measurable f and an underlying strictly stationary finite-state irreducible aperiodic Markov chain (Z_t), $\beta(\ell)$ decays geometrically—this is exactly Assumption F.8.

Block structure. Partition $\{1,\dots,n\}$ into consecutive blocks of length m,

$$\mathcal{B}_j := \{(j-1)m+1,\dots,jm\}, \qquad j = 1,\dots,J, \quad J := \lfloor n/m \rfloor,$$

and retain every other block: the retained block index set is $\mathcal{K} := \{1,3,5,\dots,2K-1\}$ with $K := \lfloor J/2 \rfloor$. Let

$$I_j := \sum_{t\in\mathcal{B}_j} \zeta_t \in \{0,1,\dots,m\} \qquad \text{and} \qquad I_{\text{kept}} := \sum_{j\in\mathcal{K}} I_j$$

denote the per-block count and the total count over retained blocks. Under stationarity, $\mathbb{E}[I_{\text{kept}}] = mK\pi$.

Shortfall. For a tolerance parameter $\theta \in (0,1)$, the *shortfall event* is $\{I_{\text{kept}} < \theta\, mK\pi\}$: the kept count falls below a θ-fraction of its mean. The next theorem bounds the probability of this event in terms of the block length m (controlling the mixing slack via $\beta(m)$) and the retained block count K (controlling concentration).

Theorem F.1 (Shortfall bound)**.** *With the setup above, for any $\theta \in (0,1)$ and $Q := \theta mK\pi$,*

$$\mathbb{P}(I_{\text{kept}} < Q) \leq mK\beta(m) + \exp\big(-c(1-\theta)^2 K\pi\big),$$

where one may take $c = 1/2$. No integrality assumption on Q is needed.

Proof. See Appendix H. □

Remark F.2 (Separating coupling error and concentration). The two terms in the shortfall bound encode a deliberate trade. Bernstein-type concentration inequalities for Markov chains (e.g. via the spectral gap) typically take the form

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{n} \zeta_t - n\pi\Big| \geq \eta\, n\pi\Big) \;\leq\; 2\exp\Big(-c\,\eta^2\,\frac{n\pi}{\tau_{\text{mix}}}\Big),$$

i.e. the mixing time τ_{mix} enters *multiplicatively* in the exponent. Translating this into a sample-complexity statement gives $n \gtrsim \tau_{\text{mix}} \log(1/\varepsilon)/\pi$: the mixing time multiplies the budget.

Theorem F.1 separates cross-block and within-block correlation:

- *Cross-block correlation* is handled by Berbee's coupling (Appendix H): the K retained blocks are coupled to an independent copy at total variation cost $mK\beta(m)$. This is the first (*additive*) term in the bound. After coupling, the kept block counts are honestly independent.

- *Within-block correlation* is handled by an extremal reduction (Appendix H): convex functions of each correlated block count $I_j \in [0, m]$ are bounded in expectation by their two-point counterparts on $\{0, m\}$ with the same mean. Applying this to the exponential gives the probability bound (the conclusion of Appendix H). The result is a sum of independent Bernoullis, to which a textbook Chernoff bound applies—producing the second term, with exponent $K\pi$. Here $K = \lfloor n/(2m) \rfloor$, so the block length still affects concentration.

For geometrically mixing chains ($\beta(\ell) \le C_\beta \rho^\ell$), choosing $m = \Theta(\tau_{\mathrm{mix}} \log n)$ with a sufficiently large constant makes the coupling slack negligible. The resulting run-length bound equation 24 contains a term of order $\tau_{\mathrm{mix}} \log n$ for exploration; see equation 25. This decomposition does not establish a sharper bound than direct Markov concentration.

F.3 How the shortfall bound is used

The shortfall theorem controls the probability that a *single event* appears too few times in the retained blocks. To turn it into Assumption B.1, one needs a model-specific representation of each relevant row/column event as an event to which Theorem F.1 applies.

Assumption F.3 (Single-stratum blocking hypotheses)**.** Fix $u \in \mathcal{U}$ and adopt the keep-one/skip-one scheme of §F (so $|\mathcal{S}| = mK$).

(i) *(Shortfall applicability)* For every row event $A_a := \{U_t = u,\, X_{t-\delta} = a\}$, $a \in [r]$, and every column event $B_b := \{U_t = u,\, Y_t = b\}$, $b \in [c]$, there is an underlying stationary blocked process to which Theorem F.1 applies with the same parameters (m, K) and the same mixing coefficient bound $\beta(m)$.

(ii) *(Positivity)* $\pi^X_{\min} := \min_{1 \le a \le r} \mathbb{P}(A_a) > 0$ and $\pi^Y_{\min} := \min_{1 \le b \le c} \mathbb{P}(B_b) > 0$; write $\pi_{\min} := \min(\pi^X_{\min}, \pi^Y_{\min})$.

(iii) *(Rate)* $mK \to \infty$, $mK\,\beta(m) \to 0$, and $K\pi_{\min} \ge C_1 \log\big((r+c)/\varepsilon_n\big)$ for some $\varepsilon_n \downarrow 0$ and a sufficiently large absolute constant C_1.

Proposition F.4 (Route from blocking to Assumption B.1)**.** *Under Assumption F.3, Assumption B.1 holds with*

$$\varepsilon_{\mathrm{row}} := \frac{\pi^X_{\min}}{2}, \qquad c_0 := \frac{\pi^Y_{\min}}{2}.$$

Proof. Apply Theorem F.1 with $\theta = \frac{1}{2}$ separately to each row event A_a and each column event B_b.

For each $a \in [r]$, Theorem F.1 gives

$$\mathbb{P}\left(N_{a\cdot} < \frac{1}{2}\, mK\, \mathbb{P}(A_a) \right) \le mK\, \beta(m) + \exp\big(- cK\mathbb{P}(A_a)/4 \big).$$

Since $\mathbb{P}(A_a) \ge \pi^X_{\min} \ge \pi_{\min}$,

$$\mathbb{P}\left(N_{a\cdot} < \frac{1}{2}\, mK\, \pi^X_{\min} \right) \le mK\, \beta(m) + \exp\big(- cK\pi_{\min}/4 \big).$$

Likewise, for each $b \in [c]$, since $\mathbb{P}(B_b) \ge \pi^Y_{\min} \ge \pi_{\min}$,

$$\mathbb{P}\left(N_{\cdot b} < \frac{1}{2}\, mK\, \pi^Y_{\min} \right) \le mK\, \beta(m) + \exp\big(- cK\pi_{\min}/4 \big).$$

Because $K\pi_{\min} \ge C_1 \log((r+c)/\varepsilon_n)$ and C_1 is chosen large enough, the exponential term is

$$\exp\big(- cK\pi_{\min}/4 \big) = O\left(\frac{\varepsilon_n}{r+c} \right).$$

Since also $mK\,\beta(m) \to 0$, a union bound over the $r + c$ events yields that with probability at least $1 - \varepsilon_n - o(1)$,

$$N_{a\cdot} \ge \frac{1}{2}\, mK\, \pi^X_{\min} \qquad \text{for all } a \in [r], \tag{16}$$

and

$$N_{\cdot b} \ge \frac{1}{2}\, mK\, \pi^Y_{\min} \qquad \text{for all } b \in [c]. \tag{17}$$

We now verify the three parts of Assumption B.1.

First, recalling $m_n := \min_{1\le a\le r} N_{a\cdot}$, equation 16 implies that on the high-probability event above,

$$m_n \ge \frac{1}{2}\, mK\, \pi^X_{\min}.$$

Since $mK \to \infty$, it follows that

$$m_n \xrightarrow{p} \infty.$$

Second, because $N_u = \sum_{a=1}^{r} N_{a\cdot} \le |\mathcal{S}| = mK$ under the keep-one/skip-one scheme, equation 16 also gives, for every a,

$$\frac{N_{a\cdot}}{N_u} \ge \frac{N_{a\cdot}}{mK} \ge \frac{\pi^X_{\min}}{2}$$

on the same high-probability event. Hence

$$\min_{1\le a\le r} \frac{N_{a\cdot}}{N_u} \ge \frac{\pi^X_{\min}}{2}$$

with probability $1 - o(1)$.

Third, by definition $\widehat{q}_b = N_{\cdot b}/N_u$. Using again $N_u \le mK$ and equation 17,

$$\widehat{q}_b = \frac{N_{\cdot b}}{N_u} \ge \frac{N_{\cdot b}}{mK} \ge \frac{\pi^Y_{\min}}{2}$$

for every b on the same high-probability event. Therefore

$$\min_{1\le b\le c} \widehat{q}_b \ge \frac{\pi^Y_{\min}}{2}$$

with probability $1 - o(1)$.

Thus all three requirements in Assumption B.1 hold, with

$$\varepsilon_{\mathrm{row}} = \frac{\pi^X_{\min}}{2}, \qquad c_0 = \frac{\pi^Y_{\min}}{2}.$$

□

Remark F.5 (Interpretation). This is exactly where a shortfall argument enters: it is an external device for checking table regularity, not part of the martingale proof of the null or the alternative analysis. If Assumption B.1 is available from some other source, Section F may be omitted entirely.

Sections F.4 and F.5 carry out this program under geometric β-mixing.

Assumption F.6 (Uniform blocking hypotheses)**.** Assume $\mathcal{U}$ is finite. For each $u \in \mathcal{U}$, $a \in [r]$, and $b \in [c]$, define the row and column events

$$A_{u,a} := \{U_t = u,\, X_{t-\delta} = a\}, \qquad B_{u,b} := \{U_t = u,\, Y_t = b\}.$$

Adopt the keep-one/skip-one scheme of §F (so $|\mathcal{S}| = mK$).

(i) *(Shortfall applicability)* For every event in the finite collection $\mathscr{E} := \{A_{u,a} : u \in \mathcal{U},\, a \in [r]\} \cup \{B_{u,b} : u \in \mathcal{U},\, b \in [c]\}$, there is an underlying stationary blocked process to which Theorem F.1 applies with the same parameters (m, K) and the same mixing coefficient bound $\beta(m)$.

(ii) *(Positivity)* $\pi^X_{\min} := \min_{u\in\mathcal{U}} \min_{1\le a\le r} \mathbb{P}(A_{u,a}) > 0$ and $\pi^Y_{\min} := \min_{u\in\mathcal{U}} \min_{1\le b\le c} \mathbb{P}(B_{u,b}) > 0$.

(iii) *(Rate)* $mK \to \infty$, $mK\,\beta(m) \to 0$, and $K \min(\pi^X_{\min}, \pi^Y_{\min}) \ge C_1 \log\big(|\mathcal{U}|(r+c)/\varepsilon_n\big)$ for some $\varepsilon_n \downarrow 0$ and a sufficiently large absolute constant C_1.

