
Learning Stable Digraphs from Sparse-Input Linear Structural Causal Models

Panagiotis Misiakos¹

Markus Püschel¹

¹Department of Computer Science, ETH Zurich, Zürich, Switzerland

Abstract

We propose STABLESPIN, a continuous optimization framework for learning potentially cyclic causal directed graphs (digraphs) in linear structural equation models via a spectral-radius constraint on the *absolute* adjacency matrix. Our motivation is twofold: first, stability is a natural requirement for linear structural models; second, the contractive cycles that stability induces facilitate the identifiability of digraphs. We formulate causal discovery as a likelihood-based optimization problem under a sparse-input assumption, yielding a scalable algorithm applicable to both cross-sectional and time-series data. Unlike prior continuous acyclicity constraints, our stability regularizer is cheaper to evaluate while accommodating general digraphs, yielding faster optimization in practice. Experiments on synthetic benchmarks show improved graph recovery over state-of-the-art baselines, and empirical studies on U.S., European, and Swiss equity markets reveal interpretable cyclic dependencies that persist across time windows. Our implementation is available at <https://github.com/pmisiakos/StableSpIn>.

1 INTRODUCTION

Causal discovery aims to infer the underlying causal structure of a multivariate system, typically represented as a directed graph (digraph) whose edges encode direct causal relationships between variables [Pearl, 2009]. Most existing approaches assume acyclicity, i.e., the digraph is a directed acyclic graph (DAG), and can be broadly categorized into constraint-based methods relying on conditional independence tests [Spirtes et al., 2000], score-based methods [Chickering, 2002], and functional or optimization-based approaches [Shimizu et al., 2006, Zheng et al., 2018,

Ng et al., 2020]. A major advance was introduced by Zheng et al. [2018], who characterized acyclicity through a smooth constraint, enabling gradient-based optimization for causal discovery. This continuous formulation, and subsequent extensions [Ng et al., 2020, Misiakos et al., 2023], improve scalability and empirical performance over traditional constraint-based methods and naturally extend to time-series settings [Hyvärinen et al., 2010, Sun et al., 2023, Misiakos and Püschel, 2025].

However, acyclicity is often unrealistic in real-world systems. Biological and brain networks [Sachs et al., 2005, Mooij and Heskes, 2013, Smith et al., 2011], environmental systems [Runge et al., 2019], and macroeconomic or financial systems [Sims, 1980, Lanne et al., 2017] inherently exhibit feedback mechanisms that induce cycles in the causal graph. To address this, cyclic causal discovery methods have been developed, including approaches based on conditional independence testing [Richardson, 1996a,b, Forré and Mooij, 2018], extensions handling latent confounders [Hytinen et al., 2012, Mooij and Heskes, 2013, Améndola et al., 2020], and methods leveraging interventional data [Rothenhäusler et al., 2015, Sethuraman et al., 2023]. Despite these advances, existing approaches lack a structural constraint tailored to cyclic graphs that enables a continuous optimization framework analogous to NOTEARS for DAGs [Zheng et al., 2018]. In this work, we close this gap by introducing a stability-based structural constraint for (potentially cyclic) digraphs within a continuous optimization framework, yielding substantial improvements in computational efficiency and accuracy compared to prior work.

Our approach builds on linear structural equation models (SEMs) [Peters et al., 2017]¹, in which each variable is expressed as a linear combination of its direct causes plus a non-Gaussian noise term. In this setting, Lacerda et al.

¹Throughout this work, we restrict attention to linear SEMs; extending stability-based causal discovery to nonlinear settings would require new identifiability results and remains an important direction for future research.

[2008] established identifiability for digraphs with vertex-disjoint, contractive cycles whose product of edge weights is less than one. More recently, Drton et al. [2025] characterized the distributional equivalence classes of digraphs via cycle reversals. While contractive cycles are sufficient for identifiability, enforcing contractiveness directly through a continuous constraint is technically demanding. Instead, we adopt a more natural structural condition for SEMs: stability of the adjacency matrix, i.e., convergence of its powers to zero (nilpotency in the limit). Stability is equivalent to contractive cycles in two important cases: disjoint cycles or positive edge weights, but neither condition implies the other in general. To ensure contractiveness beyond these special cases, we impose *absolute stability*, which guarantees contractive cycles for arbitrary digraphs and yields a continuous optimization constraint. As with acyclicity in DAG-based causal discovery, absolute stability is an assumption on the unknown data-generating process rather than a property testable from the observations; StableSpIn therefore imposes it as a constraint during optimization, as NOTEARS does for acyclicity.

Stability assumptions have appeared in prior work. For example, Jin et al. [2025] propose a Bayesian approach for learning causal digraphs from time-series data under stability, but without explicitly promoting contractive cycles. In econometrics, Lanne et al. [2017] establish identifiability only up to impulse-response equivalence and do not impose stability constraints on the underlying causal digraph. Moreover, [Lacerda et al., 2008, Drton et al., 2025] neither incorporate stability as an optimization constraint nor address time-series settings. We extend this line of work with STABLESPIN, a continuous optimization framework for learning stable causal digraphs in linear structural equation models under a sparse-input assumption [Misiakos and Püschel, 2025], for both cross-sectional and time-series data. StableSpIn inherits the sparse-input, likelihood-based formulation of SpinSVAR but replaces its acyclicity constraint with a stability-based one, enabling the identification and estimation of cyclic digraphs and requiring a distinct theoretical analysis (identifiability and consistency for cyclic models).

Our main contributions are as follows:

- We introduce a spectral-radius constraint on the *absolute* adjacency matrix that guarantees stability of the digraph and contractive cycles.
- We formulate STABLESPIN for causal discovery in linear structural equation models as a continuous likelihood-based optimization problem that combines the sparse-input assumption with a stability regularizer. Our spectral stability constraint is evaluated in $\mathcal{O}(d^2)$ per iteration, replacing the $\mathcal{O}(d^3)$ matrix-exponential acyclicity constraint of NOTEARS [Zheng et al., 2018]. While the overall per-iteration cost re-

mains $\mathcal{O}(Td^2 + d^3)$, this cheaper constraint yields substantial speedups in practice. We further establish consistency of the proposed estimator.

- Synthetic experiments on static and time-series data demonstrate consistent improvements over state-of-the-art baselines in both estimation accuracy and runtime, with scalability to digraphs up to 10,000 nodes. We validate our method on U.S., European, and Swiss equity markets, uncovering interpretable cyclic dependencies that persist across distinct time windows.

2 BACKGROUND: DIGRAPH MODELS

We denote with $\mathcal{G}(V, \mathbf{B})$ a digraph with node set $V = \{1, 2, \dots, d\}$ and weighted adjacency matrix $\mathbf{B}$, where $\mathbf{B}_{ij}$ represents the direct influence of node i on node j . Our digraph model allows for arbitrary cyclic structures in $\mathbf{B}$, but excludes self-loops [Lacerda et al., 2008, Drton et al., 2025]. Doing so generalizes the assumptions commonly imposed in prior work, such as acyclicity [Shimizu et al., 2006] or the restriction to vertex-disjoint directed cycles [Lacerda et al., 2008]. In our setting, the only structural assumption imposed on $\mathbf{B}$ is (absolute) stability, which we formalize later and which implies stability of $\mathbf{B}$ and contractive cycles whose product is less than one.

For the data, we use $\mathbf{x} \in \mathbb{R}^{1 \times d}$ to denote a (row) observation vector associated with $\mathcal{G}$, where x_i corresponds to node i .

2.1 LINEAR STRUCTURAL CAUSAL MODELS

For data on digraphs, we consider two widely used linear structural causal models [Lacerda et al., 2008, Hyvärinen et al., 2010]: the linear structural equation model (SEM) for cross-sectional data and the linear structural vector autoregression (SVAR) for time-series data.

Linear SEM Given a digraph $\mathcal{G}(V, \mathbf{B})$, a data vector $\mathbf{x}$ is said to follow a linear SEM if

$$\mathbf{x} = \mathbf{x}\mathbf{B} + \mathbf{s}, \quad (1)$$

where $\mathbf{s}$ denotes the vector of structural shocks. By stacking N independent observations $\mathbf{x}_1, \dots, \mathbf{x}_N$ into a data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, the SEM can be written in matrix form as

$$\mathbf{X} = \mathbf{X}\mathbf{B} + \mathbf{S}, \quad (2)$$

where $\mathbf{S} \in \mathbb{R}^{N \times d}$ collects all the structural shocks.

SVAR The time-series analogue of the linear SEM is the structural vector autoregression (SVAR). Formally, an $\text{SVAR}(k)$ is defined by an instantaneous effect matrix $\mathbf{B}_0$ and lagged-effect matrices $\mathbf{B}_1, \dots, \mathbf{B}_k \in \mathbb{R}^{d \times d}$, and models a d -dimensional time series $(\mathbf{x}_t)_{t=0}^{T-1}$ according to

$$\mathbf{x}_t = \mathbf{x}_t \mathbf{B}_0 + \mathbf{x}_{t-1} \mathbf{B}_1 + \dots + \mathbf{x}_{t-k} \mathbf{B}_k + \mathbf{s}_t, \quad (3)$$

where $\mathbf{s}_t$ denotes the structural shocks. Intuitively, $(\mathbf{B}_\tau)_{ij}$ captures the influence of node i on node j after τ time steps. The SVAR parameters $\mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_k$ can be collected as block rows in a matrix $\mathbf{W} \in \mathbb{R}^{d(k+1) \times d}$, which defines the so-called window graph. Stacking the observations $\mathbf{x}_t$ as rows yields the data matrix $\mathbf{X} \in \mathbb{R}^{T \times d}$. Similarly, defining the lagged design vectors $\mathbf{x}_{t,\text{past}} = (\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-k})$ for $t \in [T]$ and stacking them produces $\mathbf{X}_{\text{past}} \in \mathbb{R}^{T \times d(k+1)}$. With this notation, the SVAR model (3) can be written compactly in matrix form as

$$\mathbf{X} = \mathbf{X}_{\text{past}} \mathbf{W} + \mathbf{S}. \quad (4)$$

Note that (2) coincides with (4) when $k = 0$. An SVAR can therefore be viewed as a linear SEM augmented with autoregressive temporal dynamics. The main semantic distinction between the two settings lies in the sampling scheme. For the SEM (2), $\mathbf{X} \in \mathbb{R}^{N \times d}$ consists of N independent samples, whereas for the SVAR (4), $\mathbf{X} \in \mathbb{R}^{T \times d}$ represents a time-series of length T . While the latter samples are not independent in the strict sense, their dependence weakens as the temporal separation increases, when the SVAR is stable, as explained in Appendix A.2.

Sparse input Intuitively, the structural shocks $\mathbf{S}$ represent unobserved events that propagate through space (according to $\mathbf{B}_0$) and time t (according to $\mathbf{B}_1, \dots, \mathbf{B}_k$) to generate $\mathbf{X}$ via (3). We adopt the viewpoint of [Misiakos et al., 2023], interpreting these shocks as external inputs to the SVAR and assuming that they exhibit a sparse structure. The structural shocks $\mathbf{s}_t$ consist of i.i.d. random variables with sparse structure, modeled with a Laplace(0, β) distribution. Sparse shocks correspond to intermittent external events that affect only a subset of variables at a given time; in financial markets, for example, large market-moving events occur sporadically and may directly impact only a subset of assets. Such variability facilitates disentangling causal effects and is a key ingredient underlying identifiability and empirical performance.

2.2 STABILITY OF LINEAR MODELS

We focus on linear models that are stable in the bounded-input bounded-output (BIBO) sense, i.e., bounded structural shocks $\mathbf{S}$ induce bounded observations $\mathbf{X}$. We distinguish three related notions of stability. First, we consider stability of the SEM, which means convergence of its powers to zero (nilpotency in the limit). Second, we consider *cycle contraction*, which constrains the amplification along directed cycles and will later play a key role for identifiability. Finally, we discuss stability of the SVAR, which ensures that time-series trajectories remain bounded.

Stability of an SEM In the SEM setting (1), the model is well-defined whenever $\mathbf{I} - \mathbf{B}$ is invertible, in which case

$$\mathbf{x} - \mathbf{x}\mathbf{B} = \mathbf{s} \iff \mathbf{x} = \mathbf{s}(\mathbf{I} - \mathbf{B})^{-1}. \quad (5)$$

Thus, invertibility of $\mathbf{I} - \mathbf{B}$ is the only algebraic requirement for the SEM to be well-defined and BIBO stable. In particular, when $\mathbf{B}$ corresponds to a DAG, this condition holds trivially, since the inverse admits the finite expansion $(\mathbf{I} - \mathbf{B})^{-1} = \mathbf{I} + \mathbf{B} + \dots + \mathbf{B}^{d-1}$ [Misiakos et al., 2023]. However, in many real-world systems, SEMs arise as steady-state limits of underlying dynamical processes [Lacerda et al., 2008]; e.g., in a river network where $\mathbf{s}$ is the pollution inflow at each node and $\mathbf{x}$ the accumulated pollution, the SEM (1) describes the system only after levels reach equilibrium. Throughout this work, we therefore assume that observations are collected in a regime that is approximately stationary at the sampling resolution, so that an equilibrium representation is meaningful.

More formally, the SEM (1) can be obtained as the stationary limit of the VAR(1) recursion

$$\mathbf{x}_t = \mathbf{x}_{t-1}\mathbf{B} + \mathbf{s}, \quad (6)$$

in the sense that $\lim_{t \rightarrow \infty} \mathbf{x}_t = \mathbf{x}$. This convergence holds if and only if the matrix $\mathbf{B}$ is stable, that is, if $\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}$.

Stability therefore emerges as a natural and sufficient condition for an SEM to be realizable as the equilibrium of an underlying linear dynamical system. Moreover, stability can be characterized through the spectral radius $\rho(\mathbf{B})$ of the adjacency matrix, defined as the magnitude of its largest eigenvalue. The following lemma summarizes this condition; a proof is provided in Appendix A.1.

Lemma 2.1. $\mathbf{B}$ is stable iff $\rho(\mathbf{B}) < 1$. In this case, the inverse admits the representation $(\mathbf{I} - \mathbf{B})^{-1} = \sum_{k=0}^{\infty} \mathbf{B}^k$.

Contractive cycles For digraphs with vertex-disjoint cycles, stability is equivalent to requiring *cycle contraction*: the product of edge weights along every directed cycle is strictly less than one [Lacerda et al., 2008]. Formally, a digraph with adjacency matrix $\mathbf{B}$ is said to have contractive cycles if, for every directed cycle C , $\prod_{(i_j, i_{j+1}) \in C} |b_{i_j i_{j+1}}| < 1$. This condition ensures that the effect of repeatedly traversing any cycle decays geometrically rather than amplifying. In general, cycle contraction and digraph stability are not equivalent and neither property implies the other: a digraph may have all cycles contractive yet be unstable due to interactions across multiple cycles, or be stable while containing individual cycles whose edge-weight products exceed one.

Example 2.2. Figure 1 illustrates this distinction using two digraphs with intersecting cycles. The first digraph $\mathbf{B}$ has two contractive cycles, each with edge-weight product 0.9, but because their effects combine into a joint gain exceeding one, $\mathbf{B}$ is unstable with $\rho(\mathbf{B}) = 1.22$. The second digraph $\tilde{\mathbf{B}}$ contains a cycle whose edge-weight product has magnitude 1.05 (non-contractive); however, this cycle carries a negative sign (red edge -2.1), so the joint gain of both cycles is -0.15 and $\tilde{\mathbf{B}}$ is stable with radius $\rho(\tilde{\mathbf{B}}) = 0.53$.

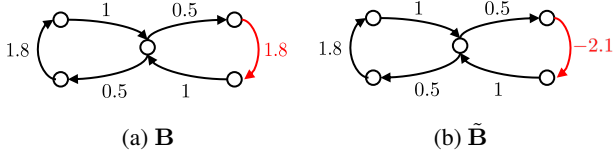


Figure 1: Example with two digraphs $\mathbf{B}, \tilde{\mathbf{B}}$, differing in one edge (red). $\mathbf{B}$ has contractive cycles but is unstable while $\tilde{\mathbf{B}}$ is stable while having a non-contractive cycle.

Nevertheless, as we show later, contractive cycles constitute a sufficient condition for establishing identifiability in digraphs. Consequently, our method enforces cycle contraction through a regularizer that uses $\rho(|\mathbf{B}|)$.

Stability of an SVAR Time series generated by an SVAR model necessarily require stability in order to prevent divergence or numerical explosion of the trajectories. The SVAR(k) model in (3) can be equivalently written as

$$\mathbf{x}_t(\mathbf{I} - \mathbf{B}_0) = \sum_{i=1}^k \mathbf{x}_{t-i} \mathbf{B}_i + \mathbf{s}_t, \quad (7)$$

where $\mathbf{I} - \mathbf{B}_0$ is assumed to be invertible. Multiplying both sides by $(\mathbf{I} - \mathbf{B}_0)^{-1}$ yields the reduced-form VAR(k) representation

$$\mathbf{x}_t = \sum_{i=1}^k \mathbf{x}_{t-i} \mathbf{A}_i + \tilde{\mathbf{s}}_t, \quad \mathbf{A}_i = \mathbf{B}_i(\mathbf{I} - \mathbf{B}_0)^{-1}. \quad (8)$$

The stability of VAR(k) processes has been extensively studied (e.g., Lütkepohl [2005]). Stability holds if and only if all eigenvalues of the companion matrix $\Phi(\mathbf{W})$, as analyzed in Appendix A.1, lie strictly inside the unit circle. Equivalently, the characteristic polynomial

$$\det\left(\mathbf{I} - \sum_{i=1}^k \mathbf{A}_i z^{-i}\right) \quad (9)$$

has no zeros for $|z| \geq 1$. However, as shown in Lemma A.5, stability of the SVAR does not impose additional constraints on the instantaneous effect matrix $\mathbf{B}_0$ beyond the invertibility of $\mathbf{I} - \mathbf{B}_0$. Consequently, SVAR stability does not restrict the admissible values of $\mathbf{B}_0$ and thus leaves the contemporaneous causal structure it encodes unconstrained. Therefore, we explicitly enforce $\rho(\mathbf{B}_0) < 1$, which in turn guarantees invertibility of $\mathbf{I} - \mathbf{B}_0$ by Lemma 2.1.

