Unitless Unrestricted Markov-Consistent SCM Generation: Better Benchmark Datasets for Causal Discovery

Rebecca J. Herman REBECCA.HERMAN@TU-DRESDEN.DE
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany, German Aerospace Center (DLR), Institute of Data Science, Jena, Germany

Jonas Wahl JONAS.WAHL@DFKI.DE
German Research Centre for Artificial Intelligence (DFKI), Berlin, Germany

Urmi Ninad URMI.NINAD@TU-BERLIN.DE
Berlin Institute of Technology (TUB), Institute of Computer Engineering and Microelectronics, Germany

Jakob Runge JAKOB.RUNGE@TU-DRESDEN.DE
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
Berlin Institute of Technology (TUB), Institute of Computer Engineering and Microelectronics, Germany

Editors: Biwei Huang and Mathias Drton

Abstract

Causal discovery aims to extract qualitative causal knowledge in the form of causal graphs from data. Because causal ground truth is rarely known in the real world, simulated data plays a vital role in evaluating the performance of the various causal discovery algorithms proposed in the literature. But recent work highlighted certain artifacts of commonly used data generation techniques for a standard class of structural causal models (SCM) that may be nonphysical, including var- and R2-sortability, where the variables' variance and coefficients of determination (R2) after regressing on all other variables, respectively, increase along the causal order. Some causal methods exploit such artifacts, leading to unrealistic expectations for their performance on real-world data. Some modifications have been proposed to remove these artifacts; notably, the internally-standardized structural causal model (iSCM) avoids varsortability and largely alleviates R2-sortability on sparse causal graphs, but exhibits a reversed R2-sortability pattern for denser graphs not featured in their work. We analyze which sortability patterns we expect to see in real data, and propose a method for drawing coefficients that we argue more effectively samples the space of SCMs. Finally, we propose a novel extension of our SCM generation method to the time series setting.

Keywords: causal discovery, benchmarking, data generation, sortability, time series

1. Introduction

Causal discovery aims to extract qualitative causal knowledge from observational (or imperfect or incomplete interventional) data (Spirtes et al., 2000; Peters et al., 2017; Pearl, 2009), and a plethora of learning algorithms using different approaches exist for both the static (Nogueira et al., 2022; Hyttinen et al., 2014) and time series settings (Runge et al., 2019, 2023; Assaad et al., 2022; Camps-Valls et al., 2023). Benchmark datasets with known causal ground truth are vital to spur method development and to compare and evaluate existing methods, but unfortunately, datasets with known ground truth are often limited to a few feature variables (e.g. the bivariate Tübingen Cause-Effect pairs by Mooij et al., 2016) because it is rare to find real multivariate data where the underlying causal mechanisms are known with high certainty (Runge et al., 2019; Brouillard et al., 2024). Hence, synthetic data have played a vital role in method development and evaluation, but if the simulated datasets do not reflect the properties of the real-world datasets on which the algorithms will be applied, then high performance during evaluation may not translate to the real world.

There are typically two preliminary steps in data generation: generation of a causal graph, and assignment of causal functions and noise variables to form a Structural Causal Model (SCM). The details of both steps can affect the performance of causal discovery algorithms. It's generally unclear what constitutes the best approach for sampling causal graphs, and usually Erdös-Rényi (ER, Erdös and Rényi, 1960) and scale-free graphs (Barabási and Albert, 1999) are tested in parallel. The first approach samples randomly over possible edges in graphs with distinguishable vertices, and the second mimics the distribution of graphs resulting from typical mechanisms of network growth. Unfortunately, neither approach is designed to reflect the datasets intentionally curated from natural phenomena by scientists, and both over-represent graph structures with more isomorphic forms. The graph isomorphism problem is computationally hard (Torán, 2004), and we do not address it here. Instead, we address the functional mechanisms for specified graphs, focusing on linear additive models with Gaussian noise, in which mechanisms consist of coefficients and noise variances.

A standard approach is to sample coefficients from some uniform distribution over a bounded interval such as $(-2, -0.5) \cup (0.5, 2)$ and set noise variances to 1 (e.g. Zheng et al., 2018). However, recent work has highlighted certain artifacts that this method of choosing "random" functional relationships leaves on the generated data in both the static (Reisach et al., 2021, 2023) and time series settings (Lohse and Wahl, 2024). Two types of artifacts have received much attention: varsortability (Reisach et al., 2021) and R2-sortability (Reisach et al., 2023), in which the sample variance of the variables or the coefficient of determination after regressing on the other variables, respectively, are related to the topological order of the variables. Many prominent benchmarking datasets, such as that used for the NeurIPS 2019 Causality 4 Climate (C4C) competition (Runge et al., 2020), are strongly characterized by varsortability, tempting some to conclude that "effect variables tend to have larger marginal variance than their causal ancestors" and "large regression coefficients may predict causal links better in practice than small p-values" (Weichwald et al., 2020), but the validity of these claims rests strongly on whether it is reasonable to expect varsortability in the real world.