Proposition F.7 (Uniform blocking route to Assumption E.11)**.** *Under Assumption F.6, Assumption E.11 holds with* $\varepsilon_{\mathrm{row}} := \pi^X_{\min}/2$ *and* $c_0 := \pi^Y_{\min}/2$*. Moreover, with probability at least* $1 - \varepsilon_n - o(1)$*, simultaneously for all* $u \in \mathcal{U}$*,* $a \in [r]$*, and* $b \in [c]$*,*

$$N_{a\cdot}(u) \ge \frac{1}{2}\, mK\, \mathbb{P}(A_{u,a}), \tag{18}$$

and

$$N_{\cdot b}(u) \ge \frac{1}{2}\, mK\, \mathbb{P}(B_{u,b}). \tag{19}$$

Proof. Apply Theorem F.1 with $\theta = \frac{1}{2}$ separately to each event in $\mathscr{E}$. For every row event $A_{u,a}$,

$$\mathbb{P}\Big(N_{a\cdot}(u) < \frac{1}{2}\, mK\, \mathbb{P}(A_{u,a})\Big) \le mK\, \beta(m) + \exp\big(-cK\mathbb{P}(A_{u,a})/4\big),$$

and likewise for every column event $B_{u,b}$,

$$\mathbb{P}\Big(N_{\cdot b}(u) < \frac{1}{2}\, mK\, \mathbb{P}(B_{u,b})\Big) \le mK\, \beta(m) + \exp\big(-cK\mathbb{P}(B_{u,b})/4\big).$$

Because $\mathcal{U}$ is finite and the probabilities are bounded below by $\pi^X_{\min}$ and $\pi^Y_{\min}$, the lower bound on $K\min(\pi^X_{\min}, \pi^Y_{\min})$ ensures that each exponential term is $O(\varepsilon_n/(|\mathcal{U}|(r+c)))$. Since $mK\,\beta(m) \to 0$, a union bound over the $|\mathcal{U}|r + |\mathcal{U}|c$ events in $\mathscr{E}$ yields equation 18–equation 19 simultaneously with probability $1 - \varepsilon_n - o(1)$.

It remains to verify Assumption E.11. First,

$$m_n(u) := \min_{1\le a\le r} N_{a\cdot}(u) \ge \frac{1}{2}\, mK\, \pi^X_{\min}$$

for every u on the high-probability event above, so

$$\min_{u\in\mathcal{U}} m_n(u) \xrightarrow{p} \infty$$

because $mK \to \infty$.

Next, for every u and a,

$$\frac{N_{a\cdot}(u)}{N_u} \ge \frac{N_{a\cdot}(u)}{|\mathcal{S}|} \ge \frac{\frac{1}{2}\, mK\, \mathbb{P}(A_{u,a})}{mK} = \frac{1}{2}\, \mathbb{P}(A_{u,a}) \ge \frac{1}{2}\, \pi^X_{\min},$$

since $N_u \le |\mathcal{S}| = mK$. Thus

$$\min_{u\in\mathcal{U}} \min_{1\le a\le r} \frac{N_{a\cdot}(u)}{N_u} \ge \frac{1}{2}\, \pi^X_{\min}$$

with probability $1 - \varepsilon_n - o(1)$.

Similarly, for every u and b,

$$\widehat{q}_b(u) = \frac{N_{\cdot b}(u)}{N_u} \ge \frac{N_{\cdot b}(u)}{|\mathcal{S}|} \ge \frac{\frac{1}{2}\, mK\, \mathbb{P}(B_{u,b})}{mK} = \frac{1}{2}\, \mathbb{P}(B_{u,b}) \ge \frac{1}{2}\, \pi^Y_{\min},$$

so

$$\min_{u\in\mathcal{U}} \min_{1\le b\le c} \widehat{q}_b(u) \ge \frac{1}{2}\, \pi^Y_{\min}$$

with probability $1 - \varepsilon_n - o(1)$.

These are exactly the three requirements in Assumption E.11. □

F.4 Explicit run-length conditions

We now specialize the schematic route of Proposition F.4 to obtain an explicit sample-size requirement for table regularity and an IID benchmark for the stratumwise test $\phi_n(u) = \mathbb{1}\{\chi^2_n(u) > c_{k,\alpha}\}$ of Corollary E.10. Throughout this subsection we assume stationarity of the joint process $(X_{t-\delta}, U_t, Y_t)$ and geometric β-mixing.

Assumption F.8 (Stationarity and geometric mixing)**.** The process $(X_{t-\delta}, U_t, Y_t)$ is strictly stationary and has β-mixing coefficients satisfying $\beta(\ell) \le C_\beta \rho^\ell$ for constants $C_\beta \ge 1$ and $\rho \in (0,1)$. Define the mixing scale $\tau_{\mathrm{mix}} := (-\log\rho)^{-1}$.

Under stationarity, define the stratum probability

$$\pi_u := \mathbb{P}(U_t = u), \tag{20}$$

and for each row $a \in [r]$ the row–stratum probability

$$\pi_{au} := \mathbb{P}(U_t = u,\ X_{t-\delta} = a). \tag{21}$$

Let $\pi_{\min} := \min_{a\in[r]} \pi_{au}$. Note that $\sum_{a=1}^r \pi_{au} = \pi_u$ and $\pi_{\min} \le \pi_u/r$.

Noncentrality requirement. Recall from Theorem E.9 that under the structural alternative, $\chi^2_n(u) = \lambda^{\mathrm{orc}}_n + O_p(\sqrt{\lambda^{\mathrm{orc}}_n})$, and from Lemma E.3 that $\lambda^{\mathrm{orc}}_n \geq \Lambda_{\min} N_u$ with probability $1 - o(1)$. Define, for degrees of freedom $k = (r-1)(c-1)$ and prescribed type-I error α and type-II error β,

$$\lambda_\star(\alpha, \beta, k) := \inf\{\lambda \geq 0 : \mathbb{P}\big(\chi^2_k(\lambda) > c_{k,\alpha}\big) \geq 1 - \beta\}, \tag{22}$$

where $\chi^2_k(\lambda)$ denotes a noncentral chi-square with k degrees of freedom and noncentrality λ, and $c_{k,\alpha}$ is the $(1-\alpha)$-quantile of the central χ^2_k. Using the i.i.d. noncentral-chi-square approximation gives the benchmark $N_u \geq \lambda_\star / \Lambda_{\min}$ for power $1 - \beta$; define the baseline sample-size benchmark

$$n_{\mathrm{iid}} := \frac{\lambda_\star(\alpha, \beta, k)}{\Lambda_{\min}}. \tag{23}$$

We use $\lambda_\star$ and n_{iid} throughout this subsection as a *benchmark* inherited from the i.i.d. case: they pin down the constant in the leading term of equation 24 by reference to the i.i.d. noncentrality at which a noncentral $\chi^2_k(\lambda_\star)$ achieves power $1 - \beta$. Our actual guarantees in Theorem F.10 are finite-run table regularity and asymptotic consistency ($\mathbb{P}_{H_1}(\phi_n = 1) \to 1$); a finite-$\beta$ power statement for our test would require the noncentral-χ^2 approximation discussed in Remark F.11, which we do not establish. The benchmark is included because it provides a useful comparison to the i.i.d. baseline and motivates the form of the bound, not because it is itself a guarantee of finite-β power.

Block parameters. In the keep-one/skip-one scheme, time is partitioned into consecutive blocks of length m, and every other block is retained. With $J = \lfloor n/m \rfloor$ total blocks and $K = \lfloor J/2 \rfloor$ retained blocks, the deterministic used-index set $\mathcal{S}$ consists of the time points inside the K retained blocks, so $|\mathcal{S}| = mK \leq n/2$.

The following lemma records the standard choice of block length that makes the coupling error negligible.

Lemma F.9 (Block-length choice)**.** *Under Assumption F.8, if*

$$m \geq \tau_{\mathrm{mix}} \log\Big(\frac{2C_\beta n}{\varepsilon}\Big),$$

then $mK\,\beta(m) \leq \varepsilon/2$.

Proof. By $mK \leq n/2$ and $\beta(m) \leq C_\beta \rho^m$,

$$mK\,\beta(m) \leq \tfrac{n}{2}\, C_\beta \rho^m.$$

The stated lower bound on m gives $\rho^m \leq \varepsilon/(nC_\beta)$, so the product is at most $\varepsilon/2$. □

Theorem F.10 (Explicit run length for regularity and the IID benchmark)**.** *Let Assumptions C.1, C.3, and F.8 hold. Suppose $\delta \geq 1$ and that the structural alternative (Assumption C.7) holds in stratum u with signal margin $\Lambda_{\min} > 0$ (Lemma E.3). Assume $p_{ab} \geq p_{\min} > 0$ for all a, b (so the positivity condition of Theorem E.9 is satisfied with $\bar{c}_0 = p_{\min}$). Fix $\alpha, \beta \in (0,1)$ and let $\varepsilon \in (0,1)$ be a failure budget for the regularity conditions. Write $R := r + c$ and choose*

$$m := \max\left\{1, \left\lceil \tau_{\mathrm{mix}} \log\Big(\frac{2RC_\beta n}{\varepsilon}\Big) \right\rceil\right\}.$$

If the total run length satisfies

$$n \;\geq\; \frac{4\,\lambda_\star(\alpha, \beta, k)}{\Lambda_{\min}\,\pi_u} \;+\; 2m\left(1 + \frac{C_1 \log\big(\frac{2(r+c)}{\varepsilon}\big)}{\pi_{\min} p_{\min}}\right), \tag{24}$$

for an absolute constant $C_1 \geq 8$, then:

(i) *with probability at least $1 - \varepsilon$, $N_{a\cdot} \geq mK\pi_{au}/2$ for every row, $\widehat{q}_b \geq \pi_{\min} p_{\min}/2$ for every column, and $N_u \geq n_{\mathrm{iid}}$;*

(ii) *as $n \to \infty$ with $mK \to \infty$, the test $\phi_n(u) = \mathbb{1}\{\chi^2_n(u) > c_{k,\alpha}\}$ is consistent under the stated alternative. Under the corresponding structural null and standing support assumptions, it has asymptotic size α. (A finite-β power statement is discussed in Remark F.11 below.)*

Proof. The integer block parameters satisfy $K = \lfloor n/(2m) \rfloor$ and $mK \geq n/2 - m$.

Part (i): verifying regularity. The stated choice of m, by Lemma F.9 with budget ε/R, gives $RmK\,\beta(m) \leq \varepsilon/2$. It remains to control the exponential shortfall terms for all rows and columns.

Apply Theorem F.1 to each of the r row events $\{U_t = u, X_{t-\delta} = a\}$ (with stationary probability π_{au}) and to each of the c column events $\{U_t = u, Y_t = b\}$. Under the structural alternative, the column-event probabilities satisfy $\mathbb{P}(U_t = u, Y_t = b) = \sum_a \pi_{au} p_{ab} \geq \pi_{\min} p_{\min}$. Taking $\theta = 1/2$ in Theorem F.1 and allocating failure probability $\varepsilon/(2(r+c))$ to each event, the exponential term is at most $\varepsilon/(2(r+c))$ provided

$$K \pi_{\min} p_{\min} \;\geq\; C_1 \log\Big(\frac{2(r+c)}{\varepsilon}\Big),$$

which follows from equation 24 and $K \geq n/(2m) - 1$. Here Theorem F.1 permits $c = 1/2$, so $C_1 \geq 8$ suffices. A union bound over the $r + c$ events gives simultaneous lower bounds on $N_{a\cdot}$ and $N_{\cdot b}$ with probability at least $1 - \varepsilon/2 - RmK\,\beta(m) \geq 1 - \varepsilon$. On this event $N_{a\cdot} \geq \frac{1}{2} mK \pi_{au}$ for all rows. The row-fraction bound $N_{a\cdot}/N_u \geq \pi_{\min}/2 =: \varepsilon_{\text{row}}$ (using $N_u \leq |\mathcal{S}| = mK$ under keep-one/skip-one) and the column-marginal bound $\widehat{q}_b \geq c_0$ with $c_0 := \pi_{\min} p_{\min}/2$ follow from the same shortfall lower bounds (cf. Proposition F.4). Summing the row bounds and using $mK \geq n/2 - m$ gives

$$N_u \geq \tfrac{1}{2} mK \pi_u \geq (n/4 - m/2)\pi_u \geq n_{\text{iid}},$$

by equation 24. This proves (i).