3 METHODOLOGY

In this section, we establish identifiability for digraphs with contractive cycles, introduce a regularizer that promotes cycle contraction via $\rho(|\mathbf{B}|)$, formulate a maximum likelihood estimator (MLE) for learning the window graph $\mathbf{W}$ from data $\mathbf{X}$, and analyze the consistency of our estimator.

Identifiability Valid causal discovery from observational data requires assumptions beyond model fitting: that the data-generating process follows a structural causal model, causal sufficiency (no latent confounders), and identifiability, i.e., that the observational distribution uniquely determines the underlying causal digraph.

We show that for the structural models in (2) and (4), cycle contraction is sufficient to guarantee identifiability. We first establish identifiability of the adjacency matrix $\mathbf{B}$ in the SEM setting (Theorem 3.1), and then extend the result to $\mathbf{B}_0$ and the whole window graph $\mathbf{W}$ of the SVAR model (Theorem 3.3). Our criterion applies to generic digraphs and strictly generalizes classical identifiability results for DAGs [Shimizu et al., 2006] and digraphs with disjoint stable cycles [Lacerda et al., 2008], under non-Gaussian noise.

Theorem 3.1. *Consider data $\mathbf{X}$ generated via an SEM (2) with non-Gaussian structural shocks $\mathbf{S}$ and a digraph $\mathbf{B}$ whose cycles are contractive. Then $\mathbf{B}$ is identifiable from the distribution of $\mathbf{X}$.*

Proof sketch. The proof builds on the characterization of distributional equivalence classes for cyclic linear SEMs with non-Gaussian noise by Drton et al. [2025], which shows that equivalent digraphs are related through cycle reversals. Crucially, this equivalence characterization holds for *arbitrary* digraphs—not only the disjoint-cycle graphs targeted by their discovery algorithm—so our result establishes identifiability for general digraphs. Among all graphs in an equivalence class, at most one has all cycles contractive. Therefore, contractiveness uniquely selects $\mathbf{B}$ within its equivalence class, establishing identifiability. A complete proof is provided in Theorem A.7. $\square$

Motivated by the requirement that an SEM admits a meaningful realization as the equilibrium of a dynamical system (6), and by the sufficiency of cycle contraction for identifiability (Theorem 3.1), we introduce a structural constraint for digraphs to be used as a regularizer in our continuous optimization framework. While stability follows from requiring the spectral radius of $\mathbf{B}$ to be less than one (Lemma 2.1) and identifiability from contractive cycles, we enforce both properties simultaneously through a single constraint on the spectral radius of the entrywise absolute adjacency matrix $|\mathbf{B}|$. Note that the idea of using positive entries was introduced in NOTEARS [Zheng et al., 2018] (which uses squared weights) within the acyclicity constraint. If one were to use the raw weights, certain cycles could cancel through sign patterns and satisfy the constraint even though the digraph is not acyclic. We refer to this unified condition as *absolute stability* and formalize it in Lemma 3.2. A complete proof is provided in Appendix A.1.

Lemma 3.2. *If $\rho(|\mathbf{B}|) < 1$, then $\mathbf{B}$ is stable and all directed cycles in $\mathbf{B}$ are contractive.*

Having established identifiability for SEMs under cycle contraction, we now extend the result to the time-series setting. This extension is nontrivial, since the observational distribution depends on both the contemporaneous structure $\mathbf{B}_0$ and the lagged effects $\mathbf{B}_1, \dots, \mathbf{B}_k$. Following the VAR-LiNGAM approach, we first regress out the lagged components, yielding non-Gaussian residuals that satisfy a linear SEM. Theorem 3.1 can then be applied to identify $\mathbf{B}_0$, from which the full window graph $\mathbf{W}$ is recovered. We therefore conclude that the same structural conditions suffice to identify the window graph of an SVAR (4). A complete proof is provided in Theorem A.8.

Theorem 3.3. *Consider $\mathbf{X}$ generated by an SVAR(k) (4) with non-Gaussian structural shocks $\mathbf{S}$ and a contemporaneous matrix $\mathbf{B}_0$ that has contractive cycles. Then the window graph $\mathbf{W}$ is identifiable from the distribution of $\mathbf{X}$.*

Maximum likelihood estimator Having established identifiability of $\mathbf{W}$ under cycle contraction, we now focus on its efficient estimation. We adopt a maximum likelihood framework, which provides a principled approach for estimating $\mathbf{W}$ by maximizing the likelihood of the observed data $\mathbf{X}$ generated via (4). Specifically, an estimate $\widehat{\mathbf{W}}$ can be obtained by maximizing the data likelihood, or equivalently the log-likelihood $\mathcal{L}(\mathbf{W}, \beta; \mathbf{X}) = \log f_{\mathbf{X}}(\mathbf{X} | \mathbf{W}, \beta)$, yielding

$$\widehat{\mathbf{W}} = \arg \max_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{W}, \beta; \mathbf{X}). \quad (10)$$

When $\mathbf{S}$ in (4) is modeled using a Laplace(0, β) distribution, the MLE admits the equivalent optimization-based formulation [Misiakos and Püschel, 2025]:

$$\begin{aligned} \widehat{\mathbf{W}} &= \arg \max_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{W}, \widehat{\beta}; \mathbf{X}) \\ &= \arg \min_{\mathbf{W} \in \mathcal{W}} \log \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 - \frac{1}{d} \log |\det(\mathbf{I} - \mathbf{B}_0)|. \end{aligned} \quad (11)$$

The set $\mathcal{W}$ consists of window graphs whose contemporaneous matrix $\mathbf{B}_0$ satisfies absolute stability, i.e., $\rho(|\mathbf{B}_0|) < 1$ as characterized in Lemma 3.2, and whose lagged components correspond to a stable SVAR in the sense of (9). By Theorem 3.3, the window graph $\mathbf{W}$ is identifiable from the observational distribution of $\mathbf{X}$. Consequently, in the infinite-sample regime where the empirical log-likelihood converges to its population counterpart, the optimizer $\widehat{\mathbf{W}}$ coincides with the ground-truth window graph $\mathbf{W}$.

Estimator consistency The population likelihood considers the infinite-sample regime, whereas practical applications necessarily operate with finite data. A key statistical guarantee in this setting is consistency, namely that the estimator $\widehat{\mathbf{W}}$ converges to the true window graph $\mathbf{W}$ as the sample size N increases for SEMs, or as the time horizon grows T for SVARs. We establish consistency of the proposed estimator in Theorem 3.4. Note that our result extends

prior work [Misiakos and Püschel, 2025] in two important directions. First, it accommodates cyclic contemporaneous structures through the inclusion of cyclic $\mathbf{B}_0$. Second, the proof relies on the ergodicity of the time series and therefore does not require multiple independent samples.

Theorem 3.4. *The maximum log-likelihood estimator defined in (11) is consistent under the following assumptions:*

- *The search space of window graphs is restricted to stable matrices $\mathcal{W} = \{\mathbf{W} \in [-b, b]^{d(k+1) \times d} : \rho(|\mathbf{B}_0|) \leq r, \rho(\Phi(\mathbf{W})) \leq r\}$.*
- *The Laplace scale parameter satisfies $\beta \in [a, b]$ with $0 < a \leq b$.*

Proof sketch. The proof proceeds in three main steps. First, compactness of the parameter space $\mathcal{W}$ is established via the boundedness of $\mathbf{W}$ and β , together with continuity of the spectral radius. Second, identifiability of $(\mathbf{W}, \beta)$ follows from Theorem 3.3. Finally, under these conditions, Theorem 2.5 of Newey and McFadden [1994] applies, yielding consistency of $\widehat{\mathbf{W}}$ via a strong uniform law of large numbers (ULLN). The strong ULLN relies on ergodicity of the underlying time series, which for our stable SVAR follows from strong mixing; intuitively, mixing ensures that temporal dependencies decay over time, so that observations far apart behave approximately independently. A complete proof is provided in Appendix A.6. $\square$

Methodology: StableSpIn Having established identifiability and consistency of the MLE, we consider its practical optimization. Since directly solving (10) is challenging, we adopt a Lagrangian relaxation that combines the likelihood term with penalties promoting sparsity, absolute stability, and the absence of self-loops. Specifically, we optimize

$$\begin{aligned} \widetilde{\mathbf{W}} &= \arg \min_{\mathbf{W} \in \mathbb{R}^{d(k+1) \times d}} \log \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \\ &\quad - \frac{1}{d} \log |\det(\mathbf{I} - \mathbf{B}_0)| + \lambda_1 \|\mathbf{W}\|_1 \\ &\quad + \lambda_2 \cdot \max(\rho(|\mathbf{B}_0|) - 1, 0) + \lambda_3 \cdot \text{tr}(\mathbf{B}_0 \odot \mathbf{B}_0). \end{aligned} \quad (12)$$

The ℓ_1 penalty promotes sparsity in the window graph, the term $\max(\rho(|\mathbf{B}_0|) - 1, 0)$ softly enforces absolute stability by penalizing violations of the spectral-radius, while $\text{tr}(\mathbf{B}_0 \odot \mathbf{B}_0)$ discourages self-loops. We constrain only $\mathbf{B}_0$, as identifiability requires (Theorems 3.1 and 3.3); the lagged blocks are fit by regression, and full-SVAR stability is ensured at data generation by bounding edge weights and node degrees. The penalty activates only above $\rho(|\mathbf{B}_0|) = 1$, but a large enough λ_2 keeps the recovered radius strictly below one.

Implementation and time complexity We approximate the spectral radius of the nonnegative matrix $|\mathbf{B}_0|$ using the Perron-Frobenius power method (see Exercise 8.5.P16 in Horn and Johnson [2012]). Starting from a normalized positive vector $\mathbf{x}^{(0)}$, for $k = 0, \dots, M - 1$ we iterate

$$\mathbf{y}^{(k+1)} = |\mathbf{B}_0| \mathbf{x}^{(k)}, \mathbf{x}^{(k+1)} = \frac{\mathbf{y}^{(k+1)}}{\mathbf{1}^\top \mathbf{y}^{(k+1)}}. \quad (13)$$

The normalization scalar $\mathbf{1}^\top \mathbf{y}^{(k+1)}$ provides an estimate of the Perron root and converges to $\rho(|\mathbf{B}_0|)$ when $|\mathbf{B}_0|$ represents a connected digraph. Each iteration requires a matrix–vector multiplication, so for M iterations the stability constraint costs $\mathcal{O}(Md^2)$, replacing the cubic-time matrix exponential [Moler and Van Loan, 2003] of the NOTEARS acyclicity constraint [Zheng et al., 2018] used in prior work [Ng et al., 2020, Misiakos and Püschel, 2025]. The full per-iteration cost is $\mathcal{O}(Td^2 + d^3)$, where the $\mathcal{O}(Td^2)$ data term dominates whenever $d \lesssim T$ (e.g., our real data has $d = 97 \ll T = 1220$); the $\mathcal{O}(d^3)$ log-determinant $\log|\det(\mathbf{I} - \mathbf{B}_0)|$ for cyclic $\mathbf{B}_0$ is thus not a separate bottleneck and remains tractable to $d = 10,000$. Hence StableSpIn is asymptotically faster than combinatorial cyclic methods and, while matching the per-iteration order of continuous DAG optimizers, is faster in practice because the power iteration and a single factorization-based log-determinant [Golub and Van Loan, 2013] are cheaper and more GPU-amenable than the matrix exponential; the sparse-input assumption further improves sample efficiency and accuracy. A power-series log-determinant estimate (valid since $\rho(|\mathbf{B}_0|) < 1$) is available but proved less accurate, so we evaluate it exactly.

Remark 3.5. Our methodology and theory for SVAR time-series applies equally to cross-sectional SEM data. One can simply replace $\mathbf{W}$ with $\mathbf{B}$, $\mathbf{X}_{\text{past}}$ by $\mathbf{X}$ in (12), and the time index t by the sample index n in Theorem 3.4.

4 RELATED WORK

Acyclic causal discovery Causal discovery in DAGs has been extensively studied and can be broadly categorized into constraint-based, score-based, noise-based, and optimization-based approaches. Constraint-based methods, such as PC [Spirtes et al., 2000], infer causal structure from conditional independence relations, while score-based methods, including GES [Chickering, 2002], search over DAG space using likelihood-based criteria. Noise-based approaches exploit asymmetries in the data-generating process, most notably through non-Gaussianity as in LiNGAM [Shimizu et al., 2006], and have also been extended to time-series settings via VAR-LiNGAM [Hyvärinen et al., 2010]. A key theme underlying recent progress is the use of continuous constraints and optimization-based formulations, which substantially improve scalability compared to combinatorial search. In particular, Zheng

et al. [2018] introduced NOTEARS, which characterizes acyclicity through a smooth constraint and enables efficient gradient-based learning. Subsequent methods built on this framework to improve the acyclicity characterization itself, e.g., via a log-determinant formulation [Bello et al., 2022], and to formulate likelihood-based approaches under Gaussian [Ng et al., 2020] or sparse structural shocks [Misiakos et al., 2023, Misiakos and Püschel, 2025].

Cyclic causal discovery Causal discovery with feedback loops has been studied using constraint-based, score-based, and likelihood-based approaches; however, unlike the acyclic case, the literature largely lacks continuous structural constraints capable of characterizing general cyclic digraphs, which limits scalability. Early work focused on constraint-based methods such as CCD and its extensions [Richardson, 1996a, Hyttinen et al., 2012, Forré and Mooij, 2018], which rely on conditional independence testing and combinatorial search. Score- and likelihood-based alternatives, including DGLearn and mixed-graph search [Ghassami et al., 2020, Améndola et al., 2020, Mooij and Heskes, 2013], remain computationally demanding and are typically restricted to small digraphs. Other approaches exploit additional structure, such as shift interventions or domain-specific assumptions [Rothenhäusler et al., 2015, Sanchez-Romero et al., 2019, Arab et al., 2025]. Overall, existing methods are either combinatorial or highly specialized, primarily target cross-sectional data, and typically do not impose or exploit stability of the underlying cyclic system.

Non-Gaussian SVAR-based models Early work on non-Gaussian cyclic causal discovery was initiated by Lacerda et al. [2008], who extended LiNGAM to cyclic settings under a contractive cycles assumption, but with restricted focus to vertex-disjoint cycles and cross-sectional data. Subsequent work refined this by targeting local causal dependencies [Dai et al., 2024]. More recently, Drton et al. [2025] provided a complete characterization of distribution-equivalence classes for linear non-Gaussian cyclic SEMs via cycle reversals, and proposed a moment-based method for learning disjoint cyclic digraphs, again focusing on cross-sectional data. This equivalence characterization was very recently extended to settings with latent confounders [Dai et al., 2026]. Another related approach is Two-Step [Sanchez-Romero et al., 2019], which combines adjacency search with ICA-based estimation and accommodates both cyclic causal structures and latent confounders. Time-series extensions largely adopt different modeling assumptions. Roy et al. [2023] study cyclic dependencies in EEG data using Bayesian functional models, assuming stability of the underlying operator but without enforcing contractive cycles in the optimization. Similarly, BayesDCG [Jin et al., 2025] imposes stability in a Bayesian time-series framework but relies on instrumental variables for identifiability. A related econometric literature studies

SVARs under independent non-Gaussian shocks [Lanne et al., 2017, Gouriéroux et al., 2017], achieving ICA-type identifiability of the contemporaneous matrix up to permutation and scaling, which suffices for identifying impulse responses but does not recover the causal digraph. Overall, although these approaches share with our method the use of non-Gaussian structural shocks and the exclusion of self-loops, none explicitly enforce contractive cycles. Among existing methods, SpinSVAR [Misiakos and Püschel, 2025] is most closely related to our work, as it combines the sparse-input assumption with continuous optimization for time-series causal discovery. However, it is restricted to DAGs through a NOTEARS-style acyclicity constraint, whereas StableSpIn employs a stability-based constraint that permits cyclic digraphs. This changes fundamentally the class of admissible causal structures and the associated theory.

5 EXPERIMENTS

We conduct synthetic and real-data experiments to evaluate the performance of StableSpIn in recovering a digraph $\mathbf{B}$ from data $\mathbf{X}$ generated via a stable structural model. For time-series data, StableSpIn instead learns a window graph $\mathbf{W}$ corresponding to the underlying stable SVAR model.

Baselines We consider the following baselines. Among conditional independence testing methods, we include CCD [Richardson, 1996a] and FRP [Yi and Lee, 2024]. Among distributional approaches, we include DGlearn [Ghassami et al., 2020] and Backshift [Rothenhäusler et al., 2015], both designed for static data. We further consider BayesDCG [Jin et al., 2025] for temporal data, as well as the fMRI-inspired methods FASK [Sanchez-Romero et al., 2019] and CallTiF [Arab et al., 2025], adapted to static and temporal settings respectively. Most relevant to our assumptions are methods based on non-Gaussian structural shocks, namely LiNG-D [Lacerda et al., 2008] for static disjoint cycles, VARLiNGAM [Hyvärinen et al., 2010] for DAGs in time-series, SparseRC [Misiakos et al., 2023] for DAGs generated with SEMs with sparse shocks and its temporal extension SpinSVAR [Misiakos and Püschel, 2025], DisjointCycles [Drton et al., 2025] for static cyclic models, and NGML [Lanne et al., 2017] for non-Gaussian SVARs employing maximum likelihood. Methods originally developed for static data are extended to time-series settings by first estimating a VAR model, extracting the residuals, and then applying the static method to the resulting non-Gaussian residuals. Although this procedure can introduce additional estimation error, it enables these methods to operate in temporal contexts. We also apply DAG learning methods (e.g., SparseRC, VARLiNGAM) without modification, even when the acyclicity assumption may be violated, allowing us to assess the impact of misspecified assumptions. Hyperparameters are selected via grid search on synthetic experiment with 20 nodes and are reported in Appendix B.2.