A couple alternative SCM generation methods have been proposed that attempt to avoid these artifacts (e.g. Mooij et al., 2020; Squires et al., 2022), including the internally-standardized structural causal model (iSCM; Ormaniec et al., 2024), which computationally standardizes every variable before generating the next. This removes varsortability and greatly reduces R2-sortability, as evidenced by numerical experiments on a certain subset of causal graphs. But their method of drawing causal parameters still limits the relative magnitudes of causal coefficients and noise terms, which may affect the absolute and relative performance of causal discovery methods in a nonphysical way.

We assert here that varsortability is nonphysical, but show that any data that is modeled by an SCM *should* feature a mild R2-sortability tendency *opposite* to that of standard data generation techniques. We argue that an unbiased SCM generation process should be *unitless*, *Markov-consistent* (remote structural properties are not associated with local functional parameters), and *unrestricted* (the relative magnitudes of parameters can be arbitrarily large or small; see Section 3.4), and propose a distribution from which to choose the coefficients and noise variances that satisfies these properties. A final contribution of our work is a novel extension of our SCM generation method to the time series setting, which enables benchmarking on realistic and theoretically grounded data generation models for data from time-dependent systems that ubiquitously occur across the sciences.

2. Foundations

2.1. Linear Additive Models with Gaussian Noise

2.1.1. STRUCTURAL CAUSAL MODELS

An acyclic linear additive Gaussian Structural Causal Model (SCM) with $N \in \mathbb{N}$ variables can be fully specified by the set of structural assignments

$$\left\{ X_i := \sum_j a_{ji} X_j + U_i,\ U_i \sim \mathcal{N}(m_i, s_i) \right. \tag{1}$$

for $i = 1, \ldots, N$, vectors $\boldsymbol{m} \in \mathbb{R}^N$ and $\boldsymbol{s} \in (\mathbb{R}^+)^N$, and upper diagonal matrix $A = (a_{ji}) \in \mathbb{R}^{N \times N}$ (an *intervention* on X_j consists of changing the right-hand side of its structural assignment). We assign the indices i consistent with the topological order, so that each *endogenous* variable X_i is directly affected by a subset of its *predecessors* $X_{<i} = \{X_j | j < i\}$ which we call its *parents* $pa(X_i) = \{X_j | a_{ji} \neq 0\} \subseteq X_{<i}$. X_i is affected by $X_j \in pa(X_i)$ via a linear causal effect a_{ji} where $a_{ji} = 0\ \forall j \geq i$ enforces acyclicity, and by noise drawn from independent normal distributions $\mathcal{N}(m_i, s_i)$ with mean m_i and standard deviation s_i. Any dependence between X_i and its non-parent predecessors is destroyed by conditioning on the parents of X_i: $X_i \perp\!\!\!\perp X_{<i} \setminus pa(X_i) | pa(X_i)$, known as the *local Markov property*. The SCM can be represented graphically by listing the $\{X_i\}$ and connecting $X_j \to X_i$ if and only if $X_j \in pa(X_i) \Leftrightarrow a_{ji} \neq 0$, and $X_j \in an(X_i) \subseteq X_{<i+1}$ is referred to as an *ancestor* of X_i if there is a directed path $X_j \to \cdots \to X_i$ of any length. The graph $\mathcal{G} = (V, E)$ is notated as an ordered pair with a list of nodes $V = \{X_i | i = 1, \ldots, N\}$ and a matrix of oriented adjacencies $E = \mathbb{1}(a_{ji} \neq 0) \in \{0, 1\}^{N \times N}$. The SCM may be standardized by subtracting the mean $\mu_i = \mathbb{E}[X_i] = \sum_j a_{ji} \mu_j + m_i$ and dividing by the standard deviation $\sigma_i = (\mathbb{E}[X_i^2] - \mu_i^2)^{1/2}$ for each variable X_i, yielding the *Unitless SCM* (see Definition 2)

$$\left\{ \hat{X}_i := \sum_j \hat{a}_{ji} \hat{X}_j + \hat{U}_i, \hat{U}_i \sim \mathcal{N}(0, \hat{s}_i) \right. \tag{2}$$

subject to the constraint $\sum_{jk} \hat{a}_{ji} \hat{a}_{ki} \hat{\rho}_{jk} + \hat{s}_i^2 = 1$, where $\hat{a}_{ji} := \frac{\sigma_j}{\sigma_i} a_{ji}$ and $\hat{s}_i := \frac{s_i}{\sigma_i}$ are unitless, and $\hat{\rho}_{jk} = \mathbb{E}[\hat{X}_j \hat{X}_k] = \mathbb{E}\left[\frac{X_j - \mu_j}{\sigma_j} \frac{X_k - \mu_k}{\sigma_k} \right] = \rho_{jk}$ are elements of the correlation matrix for $\{X_i\}$.