Part (ii): size and consistency. By equation 5, $N_u/(mK) \to \pi_u > 0$ and the row and column fractions have positive limits. Thus Assumption B.1 holds as $mK \to \infty$, independently of the fixed finite-run failure budget ε. Under the null, Theorem D.1 gives asymptotic size α. Under the fixed alternative, $\Lambda_{\min} N_u \to \infty$ in probability, and Theorem E.9 gives $\chi^2_n(u) = \lambda^{\text{orc}}_n + O_p(\sqrt{\lambda^{\text{orc}}_n})$ with $\lambda^{\text{orc}}_n \geq \Lambda_{\min} N_u$ with probability tending to one. Therefore $\chi^2_n(u) \xrightarrow{p} \infty$, while the critical value $c_{k,\alpha}$ is fixed, so $\mathbb{P}_{H_1}(\phi_n(u) = 1) \to 1$. □

Remark F.11 (On finite-β power and the constant 4). The additional $2m$ in equation 24 accounts for integer rounding. The factor of 4 in the leading term of equation 24 has two sources: half the blocks are discarded in the keep-one/skip-one scheme, and the shortfall threshold $\theta = 1/2$ in Theorem F.1 costs another factor of two. This guarantees $\Lambda_{\min} N_u \geq \lambda_\star(\alpha, \beta, k)$, which is the i.i.d. noncentrality at which a noncentral $\chi^2_k(\lambda_\star)$ achieves power $1 - \beta$.

Translating this i.i.d. noncentrality requirement into a finite-β power statement for our test would require a noncentral-χ^2 approximation $\chi^2_n(u) \approx \chi^2_k(\lambda^{\text{orc}}_n)$ in distribution. Theorem E.9 only delivers $\chi^2_n(u) = \lambda^{\text{orc}}_n + O_p(\sqrt{\lambda^{\text{orc}}_n})$, which is enough for consistency but not for a precise distributional statement; the cross term $2\nu_n^\top A_n Z_n$ in the proof of Theorem E.9 is of order $\sqrt{\lambda^{\text{orc}}_n}$, and a noncentral-$\chi^2$ approximation would require a separate distributional argument under appropriate local alternatives. We do not pursue this here; the consistency statement above suffices for the validity of the test.

F.5 Comparison to the i.i.d. case and the role of mixing time

The bound equation 24 consists of two terms with distinct roles:

Signal-accumulation term. The first term,

$$\frac{4\,\lambda_\star(\alpha, \beta, k)}{\Lambda_{\min}\,\pi_u} \;=\; \frac{4\,n_{\text{iid}}}{\pi_u},$$

is the number of time steps required so that the in-stratum sample size N_u reaches the i.i.d. benchmark n_{iid}, i.e. the sample size suggested by the i.i.d. noncentral-chi-square approximation for power $1 - \beta$. As discussed in Remark F.11, our test is only guaranteed to be consistent here, so this should be read as a benchmark comparison rather than a finite-β power statement for our test. Relative to the i.i.d. baseline n_{iid}, the term is inflated by $1/\pi_u$ (stratum rarity: only a π_u-fraction of time steps fall in stratum u) and by a factor of 4 ($\times 2$ from discarding half the blocks in the keep-one/skip-one scheme, and $\times 2$ from the shortfall threshold $\theta = 1/2$ in Theorem F.1).

Mixing/regularity term. The second term,

$$2m\left(1 + \frac{C_1 \log\big(\frac{2(r+c)}{\varepsilon}\big)}{\pi_{\min} p_{\min}}\right),$$

ensures that the block length m is large enough for the coupling argument and that there are enough retained blocks for the shortfall bound. This term depends on the mixing time τ_{mix} only through $\tau_{\text{mix}} \log n$, up to integer rounding; the dependence on the mixing scale is linear. The factor $1/(\pi_{\min} p_{\min})$ ensures that even the rarest row and column events receive enough visits.

Combined scaling. Setting ε to a small constant, the total run length satisfies

$$n \;=\; \underbrace{\frac{4n_{\mathrm{iid}}}{\pi_u}}_{\text{signal}} \;+\; \underbrace{O\Big(\frac{1+\tau_{\mathrm{mix}}\,\log n}{\pi_{\min}p_{\min}}\Big)}_{\text{mixing/regularity}}, \tag{25}$$

up to constants depending on $C_\beta, \varepsilon, c, r, c_0$. Recall that $\pi_{\min} \le \pi_u/r$ in general (with equality iff the row distribution conditional on $U_t = u$ is uniform); under that "balanced" regime the mixing/regularity term simplifies to $O(r(1+\tau_{\mathrm{mix}}\log n)/(\pi_u p_{\min}))$, but in the unbalanced case it scales with $1/\pi_{\min}$ and may be substantially larger. In typical regimes where the signal term dominates ($n_{\mathrm{iid}} \gg \pi_u(1+\tau_{\mathrm{mix}}\log n)/(\pi_{\min}p_{\min})$), the run length is controlled by n_{iid}/π_u and the mixing-time contribution is an additive correction, linear in the mixing scale and logarithmic in n.

Remark F.12 (Heuristic near-optimality of the scaling). We do not establish a formal minimax lower bound for our setting. However, both multiplicative factors in the signal term are heuristically unavoidable, by analogy with the i.i.d. case:

(i) *The $1/\pi_u$ factor.* Only a π_u-fraction of observations fall in stratum u, and the in-stratum sample size N_u is what controls detectability. Accumulating the benchmark n_{iid} in-stratum observations requires n_{iid}/π_u total observations in expectation in the i.i.d. sampling comparison. A formal lower bound for our dependent setting would require a separate minimax analysis, which we do not pursue.

(ii) *The mixing-time floor.* Even with stationary initialization, slowly switching models can take order τ_{mix} steps to visit an informative stratum. This motivates an exploration term but is not a universal lower bound.

A positive signal margin $\Lambda_{\min} > 0$ is the sufficient separation condition used in our consistency proof. Theorem E.9 does not assert a converse when this condition fails.

G Helpful Lemmas

G.1 Martingale limit tools

Lemma G.1 (Martingale CLT with deterministic limiting variation)**.** *For each n, let $(\Delta M_{n,t}, \mathcal{F}_t)_{1\le t\le n}$ be martingale differences in $\mathbb{R}^d$, and set*

$$M_n := \sum_{t=1}^{n} \Delta M_{n,t}, \qquad Q_n := \sum_{t=1}^{n} \mathbb{E}[\Delta M_{n,t}\Delta M_{n,t}^\top \mid \mathcal{F}_{t-1}].$$

Let $\mathcal{S}_n \subseteq \{1,\dots,n\}$ be deterministic with $s_n := |\mathcal{S}_n| \to \infty$. Suppose $\|\Delta M_{n,t}\|_2 \le B\mathbb{1}\{t \in \mathcal{S}_n\}$ for a fixed finite B, and $Q_n/s_n \to V$ in probability for a deterministic positive definite matrix V. Then

$$M_n/\sqrt{s_n} \Rightarrow \mathcal{N}(0, V), \qquad Q_n^{-1/2} M_n \Rightarrow \mathcal{N}(0, I_d).$$

The studentized vector can be defined arbitrarily when Q_n is singular; that event has probability tending to zero.

Proof. For any fixed vector v, the differences $v^\top \Delta M_{n,t}/\sqrt{s_n}$ have maximum absolute value at most $B\|v\|_2/\sqrt{s_n} \to 0$. Their conditional Lindeberg sums are therefore eventually zero, and their total conditional variance tends to $v^\top V v$. The scalar martingale central limit theorem (Hall & Heyde, 1980, Ch. 3) and the Cramer–Wold device give the first limit. Since $(Q_n/s_n)^{-1/2} \to V^{-1/2}$ in probability, Slutsky's theorem gives the second limit. □

Remark G.2. The deterministic positive definite limit of Q_n/s_n is essential to this argument. Divergence of the smallest eigenvalue of Q_n alone is not the hypothesis used here.

Lemma G.3 (Self-normalized bound for scalar martingales)**.** *Let $(M_n, \mathcal{F}_n)$ be a scalar martingale starting at zero with increments satisfying $|\Delta M_t| \le 1$ a.s. Let V_n be a nonnegative $\mathcal{F}_n$-measurable random variable such that*

$$\langle M\rangle_n \le V_n \quad \text{a.s.}$$

Assume that, for a deterministic sequence $s_n \to \infty$ and a constant $v > 0$, $V_n/s_n \xrightarrow{p} v$ and $\mathbb{E}[V_n] = O(s_n)$. Then $M_n = O_p(\sqrt{V_n})$. The same statement holds for a triangular array of martingale differences.

Proof. Orthogonality gives $\mathbb{E}[M_n^2] = \mathbb{E}[\langle M\rangle_n] \le \mathbb{E}[V_n] = O(s_n)$, so Chebyshev's inequality gives $M_n = O_p(\sqrt{s_n})$. Since $V_n/s_n \to v > 0$ in probability, $M_n = O_p(\sqrt{V_n})$. □

G.2 Auxiliary linear-algebra facts

Lemma G.4 (Gaussian quadratic form with (possibly random) projection)**.** *Let $\mathcal{G}$ be a sigma-field. Let A be a $\mathcal{G}$-measurable random matrix such that A is a.s. symmetric and idempotent ($A^2 = A$) with $\operatorname{rank}(A) = k$ a.s. Let $Z \sim \mathcal{N}(0, I_d)$ be independent of $\mathcal{G}$. Then conditionally on $\mathcal{G}$,*

$$Z^\top AZ \;\sim\; \chi^2_k,$$

and hence unconditionally $Z^\top AZ \sim \chi^2_k$.

Proof. Fix an outcome of $\mathcal{G}$ so that A is deterministic, symmetric, idempotent with rank k. There exists an orthogonal matrix Q such that $A = Q^\top \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} Q$. By orthogonal invariance, $QZ \sim \mathcal{N}(0, I_d)$, and therefore

$$Z^\top AZ = (QZ)^\top \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} (QZ) = \sum_{i=1}^{k} (QZ)_i^2 \sim \chi^2_k.$$

Since this holds for every $\mathcal{G}$-outcome, it holds conditionally on $\mathcal{G}$. The unconditional law is obtained via the law of total probability: since the conditional distribution χ^2_k does not vary with $\mathcal{G}$ (crucially, the rank k is constant a.s.), the unconditional distribution is also χ^2_k. □

H Proof of the shortfall bound

This appendix provides a complete proof of Theorem F.1. The argument is self-contained and relies on Berbee's coupling for β-mixing sequences, together with convexity and a Poisson–binomial tail bound.

H.1 Setup and notation

We retain the abstract setup of §F: $(\zeta_t)_{t\geq 1}$ is a strictly stationary $\{0,1\}$-valued process with $\pi := \mathbb{E}[\zeta_1]$ and β-mixing coefficients $\beta(\ell)$ as defined in §F. For our application, $\zeta_t = \mathbb{1}\{Z_t \in A\}$ for some event A in an underlying strictly stationary process (Z_t) (e.g. a row, column, or row–stratum event derived from $(X_{t-\delta}, U_t, Y_t)$ in Assumption F.8); the proof below depends only on the σ-fields generated by (ζ_t) itself.