Metrics As performance metrics, we report the normalized structural Hamming distance (nSHD), defined as the number of incorrectly recovered edges (false positives and false negatives) divided by the total number of ground-truth edges, as well as runtime in seconds. Additional metrics, including parameter recovery measured via normalized mean squared error (NMSE), are provided in Appendix B.1. For real data, where no ground truth is available, we qualitatively assess the learned digraphs through visualization and analysis of the cycles persistence and length distribution.

5.1 SYNTHETIC EXPERIMENTS

Data generation We consider synthetic experiments for both static and time-series data generation. For time-series experiments, we use SVAR(2) models throughout. For the static scenario, we generate a random digraph with average degree 5 per node and generate a weighted $\mathbf{B}$ with weights from $[-0.5, -0.1] \cup [0.1, 0.5]$ which ensures contractive cycles. Additionally, we normalize by $\rho(\mathbf{B})$ to ensure stability. The data $\mathbf{X}$ are generated via (2). To satisfy a sparse model for structural shocks we consider the simple scenario of sparse Bernoulli structural shocks $\mathbf{S}$ [Misiakos et al., 2023], namely an entry at $\mathbf{S}$ is a nonzero uniformly chosen weight from $[-1, -0.1] \cup [0.1, 1]$ with probability $p = 0.05$ and zero otherwise. The digraph generation is the same for the SVAR (4): we generate $\mathbf{B}_0$ in the same manner and the lagged matrices $\mathbf{B}_i$ with average degree 2, ensuring stability by discarding any time series that overflows. As a special case for these setups we further consider scenarios where $\mathbf{B}_0$ (or $\mathbf{B}$) has disjoint cycles or is a DAG, shown in Appendix B.1.

The synthetic experiments are intentionally designed to match the sparse-input setting for which StableSpIn is developed. Their purpose is therefore to evaluate performance under the assumptions required by the theory rather than to demonstrate universal superiority across all data-generating mechanisms. We also note that the structural shocks are generated from a symmetric distribution, which may disadvantage methods that rely on distributional asymmetry for edge orientation, such as FASK [Sanchez-Romero et al., 2019]. To assess robustness, Appendix B.1.5 further considers settings that violate the sparse-input assumption.

Results We report synthetic results in Fig. 2. For each setting, we generate 5 independent random instances and report the mean (solid line) and standard deviation (shaded area). In the first two experiments (Fig. 2a and Fig. 2b), we evaluate all methods as the number of nodes d increases from 20 to 5,000. Performance is assessed both in terms of accuracy (nSHD, first row of Fig. 2) and efficiency (runtime, second row), with a time-out of 5,000 seconds per run. Most methods are either inaccurate—often reaching or exceeding the failure threshold $\text{nSHD} = 1$ (achievable

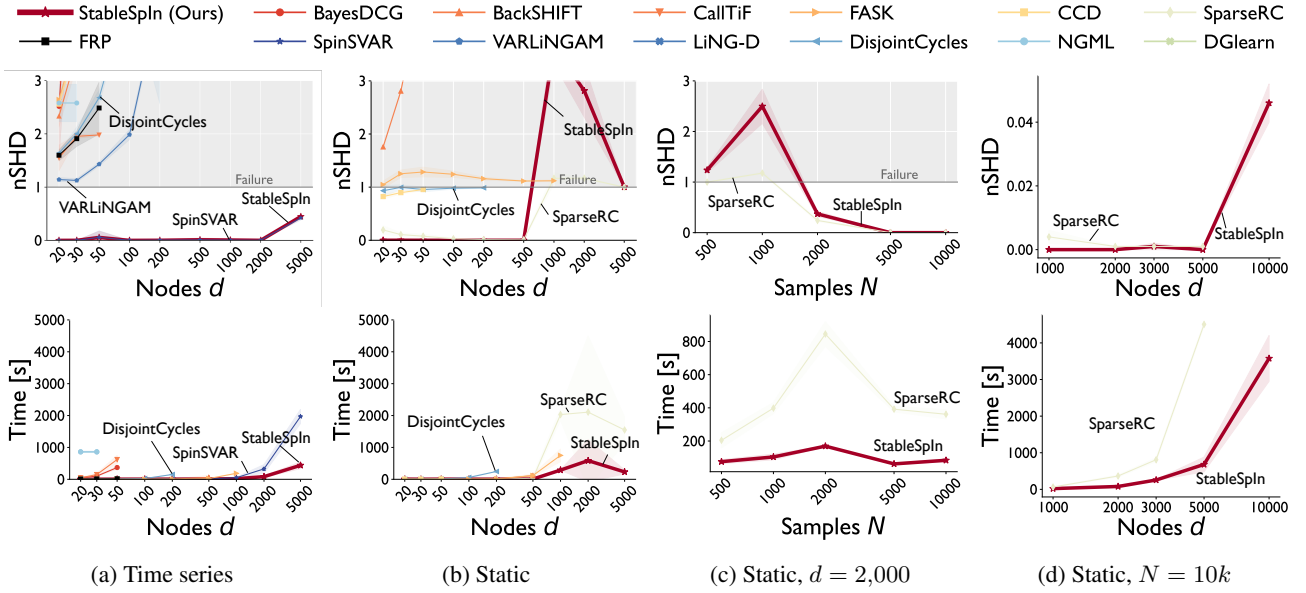


Figure 2: First row: nSHD (lower is better); second row: runtime. (a) SVAR time series with $T = 1,000$ and varying nodes d , (b) SEM data with $N = 1,000$ samples and varying d , (c) SEM data for $d = 2,000$ and varying N and (d) $N = 10k$ samples for larger digraphs. Any non-reported point implies a time-out (defined as execution time $> 5,000s \approx 1:23h$).

by predicting an empty digraph)—or fail to scale to large digraphs. Their degradation is largely due to our specialized data setup (sparse structural shocks) and the presence of general cyclic structures, which are not supported by some methods (e.g., DisjointCycles and LiNG-D). In particular, LiNG-D performs well only for digraphs with disjoint cycles (Appendix B.1), and fails or times out for $d \geq 50$. DisjointCycles also fails under our experimental setup.

We therefore focus on large digraphs and restrict attention to the two strongest baselines, SpinSVAR and SparseRC. Since both show similar trends, we present SparseRC here and defer SpinSVAR to Appendix B.1. In Fig. 2b, SparseRC makes a few errors on small digraphs, while StableSpIn achieves perfect recovery (nSHD = 0); both fail for $d \geq 1,000$. In Fig. 2c, both methods fail for $N = 500$ or $1,000$ samples and begin to recover the digraph for $N \geq 2,000$. Here, the relatively large sample requirements are not unexpected: $\mathbf{W}$ contains $d^2(k+1)$ parameters, making reliable estimation challenging in low-sample regimes. Nevertheless, the sparse-input assumption helps reduce the effective complexity and facilitates disentangling causal effects, which can lower sample requirements compared to fully dense settings. StableSpIn remains exact, whereas SparseRC misses 13–18 edges. Although both are reasonably accurate, StableSpIn achieves perfect recovery using roughly 1/8 of SparseRC’s runtime, owing to the low complexity of the stability constraint. Finally, in Fig. 2d, only StableSpIn completes within the 5,000-second limit; SparseRC times out. StableSpIn scales up to 5,000 nodes with zero errors and attains nSHD = 6% at 10,000 nodes. To our knowledge, this is the first cyclic causal discovery method capable of

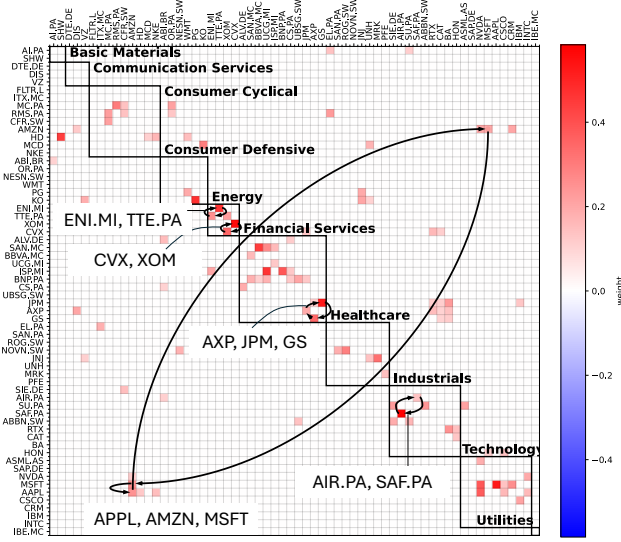
effectively learning cyclic digraphs at this scale.

Additional experiments are presented in Appendix B.1, covering an analogous time-series setting, results for SpinSVAR, denser digraphs with a larger number of cycles, and special cases such as disjoint cycles and DAGs. Appendix B.1.5 further investigates the effect of sparsity, ranging from fully dense Gaussian noise to the fully sparse non-Gaussian setting considered in our main experiments.

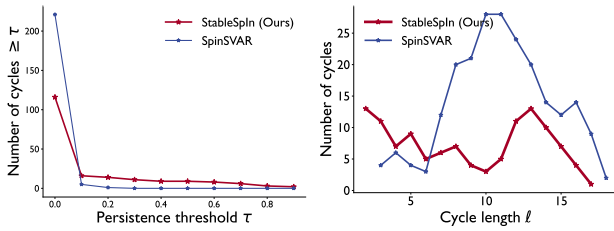
5.2 REAL EXPERIMENTS

Stock market data We evaluate our method on real data from the US and European stock market, using daily log-returns from 2020–2025 for the 100 largest stocks by market capitalization. After cleaning and retaining stocks with sufficient coverage over this period, we obtain 97 stocks and 1,220 time steps, matching the regime where StableSpIn performed reliably in the synthetic experiments. The data are strongly non-Gaussian: the Shapiro–Wilk test rejects normality at significance level $p < 0.01$ for all stocks. Results are shown in Fig. 3. We emphasize that, in the absence of ground truth and under the likely presence of latent confounders, this experiment should be viewed as exploratory rather than a validation of causal correctness. Our goal is to assess whether StableSpIn recovers structured and reproducible cyclic patterns in a real-world setting.

Cyclic causal discovery We apply StableSpIn and the time-series variant of the strongest synthetic baseline, SpinSVAR. The StableSpIn adjacency matrix (Fig. 3a) re-



(a) StableSpIn learned B_0



(b) Cycle persistence and length statistics

Figure 3: Real stock data: (a) StableSpIn B_0 adjacency matrix. (b) Left: number of cycles exceeding persistence threshold τ , right: distribution of cycle lengths.

veals clear clustering by financial sector, with 74% of edges being intra-country. SpinSVAR results (Appendix) show that both methods recover meaningful dependency structures. We argue, however, that StableSpIn captures stable and interpretable cyclic structures, whereas SpinSVAR tends to detect spurious ones. Figure 3a shows the five highest-weighted cycles found by StableSpIn. To assess robustness, we split the data into 16 non-overlapping windows, reran both methods, and measured the persistence of the initially discovered cycles, defined as the fraction of windows in which they reappear. Notably, length-2 cycles (mutual edges $i \leftrightarrow j$) such as (ENI.MI $\leftrightarrow$ TTE.PA, 100%), (CVX $\leftrightarrow$ XOM, 94%), and (AMZN $\leftrightarrow$ MSFT, 75%) remain highly persistent, indicating stable feedback relations. Aggregate persistence statistics are shown in the upper plot of Fig. 3b: for each threshold $\tau \in [0, 100]\%$, we count cycles present in at least a fraction τ of windows.

A further distinction appears in the cycle-length distribution. SpinSVAR enforces acyclicity via the soft constraint $\text{tr}(e^{B \odot B}) - d$, which is never exactly zero. In the series expansion of the matrix exponential, longer cycles are penalized less due to the $l!$ factor in the denominator. As a

result, SpinSVAR suppresses short feedback loops and instead detects long, community-like cycles that likely reflect latent confounding rather than direct causality. For example, one of its top cycles links six major European banks across countries (BBVA.MC, SAN.MC, UCG.MI, ISP.MI, INGA.AS, BNP.PA), plausibly reflecting common latent factors rather than direct causal interactions. The lower plot of Fig. 3b highlights this contrast: StableSpIn identifies cycles across a range of lengths, while SpinSVAR concentrates on longer cycles. Overall, StableSpIn detects more persistent cycles spread across meaningful lengths; while not evidence of ground-truth causal recovery, this indicates that it extracts reproducible cyclic structure despite substantial confounding and measurement noise.

6 LIMITATIONS

While our method improves upon prior work, it has several limitations. First, it is designed for sparse, non-Gaussian structural shocks and performs best when the digraph is sparse, so its applicability may be limited when shocks are dense or continuously active. Second, reliable estimation may require thousands of samples, though this is expected in high-dimensional settings where the window graph contains $d^2(k+1)$ parameters. Third, the method assumes no latent confounders, and in time-series settings no temporal undersampling, where causal interactions occurring between observation times may induce spurious dependencies. Consequently, the real-data results in Section 5.2 should be interpreted as exploratory rather than a validation of causal correctness; establishing true causality requires further validation and, ideally, interventional data. Finally, both stability and the sparse-input assumption are modeling assumptions rather than testable properties of the data (the unobserved shocks make sparsity hard to validate); if either is violated, the recovered graph should be interpreted with caution.

7 CONCLUSION

We proposed StableSpIn, a method for learning cyclic causal dependencies in linear structural models with sparse structural shocks. Our approach introduces a novel continuous constraint that characterizes causal digraphs through stability, quantified via the spectral radius of the absolute adjacency matrix. The constraint approximates the spectral radius using iterative vector multiplication, thereby avoiding the cubic-time complexity of the NOTEARS acyclicity constraint. This results in substantial computational gains over prior baselines and enables accurate causal discovery in digraphs with up to 10,000 nodes, which has not previously been achieved for cyclic models. Empirical results on stock market data further demonstrate that StableSpIn recovers meaningful and persistent cyclic dependencies across time windows, highlighting its practical relevance and robustness.

References

- Carlos Améndola, Philipp Dettling, Mathias Drton, Federica Onori, and Jun Wu. Structure learning for cyclic linear causal models. In *Conference on Uncertainty in Artificial Intelligence*, pages 999–1008. PMLR, 2020.
- Fahimeh Arab, AmirEmad Ghassami, Hamidreza Jamalabadi, Megan AK Peters, and Erfan Nozari. Whole-brain causal discovery using fMRI. *Network Neuroscience*, 9(1):392–420, 2025.
- Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.
- R.C. Bradley. *Introduction to Strong Mixing Conditions*. Number v. 1 in Introduction to Strong Mixing Conditions. Kendrick Press, 2007.
- Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer science & business media, 1991.
- David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- Haoyue Dai, Ignavier Ng, Yujia Zheng, Zhengqing Gao, and Kun Zhang. Local causal discovery with linear non-gaussian cyclic models. In *International Conference on Artificial Intelligence and Statistics*, pages 154–162. PMLR, 2024.
- Haoyue Dai, Immanuel Albrecht, Peter Spirtes, and Kun Zhang. Distributional equivalence in linear non-gaussian latent-variable cyclic causal models: Characterization and learning. In *International Conference on Learning Representations (ICLR)*, 2026.
- Mathias Drton, Marina Garrote-López, Niko Nikov, Elina Robeva, and Y Samuel Wang. Causal discovery for linear non-gaussian models with disjoint cycles. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.
- Patrick Forré and Joris M Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pages 269–278, 2018.
- AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International conference on machine learning*, pages 3494–3504. PMLR, 2020.
- Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.
- Christian Gouriéroux, Alain Monfort, and Jean-Paul Renne. Statistical inference for independent component analysis: Application to structural var models. *Journal of Econometrics*, 196(1):111–126, 2017.
- Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.
- 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.
- Wei Jin, Yang Ni, Amanda B Spence, Leah H Rubin, and Yanxun Xu. Directed cyclic graphs for simultaneous discovery of time-lagged and instantaneous causality from longitudinal data using instrumental variables. *Journal of Machine Learning Research*, 26(22):1–62, 2025.
- Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 366–374, 2008.
- Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2019.
- Markku Lanne, Mika Meitz, and Pentti Saikkonen. Identification and estimation of non-Gaussian structural vector autoregressions. *Journal of Econometrics*, 196(2):288–304, 2017.
- Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.
- Panagiotis Misiakos and Markus Püschel. SpinSVAR: Estimating Structural Vector Autoregression Assuming Sparse Input. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.
- Panagiotis Misiakos, Chris Wendler, and Markus Püschel. Learning DAGs from Data with Few Root Causes. *Advances in Neural Information Processing Systems*, 36, 2023.
- Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic processes and their applications*, 29(2):309–315, 1988.

- Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.
- Joris M Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 431–439, 2013.
- Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.
- Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the Role of Sparsity and DAG Constraints for Learning Linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- Benedikt M Pötscher and Ingmar R Prucha. A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica: Journal of the Econometric Society*, pages 675–683, 1989.
- Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461, 1996a.
- Thomas Richardson. A polynomial-time algorithm for deciding markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 462–469, 1996b.
- Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. *Advances in neural information processing systems*, 28, 2015.
- Saptarshi Roy, Raymond KW Wong, and Yang Ni. Directed cyclic graph for causal discovery from multivariate functional data. *Advances in neural information processing systems*, 36:42762–42774, 2023.
- 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.
- Karen Sachs, Omar Perez, Dana Pe’er, Douglas A Lauenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.
- Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.
- Muralikrishna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6371–6387. PMLR, 2023.
- Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL <http://jmlr.org/papers/v7/shimizu06a.html>.
- Christopher A Sims. Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered, 1980.
- Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for FMRI. *Neuroimage*, 54(2):875–891, 2011.
- Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.
- Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTARS: Learning Nonparametric DBNs With Prior Knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 1942–1964. PMLR, 2023.
- Soheun Yi and Sanghack Lee. Filter, Rank, and Prune: Learning Linear Cyclic Gaussian Graphical Models. In *International Conference on Artificial Intelligence and Statistics*, pages 1135–1143. PMLR, 2024.
- Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Learning Stable Cyclic Graphs from Linear Structural Models (Supplementary Material)

Panagiotis Misiakos¹

Markus Püschel¹

¹Department of Computer Science, ETH Zurich, Zürich, Switzerland

A THEORETICAL ANALYSIS

A.1 STABILITY

Lemma A.1. *For any initial condition $\mathbf{x}_0$, the limit $\lim_{t \rightarrow \infty} \mathbf{x}_t$ exists and equals the fixed-point solution $\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{s}$ of (1) if and only if $\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}$.*

Proof. From (6), we obtain by unrolling the recursion

$$\mathbf{x}_t = \mathbf{x}_0 \mathbf{B}^t + \mathbf{s} (\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^{t-1}).$$

We first prove the forward direction. Assume $\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}$. Then from Horn and Johnson [2012] (Th. 5.6.12), $\rho(\mathbf{B}) < 1$, and consequently

$$(\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^{t-1}) \rightarrow (\mathbf{I} - \mathbf{B})^{-1}.$$

Therefore,

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbf{x}_t &= \lim_{t \rightarrow \infty} \mathbf{x}_0 \mathbf{B}^t + \mathbf{s} (\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^{t-1}) \\ &= \mathbf{0} + \mathbf{s} (\mathbf{I} - \mathbf{B})^{-1}. \end{aligned}$$

For the reverse direction, assume that $\mathbf{x}_t \rightarrow \mathbf{x}$. Consider two arbitrary initial conditions $\mathbf{x}_0$ and $\mathbf{x}'_0$. The corresponding trajectories satisfy

$$\mathbf{x}_t - \mathbf{x}'_t = (\mathbf{x}_0 - \mathbf{x}'_0) \mathbf{B}^t.$$

Since both sequences converge, it must hold that $\lim_{t \rightarrow \infty} (\mathbf{x}_0 - \mathbf{x}'_0) \mathbf{B}^t = \mathbf{0}$. Because the vector $(\mathbf{x}_0 - \mathbf{x}'_0)$ is arbitrary, this implies $\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}$, as required. $\square$

Lemma A.1 motivates the following definition of stable digraphs, which has also been adopted in prior work Lacerda et al. [2008].

Definition A.2. A digraph $\mathcal{G}(V, \mathbf{B})$ is said to be *stable* if

$$\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}.$$

Next, we show that this notion of stability is equivalent to the classical BIBO stability of linear dynamical systems. In particular, we provide a necessary and sufficient characterization of stability in terms of the eigenvalues of $\mathbf{B}$.

Lemma A.3. *BIBO stability of the VAR(1) process (6) is equivalent to $\rho(\mathbf{B}) < 1$, which is further equivalent to*

$$\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}.$$

In this case, the inverse admits the representation

$$(\mathbf{I} - \mathbf{B})^{-1} = \sum_{k=0}^{\infty} \mathbf{B}^k.$$

Proof. The transfer function associated with (6) is

$$H(z) = (\mathbf{I} - \mathbf{B}z^{-1})^{-1}.$$

BIBO stability holds if and only if all poles of $H(z)$ lie strictly inside the unit circle Lütkepohl [2005]. Equivalently,

$$\det(\mathbf{I} - \mathbf{B}z^{-1}) = 0 \Leftrightarrow \det(z\mathbf{I} - \mathbf{B}) = 0 \Rightarrow |z| < 1.$$

This condition is equivalent to all eigenvalues of $\mathbf{B}$ having modulus strictly less than 1, that is, $\rho(\mathbf{B}) < 1$. Moreover, by Horn and Johnson [2012] (Thm. 5.6.12), $\rho(\mathbf{B}) < 1$ if and only if $\lim_{t \rightarrow \infty} \mathbf{B}^t = \mathbf{0}$, and by Cor. 5.6.16, it implies that $\mathbf{I} - \mathbf{B}$ is invertible and admits the desired representation $\square$

Finally, we prove that our chosen regularizer, motivated via stability and in particular promoting stability in $|\mathbf{B}|$, also promotes cycle contraction.

Lemma A.4. *If $\rho(|\mathbf{B}|) < 1$, then $\mathbf{B}$ is stable and has contractive cycles.*

Proof. From Horn and Johnson [2012][Th. 8.1.18] we have that $\rho(\mathbf{B}) \leq \rho(|\mathbf{B}|) < 1$, thus $\mathbf{B}$ is stable from Lemma 2.1.

Moreover, since $\rho(|\mathbf{B}|) < 1$ from Lemma 2.1 it follows that $\lim_{t \rightarrow \infty} |\mathbf{B}|^t = \mathbf{0}$, which implies that $|\mathbf{B}|_{ii}^t$ goes to 0. $|\mathbf{B}|_{ii}^t$ represents the sum of the absolute weights of all directed closed walks of length t from node i to itself.

Now assume in contrary that there is an unstable cycle C_0 of length m ,

$$|\mathbf{B}_{C_0}| = \prod_{(i_j, i_{j+1}) \in C_0} |b_{i_j i_{j+1}}| \geq 1.$$

Take any node $i \in C_0$. Then

$$|\mathbf{B}|_{ii}^{tm} = \sum_{\text{len}(C)=tm} |\mathbf{B}_C| > \left(\prod_{(i_j, i_{j+1}) \in C_0} |b_{i_j i_{j+1}}| \right)^t \geq 1,$$

absurd. Therefore, every directed cycle is contractive. $\square$

The next lemma shows that SVAR stability doesn't impose any restriction on $\mathbf{B}_0$.

Lemma A.5. *Fix any $\mathbf{B}_0 \in \mathbb{R}^{d \times d}$ such that $\mathbf{I} - \mathbf{B}_0$ is invertible. Stability of an SVAR(k) model with instantaneous matrix $\mathbf{B}_0$ is equivalent to stability of its reduced-form VAR(k) representation, and therefore imposes no additional restrictions on $\mathbf{B}_0$ beyond invertibility of $\mathbf{I} - \mathbf{B}_0$.*

Proof. Stability of an SVAR(k) is determined by the roots of (9) and hence depends only on the reduced-form coefficients $(\mathbf{A}_1, \dots, \mathbf{A}_k)$. The mapping

$$(\mathbf{B}_1, \dots, \mathbf{B}_k) \longleftrightarrow (\mathbf{A}_1, \dots, \mathbf{A}_k)$$

is bijective for any fixed $\mathbf{B}_0$ with $\mathbf{I} - \mathbf{B}_0$ invertible. Consequently, for any stable reduced-form VAR(k), and for any such $\mathbf{B}_0$, there exists a unique SVAR representation inducing identical dynamics. Therefore, imposing stability on the SVAR constrains only the lag dynamics and does not restrict $\mathbf{B}_0$ beyond invertibility of $\mathbf{I} - \mathbf{B}_0$. $\square$

Companion matrix form Another form for the SVAR comes from the companion matrix. If we rewrite the SVAR in its reduced-form VAR(k) representation (8):

$$\mathbf{x}_t = \sum_{i=1}^k \mathbf{x}_{t-i} \mathbf{A}_i + \tilde{\mathbf{s}}_t, \quad \mathbf{A}_i = \mathbf{B}_i (\mathbf{I} - \mathbf{B}_0)^{-1}, \quad (14)$$

we can define the companion matrix $\Phi = \Phi(\mathbf{W})$ as

$$\Phi = \begin{pmatrix} \mathbf{B}_1 (\mathbf{I} - \mathbf{B}_0)^{-1} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{B}_2 (\mathbf{I} - \mathbf{B}_0)^{-1} & \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{k-1} (\mathbf{I} - \mathbf{B}_0)^{-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} \\ \mathbf{B}_k (\mathbf{I} - \mathbf{B}_0)^{-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{A}_2 & \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{k-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} \\ \mathbf{A}_k & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \quad (15)$$

Considering the sequence $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-k+1})$ and $\mathbf{n}_t = (\mathbf{s}_t, 0, \dots, 0)$ then

$$\mathbf{z}_t = \mathbf{z}_{t-1} \Phi + \mathbf{n}_t \quad (16)$$

Assuming stability of the SVAR, which is equivalent to $\rho(\Phi) < 1$ [Lütkepohl, 2005] we have the following equation:

$$\mathbf{z}_t = \sum_{k=0}^{\infty} \mathbf{n}_{t-k} \Phi^k \quad (17)$$

$\rho(\Phi) < 1$ implies stability, which means absolute summability:

$$\sum_{k=0}^{\infty} \|\Phi^k\| < \infty \quad (18)$$

Impulse Response Functions. Impulse response functions (IRFs) quantify the effect of a one-time unit structural shock in component i of $\mathbf{s}_t \in \mathbb{R}^d$ on the evolution of the time series $\mathbf{x}_t$. For an SVAR, IRFs describe the propagation of structural shocks across time. Given the reduced-form VAR(k) representation (8), the IRF is given by the sequence of matrices $\{\Psi_h\}_{h \geq 0}$ defined recursively as

$$\Psi_0 = \mathbf{I}, \quad \Psi_h = \sum_{i=1}^{\min(h,k)} \Psi_{h-i} \mathbf{A}_i. \quad (19)$$

For a given structural shock $\mathbf{s} \in \mathbb{R}^d$, the impulse response at horizon h is the vector $\mathbf{x}_{t+h} = \mathbf{s} \Psi_h$. This yields the convolution representation of (8)

$$\mathbf{x}_t = \sum_{h=0}^{\infty} \tilde{\mathbf{s}}_{t-h} \Psi_h, \quad (20)$$

which corresponds to the classical signal-processing input–output relation. The existence and summability of the IRF sequence require stability of the underlying VAR dynamics. For a stable SVAR we have that:

$$\sum_{k=0}^{\infty} \|\Psi_k\| < \infty \quad (21)$$

Example As a simple illustrative case, consider a VAR(1) process (6). In this case, the IRF recursion (19) admits the closed-form solution

$$\Psi_h = \mathbf{B}^h, \quad h \geq 0,$$

and the response to a one-time structural shock $\mathbf{s}$ is given by $\mathbf{x}_{t+h} = \mathbf{s} \mathbf{B}^h$. Under stability, i.e., $\rho(\mathbf{B}) < 1$, the convolution series converges and yields the SEM expansion

$$\mathbf{x}_t = \sum_{h=0}^{\infty} \mathbf{s}_{t-h} \mathbf{B}^h = \sum_{h=0}^{\infty} \mathbf{s} \mathbf{B}^h = \mathbf{s} (\mathbf{I} - \mathbf{B})^{-1}.$$

If the underlying digraph is acyclic, then $\mathbf{B}^h = \mathbf{0}$ for all sufficiently large h , and the SEM exhibits a finite impulse response (FIR). In contrast, cyclic digraphs give rise to infinite impulse response (IIR) behavior, with contributions from arbitrarily long propagation paths.

According to the above discussion, we focus in the sequel on estimating the parameter $\mathbf{B}$ from data $\mathbf{X}$ generated via an SEM (2) and a window graph $\mathbf{W}$ from time series $\mathbf{X}$ generated via an SVAR (4). We assume stability of the SVAR primarily for numerical reasons, while the stability of the SEM and the assumption of cycle contraction in the instantaneous component $\mathbf{B}_0$, will be instrumental in establishing identifiability of the unknown SEM and SVAR parameters.

A.2 STRICT STATIONARITY AND ERGODICITY

Strict stationarity means that the joint distribution of the data on any n consecutive time points is the same for every t . We may prove this by using the impulse response formulation of the SVAR (20). Just notice that $\mathbf{s}_t$ are iid distributed and therefore in terms of distribution:

$$\mathbf{x}_t = \sum_{k=0}^{\infty} \mathbf{s}_{t-k} \Psi_k \stackrel{d}{=} \sum_{k=0}^{\infty} \mathbf{s}_{t+h-k} \Psi_k = \mathbf{x}_{t+h} \quad (22)$$

which implies that $(\mathbf{x}_t, \mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+l})$ and $(\mathbf{x}_{t+h}, \mathbf{x}_{t+h+1}, \dots, \mathbf{x}_{t+h+l})$ have the same distribution.

Moreover, a well known result for our stable time-series ([Brockwell and Davis, 1991], p.379) is that it is ergodic. This by definition means that $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t = \mathbb{E}[\mathbf{x}_t]$. In fact Makkadem [1988] show that an ARMA process is strongly mixing which further implies ergodicity Bradley [2007]. Intuitively, mixing means that the dependence between time samples weakens as the temporal distance between them increases. In simple words, samples become more independent as time passes by.

We summarize our result as the following lemma which can be used as a substitute assumption on our MLE consistency proof later.

Lemma A.6. *A stable SVAR (4) is strictly stationary, strongly mixing and ergodic.*

A.3 IDENTIFIABILITY

Theorem A.7. *Consider data $\mathbf{X}$ generated via (2) with non-Gaussian structural shocks $\mathbf{S}$ and a digraph $\mathbf{B}$ with contractive cycles. Then $\mathbf{B}$ is identifiable from the distribution of $\mathbf{X}$.*

Proof. We rely on the characterization of Drton et al. [2025] for the equivalence classes of digraphs with non-Gaussian noise. Assume, for contradiction, that $\mathbf{B}$ is not identifiable. Then there exists $\mathbf{B}'$ with $\mathbf{B}' \neq \mathbf{B}$ such that

$$(\mathbf{I} - \mathbf{B})^{-1} \mathbf{S} \stackrel{d}{=} (\mathbf{I} - \mathbf{B}')^{-1} \mathbf{S}' \Rightarrow (\mathbf{I} - \mathbf{B}') (\mathbf{I} - \mathbf{B})^{-1} = \mathbf{P}\mathbf{D},$$

where $\mathbf{P}$ is a permutation and $\mathbf{D}$ is a dilation matrix, according to an implication of Darrois–Skitovich theorem and Drton et al. [2025]. Rearranging yields

$$(\mathbf{I} - \mathbf{B}) \mathbf{P}\mathbf{D} = (\mathbf{I} - \mathbf{B}'),$$

Let the entries of $(\mathbf{I} - \mathbf{B})$ be $-b_{ij}$ with $b_{ii} = -1$ and respectively $(\mathbf{I} - \mathbf{B}')$ be $-b'_{ij}$ again with $b'_{ii} = -1$.

If we consider the permutation $\pi(i)$ imposed by the permutation matrix $\mathbf{P}$, then the entries of $(\mathbf{I} - \mathbf{B}) \mathbf{P}\mathbf{D}$ are $-b_{i\pi(j)} d_{jj} = -b'_{ij}$. On the diagonal $i = j$ we get:

$$d_{ii} = \frac{-1}{b_{i\pi(i)}}$$

for $i \neq \pi(i)$ which implies that necessarily $b_{i\pi(i)} \neq 0$ and so there is an edge from i to $\pi(i)$.

Then looking at a permutation cycle: $i = \pi(j)$:

$$b'_{ij} = b_{i\pi(j)} d_{jj} = -\frac{b_{i\pi(j)}}{b_{i\pi(i)}} = \frac{1}{b_{i\pi(i)}} \Leftrightarrow b'_{i\pi^{-1}(i)} = \frac{1}{b_{i\pi(i)}}$$

Thus a permutation cycle $(i_1, \dots, i_k)$ such that $i_{l+1} = \pi(i_l)$ imposes a cycle in $\mathbf{B}$ and a reverse cycle in $\mathbf{B}'$ with reciprocal weights. Since all cycles in $\mathbf{B}$ are contractive, this implies that $\mathbf{B}'$ has a non-contractive cycle, a contradiction. Thus, among the equivalence class only $\mathbf{B}$ has contractive cycles, and therefore it is identifiable. $\square$

Theorem A.8. *Consider data $\mathbf{X}$ generated by an SVAR(k) (4) with non-Gaussian structural shocks $\mathbf{S}$ and a contemporaneous matrix $\mathbf{B}_0$ that has contractive cycles. Then the window graph $\mathbf{W}$ is identifiable from the distribution of $\mathbf{X}$.*

Proof. The proof follows the VAR-LiNGAM approach. Multiplying the SVAR(k) equation by $(\mathbf{I} - \mathbf{B}_0)^{-1}$ yields the reduced-form VAR(k) (8). Under standard regularity conditions, the reduced-form coefficients $\mathbf{A}_1, \dots, \mathbf{A}_k$ and residuals $\tilde{\mathbf{s}}_t$ are identifiable from the joint distribution of $\{\mathbf{x}_t\}$ via population least squares.

The reduced-form residuals satisfy the linear SEM

$$\tilde{\mathbf{s}}_t(\mathbf{I} - \mathbf{B}_0) = \mathbf{s}_t,$$

with independent non-Gaussian noise components. Since $\mathbf{B}_0$ has contractive cycles, Theorem 3.1 applies and implies that $\mathbf{B}_0$ is identifiable from the residuals $\tilde{\mathbf{s}}_t$.