2.1.2. STRUCTURAL VECTOR AUTOREGRESSIVE MODELS

For time series, we cannot assume that samples from V are independent. Formally, we can include a node $X_i(t)$ for every variable at every time, and allow a lagged causal effect $X_j(t - \tau) \to X_i(t)$ for any $\tau \in \mathbb{N}$. The resulting infinite graph is called a time series (Runge et al., 2012) or full time graph (Peters et al., 2017). Complete models often require contemporaneous dependencies as well (where $\tau = 0$), and in practice, $\tau \leq \tau_{\max}$ is bounded by some maximum lag $\tau_{\max} \in \mathbb{Z}^+$.

Without further assumptions, we would need a causal effect coefficient $a_{ji,t}(\tau)$ for every pair $(X_j(t - \tau), X_i(t))$ of X_i and X_j at time t and lag $\tau \in 0, \ldots, \tau_{\max}$, as well as means $m_{i,t}$ and standard deviations $s_{i,t}$ for the noise distributions $U_i(t)$ for every variable X_i at every time t. Under the assumption of *causal stationarity*, $a_{ji,t}(\tau) = a_{ji}(\tau)$, $m_{i,t} = m_i$, and $s_{i,t} = s_i\ \forall i, j, t, \tau$, giving

$$\left\{ X_i(t) := \sum_{j=1}^{N} \sum_{\tau=0}^{\tau_{\max}} a_{ji}(\tau) X_j(t - \tau) + U_i(t),\ U_i \sim \mathcal{N}(m_i, s_i) \right. \tag{3}$$

This can be represented with a *summary causal graph* with nodes $\{X_i | i \in 1, \ldots, N\}$ and an edge $X_j \to X_i$ if $X_j(t - \tau) \in pa(X_i(t))$ for some $\tau \in 0, \ldots, \tau_{\max}$. This process may contain causal effects from a variable to itself (when $i = j$), and though we restrict our analysis to acyclic time series graphs where $a_{ji}(0) = 0\ \forall j \geq i$, *unrolled* cross-dependencies that lead to a cyclic summary graph ($X_j(t - \tau) \to X_i(t)$ and $X_i(t - \nu) \to X_j(t)$, with $\tau > 0,\ \nu \geq 0$) may appear.

With linear additive causal effects $A(\tau) = (a_{ji})(\tau) \in \mathbb{R}^{N\times N\times\tau}$ and Gaussian noise, such a model would constitute a *Structural Vector Autoregressive* model (SVAR; Lütkepohl, 2005a) $\mathsf{A}\mathbf{X}(t) = \sum_{\tau=1}^{\tau_{\max}} \mathsf{A}A^*(\tau)\mathbf{X}(t-\tau) + \mathbf{U}(t)$ where $\mathsf{A} = A(0)^{-1} + \mathbf{I}_N$ and $A^*(\tau) = \mathsf{A}^{-1}A(\tau)$. This process is *stable* if $\det(\mathsf{A} - \sum_{\tau=1}^{\tau_{\max}} \mathsf{A}A^*(\tau)z^\tau) = \det(A(0)^{-1} + \mathbf{I}_N - \sum_{\tau=1}^{\tau_{\max}} A(\tau)z^\tau) = 0$ has roots with magnitude less than 1 (extended from Lütkepohl, 2005b). Stability is a sufficient condition for *statistical stationarity*, meaning that the mean and (co)variance is equivalent conditioned on any time (but not on previous observations). The standardized form of the SVAR model is

$$\left\{\hat{X}_i(t) = \sum_{j=1}^{N}\sum_{\tau=0}^{\tau_{\max}} \hat{a}_{ji}(\tau)\hat{X}_j(t-\tau) + \hat{U}_i(t),\ \hat{U}_i \sim \mathcal{N}(0, \hat{s}_i) \right. \tag{4}$$

where $\hat{a}_{ji}(\tau) = \frac{\sigma_j}{\sigma_i}a_{ji}(\tau)$ and $\hat{s}_i = \frac{s_i}{\sigma_i}$, constrained by $\sum_{jk}\sum_{\tau\nu}\hat{a}_{ji}(\tau)\hat{a}_{ki}(\nu)\hat{\rho}_{jk}(\tau-\nu) + \hat{s}_i^2 = 1$ where $\hat{\rho}_{jk}(\omega) = \mathbb{E}[\hat{X}_j(t-\omega)\hat{X}_k(t)] = \mathbb{E}\left[\frac{X_j(t-\omega)-\mu_j}{\sigma_j}\frac{X_k(t)-\mu_k}{\sigma_k}\right]$ are lagged correlations for $\{X_i\}$.