Partition $\{1, \dots, n\}$ into consecutive blocks of length m,

$$\mathcal{B}_j := \{(j-1)m + 1, \dots, jm\}, \qquad j = 1, \dots, J.$$

Let

$$I_j := \sum_{t \in \mathcal{B}_j} \zeta_t \in \{0, 1, \dots, m\}$$

be the block counts. Retain every other block, with index set

$$\mathcal{K} := \{1, 3, 5, \dots, 2K-1\}, \qquad K = \lfloor J/2 \rfloor,$$

and define the kept total

$$I_{\text{kept}} := \sum_{j \in \mathcal{K}} I_j.$$

Fix a target $Q \in [0, mK]$ and define the shortfall

$$L := (Q - I_{\text{kept}})_+.$$

H.2 Berbee coupling across blocks

Because the retained blocks are separated by at least one full block of length m, their σ-fields are separated by at least m time steps. Let $\mathcal{F}_j := \sigma(\zeta_t : t \in \mathcal{B}_j)$.

By Berbee's coupling theorem for absolutely regular sequences Berbee (1979); Doukhan et al. (1995), there exist random variables $\{\tilde{I}_j : j \in \mathcal{K}\}$ such that:

(i) each $\tilde{I}_j$ has the same marginal distribution as I_j;

(ii) the family $\{\tilde{I}_j : j \in \mathcal{K}\}$ is independent;

(iii) $\mathbb{P}(I_j \neq \tilde{I}_j) \leq \beta(m)$ for all $j \in \mathcal{K}$.

Define $\tilde{I}_{\text{kept}} := \sum_{j\in\mathcal{K}} \tilde{I}_j$.

H.3 Lipschitz control of the shortfall

The map $\phi(x) := (Q - x)_+$ is convex and 1-Lipschitz, i.e., $|\phi(x) - \phi(y)| \leq |x - y|$. Therefore,

$$|L - (Q - \tilde{I}_{\text{kept}})_+| \leq |I_{\text{kept}} - \tilde{I}_{\text{kept}}|.$$

Taking expectations gives

$$\mathbb{E}[L] \leq \mathbb{E}[(Q - \tilde{I}_{\text{kept}})_+] + \mathbb{E}[|I_{\text{kept}} - \tilde{I}_{\text{kept}}|].$$

Since $|I_j - \tilde{I}_j| \leq m$, by Berbee coupling property,

$$\mathbb{E}[|I_{\text{kept}} - \tilde{I}_{\text{kept}}|] \leq \sum_{j\in\mathcal{K}} m\,\mathbb{P}(I_j \neq \tilde{I}_j) \leq mK\,\beta(m).$$

Hence

$$\mathbb{E}[L] \leq mK\,\beta(m) + \mathbb{E}[(Q - \tilde{I}_{\text{kept}})_+]. \tag{26}$$

H.4 Extremal reduction to Bernoulli blocks

Fix $j \in \mathcal{K}$. Conditional on the other blocks, the function $i \mapsto (Q - (s + i))_+$ is convex on $[0, m]$ for any fixed s. By Jensen's inequality applied to the convex order,

$$\mathbb{E}[(Q - (s + \tilde{I}_j))_+ \mid s] \leq \mathbb{E}[(Q - (s + \tilde{I}_j^\star))_+ \mid s],$$

where $\tilde{I}_j^\star \in \{0, m\}$ satisfies

$$\mathbb{P}(\tilde{I}_j^\star = m) = \frac{\mathbb{E}[\tilde{I}_j]}{m}.$$

Iterating this replacement over all $j \in \mathcal{K}$ yields

$$\mathbb{E}[(Q - \tilde{I}_{\text{kept}})_+] \leq \mathbb{E}[(Q - \tilde{I}^\star_{\text{kept}})_+],$$

where

$$\tilde{I}^\star_{\text{kept}} = \sum_{j\in\mathcal{K}} \tilde{I}_j^\star = m \sum_{j\in\mathcal{K}} Y_j,$$

with independent Bernoulli variables $Y_j \sim \text{Bernoulli}(p_j)$ and $p_j := \mathbb{E}[\tilde{I}_j]/m$.

H.5 Poisson–binomial tail bound

Let $B := \sum_{j\in\mathcal{K}} Y_j$ and $P_\Sigma := \mathbb{E}[B] = \sum_{j\in\mathcal{K}} p_j$. Then $\tilde{I}^\star_{\text{kept}} = mB$. For any $\theta \in (0, 1)$, set $Q = \theta m P_\Sigma$. We have

$$(Q - \tilde{I}^\star_{\text{kept}})_+ \leq Q\,\mathbb{1}\{B \leq \theta P_\Sigma\}.$$

Therefore

$$\mathbb{E}[(Q - \tilde{I}^\star_{\text{kept}})_+] \leq Q\,\mathbb{P}(B \leq \theta P_\Sigma).$$

A Chernoff bound for Poisson–binomial variables yields

$$\mathbb{P}(B \leq \theta P_\Sigma) \leq \exp\bigl(-c(1 - \theta)^2 P_\Sigma\bigr)$$

for an absolute constant $c > 0$. Consequently,

$$\mathbb{E}[(Q - \tilde{I}^\star_{\text{kept}})_+] \leq Q\,\exp\bigl(-c(1 - \theta)^2 P_\Sigma\bigr)\,. \tag{27}$$

H.6 CONCLUSION

Combining equation 26 and equation 27 gives

$$\mathbb{E}[L] \le mK\,\beta(m) + Q\,\exp\!\big(-c(1-\theta)^2 P_\Sigma\big)\,.$$

For the probability bound, including noninteger Q, apply the same convex endpoint bound to the exponential instead of the shortfall. Under stationarity $\mathbb{E}[\tilde{I}_j/m] = \pi$, and for every $h > 0$,

$$\mathbb{E}[e^{-h\tilde{I}_j/m}] \le 1 - \pi + \pi e^{-h}.$$

Independence of the coupled blocks and Markov's inequality imply

$$\mathbb{P}(\tilde{I}_{\text{kept}} < \theta m K \pi) \le \exp\{K\pi(h\theta + e^{-h} - 1)\}.$$

Taking $h = -\log\theta$ and using $1 - \theta + \theta\log\theta \ge (1-\theta)^2/2$ gives

$$\mathbb{P}(\tilde{I}_{\text{kept}} < \theta m K \pi) \le \exp\!\big(-\tfrac{1}{2}(1-\theta)^2 K\pi\big)\,.$$

The original and coupled totals agree unless at least one block coupling fails, an event of probability at most $K\beta(m) \le mK\beta(m)$. Adding this error proves Theorem F.1, with $c = 1/2$. □

I FULL ALGORITHM: PC-OP

For completeness, we show the detailed first step in a complete version of the PC-OP.

Algorithm 2 PC with One-Parent-Set (PC-OP) Algorithm for Time Series

Input: Dataset D, significance level α, Markov Order k
Output: Estimated causal graph G
Step 1: Find parents' supersets of each variable
Initialize G a complete graph as $G_{i,j} = 1, i \neq j$ and $G_{i,j} = 0, i = j$, with nodes $\{V_i\}_{i=1}^k$.
Let $n = 0$.
repeat
  for all pairs of adjacent variables (X, Y) in G **do**
    Let $Adj(X)$ be the set of vertices adjacent to X (excluding Y).
    if $|Adj(X)| \ge n$ **then**
      for all subsets $\mathbf{S} \subseteq Adj(X)$ such that $|\mathbf{S}| = n$ **do**
        if $X \perp Y \mid \mathbf{S}$ **then**
          Remove edge $X \to Y$ from G.
          Record $\mathbf{S}$ as the separating set $Sepset(X, Y)$.
          break to next pair (X, Y).
        end if
      end for
    end if
  end for
  $n = n + 1$.
until all pairs (X, Y) satisfy $|Adj(X)| < n$
{ Step 2a (Optional): verify independence needs a different threshold (higher).
$\alpha' \leftarrow \frac{\alpha}{\text{the number of missing edges}}$
for each pair of variables (X, Y) not adjacent in G **do**
  P ← Parent set of Y in G, excluding X
  test $X \perp Y \mid P$ is indeed true
end for
}
Step 2: verify dependence with two-sided tests
for each pair of variables (X, Y) adjacent in G **do**
  P ← Parent set of Y in G, excluding X
  test $X \not\perp Y \mid P$ is indeed true with α'
end for
Return G

Difference from PCMCI Our algorithm presents two major differences from the existing PCMCI algorithm (Runge et al., 2019). First, the 2nd step of our algorithm is conditioned on the parental set of Y only, instead of both X and Y in similar algorithms. This reduces the size of the conditioned set, reducing the sample requirements and improving the power of tests. Secondly, we adjust the α values, according to our analysis in previous section, to obtain more accurate test results. Our analysis establishes asymptotic validity of the CI tests for a fixed parental conditioning set; soundness and completeness of the full procedure with estimated conditioning sets are not established here.

J Detailed Experiment Settings and Results

We conduct a comprehensive set of experiments on synthetic categorical time series data to evaluate the proposed PC-OPalgorithm and compare it against several established causal discovery baselines. Our experimental design is structured to isolate the effects of key data-generating factors—including the number of variables, sample size, lag order, and graph density—on each method's ability to recover the true underlying causal graph.

J.1 Synthetic Dataset

We generate synthetic categorical time series from a structural time series model composed of two causal graph components and a set of conditional probability tables (CPTs). The data-generating process proceeds as follows.

Causal graph construction. For each experimental condition, a ground-truth causal graph is constructed using a set of *inter-slice* adjacency matrices $\{A_{\text{inter}}^{(k)}\}_{k=1}^{L}$, each in $\{0,1\}^{d\times d}$, where L denotes the Markov order (i.e., the maximum time lag). Entry $A_{\text{inter}}^{(k)}[i,j]=1$ indicates a directed causal edge from variable X_i at time $t-k$ to variable X_j at time t. Each inter-slice matrix is generated by selecting, for each target variable (column), a fixed number of parents uniformly at random from among all d candidate cause variables. The number of selected parents per variable per lag is controlled by the inter-slice in-degree parameter.

Conditional probability tables. Given the combined parent structure defined by A_{intra} and $\{A_{\text{inter}}^{(k)}\}_{k=1}^{L}$, a conditional probability table (CPT) is generated for each variable X_j. For every possible configuration of parent values, a probability vector over the C categories of X_j is drawn independently from a symmetric Dirichlet distribution with concentration parameter $\alpha = 1.0$. This choice ensures that the conditional distributions are non-degenerate and span a diverse range of dependency structures.

Time series sampling. Given the causal graph and the CPTs, a categorical time series of length $T = S \times d$ is generated, where S is a user-specified sample-size parameter and d is the number of variables. This scaling ensures that the effective number of observations per variable grows proportionally with S. The first L time steps are initialized by drawing each variable independently and uniformly from its categorical support $\{0,1,\ldots,C-1\}$. For each subsequent time step $t \geq L$, every variable $X_j^{(t)}$ is sampled from its CPT conditioned on the realized values of its intra-slice parents at time t and its inter-slice parents at times $t-1,\ldots,t-L$. Variables within a time slice are sampled in an order consistent with the topological sort of A_{intra}, ensuring that intra-slice parent values are available when needed.

Experimental factors. We systematically vary the following data-generating parameters across experimental conditions to study their effect on causal discovery performance:

- **Number of variables** (d): We consider $d \in \{5, 10, 20, 30\}$ to assess scalability with respect to the dimensionality of the system.
- **Sample size** (S): We test $S \in \{10, 100, 500, 1000\}$, corresponding to total time series lengths $T = S \times d$, to evaluate performance under limited and abundant data regimes.
- **Markov order** (L): We vary $L \in \{1, 2, 3\}$ to examine how the depth of temporal dependencies affects recovery accuracy.
- **Inter-slice in-degree** (k_{inter}): We set $k_{\text{inter}} \in \{1, 2, 3\}$ to control the density of lagged causal connections.
- **Variable cardinality** (C): Each variable takes values in $\{0,\ldots,C-1\}$ with $C \in \{2, 3, 4\}$, allowing us to study the effect of increasing the state space on the reliability of conditional independence testing.