Finally, since $(\mathbf{I} - \mathbf{B}_0)$ is invertible and $\mathbf{A}_i$ is identified, each lag matrix is uniquely recovered as

$$\mathbf{B}_i = \mathbf{A}_i(\mathbf{I} - \mathbf{B}_0), \quad i = 1, \dots, k,$$

which establishes the identifiability of $\mathbf{W}$. $\square$

Theorem A.9. *Consider the time-series model (3) with structural shocks $\mathbf{s}_t$ whose entries are i.i.d. Laplace($0, \beta$) with scale $\beta > 0$. Then the matrices $\mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_k \in \mathbb{R}^{d \times d}$ and the scale parameter β are identifiable from the distribution of the time-series data $\mathbf{X}$.*

Proof. From Theorem A.8, the window graph $\mathbf{W}$ (and hence $\mathbf{B}_0, \mathbf{B}_1, \dots, \mathbf{B}_k$) is identifiable from the distribution of $\mathbf{X}$. Once $\mathbf{W}$ is identified, the structural shocks $\mathbf{S} = \mathbf{X} - \mathbf{X}_{\text{past}}\mathbf{W}$ are a deterministic function of the data. Since their entries are i.i.d. Laplace($0, \beta$) and $\mathbb{E}[|Z|] = \beta$ for $Z \sim \text{Laplace}(0, \beta)$, the scale is recovered as

$$\beta = \mathbb{E}[|\mathbf{S}_{t,j}|]. \quad (23)$$

Therefore $(\mathbf{W}, \beta)$ is identifiable from the distribution of $\mathbf{X}$. $\square$

A.4 MAXIMUM LIKELIHOOD ESTIMATOR COMPUTATION

SVAR as a Linear SEM An SVAR can also be written as a linear structural equation model (SEM), which is the analogous model for generating linear static DAG data. For this we require the unrolled digraph $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \dots & \mathbf{B}_k & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_0 & \mathbf{B}_1 & & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \mathbf{B}_k \\ & & & \ddots & \ddots & \vdots \\ \mathbf{0} & & \dots & \mathbf{0} & \mathbf{B}_0 & \mathbf{B}_1 \\ \mathbf{0} & \mathbf{0} & & \dots & \mathbf{0} & \mathbf{B}_0 \end{pmatrix}. \quad (24)$$

Then by considering the single-row vector $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{T-1}) \in \mathbb{R}^{1 \times dT}$ we have that

$$\mathbf{x} = \mathbf{x}\mathbf{A} + \mathbf{s}, \quad (25)$$

Estimator computation Here we compute the MLE assuming that each entry of the structural shocks $\mathbf{S} \in \mathbb{R}^{T \times d}$ follows independently a Laplace distribution $\text{Laplace}(0, \beta)$. The multivariate probability density function of $\mathbf{S}$ is:

$$f_S(\mathbf{S}) = \prod_{t,j} \frac{1}{2\beta} e^{-\frac{|\mathbf{s}_{t,j}|}{\beta}} \quad (26)$$

Solving with respect to $\mathbf{x}$ equation (25) gives $\mathbf{x} = \mathbf{s}(\mathbf{I} - \mathbf{A})^{-1}$ where $\mathbf{A} \in \mathbb{R}^{dT \times dT}$ is the unrolled DAG matrix of $\mathbf{W}$ according to (24). Here $\mathbf{x}$ and $\mathbf{s}$ represent $1 \times dT$ dimensional vectors. This interchange doesn't affect the probability distribution. Using this linear transformation the probability density function (pdf) of $\mathbf{X}$, becomes

$$\begin{aligned} f_X(\mathbf{X}|\mathbf{W}, \beta) &= \frac{f_S(\mathbf{x}(\mathbf{I} - \mathbf{A}))}{\left| \det \left((\mathbf{I} - \mathbf{A})^{-1} \right) \right|} \\ &= |\det(\mathbf{I} - \mathbf{A})| \prod_{t,j} \frac{1}{2\beta} e^{-\frac{|\mathbf{x}_{t,j} - \mathbf{x}_{t,\text{past}} \mathbf{W}_{:,j}|}{\beta}} \\ &= \left| \det(\mathbf{I} - \mathbf{B}_0) \right|^T \prod_{t,j} \frac{1}{2\beta} e^{-\frac{|\mathbf{x}_{t,j} - \mathbf{x}_{t,\text{past}} \mathbf{W}_{:,j}|}{\beta}} \\ &= |\det(\mathbf{I} - \mathbf{B}_0)|^T \frac{1}{(2\beta)^{dT}} e^{-\frac{\|\mathbf{x} - \mathbf{x}_{\text{past}} \mathbf{W}\|_1}{\beta}} \end{aligned}$$

Given $\mathbf{X}$ the the log-likelihood for the window graph can be computed as:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \beta; \mathbf{X}) &= \log f_X(\mathbf{X}|\mathbf{W}, \beta) \\ &= T \log |\det(\mathbf{I} - \mathbf{B}_0)| - Td \log(2\beta) - \frac{1}{\beta} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1. \end{aligned} \quad (27)$$

In what follows, for simplicity of notation we will skip the parameter β and will use $\mathcal{L}(\mathbf{W}, \beta; \mathbf{X})$ and $\mathcal{L}(\mathbf{W}; \mathbf{X})$ interchangeably.

A.5 MLE CONSISTENCY BACKGROUND

We proceed by analyzing the prior theorems that we will use to prove our results. First, denote with $L(\mathbf{W})$ the population log-likelihood [Lachapelle et al., 2019, Newey and McFadden, 1994], defined as:

$$L(\mathbf{W}, \beta) = \mathbb{E}_{\mathbf{W}^*, \beta^*} [\mathcal{L}(\mathbf{W}, \beta; \mathbf{X})]. \quad (28)$$

Note that we use $\mathcal{L}(\mathbf{W}, \beta; \mathbf{X})$ and $\mathcal{L}(\mathbf{W}; \mathbf{X})$ interchangeably, as well as $L(\mathbf{W}, \beta)$ and $L(\mathbf{W})$.

In essence, the population log-likelihood is the expected value of the log-likelihood function computed with the probability density $f_X(\mathbf{X}|\mathbf{W}^*, \beta^*)$ with parameters the ground truth window graph $\mathbf{W}^*$ and parameter β^* . This expected value is computed over the distribution of $\mathbf{X}$, parametrized assuming ground truth parameters $\mathbf{W}^*, \beta^*$. Formally, we have:

$$L(\mathbf{W}, \beta) = \mathbb{E}_{\mathbf{W}^*, \beta^*} [\mathcal{L}(\mathbf{W}, \beta; \mathbf{X})] = \int \mathcal{L}(\mathbf{W}, \beta; \mathbf{X}) f_X(\mathbf{X}|\mathbf{W}^*, \beta^*) d\mathbf{X}$$

The population likelihood has a unique maximizer at the ground truth window graph if it is identifiable. This statement follows from the following Lemma A.10 which is proved as Lemma 2.2 Newey and McFadden [1994].

Lemma A.10. *Assume that the ground truth window graph $\mathbf{W}^*$ and parameter β^* are identifiable from the data distribution. This means, that for $(\mathbf{W}, \beta) \neq (\mathbf{W}^*, \beta^*)$ it is true that $f_X(\mathbf{X}|\mathbf{W}, \beta) \neq f_X(\mathbf{X}|\mathbf{W}^*, \beta^*)$. Then, the population likelihood $L(\mathbf{W}, \beta)$ has unique maximum at the true window graph $\mathbf{W}^*$ and true β^* .*

A crucial result for our MLE consistency, is the strong version of the uniform law of large numbers. In its strong version the uniform law of large numbers doesn't require independence of the data samples, but rather to be generated via a strictly stationary and ergodic process Newey and McFadden [1994], Pötscher and Prucha [1989].

Lemma A.11 (Strong Uniform Law of Large Numbers). *Consider that the log-likelihood function $\mathcal{L}(\mathbf{W}; \mathbf{X})$, $\mathbf{W} \in \mathcal{W}$ satisfy the following conditions.*

- The time series $\{\mathbf{x}_t\}$ is strictly stationary and ergodic.
- $\mathcal{W}$ is a compact space.
- $\mathcal{L}(\mathbf{W}; \mathbf{X})$ is continuous at each $\mathbf{W} \in \mathcal{W}$ with probability 1.
- There exists dominating function $D(\mathbf{X})$ such that $|\mathcal{L}(\mathbf{W}; \mathbf{X})| \leq D(\mathbf{X})$ for all $\mathbf{W} \in \mathcal{W}$ and $\mathbb{E}_{\mathbf{W}^*}[D(\mathbf{X})] < \infty$.

Then the the empirical average log-likelihood converges uniformly in probability to the population likelihood $L(\mathbf{W})$:

$$\sup_{\mathbf{W} \in \mathcal{W}} \left| \frac{1}{T} \sum_{t=1}^T \mathcal{L}(\mathbf{x}_t; \mathbf{W}) - L(\mathbf{W}) \right| \xrightarrow{p} 0 \quad (29)$$

We now present Theorem A.12, which establishes the consistency of the maximum likelihood estimator (MLE). This theorem is based on a set of sufficient assumptions for ensuring MLE consistency. For completeness, we include a detailed proof of Theorem A.12, leveraging Lemma A.11.

Theorem A.12. *Consider that the average log-likelihood function $L_t(\mathbf{W}) = \frac{1}{T} \sum_{t=1}^T \mathcal{L}(\mathbf{x}_t; \mathbf{W})$ and population $L(\mathbf{W})$ satisfy the following conditions for $\mathbf{W} \in \mathcal{W}$:*

- $\mathbf{W}^* = \arg \max_{\mathbf{W} \in \mathcal{W}} L(\mathbf{W})$ is identifiable from the data.
- The conditions of Lemma A.11 are satisfied.

Then, if the maximum of $L_t(\mathbf{W})$ is achieved at $\widehat{\mathbf{W}}_t$ then $\widehat{\mathbf{W}}_t$ converges uniformly to $\mathbf{W}^*$.

Proof. We repeat the proof of Newey and McFadden [1994][Theorem 2.1] for our scenario. From the identifiability assumption, Lemma A.10 implies that $\mathbf{W}^*$ is the unique and global maximizer of $L(\mathbf{W})$. Given that the conditions of Lemma A.11 are satisfied $L_t(\mathbf{W})$ converges uniformly in probability to $L(\mathbf{W})$. We will leverage the compactness of the space $\mathcal{W}$ to show that their maxima satisfy

$$\widehat{\mathbf{W}}_t \xrightarrow{p} \mathbf{W}^* \quad (30)$$

From the uniform convergence it follows that with probability approaching 1 for any ϵ (or $\epsilon/3$ as we use next):

$$|L_t(\mathbf{W}) - L(\mathbf{W})| < \epsilon \Leftrightarrow L(\mathbf{W}) - \epsilon < L_t(\mathbf{W}) < L(\mathbf{W}) + \epsilon, \forall \mathbf{W} \in \mathcal{W}. \quad (31)$$

Since by definition $L_t(\mathbf{W})$ is a continuous function and $\mathcal{W}$ is compact it takes a maximum value at point $\widehat{\mathbf{W}}_t$. Since $L_t(\widehat{\mathbf{W}}_t) \geq L_t(\mathbf{W}^*)$ the maximum would satisfy for any $\epsilon > 0$

$$L_t(\widehat{\mathbf{W}}_t) > L_t(\mathbf{W}^*) - \epsilon/3. \quad (32)$$

This in combination with (31) would imply

$$L(\widehat{\mathbf{W}}_t) > L_t(\widehat{\mathbf{W}}_t) - \epsilon/3 > L_t(\mathbf{W}^*) - 2\epsilon/3 > L(\mathbf{W}^*) - \epsilon. \quad (33)$$

In essence we have proved that $L(\widehat{\mathbf{W}}_t)$ can get arbitrarily close to $L(\mathbf{W}^*)$. This in turn gives that $\widehat{\mathbf{W}}_t$ approaches $\mathbf{W}^*$ with probability 1 as $t \rightarrow \infty$. Indeed, if we consider any open interval $\mathcal{I}$ containing $\mathbf{W}^*$, then $\mathcal{W} \cap \mathcal{I}^c$ is compact and we can compute

$$M = \sup_{\mathbf{W} \in \mathcal{W} \cap \mathcal{I}^c} L(\mathbf{W}) < L(\mathbf{W}^*) \quad (34)$$

Note that by Lemma A.11 $L(\mathbf{W})$ is continuous, so the supremum is a finite value. If we choose $\epsilon = L(\mathbf{W}^*) - M$ then:

$$L(\widehat{\mathbf{W}}_t) > L(\mathbf{W}^*) - \epsilon = M \quad (35)$$

Thus $\widehat{\mathbf{W}}_t \in \mathcal{I}$ which concludes the proof. $\square$

A.6 LAPLACIAN MLE CONSISTENCY FOR STABLE DIGRAPHS

We will now show that the MLE computed at (27) satisfies the requirements of Theorem A.12, including those for the strong uniform Law of large numbers Lemma A.11, for consistency. Practically, this result implies that as the length T of time-series data $\mathbf{X} = \{\mathbf{x}_t\}_t$ increases, the maximizer $\widehat{\mathbf{W}}$ of the log-likelihood function $\mathcal{L}(\mathbf{W}, \beta; \mathbf{X})$ converges to the maximizer $\mathbf{W}^*$ of the population likelihood $L(\mathbf{W}, \beta)$. To begin with we show the compactness of our search space.

Lemma A.13. *For a given radius $0 < r < 1$ the set of stable window graphs $\mathcal{W}$*

$$\mathcal{W} = \left\{ \mathbf{W} \in [-b, b]^{d(k+1) \times d} : \rho(|\mathbf{B}_0|) \leq r, \rho(\mathcal{A}(\mathbf{W})) \leq r \right\},$$

where $\mathcal{A}(\mathbf{W})$ denote the (companion) state-transition matrix of the associated reduced-form VAR, is compact.

Proof. Consider the hypercube

$$\mathcal{C} := [-b, b]^{d(k+1) \times d}.$$

It is compact as a finite product of compact intervals.

The map $\mathbf{W} \mapsto \mathbf{B}_0$ (reshaping / selecting the contemporaneous block) is linear and hence continuous. The entrywise absolute value $\mathbf{B}_0 \mapsto |\mathbf{B}_0|$ is continuous, and the spectral radius map $X \mapsto \rho(X)$ is continuous on $\mathbb{R}^{d \times d}$. Therefore the composition

$$g(\mathbf{W}) := \rho(|\mathbf{B}_0|)$$

is continuous. Hence

$$\{\mathbf{W} \in \mathcal{C} : \rho(|\mathbf{B}_0|) \leq r\} = g^{-1}((-\infty, r])$$

is closed in $\mathcal{C}$.

For any $\mathbf{W}$ with $\rho(|\mathbf{B}_0|) \leq r < 1$, we have $\rho(\mathbf{B}_0) \leq \rho(|\mathbf{B}_0|) < 1$, so $\mathbf{I} - \mathbf{B}_0$ is invertible. Thus the reduced-form coefficients (obtained from $(\mathbf{B}_0, \dots, \mathbf{B}_k)$ via matrix inversion and multiplication) and the companion matrix $\mathcal{A}(\mathbf{W})$ are well-defined. Matrix inversion and multiplication are continuous on the set of invertible matrices, hence $\mathbf{W} \mapsto \mathcal{A}(\mathbf{W})$ is continuous on $\{\mathbf{W} : \rho(|\mathbf{B}_0|) \leq r\}$. By continuity of $\mathbf{W} \mapsto \mathcal{A}(\mathbf{W})$ and continuity of spectral radius, the map

$$h(\mathbf{W}) := \rho(\mathcal{A}(\mathbf{W}))$$

is continuous on $\{\mathbf{W} : \rho(|\mathbf{B}_0|) \leq r\}$, hence on $\mathcal{C}$ after restriction. Therefore

$$\{\mathbf{W} \in \mathcal{C} : \rho(\mathcal{A}(\mathbf{W})) \leq r\} = h^{-1}((-\infty, r])$$

is closed in $\mathcal{C}$.

Therefore, we have written $\mathcal{W}$ as the intersection of $\mathcal{C}$ with two closed sets:

$$\mathcal{W} = \mathcal{C} \cap \{\rho(|\mathbf{B}_0|) \leq r\} \cap \{\rho(\mathcal{A}(\mathbf{W})) \leq r\}.$$

Hence $\mathcal{W}$ is closed in the compact set $\mathcal{C}$, and therefore compact. $\square$

Now using this Lemma we are ready to prove our consistency result.

Theorem A.14. *The maximum log-likelihood estimator of (27) satisfies the conditions of Theorem A.12 and thus is consistent under the following assumptions:*

- The search space of window graphs is $\mathcal{W} = \left\{ \mathbf{W} \in [-b, b]^{d(k+1) \times d} : \rho(|\mathbf{B}_0|) \leq r, \rho(\mathcal{A}(\mathbf{W})) \leq r \right\}$.
- The Laplacian parameter $\beta \in [a, b]$ is bounded with $0 < a \leq b^1$.

¹Any fixed $a > 0$ suffices; the lower bound is used only for compactness of the parameter space and to bound $1/\beta \leq 1/a$ in the dominating function.