Each experimental condition is replicated across five independent random seeds (42, 123, 2021, 3407, 9999) to ensure robustness of the reported results.

J.2 PC-OPIMPLEMENTATION DETAILS

PC-OP introduces three variants that selectively modify the conditioning set in this test, controlled by an orientation parameter:

- **PC-OP** or **PC-OP-Y**: The conditioning set includes only the parents of the target variable Y_j. This variant tests whether the candidate cause adds explanatory power beyond the target's other identified parents. This variant serves as the base algorithm **PC-OP** in other parts of the experiments.
- **PC-OP-X**: The conditioning set includes only the time-shifted parents of the putative cause X_i. This variant removes the target variable's own parents from the conditioning set, testing whether the candidate cause carries information about the target beyond what is explained by its own lagged history.
- **PC-OP-M**: The conditioning set is chosen as the minimal set of the two parent sets—the parents of Y_j or the time-shifted parents of X_i—whichever yields a lower-dimensional conditioning set. This adaptive strategy balances conditioning completeness against the statistical cost of high-dimensional conditioning.

Combined with the optional independency confirmation step, these three orientation strategies yield six PC-OP variants in total: `PC-OP`, `PC-OP+`, `PC-OP-X`, `PC-OP-X+`, `PC-OP-M`, and `PC-OP-M+`, where the base names correspond to the Y, X, and smallest orientation strategies, and the "+" suffix indicates that independency confirmation step is enabled.

J.3 CONDITIONAL INDEPENDENCE TEST

All constraint-based methods in our experiments (PC-OP, PCMCI) employ a chi-squared test of conditional independence specifically designed for categorical data. Given a candidate edge $X_i^{t-\tau} \to Y_j^t$ and a conditioning set Z, the test proceeds by constructing, for each realization z of Z, a two-way contingency table between $X_i^{t-\tau}$ and Y_j^t. The Pearson chi-squared statistic is then computed as $\chi^2 = \sum_z \sum_{x,y} \frac{\left(O_{xy|z} - E_{xy|z}\right)^2}{E_{xy|z}}$, where $O_{xy|z}$ and $E_{xy|z}$ denote the observed and expected cell counts under the null hypothesis of conditional independence, respectively. The summation over expected frequencies excludes cells where the expected count is zero.

The degrees of freedom are adjusted following the method of Bishop et al. Bishop et al. (1975) to account for structural zeros in the contingency table. Specifically, for each conditioning stratum z, the effective degrees of freedom are computed as $(r_z - 1 - r_z^0) \times (c_z - 1 - c_z^0)$, where r_z and c_z are the numbers of rows and columns of the contingency table, and r_z^0 and c_z^0 are the numbers of rows and columns whose marginals are entirely zero. The total degrees of freedom are obtained by summing across all strata, with a minimum of one. The p-value is computed from the survival function of the chi-squared distribution with the adjusted degrees of freedom. This test is asymptotically valid. We use the python toolbox TIGRAMITE for existing test implementations.

J.4 BASELINE METHODS

We compare PC-OPagainst a diverse set of established time series causal discovery algorithms, spanning constraint-based, score-based, and neural approaches. The baselines are as follows.

PCMCI (Runge et al., 2019). A two-stage constraint-based method that combines the PC algorithm for condition selection with the MCI test for edge significance. PCMCI serves as the most direct comparator to PC-OP, differing only in the conditioning strategy used during the MCI step. We use the same hyperparameters as for PC-OP: $\alpha_{\text{PC}} = 0.2$ for the condition-selection stage and $\alpha_{\text{MCI}} = 0.05$ for the MCI significance threshold.

Multivariate Granger causality (Granger-MV) (Granger, 1969). A classical approach that tests for Granger-causal relationships by fitting vector autoregressive (VAR) models and evaluating whether past values of a candidate cause improve the prediction of a target variable beyond what is achievable from the target's own history and the histories of all other variables. We set the significance level to 0.05 and the maximum lag to match the true Markov order L of each experimental condition.

DYNOTEARS (Pamfil et al., 2020). A score-based method that formulates causal discovery as a continuous optimization problem over the space of weighted adjacency matrices, subject to an acyclicity constraint enforced via the NOTEARS augmented Lagrangian framework, extended to the temporal setting. The method simultaneously estimates both contemporaneous and time-lagged causal effects. We use regularization parameters $\lambda_w = \lambda_a = 0.1$, an

acyclicity tolerance of $h_{\text{tol}} = 10^{-8}$, a maximum of 100 optimization iterations, and an edge weight threshold of 0.1 for sparsification.

NTS-NOTEARS (Sun et al., 2021). A neural extension of the NOTEARS framework that employs locally connected layers to model nonlinear temporal causal relationships while maintaining the continuous acyclicity constraint. The number of lags is set to the true Markov order L.

TCDF (Nauta et al., 2019). An attention-based method that uses temporal convolutional neural networks to detect causal relationships in time series. Causal links are inferred from the learned attention weights, which indicate whether past values of one variable are relevant for predicting another. Default hyperparameters are used.

Multivariate Granger causality, TCDF, NTS-NOTEARS, and DYNOTEARS do not rely on an external conditional independence test, as they employ fundamentally different inference strategies (regression-based, attention-based, or continuous optimization-based). Their causal graphs are obtained directly from the learned model parameters.

J.5 Evaluation Protocol

Graph representation. To compare across all methods, learned causal graphs are converted to a common representation. Methods that produce symbolic arrow-type outputs (e.g., PCMCI, PC-OP) are first converted to binary adjacency tensors of shape $d \times d \times (L{+}1)$, where the third dimension indexes the time lag from 0 (contemporaneous) to L (maximum lag). All causal graph tensors, both learned and ground truth, are then unrolled into a two-dimensional binary adjacency matrix of dimension $d(L{+}1) \times d(L{+}1)$. This unrolled representation arranges the variables in temporal blocks $[X_{t-L}, \ldots, X_{t-1}, X_t]$ and places directed edges from the appropriate lag block to the current-time block, enabling standard graph-comparison metrics to be applied directly.

Metrics. We evaluate the quality of recovered causal graphs using two standard metrics:

- **Structural Hamming Distance (SHD)**: The total number of edge additions, deletions, and reversals required to transform the learned graph into the ground-truth graph. Lower values indicate better structural recovery.
- **F1 score**: The harmonic mean of precision and recall, providing a single balanced measure of edge recovery quality. Precision is defined as the fraction of edges in the learned graph that are present in the ground truth, defined as $\text{TP}/(\text{TP} + \text{FP})$, where TP denotes true positive edges and FP denotes false positives. Recall is defined as the fraction of ground-truth edges that are recovered by the learned graph, defined as $\text{TP}/(\text{TP}+\text{FN})$, where FN denotes false negatives.

For each experimental condition, all four metrics are computed independently for each of the five random seeds. We report the mean and standard error across replications.

Hyperparameter summary. Table 6 summarizes the key hyperparameters used across all methods.

J.6 Direct Validation of the Conditional Independence Test

The graph-recovery experiments in the remainder of this section evaluate the CI test only *indirectly*: the recovered skeleton is the product of many CI tests filtered through the PC condition-selection loop, the orientation rules, and the multiple-testing corrections of Stage 2. To validate Theorem 3.1 directly, we isolate a *single* conditional independence test and measure the two quantities the theorem characterizes: (i) the null calibration, i.e., whether the empirical Type-I error matches the nominal level $\alpha = 0.05$, and (ii) the power under alternatives of increasing strength.

Design. We construct a small dynamic ground-truth graph over three variables $\{X, Y, Z\}$ at Markov order $L = 1$, matching the purely time-lagged regime of our main experiments ($k_{\text{intra}} = 0$). The variable Z acts as a lagged confounder, $Z^{t-1} \to X^t$ and $Z^{t-1} \to Y^t$, and each of X and Y carries an autoregressive self-loop, $X^{t-1} \to X^t$ and $Y^{t-1} \to Y^t$. Under the *null* graph there is no edge into Y from X, so

$$X^{t-1} \perp\!\!\!\perp Y^t \;\Big|\; \widetilde{\mathcal{P}}(Y^t) \setminus \{X\} = \{Y^{t-1}, Z^{t-1}\} \tag{28}$$

holds by construction; under the *alternative* graph we add the edge $X^{t-1} \to Y^t$ with a controllable effect size $\beta \in [0, 1]$, where the conditional distribution of Y^t is a β-weighted blend between its X-independent base distribution ($\beta = 0$) and an X-dependent shift ($\beta = 1$). The conditioning set is computed from the known graph, so the null is a genuine d-separation and $\beta = 0$ recovers exact conditional independence. Conditional probability tables are drawn as in Section J.1 (symmetric Dirichlet, $\alpha_{\text{Dir}} = 1$), with different cardinality Cs.

Table 6: Summary of method hyperparameters.

Method	Parameter	Value
PC-OP/ PCMCI	PC-stage threshold (α_{PC})	0.2
	MCI-stage threshold (α_{MCI})	0.05
	Conditional independence test	Chi-squared or G-test
Granger-MV	Significance level	0.05
	Maximum lag	L (true Markov order)
DYNOTEARS	Regularization (λ_w, λ_a)	0.1
	Acyclicity tolerance (h_{tol})	10^{-8}
	Max iterations	100
	Edge threshold	0.1
NTS-NOTEARS	Number of lags	L (true Markov order)
TCDF	All parameters	Default

Effect size β. The effect size $\beta \in [0, 1]$ is the single knob that interpolates between the null and the alternative, and it is what generates the two regimes reported in the three tables. Concretely, let $p_0(\cdot \mid Y^{t-1}, Z^{t-1})$ be the base conditional law of Y^t that ignores X^{t-1}, obtained by averaging the Dirichlet-drawn CPT rows over the states of X^{t-1} (this makes $Y^t \perp\!\!\!\perp X^{t-1}$ hold *exactly* for that parent configuration). We then define an X-dependent perturbation as a cyclic shift of this base distribution by the value of X^{t-1}, and mix the two:

$$p_\beta\big(Y^t = y \mid X^{t-1} = x, Y^{t-1}, Z^{t-1}\big) \;=\; (1-\beta)\, p_0(y) \;+\; \beta\, p_0\big((y-x) \bmod C\big), \tag{29}$$

so that β is a genuine effect size: at $\beta = 0$ the X-term vanishes and Y^t is conditionally independent of X^{t-1} (exact null), while increasing β makes Y^t's distribution depend ever more strongly on X^{t-1}, up to a full cyclic relabeling at $\beta = 1$. Every row of equation 29 remains a valid probability vector for all β.