Proof. It suffices to fulfill one-by-one the requirements of Theorem A.12. First, the identifiability of the ground truth $\mathbf{W}^*$ and β^* follows from Appendix A.3. Secondly, the SVAR generated time series $\mathbf{X} = \{\mathbf{x}_t\}_t$ are strictly stationary and ergodic as shown in Lemma A.6. Also, $(\mathbf{W}, \beta) \in \mathcal{W} \times [a, b]$ is compact because the space $\mathcal{W}$ of stable digraphs $\mathbf{W}$ is compact from Lemma A.13 and $[a, b]$ is compact. Moreover, the log-likelihood

$$\mathcal{L}(\mathbf{W}, \beta; \mathbf{X}) = T \log |\det(\mathbf{I} - \mathbf{B}_0)| - Td \log(2\beta) - \frac{1}{\beta} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \quad (36)$$

is continuous at $(\mathbf{W}, \beta)$.

It remains to bound $|\mathcal{L}(\mathbf{W}; \mathbf{X})|$ from some function $D(\mathbf{X})$ with finite expectation. For this we compute:

$$\begin{aligned} |\mathcal{L}(\mathbf{W}; \mathbf{X})| &= |\log f_X(\mathbf{X}|\mathbf{W}, \beta)| \\ &= \left| T \log |\det(\mathbf{I} - \mathbf{B}_0)| - Td \log(2\beta) - \frac{1}{\beta} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \right| \\ &\leq T \log |\det(\mathbf{I} - \mathbf{B}_0)| + |Td \log(2\beta)| + \left| \frac{1}{\beta} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \right| \\ &\leq dT \log(1+r) + Td \max\{|\log(2a)|, |\log(2b)|\} + \left| \frac{1}{\beta} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \right| \\ &\leq C_1 + \frac{1}{a} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 \\ &\leq C_1 + C_2 \|\mathbf{X}\|_1 \end{aligned}$$

Where we used that $\mathbf{I} - \mathbf{B}_0$ has eigenvalues $(1 - \lambda)$ where λ is an eigenvalue of $\mathbf{B}_0$, $\lambda \in [-r, r]$. Therefore, $|1 - \lambda| < 1 + r$ and

$$\det(\mathbf{I} - \mathbf{B}_0) < (1 + r)^d \Leftrightarrow T \log |\det(\mathbf{I} - \mathbf{B}_0)| < dT \log(1 + r).$$

Additionally, we used the bound:

$$\begin{aligned} \|\mathbf{X} - \mathbf{X}_{\text{past}} \mathbf{W}\|_1 &\leq \|\mathbf{X}\|_1 + \|\mathbf{X}_{\text{past}}\|_1 \|\mathbf{W}\|_1 \\ &\leq \|\mathbf{X}\|_1 + (k+1) \|\mathbf{X}\|_1 \cdot r \\ &= [r(k+1) + 1] \|\mathbf{X}\|_1 = C_2 \|\mathbf{X}\|_1. \end{aligned}$$

Now one can conclude the Theorem by observing that $\|\mathbf{X}\|_1$ has finite expectation, due to the stability of SVAR. In that case we may use the impulse response formulation (20) to derive:

$$\begin{aligned} \mathbb{E}[\|\mathbf{X}\|_1] &= \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{x}_t\|_1] \\ &= \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{k=0}^{\infty} \mathbf{s}_{t-k} \Psi_k \right\|_1 \right] \\ &\leq \sum_{t=0}^{T-1} \mathbb{E} \left[\sum_{k=0}^{\infty} \|\mathbf{s}_{t-k}\|_1 \|\Psi_k\|_1 \right] \\ &\leq \sum_{t=0}^{T-1} \sum_{k=0}^{\infty} \|\Psi_k\|_1 \mathbb{E}[\|\mathbf{s}_{t-k}\|_1] \\ &\leq \sum_{t=0}^{T-1} \sum_{k=0}^{\infty} \|\Psi_k\|_1 \beta d \\ &\leq \sum_{t=0}^{T-1} \beta d \sum_{k=0}^{\infty} \|\Psi_k\|_1 < \infty \end{aligned}$$

Where, we used that for a single Laplacian variable $\mathbf{s}_{t,i} \sim \text{Laplace}(0, \beta)$ we have that $\mathbb{E}[|\mathbf{s}_{t,i}|] = \beta$ and $\mathbb{E}[\|\mathbf{s}_t\|_1] = \sum_{i=1}^d \mathbb{E}[|\mathbf{s}_{t,i}|] = \beta d$. $\square$

B EXPERIMENTS

B.1 ADDITIONAL EXPERIMENTS

B.1.1 Generic Digraphs

Here we show the rest of our experiments for generic (potentially cyclic) digraphs.

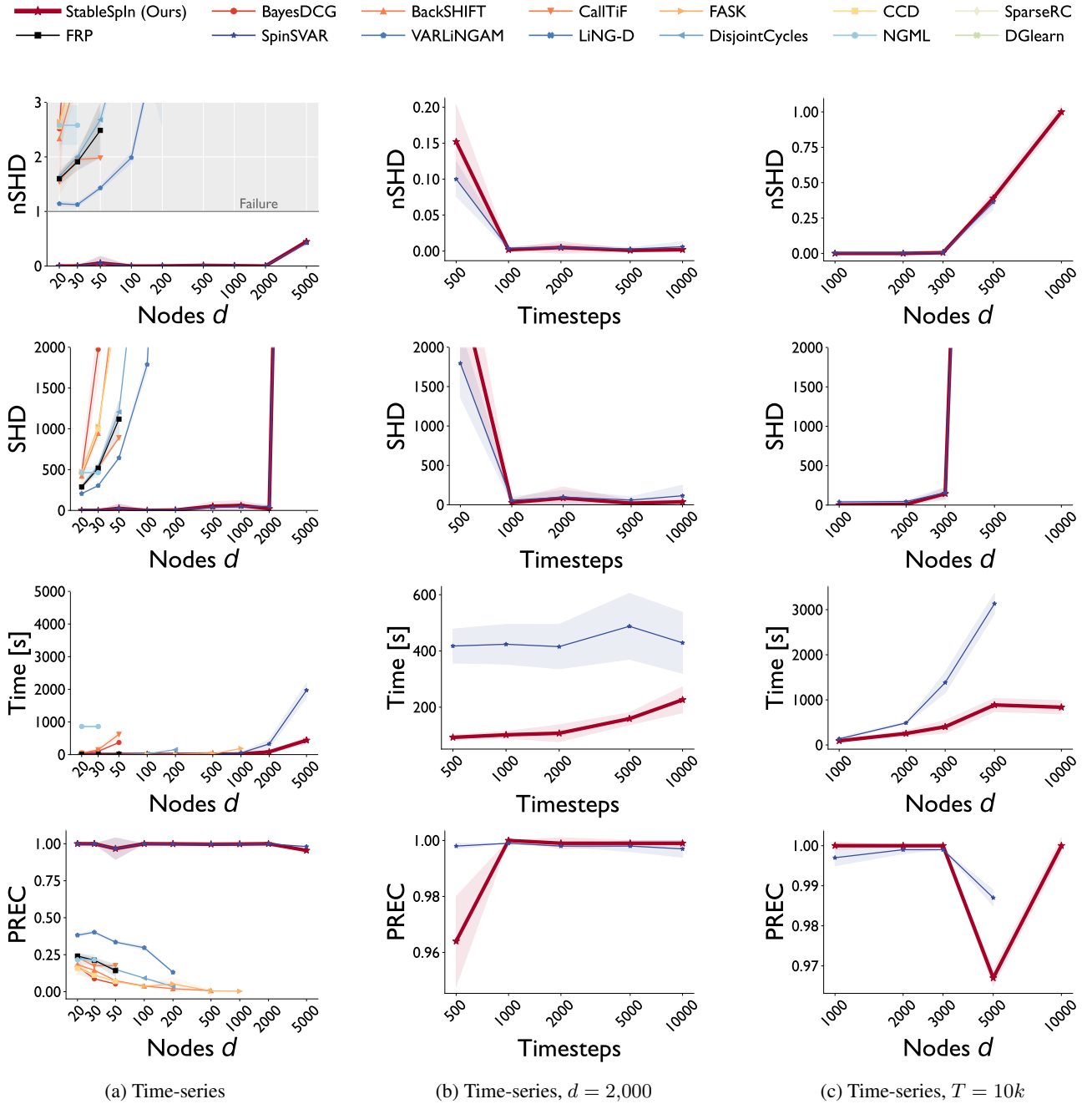


Figure 4: Generic (potentially cyclic) digraphs, time-series data: nSHD, SHD, runtime, and precision (top to bottom row).

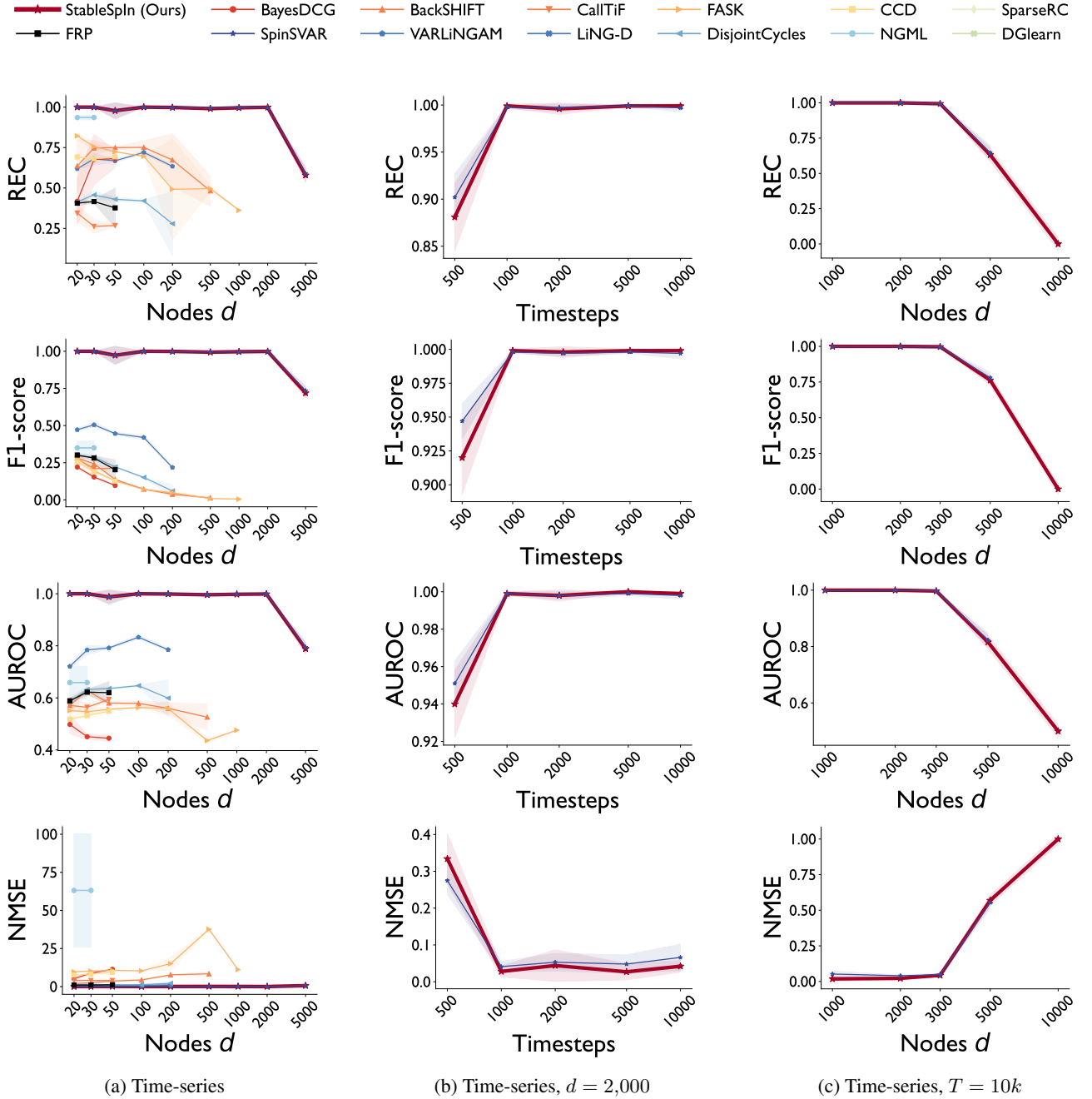


Figure 5: Generic (potentially cyclic) digraphs, time-series data: recall, F1, AUROC, and NMSE (top to bottom row).

B.1.2 Digraphs with Disjoint Cycles

Here we show the rest of our experiments for digraphs with disjoint cycles and data generated via an SVAR or an SEM.

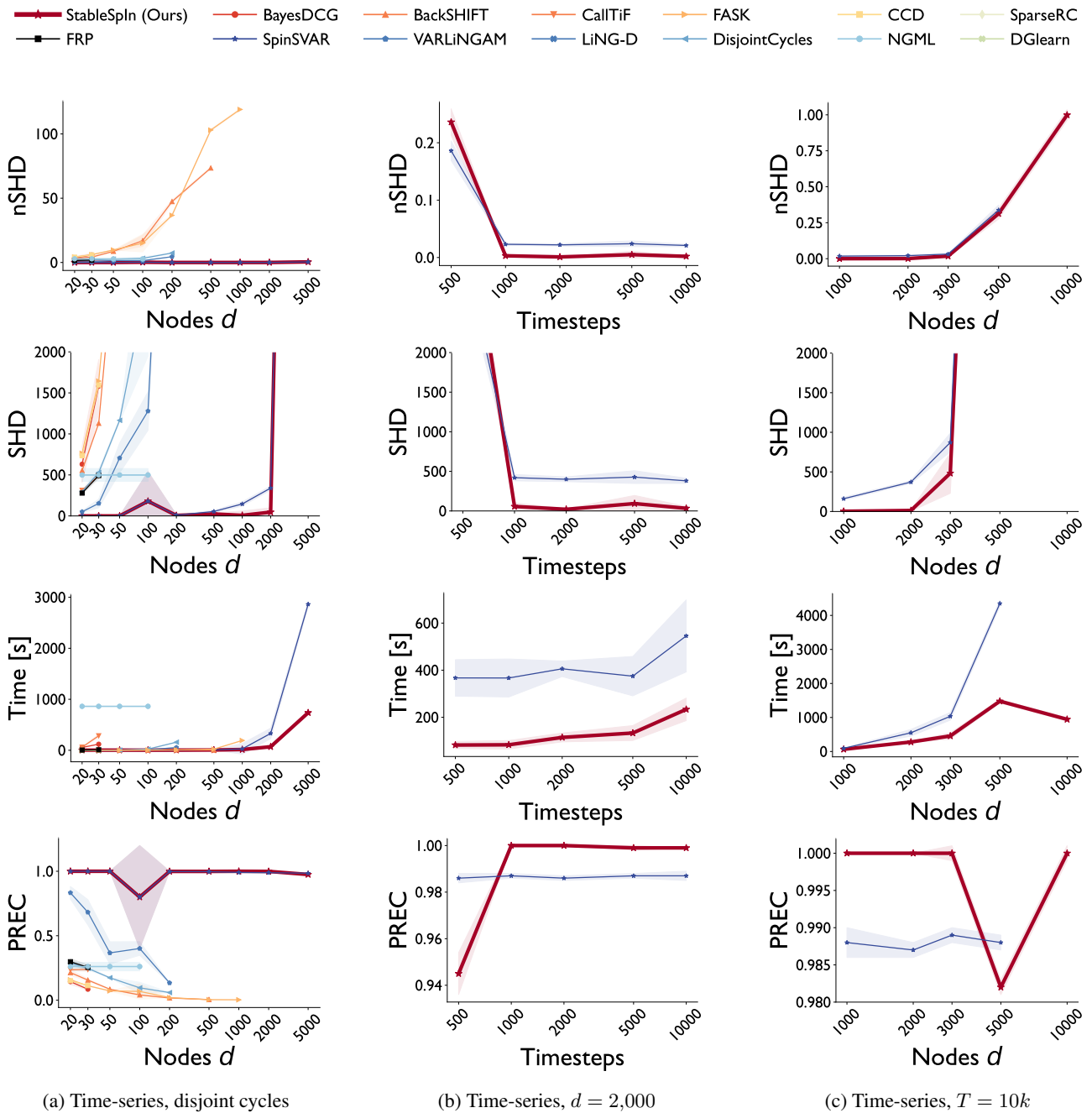
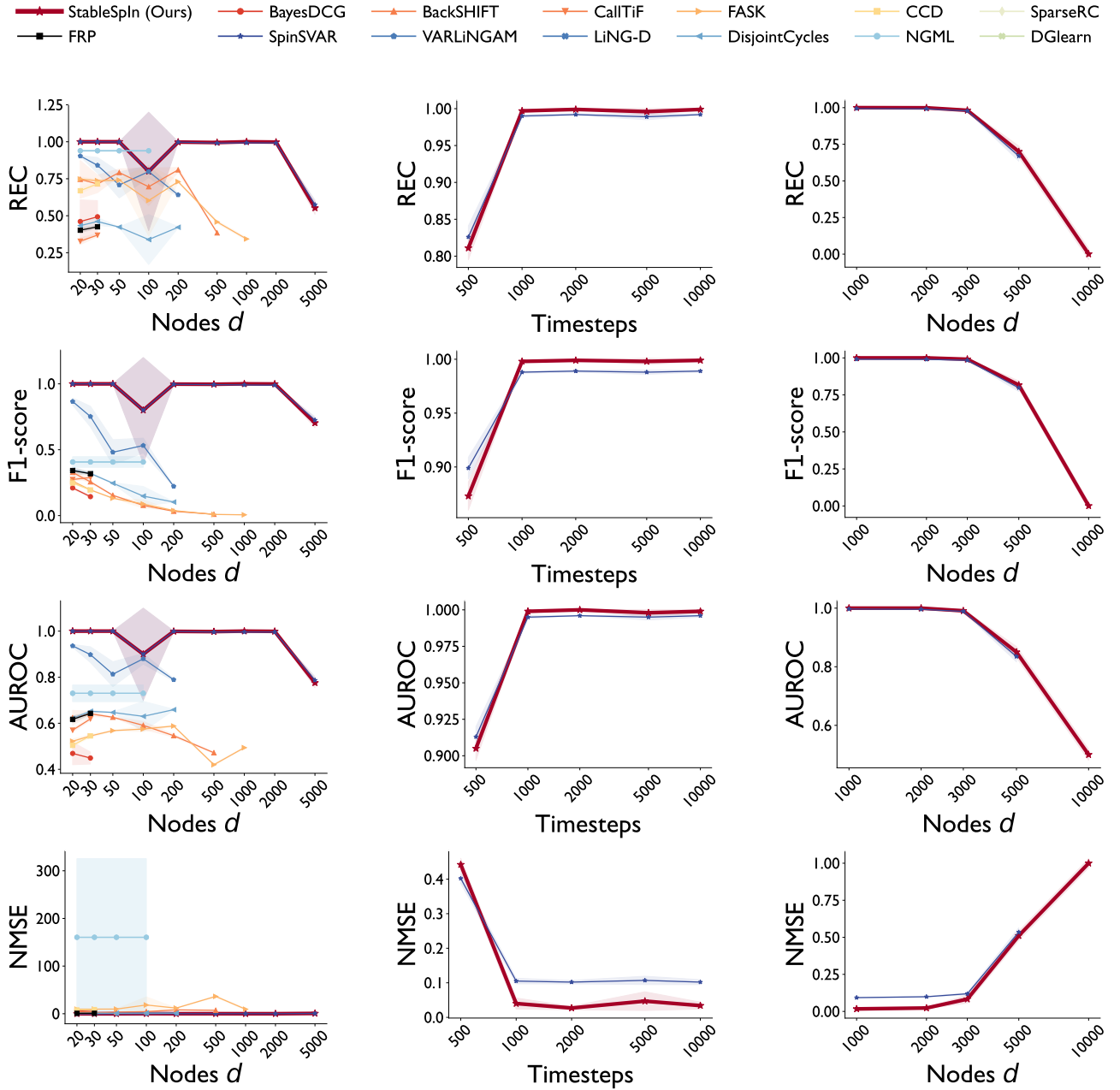


Figure 6: Digraphs with disjoint cycles, time-series (SVAR) data: nSHD, SHD, runtime, and precision (top to bottom row).