We stress that the convex combination in equation 29 does *not* make Y^t a continuous variable. The mixing acts on the *probability mass function* of Y^t, not on its values: both $p_0(\cdot)$ and its cyclic shift $p_0((\cdot - x) \bmod C)$ are probability vectors over the same $C = 3$ categories $\{0, 1, \ldots, C-1\}$, and their β-weighted average is again a probability vector over those same categories. For example, with $p_0 = (0.5, 0.3, 0.2)$, the shift for $x = 1$ is $(0.2, 0.5, 0.3)$, and at $\beta = 0.3$ the mixed law is $0.7\,(0.5, 0.3, 0.2) + 0.3\,(0.2, 0.5, 0.3) = (0.41, 0.36, 0.23)$. The observation Y^t is then *sampled* from this categorical distribution and remains discrete-valued. Thus β is a continuous knob on the conditional probability table, controlling how strongly x tilts the category probabilities of Y^t, while the data fed to the test stay categorical, exactly the data type consumed by the chi-squared tests in the main experiments. In this sense β plays the role that a correlation coefficient plays in a Gaussian toy example: a continuous measure of dependence strength attached to a variable whose type is unchanged. This construction is what lets a single scalar control the strength of the dependence being detected, and it drives the three tables as follows: the Type-I error table (Table 7) fixes $\beta = 0$ and sweeps n; the power-versus-n table (Table 8) fixes $\beta = 0.3$ and sweeps n; and the power-versus-β table (Table 9) fixes $n = 1000$ and sweeps $\beta \in \{0, 0.1, 0.2, 0.3, 0.5\}$. Because $\beta = 0$ is by construction identical to the null, the $\beta = 0$ column of Table 9 must and does reproduce the Type-I error, providing an internal consistency check on the design.

For each configuration we draw a fresh time series and run the CI test once, repeating over $R = 2000$ independent replicates; the empirical Type-I error (under the null) and power (under the alternative) are the fractions of the 2000 p-values below $\alpha = 0.05$. The Monte-Carlo standard error at the nominal level is $\sqrt{\alpha(1-\alpha)/R} \approx 0.005$. We report results for the proposed chi-squared test under two conditioning strategies: conditioning on the target's parents only ($Z = \widetilde{\mathcal{P}}(Y^t) \setminus \{X\}$, the Y-variant) and on both parent sets.

Null calibration. Table 7 reports the empirical Type-I error as a function of sample size (Figure 3, left). The chi-squared test is well calibrated across the entire range, staying within Monte-Carlo error of the nominal 0.05 even at $n = 200$. A quantile comparison of the null p-values against the Uniform$[0, 1]$ distribution at $n = 5000$ (Figure 4, right) shows the chi-squared p-values tracking the diagonal almost exactly, confirming correct calibration as predicted by Theorem 3.1.

Power. Tables 8 and 9 report power as a function of sample size (fixed $\beta = 0.3$) and of effect size (fixed $n = 1000$), respectively (Figure 3, right, and Figure 4). Power increases monotonically toward 1 in both n and β for every test

Table 7: Empirical Type-I error under the null (nominal $\alpha = 0.05$) at all three cardinalities. Cells are `rate(se)` over $R = 2000$ replicates, with the parenthesized digit the Monte-Carlo standard error in units of the last decimal place. The chi-squared test holds nominal size at $C = 2$ and, apart from one cell at $n = 100$, at $C = 3$; at $C = 4$ it shows a small-sample inflation that has largely cleared by $n = 1000$ and is gone by $n = 2000$.

C	Test	Cond.	n=100	n=200	n=1000	n=2000	n=5000
2	Chi-squared	Y-parents	0.051(5)	0.049(5)	0.050(5)	0.045(5)	0.053(5)
	Chi-squared	both parents	0.041(4)	0.049(5)	0.044(5)	0.054(5)	0.046(5)
3	Chi-squared	Y-parents	0.054(5)	0.055(5)	0.054(5)	0.051(5)	0.046(5)
	Chi-squared	both parents	0.062(5)	0.052(5)	0.043(5)	0.049(5)	0.052(5)
4	Chi-squared	Y-parents	0.096(7)	0.100(7)	0.057(5)	0.052(5)	0.056(5)
	Chi-squared	both parents	0.105(7)	0.099(7)	0.071(6)	0.052(5)	0.051(5)

Table 8: Power under the alternative as a function of sample size n, at fixed effect size $\beta = 0.3$. Cells are `rate(se)` over $R = 2000$ replicates. Comparing across the three blocks isolates the cardinality effect at fixed n.

C	Test	Cond.	n=200	n=500	n=1000	n=2000	n=5000
2	Chi-squared	Y-parents	0.281(10)	0.553(11)	0.713(10)	0.861(8)	0.958(5)
	Chi-squared	both parents	0.276(10)	0.518(11)	0.717(10)	0.852(8)	0.964(4)
3	Chi-squared	Y-parents	0.128(7)	0.265(10)	0.508(11)	0.817(9)	0.990(2)
	Chi-squared	both parents	0.131(8)	0.254(10)	0.508(11)	0.832(8)	0.990(2)
4	Chi-squared	Y-parents	0.137(8)	0.196(9)	0.284(10)	0.537(11)	0.948(5)
	Chi-squared	both parents	0.132(8)	0.164(8)	0.272(10)	0.565(11)	0.951(5)

and conditioning strategy. Crucially, at $\beta = 0$ the measured power equals the Type-I error (≈ 0.05 for chi-squared), confirming that the alternative reduces to the null as designed. The two conditioning strategies yield near-identical calibration and power when $C = 3$ and Y-parents is slightly favored when $C = 4$, especially at lower n, validating our proposed CI tests.

J.7 Full Synthetic Experiment Results

We first compare different methods on synthetic datasets and perform ablation studies on various approaches.

Ablation Studies We first study a few variants of PC-OP, compared with PCMCI on varying number of variables d. Here we use $L = 2$, $S = 1000$, and $k_{inter} = 2$. Shown in Figure 5, conditioned on children Y's parents provides the most accurate results although both PC-OP-X and PC-OP-M are not too far behind. All these variants are better than PCMCI by a significant margin, especially when the number of variables increases.

Different Independence tests We also test PC-OP with a different independence test, G-square test. The G-squared test, or G-test, is a likelihood-ratio test used to assess whether observed categorical data deviate from expected frequencies, serving as an alternative to the chi-square test. Although in general G-test performs worse than chi-squared tests, PC-OP with G-test is still better than PCMCI, as one can see in Figure 6.

Varying Number of Variables Figure 7 shows PC-OP with all the baseline methods together with different number of variables.

Varying Number of Lags Figure 8 shows PC-OP with all the baseline methods together with different number of time lags.

Varying Number of Graph-In Degrees Figure 9 shows PC-OP with all the baseline methods together with different number of graph in-degree. Higher degree results in more dense graphs and hence more difficulty to learn the causal graphs.

Table 9: Power under the alternative as a function of effect size β, at fixed sample size $n = 1000$. Cells are `rate(se)` over $R = 2000$ replicates. At $\beta = 0$ the alternative coincides with the null by construction, and the measured rate reproduces the Type-I error within Monte-Carlo error (0.053 vs. 0.050 at $C = 2$; 0.057 vs. 0.054 at $C = 3$; 0.063 vs. 0.057 at $C = 4$; Table 7), an internal consistency check on the design.

C	Test	Cond.	β=0.0	β=0.1	β=0.2	β=0.3	β=0.5
2	Chi-squared	Y-parents	0.053(5)	0.198(9)	0.491(11)	0.712(10)	0.900(7)
	Chi-squared	both parents	0.048(5)	0.213(9)	0.510(11)	0.707(10)	0.899(7)
3	Chi-squared	Y-parents	0.057(5)	0.096(7)	0.239(10)	0.531(11)	0.915(6)
	Chi-squared	both parents	0.051(5)	0.084(6)	0.248(10)	0.536(11)	0.906(7)
4	Chi-squared	Y-parents	0.063(5)	0.087(6)	0.141(8)	0.275(10)	0.699(10)
	Chi-squared	both parents	0.061(6)	0.082(6)	0.138(8)	0.279(10)	0.711(10)

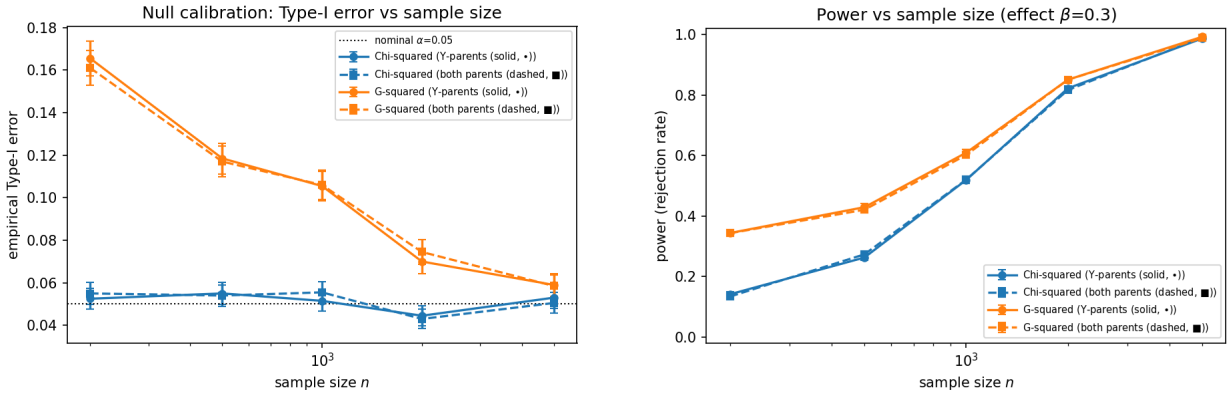


Figure 3: Direct validation of the CI test. **Left:** empirical Type-I error versus sample size; the chi-squared test tracks the nominal $\alpha = 0.05$ (dotted) at all n. **Right:** power versus sample size at $\beta = 0.3$, rising toward 1. Error bars are Monte-Carlo standard errors over $R = 2000$ replicates.

Varying Number of Samples Figure 10 shows PC-OP with all the baseline methods together with different number of time series samples.

In all the above results, our approach clearly performs the best.

J.8 Oracle ablation: isolating the conditioning set

The comparisons above confound two effects, since PCMCI and PC-OP may recover different parent sets in stage 1 and therefore condition on different variables for reasons unrelated to set *size*. To isolate the mechanism we re-run both methods with the *true* parent sets supplied as an oracle, so that stage-1 screening error is eliminated and the only remaining difference is which variables stage 2 conditions on. Table 10 reports all eight (d, L) conditions. Recall is at or within 0.02 of ceiling for both methods throughout, and exactly 1.000 for both at $L = 2$, so neither is missing edges; but PC-OP conditions on half as many variables ($|Z| = 3.9$ vs. 7.8 at $d = 20$, $L = 2$) and its false-positive rate is correspondingly lower (0.053 vs. 0.089), yielding F1 of 0.808 versus 0.715. Because recall is held fixed at ceiling and the parent sets are exact, this is an assumption-free confirmation that the benefit comes from the conditioning set itself and not from any difference in screening behaviour. We also note that the oracle Type-I error for PC-OP sits at the nominal $\alpha = 0.05$ across all eight conditions ($0.040 \sim 0.054$), whereas PCMCI's ranges up to 0.113, the finite-sample degradation of a χ^2 test with many conditioning variables and sparsely populated cells. The comparison that the design supports is the paired one at matched (d, L): in $8/8$ conditions PC-OP conditions on strictly fewer variables and has no higher false-positive rate.

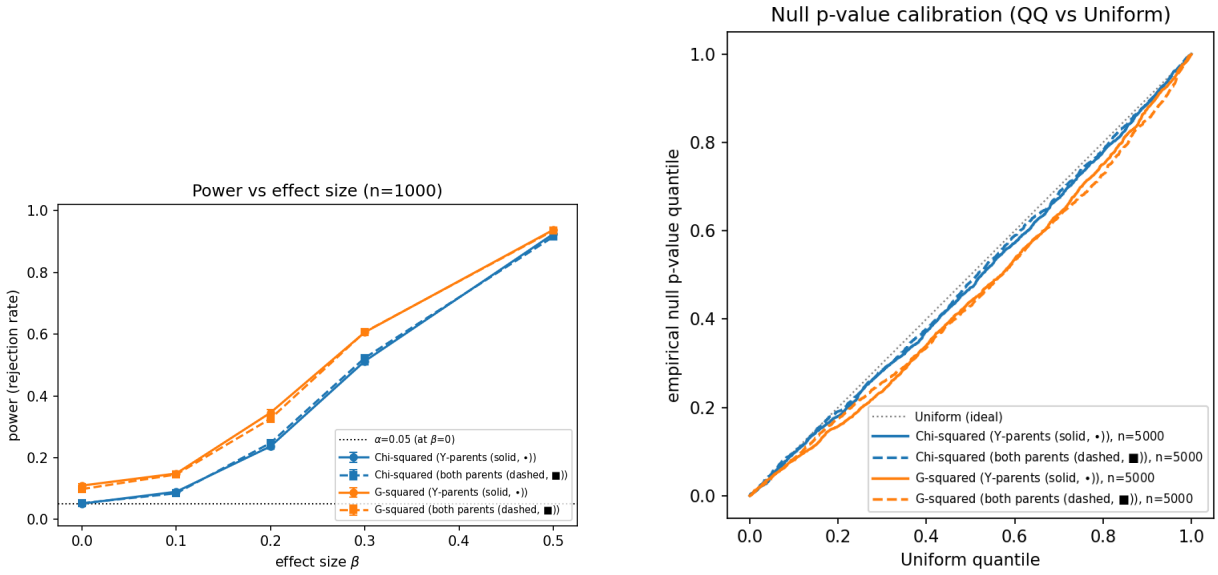
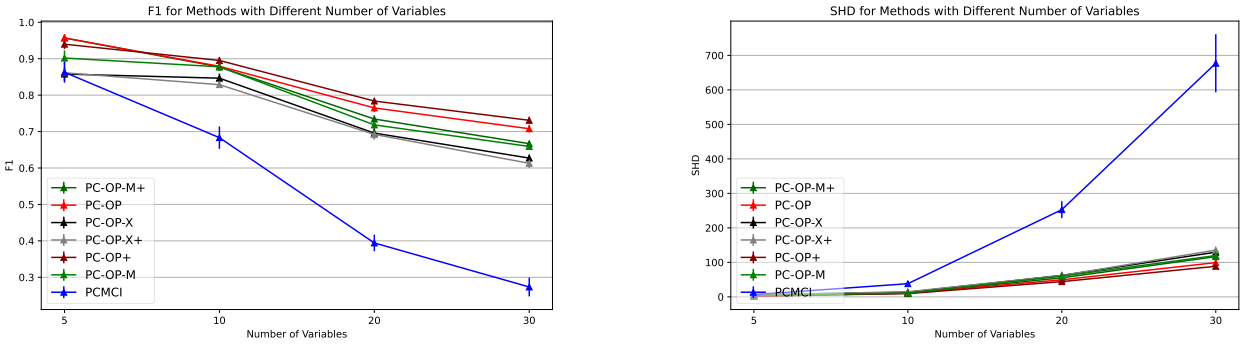
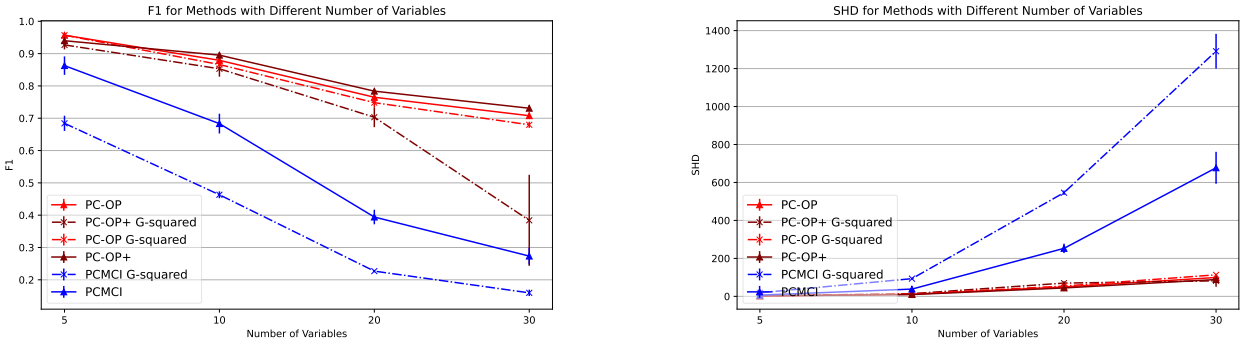


Figure 4: **Left:** power versus effect size β at $n = 1000$; power rises monotonically from the nominal level at $\beta = 0$ (dotted) toward 1. **Right:** quantile-quantile plot of the null p-values against Uniform$[0, 1]$ at $n = 5000$; the chi-squared test lies on the diagonal, confirming correct null calibration.



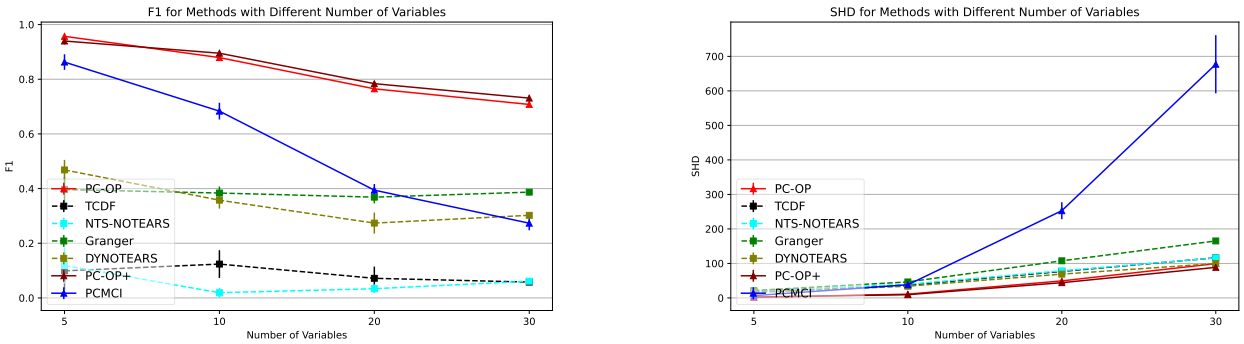
(a) F1 Score. (b) SHD.

Figure 5: Comparison of Different PC-OPVariants and PCMCI.



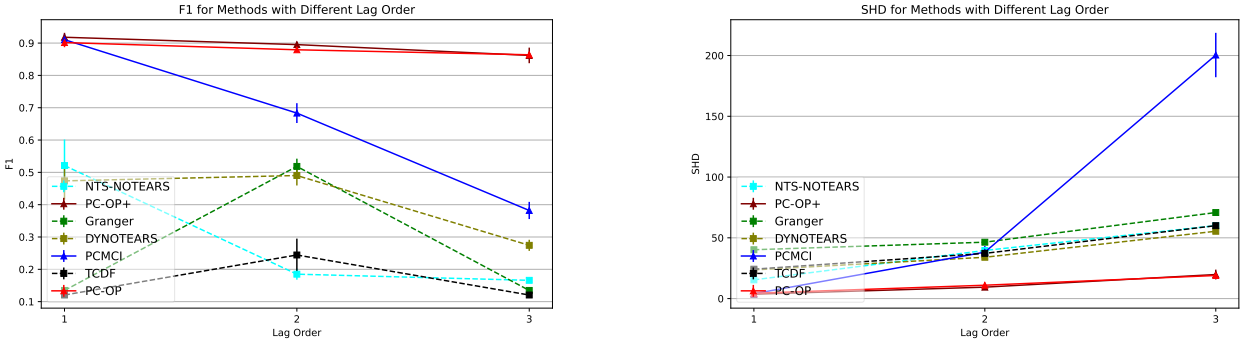
(a) F1 Score. (b) SHD.

Figure 6: Comparison of Different Methods, with Chi-squared and G-squared tests.



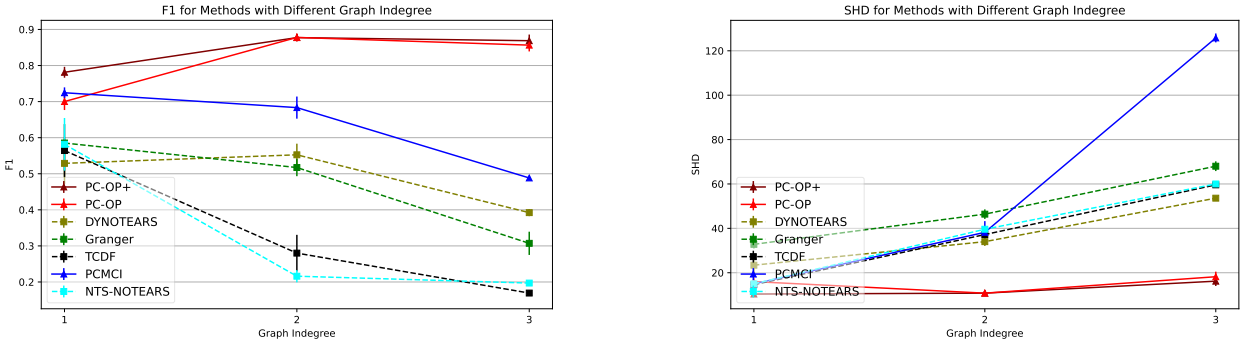
(a) F1 Score. (b) SHD.

Figure 7: Comparison of Different Methods, Varying Number of Variables.



(a) F1 Score. (b) SHD.

Figure 8: Comparison of Different Methods, Varying Maximal Lag Orders.

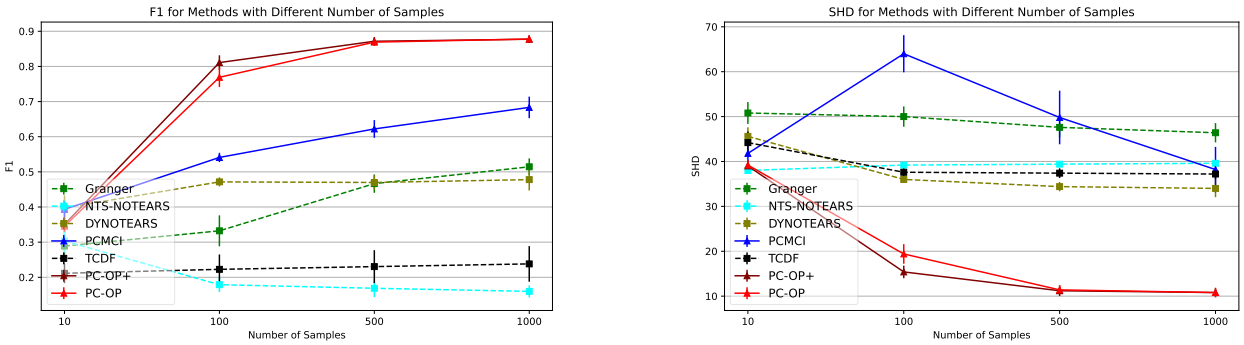


(a) F1 Score. (b) SHD.

Figure 9: Comparison of Different Methods, Varying Number of Graph In-Degrees.