(a) Time-series, disjoint cycles

(b) Time-series, $d = 2,000$

(c) Time-series, $T = 10k$

Figure 7: Digraphs with disjoint cycles, time-series (SVAR) data: recall, F1, AUROC, and NMSE (top to bottom row).

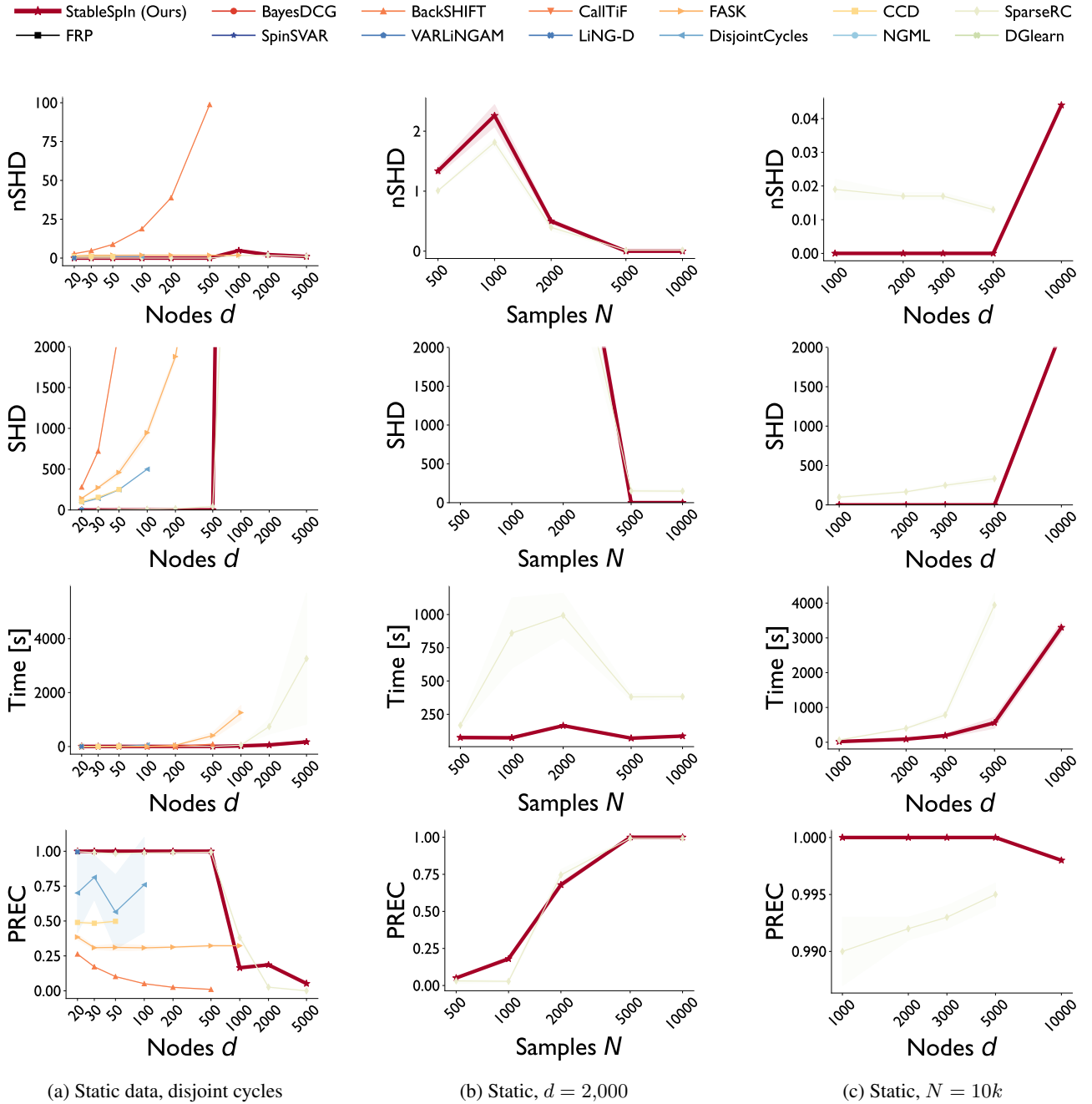


Figure 8: Digraphs with disjoint cycles, static (SEM) data: nSHD, SHD, runtime, and precision (top to bottom row).

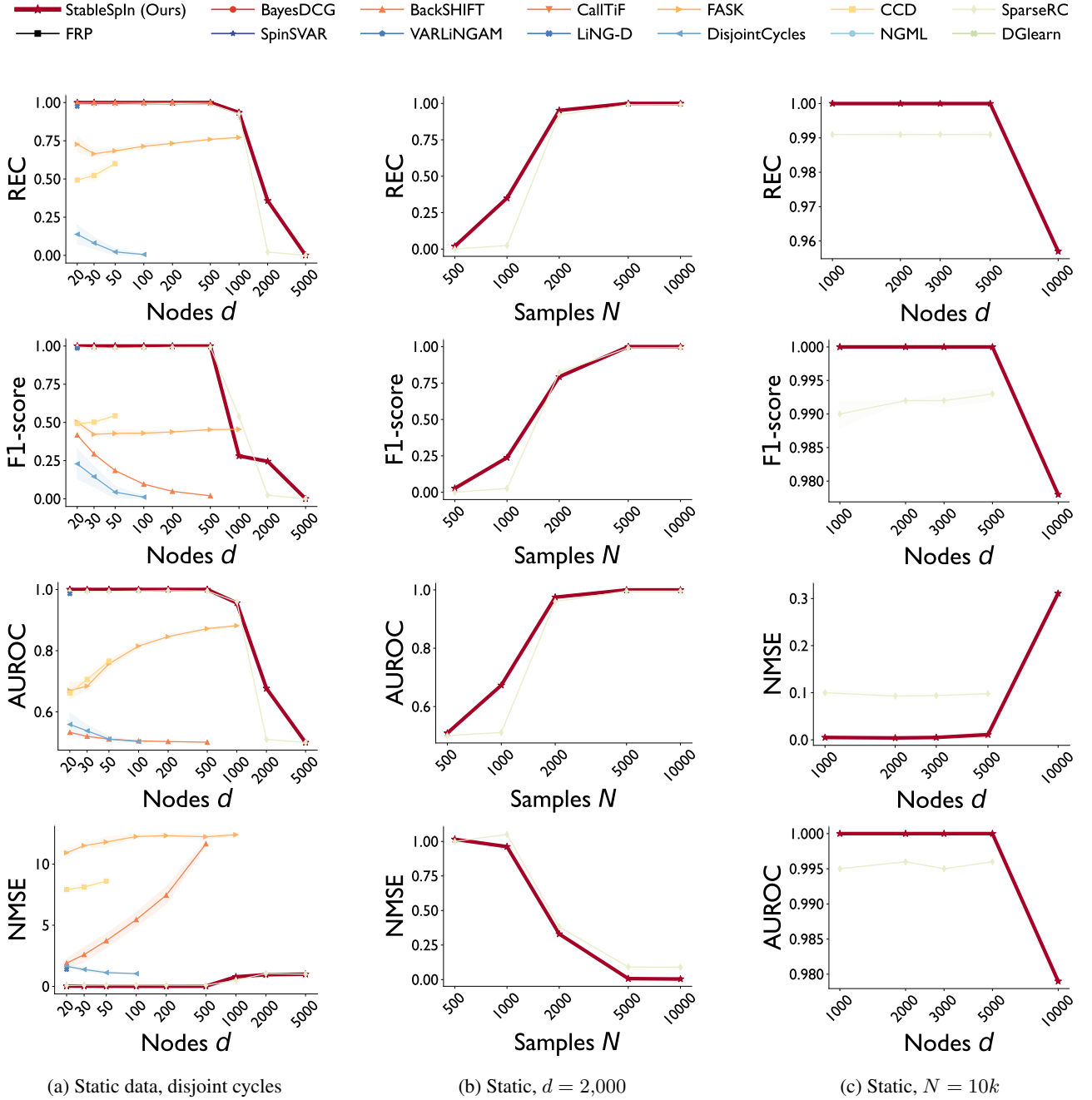
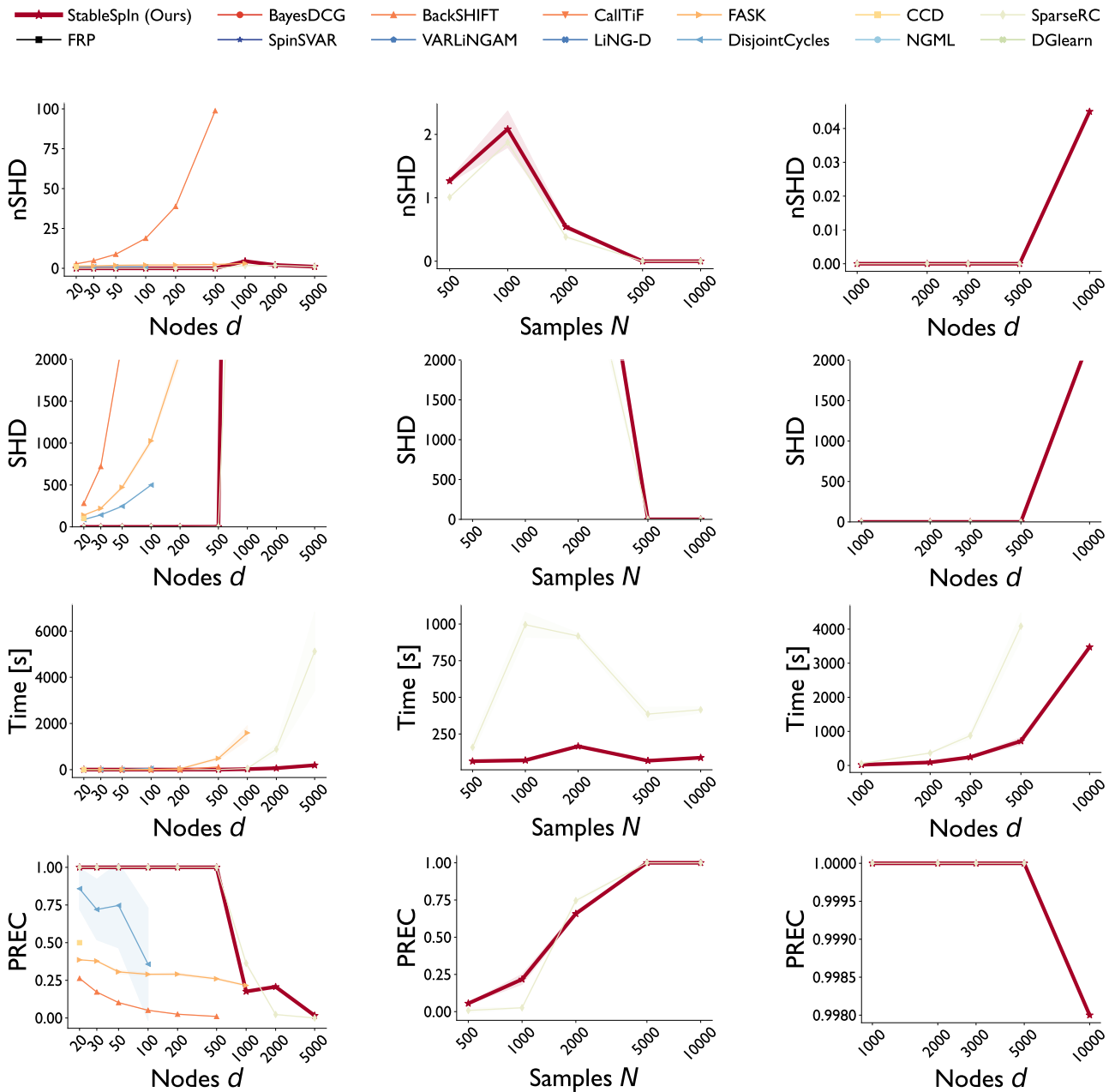


Figure 9: Digraphs with disjoint cycles, static (SEM) data: recall, F1, AUROC, and NMSE (top to bottom row).

B.1.3 Directed Acyclic graphs

For completeness, we also evaluate our method on directed acyclic graphs (DAGs). This demonstrates that it achieves performance comparable to SparseRC, while providing substantial computational speedups.



(a) Static data, directed acyclic graphs

(b) Static, $d = 2,000$

(c) Static, $N = 10k$

Figure 10: Directed acyclic graphs (DAGs), static (SEM) data: nSHD, SHD, runtime, and precision (top to bottom row).

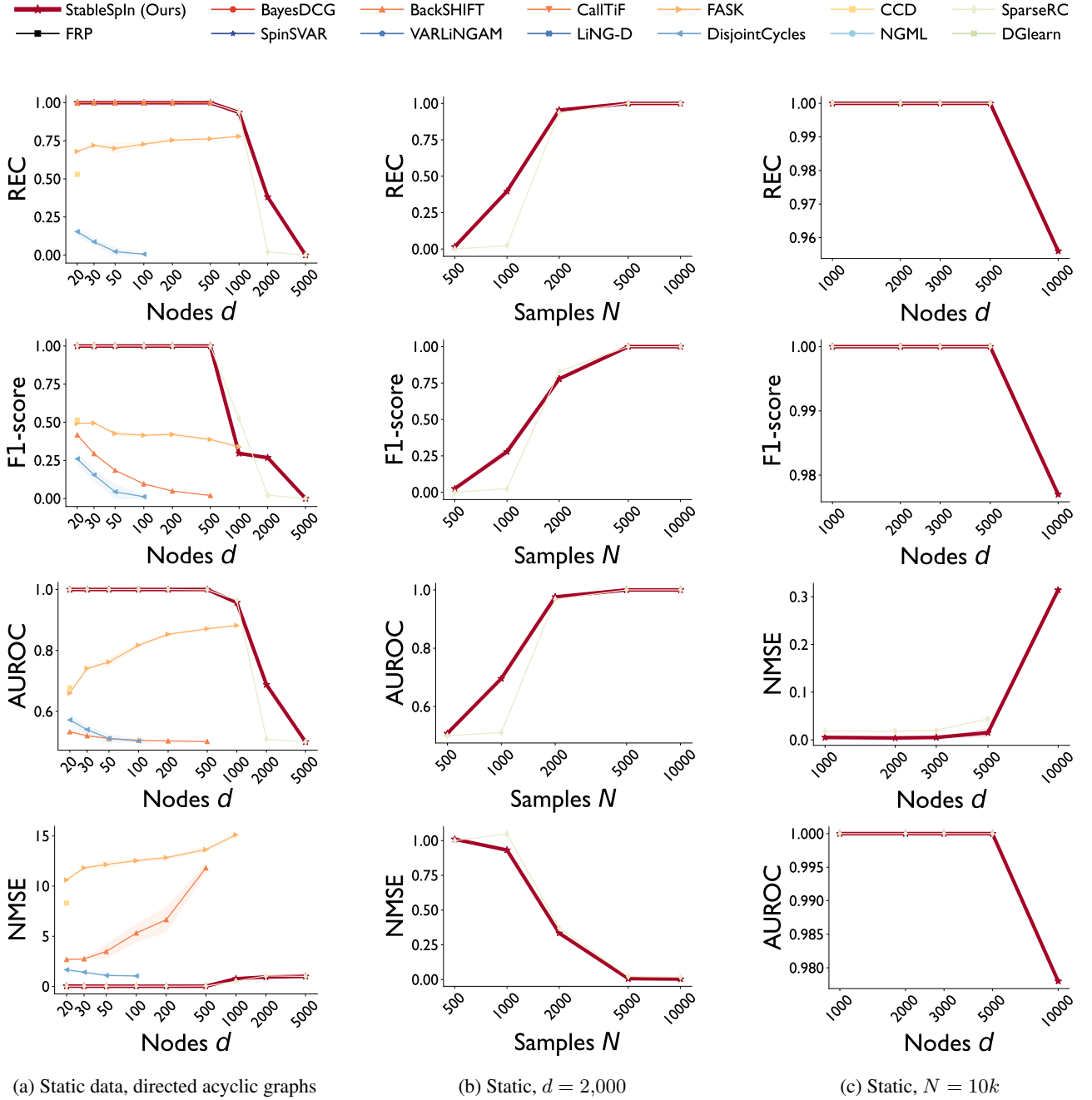


Figure 11: Directed acyclic graphs (DAGs), static (SEM) data: recall, F1, AUROC, and NMSE (top to bottom row).