J.9 Sensitivity to Significance Thresholds

A natural concern is whether the synthetic advantage above is an artifact of the shared threshold choice $(\alpha_{\text{PC}}, \alpha_{\text{MCI}}) = (0.2, 0.05)$, which might happen to suit PC-OP. Table 11 addresses this by sweeping the full 4×4 grid $\alpha_{\text{PC}} \in \{0.05, 0.1, 0.2, 0.3\}$ and $\alpha_{\text{MCI}} \in \{0.01, 0.05, 0.1, 0.2\}$ and reporting each method at its own best cell. PC-OP retains its lead under this per-method tuning, with the best F1 at $d = 30$ is 0.934 for PC-OP+ versus 0.845 for PCMCI, so the advantage is not a tuning artifact. The sweep additionally reveals a robustness benefit we did not design for: the spread between the best and worst grid cell is far smaller for PC-OP. At $d = 30$ the SHD spread is 164 for PC-OP and 177 for PC-OP+ but 1326 for PCMCI, and the F1 spread is 0.338 versus 0.695. PC-OP is therefore substantially less sensitive



(a) F1 Score.

(b) SHD.

Figure 10: Comparison of Different Methods, Varying Number of Samples.

Table 10: Oracle-parent ablation across all eight (d, L) conditions. Both methods receive the *true* parent sets, so stage-1 screening is removed as a confound and the only difference is the stage-2 conditioning set. Recall, FPR and F1 are mean $\pm$ standard error over five seeds; $|Z|$ is the mean conditioning-set size. In all eight conditions PC-OP has the strictly smaller $|Z|$ and no higher FPR.

Config	Method	Recall	FPR	F1	$\lvert Z\rvert$
d5, L=1	Oracle-PCMCI (full MCI set)	0.980 ± 0.020	0.053 ± 0.025	0.952 ± 0.014	3.60
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	0.980 ± 0.020	0.040 ± 0.016	0.961 ± 0.019	1.60
d5, L=2	Oracle-PCMCI (full MCI set)	1.000 ± 0.000	0.113 ± 0.037	0.924 ± 0.023	7.19
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.053 ± 0.013	0.962 ± 0.010	3.60
d10, L=1	Oracle-PCMCI (full MCI set)	0.990 ± 0.010	0.043 ± 0.013	0.919 ± 0.020	3.80
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.040 ± 0.009	0.927 ± 0.015	1.80
d10, L=2	Oracle-PCMCI (full MCI set)	1.000 ± 0.000	0.090 ± 0.011	0.849 ± 0.017	7.60
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.050 ± 0.003	0.909 ± 0.005	3.80
d20, L=1	Oracle-PCMCI (full MCI set)	1.000 ± 0.000	0.058 ± 0.006	0.793 ± 0.016	3.90
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.054 ± 0.003	0.804 ± 0.009	1.90
d20, L=2	Oracle-PCMCI (full MCI set)	1.000 ± 0.000	0.089 ± 0.004	0.715 ± 0.009	7.80
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.053 ± 0.006	0.808 ± 0.019	3.90
d30, L=1	Oracle-PCMCI (full MCI set)	0.997 ± 0.003	0.052 ± 0.002	0.733 ± 0.009	3.93
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.052 ± 0.003	0.733 ± 0.010	1.93
d30, L=2	Oracle-PCMCI (full MCI set)	1.000 ± 0.000	0.072 ± 0.003	0.666 ± 0.009	7.87
	Oracle-PC-OP ($\mathrm{pa}(Y) \setminus \{X\}$)	1.000 ± 0.000	0.049 ± 0.002	0.746 ± 0.008	3.93

to threshold selection, which matters in practice because the thresholds cannot be tuned against ground truth on real data.

J.10 Ablation Studies on PC-OP Variants

We study a few variants of PC-OP, compared with PCMCI on varying number of variables d. PC-OP introduces three variants that selectively modify the conditioning set in this test, controlled by an orientation parameter: 1) **PC-OP-X**: The conditioning set includes only the time-shifted parents of the putative cause X_i. 2) **PC-OP** or **PC-OP-Y**: The conditioning set includes only the parents of the target variable Y_j. 3) **PC-OP-M**: The conditioning set is chosen as the minimal set of the two parent sets—the parents of Y_j or the time-shifted parents of X_i—whichever yields a lower-dimensional conditioning set. Combined with the optional independency confirmation step, these three orientation strategies yield six PC-OPvariants in total: `PC-OP`, `PC-OP+`, `PC-OP-X`, `PC-OP-X+`, `PC-OP-M`, and `PC-OP-M+`, where the base names correspond to the Y, X, and smallest orientation strategies, and the "+" suffix indicates that independency confirmation is enabled. Here we use $L = 2$, $S = 1000$, and $k_{inter} = 2$. Shown in Figure 11, conditioned

Table 11: Sensitivity to the significance thresholds. For each method and dimension we sweep the full grid $\alpha_{\text{PC}} \in \{0.05, 0.1, 0.2, 0.3\} \times \alpha_{\text{MCI}} \in \{0.01, 0.05, 0.1, 0.2\}$ (16 cells, five seeds each) and report the *best* cell together with the spread (best minus worst) across the grid. PC-OP is both better at its optimum and dramatically more stable: at $d = 30$ the SHD spread across the grid is 164 for PC-OP versus 1326 for PCMCI, so PC-OP requires far less threshold tuning to work well.

		Best F1 (↑)		Best SHD (↓)	
d	Method	Value	Spread	Value	Spread
5	PCMCI	0.985	0.199	0.6	10.6
	PC-OP	0.980	0.058	0.8	2.6
	PC-OP+	0.990	0.064	0.4	2.8
10	PCMCI	0.937	0.456	5.4	82.6
	PC-OP	0.969	0.175	2.6	18.4
	PC-OP+	0.978	0.166	1.8	16.8
20	PCMCI	0.880	0.640	22.0	487.6
	PC-OP	0.915	0.274	15.0	75.4
	PC-OP+	0.949	0.300	8.6	78.0
30	PCMCI	0.845	0.695	44.2	1326.2
	PC-OP	0.892	0.338	29.4	164.2
	PC-OP+	0.934	0.381	17.0	177.4

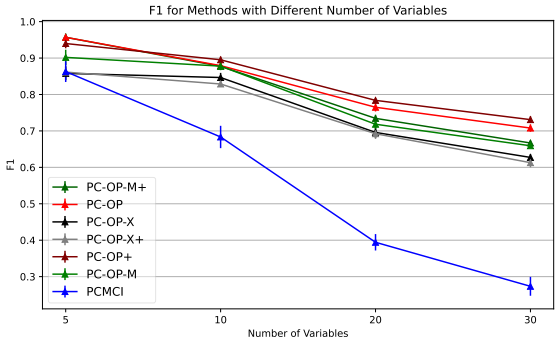


Figure 11: Comparison of Different PC-OP variants and PCMCI.

on children Y's parents provides the most accurate results although both PC-OP-X and PC-OP-M are not too far behind. All these variants are better than PCMCI by a significant margin, especially when the number of variables increases.

J.11 Full Results on CausalRivers

We evaluate on the CausalRivers benchmark (Stein et al., 2025), a real-world collection of river-discharge time series with expert-annotated ground-truth causal graphs. We report results across six sampling strategies for extracting sub-networks (`1_random`, `close`, `confounder`, `disjoint`, `random`, and `root_cause`) at two graph sizes ($d = 3$ and $d = 5$ nodes), with maximum lag $L = 3$. Each configuration is averaged over 50 randomly drawn sub-networks. Table 12 reports the AUROC score for each method and configuration.

Table 12: Graph-recovery performance on the CausalRivers benchmark, reported as AUROC (mean ± standard error over 50 sub-networks). Columns are indexed by sampling strategy and graph size (n nodes). "ParCorr", "χ^2", and "G^2" denote the conditional independence test used; PC-OP+ variants add the independency-confirmation step.

Method	1_random_3n	1_random_5n	close_3n	close_5n	confounder_3n	confounder_5n	disjoint_5n	random_3n	random_5n	root_cause_3n	root_cause_5n	**Avg.**
Granger (ParCorr)	0.716 ± 0.035	0.630 ± 0.022	0.617 ± 0.021	0.599 ± 0.020	0.609 ± 0.034	0.527 ± 0.008	0.822 ± 0.026	0.609 ± 0.024	0.606 ± 0.018	0.650 ± 0.024	0.669 ± 0.021	0.611
PCMCI (G^2)	0.771 ± 0.031	0.658 ± 0.022	0.656 ± 0.024	0.621 ± 0.018	0.625 ± 0.028	0.583 ± 0.015	0.778 ± 0.019	0.660 ± 0.023	0.631 ± 0.016	0.684 ± 0.022	0.681 ± 0.017	0.643
PCMCI (χ^2)	0.731 ± 0.028	0.651 ± 0.017	0.688 ± 0.020	0.609 ± 0.016	0.681 ± 0.035	0.585 ± 0.014	0.757 ± 0.016	0.688 ± 0.021	0.599 ± 0.016	0.696 ± 0.021	0.664 ± 0.014	0.651
PCMCI (ParCorr)	0.755 ± 0.024	0.664 ± 0.016	0.701 ± 0.014	0.630 ± 0.012	0.674 ± 0.030	0.628 ± 0.015	0.753 ± 0.021	0.714 ± 0.016	0.625 ± 0.014	0.735 ± 0.016	0.669 ± 0.012	0.672
PC-OP (G^2)	0.745 ± 0.012	0.630 ± 0.015	0.722 ± 0.013	0.634 ± 0.012	0.664 ± 0.024	0.593 ± 0.008	0.715 ± 0.013	0.707 ± 0.013	0.621 ± 0.007	0.717 ± 0.012	0.632 ± 0.009	0.661
PC-OP (χ^2)	0.755 ± 0.008	0.636 ± 0.014	0.711 ± 0.012	0.639 ± 0.011	0.698 ± 0.024	0.602 ± 0.009	0.736 ± 0.010	0.716 ± 0.011	0.623 ± 0.005	0.728 ± 0.010	0.644 ± 0.009	0.670
PC-OP (ParCorr)	0.766 ± 0.013	0.645 ± 0.015	0.723 ± 0.016	0.633 ± 0.013	0.726 ± 0.031	0.628 ± 0.013	0.736 ± 0.011	0.745 ± 0.013	0.640 ± 0.014	0.735 ± 0.013	0.659 ± 0.012	**0.686**
PC-OP+ (G^2)	0.744 ± 0.018	0.634 ± 0.015	0.725 ± 0.013	0.627 ± 0.013	0.639 ± 0.029	0.579 ± 0.011	0.745 ± 0.012	0.700 ± 0.015	0.616 ± 0.006	0.719 ± 0.015	0.642 ± 0.009	0.656
PC-OP+ (χ^2)	0.755 ± 0.011	0.640 ± 0.015	0.727 ± 0.013	0.619 ± 0.013	0.663 ± 0.028	0.574 ± 0.011	0.749 ± 0.010	0.725 ± 0.012	0.621 ± 0.012	0.737 ± 0.010	0.641 ± 0.009	0.663
PC-OP+ (ParCorr)	0.771 ± 0.015	0.657 ± 0.015	0.713 ± 0.017	0.636 ± 0.013	0.727 ± 0.028	0.616 ± 0.012	0.741 ± 0.014	0.736 ± 0.014	0.645 ± 0.015	0.720 ± 0.015	0.661 ± 0.012	0.682

† The disjoint_3n configuration is omitted: all 50 disjoint 3-node sub-networks extracted from the benchmark region degenerate to 2 effective nodes with too few ground-truth edges, so no method yields an evaluable score for it.