B.1.4 More cycles

In these experiments, we examine how increasing the number of cycles in the generated directed graph affects performance. We induce more cycles by varying the graph’s average degree (higher average degree $\rightarrow$ more edges $\rightarrow$ more potential cycles).

—* StableSpln (Ours)
 —o BayesDCG
 —△ BackSHIFT
 —▽ CalITiF
 —▶ FASK
 —□ CCD
 —◇ SparseRC
—■ FRP
—* SpinSVAR
—o VARLiNGAM
—* LiNG-D
—◀ DisjointCycles
—o NGML
—* DGLearn

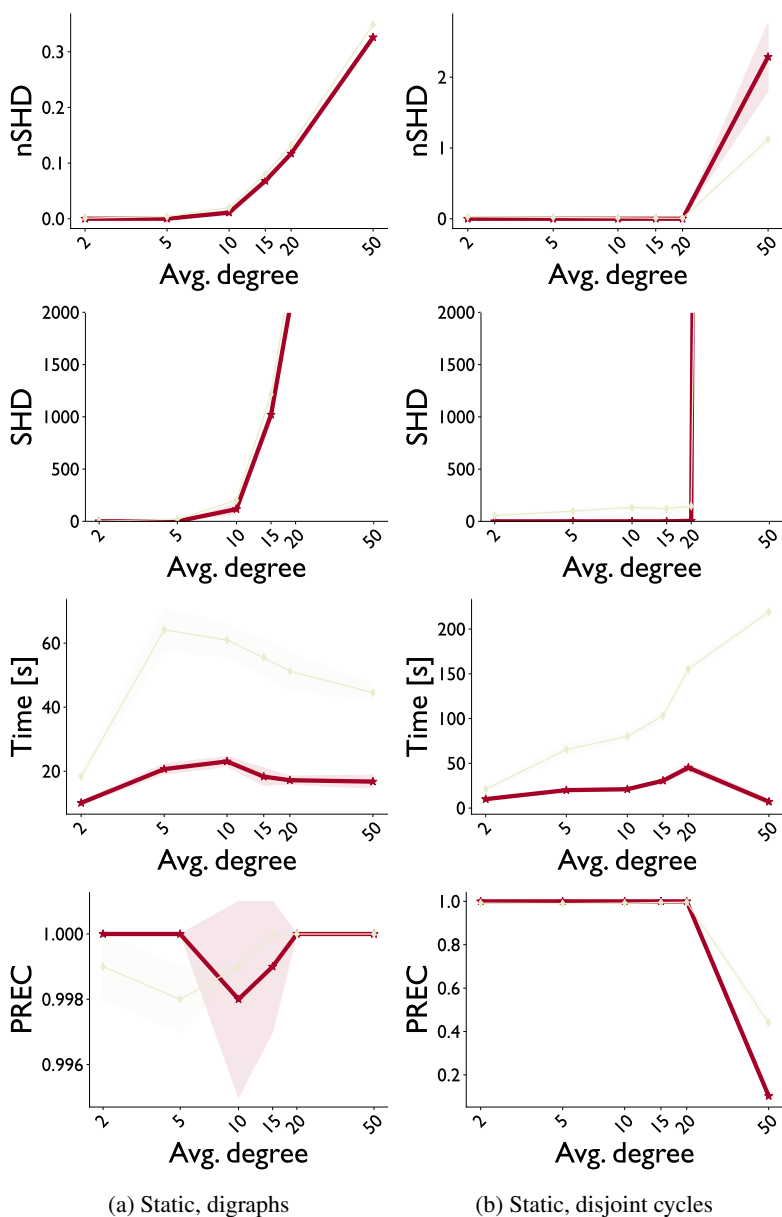


Figure 12: Increasing graph density (more edges and cycles), static (SEM) data: nSHD, SHD, runtime, and precision (top to bottom row). Left: generic digraphs; right: disjoint cycles.

—* StableSpIn (Ours) —o BayesDCG —△ BackSHIFT —▽ CalITiF —▶ FASK —□ CCD —◇ SparseRC
—■ FRP —+ SpinSVAR —● VARLiNGAM —× LiNG-D —◀ DisjointCycles —○ NGML —■ DGLearn

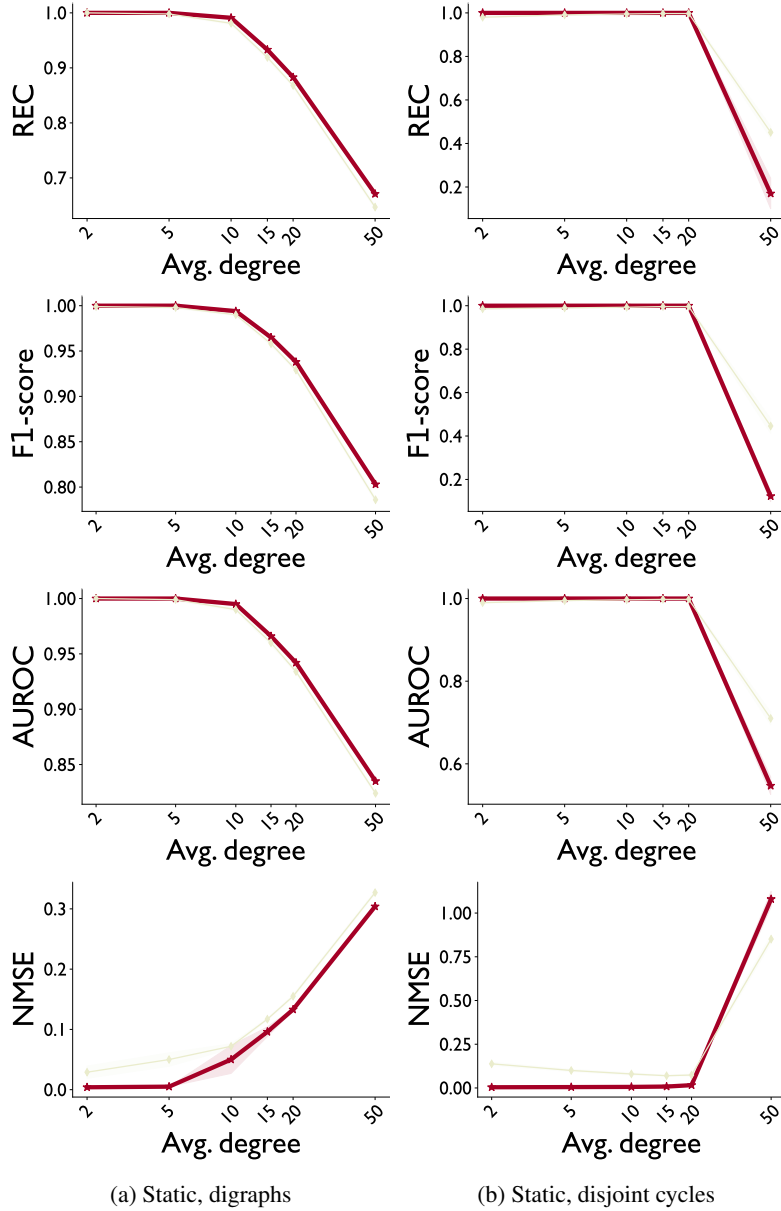


Figure 13: Increasing graph density (more edges and cycles), static (SEM) data: recall, F1, AUROC, and NMSE (top to bottom row). Left: generic digraphs; right: disjoint cycles.

B.1.5 Varying sparsity

In this experiment, we generate input shocks $\mathbf{S}$ as independent Gaussian variables $\mathcal{N}(0, 1)$, which are then subsampled via Bernoulli masking so that only a fraction remains active and the rest are set to zero. By varying the sparsity level, we smoothly transition from sparse (non-Gaussian due to zero-inflation) to fully dense Gaussian inputs. As sparsity decreases, the distribution approaches a standard Gaussian and the degree of non-Gaussianity diminishes.

We observe that StableSpIn degrades in performance as the inputs become dense and Gaussian, with a clear drop near the fully Gaussian regime (from 0.65 to 1.0). For larger graphs ($d = 500$), this degradation occurs even earlier (around 0.25

sparsity), highlighting the importance of sparse inputs for recovering more complex structures. A similar trend is observed for other methods relying on input sparsity (SparseRC and SpinSVAR), while methods not based on this assumption remain largely unaffected.

These findings are consistent with the theoretical requirement of sparsity (and non-Gaussianity).

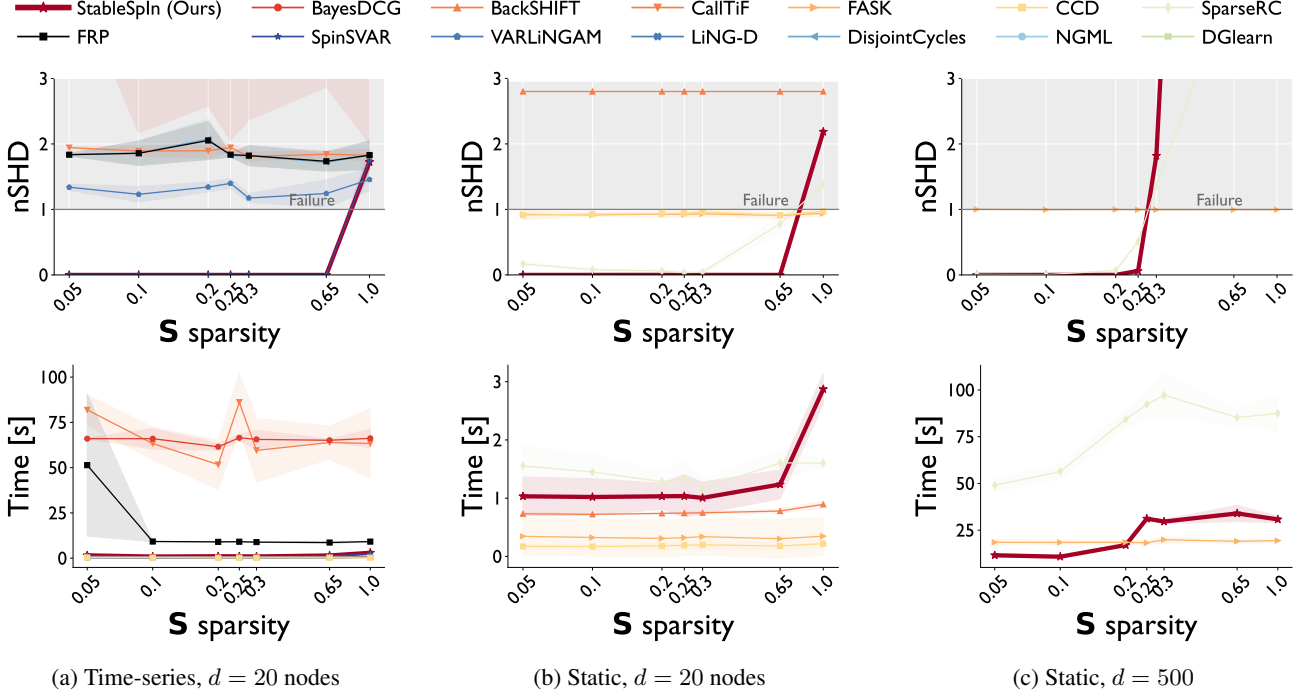


Figure 14: Effect of input sparsity. The sparsity of $\mathbf{S}$ varies from 5% to 100% non-zero entries. (a) shows the time-series setting with $d = 20$, while (b) and (c) consider static data with $d = 20$ and $d = 500$. Input shocks $\mathbf{S} \sim \mathcal{N}(0, 1)$ are subsampled via Bernoulli masking, inducing a transition from sparse, non-Gaussian (zero-inflated) to dense Gaussian inputs as the percentage of non-zero entries increases.

B.2 HYPERPARAMETER SELECTION

For both time-series and static settings, we perform a grid search on a small synthetic configuration ($N = 1$, $T = 1,000$, $d = 20$) and select the hyperparameters that achieve the best nSHD. All non-mentioned hyperparameters are left at their default values. Lag-related parameters are set equal to the ground-truth lag in the time-series case.

B.2.1 Time-series data

The resulting optimal hyperparameters are:

STABLESPIN We set $\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-3}$, $\lambda_3 = 100$, $\omega = 0.09$, and use a Laplace loss.

SPINSVAR We set $\lambda_1 = 10^{-4}$, $\lambda_2 = 0.1$, and $\omega = 0.09$.

VARLiNGAM The weight threshold is set to $\omega = 0.09$.

BayesDCG We set `pip_threshold` = 0.5, `niter` = 100, `burnin` = 20, and `thin` = 5.

CaLLTiF The significance level is set to $\alpha = 0.05$.

NGML We set the threshold $\omega = 0.09$.

FRP We use SCAD regularization with $\gamma = 3.7$, $n_inits = 5$, and $\omega = 0.09$.

BackShift We set $\omega = 0.09$, number of environments $E = 3$, enable covariance adjustment, and use $ev = 0$.

DGLearn We set $\omega = 0.09$, tabu length 4, and patience 4.

DisjointCycles We use $\alpha = 0.01$, Holm multiple-testing correction, and $\omega = 0.09$.

FASK, CCD, LiNG-D For all three methods we set $\alpha = 0.05$ and $\omega = 0.09$.

B.2.2 Static data

For static data, the same grid-search procedure is applied. The optimal hyperparameters coincide with the time-series case for all methods:

STABLESPIN $\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-3}$, $\lambda_3 = 100$, $\omega = 0.09$, Laplace loss.

SPARSERC $\lambda_1 = 10^{-4}$, $\lambda_2 = 0.1$, $\omega = 0.09$.

FRP SCAD regularization with $\gamma = 3.7$, $n_inits = 5$, $\omega = 0.09$.

BackShift $\omega = 0.09$, $E = 3$, covariance adjustment enabled, $ev = 0$.

DGLearn $\omega = 0.09$, tabu length 4, and patience 4.

DisjointCycles $\alpha = 0.01$, Holm correction, $\omega = 0.09$.

FASK, CCD, LiNG-D $\alpha = 0.05$, $\omega = 0.09$.

B.3 COMPUTE RESOURCES

All experiments were conducted on a single laptop workstation (Dell Alienware x17 R2) equipped with an 8-core CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU. Running the full suite of synthetic experiments with 5 repetitions required approximately one week of wall-clock time.

During the initial development phase, some experimental configurations failed due to implementation issues or excessive runtimes. To diagnose these cases and identify regimes in which individual methods exceeded practical computational limits, we first executed each configuration with a single repetition. Based on these preliminary runs, we fixed a global time-out threshold of 5,000 seconds for all methods, balancing computational feasibility with fair method comparison under limited resources.

B.4 CODE RESOURCES

All methods evaluated in our experiments are implemented using publicly available repositories or software packages. For each resource, we report the corresponding license information whenever it is explicitly provided by the authors.

SpinSVAR Implementation of SpinSVAR [Misiakos and Püschel, 2025] is available at <https://github.com/pmisiakos/SpinSVAR> under the CC-BY-4.0 license.

SparseRC The SparseRC [Misiakos et al., 2023] implementation is available at <https://github.com/pmisiakos/SparseRC> and is released under the MIT license.

VARLiNGAM We use the official LiNGAM implementation provided by the authors, available at <https://github.com/cdt15/lingam> (MIT license), and documented at <https://lingam.readthedocs.io/en/latest/index.html>.

LiNG-D, FASK, CCD Implementations of LiNG-D, FASK, and CCD are provided through the `py-tetrad` package, available at <https://github.com/cmu-phil/py-tetrad> under the MIT license.

DisjointCycles We use the authors' implementation from [Drton et al., 2025], available at <https://github.com/ysamwang/disjointCycles> (MIT license).

BayesDCG The BayesDCG implementation is obtained from the public repository <https://github.com/bluejw/BayesDCG>. No explicit license information is provided by the authors.

BackSHIFT BackSHIFT is provided as an R package named `backShift`, available from CRAN and developed by the authors of [Rothenhäusler et al., 2015]. The package is released under the GPL-2/GPL-3 license.

DGlearn We use the implementation provided by the authors of [Ghassami et al., 2020], available at <https://github.com/syanga/dglearn> (MIT license).

CallTiF The CallTiF implementation from [Arab et al., 2025] is available at https://github.com/nozarilab/2023Arab_CaLLTiF and is released under the GPL-3 license.

FRP We use the authors' implementation of FRP from [Yi and Lee, 2024], available at <https://github.com/soheunyi/frp>. No license information is explicitly stated.

B.5 DATA RESOURCES

US, European and Swiss stock returns The data are downloaded using *yahoofinancials* python library.