2.2. Benchmarking Data

2.2.1. EXISTING BENCHMARKS

Many real-world and synthetic causal discovery benchmarks focus on the bivariate cause-effect identification case. To name a few: the Tübingen Cause Effect Pairs (Mooij et al., 2016) cover a large number of synthetic as well as real-world pairs of static and time series data; Guyon et al. (2019) present results from a large benchmark competition, all of which is now freely available[1]; and Käding and Runge (2023) introduce a bivariate synthetic benchmark suite. But while there are some real-world multivariate static datasets used for benchmarking, such as those hosted at CMI[2] and the new customizable Causal Chamber (Gamella et al., 2024), they are few in number because it is quite challenging to obtain complete real-world ground truth causal knowledge, and thus synthetic data plays a more important role in benchmarking in the multivariate setting (Brouillard et al., 2024).

The crucial choice to make when generating random SCMs and SVAR models is what joint distribution to use for the causal coefficients and the noise standard deviations. A simple and commonly-used approach in the static setting is to draw a_{ji} and s_i independently over some uniform probability distribution. Zheng et al. (2018) and many other studies use $\mathcal{U}([-2,-0.5]\cup[0.5,2])$ for the causal coefficients and 1 for the noise standard deviations (we call this 'UVN' for 'Unit Variance Noise'), but other studies make other choices; for example, Andrews et al. (2023) use $\mathcal{U}(-1,1)$ for the causal coefficients and $\mathcal{U}(1,2)$ for the noise standard deviations. The linear-VAR datasets provided on `causeme.net` (Runge et al., 2019) as part of the NeurIPS 2019 Causality 4 Climate (C4C) competition benchmarking data (Runge et al., 2020) adapt this approach to the time series setting including a test for stability, and Lawrence et al. (2021) propose a framework for generating random time series benchmarking data with fewer assumptions about the data generation process.

Other approaches attempt to be more realistic by using real data and expert knowledge as a starting point. They often employ expert knowledge to constrain the causal graph and real data to fit the causal coefficients, and then generate simulated data from the resulting 'ground truth' SCMs. Some static examples include `causalAssembly` (Göbler et al., 2024), which is based on data measurements taken from a manufacturing assembly line, and SynTReN (Van den Bulcke et al., 2006), which mimics experimental gene expression data by randomly choosing from networks and

1. https://www.causality.inf.ethz.ch/cause-effect.php
2. https://www.cmu.edu/dietrich/causality/projects/causal_learn_benchmarks/

functional relationships crafted by experts. Unfortunately, the applicability of such pseudo-real datasets is often limited to the field from which the real data was taken, and is further limited by the properties of the finite datasets on which the pseudo-real data was based.

Pseudo-real time series data can be generated to mimic any user-provided real dataset via CausalTime (Cheng et al., 2023), and existing pseudo-real benchmarking datasets mimicking climate and weather data can be found at causeme.net. CausalTime constructs the 'ground truth' causal models by (1) using deep neural networks to fit a nonlinear AR model (NAR), (2) extracting possible tractable causal graphs using importance analysis or expert knowledge, and finally (3) extracting the part of the fitted NAR model consistent with the causal graph. CauseMe takes a similar approach: beginning with randomly-selected large climate simulations and variables, it (1) extracts major modes of variablility using PCA-Varimax (Tibau et al., 2022), (2) fits a VAR model to the data, and (3) uses the residuals to determine the magnitude of the noise. In addition to the limitations of static pseudo-real data, since both automated model-generation methods are fundamentally based on non-causal analysis, they may or may not be able to capture the desired real-world distribution of causal systems that regression-based methods would fail to fully characterize.

2.2.2. SORTABILITY

Varsortability Reisach et al. (2021) introduce the concept of *varsortability* in static graphs, which means that variance increases along the topological order. Reisach et al. (2021) define a sortability metric that ranges from 0 to 1, where values above 0.5 indicate varsortability and values below 0.5 indicate reverse-varsortability, and Lohse and Wahl (2024) extend their metric to the time series setting by applying it to summary graphs, excluding pairs of nodes that are cyclically related. Reisach et al. describe their metric as a "fraction of directed paths," but algorithmically, they only count directed paths between the same pair of nodes separately if they have *distinct lengths*, yielding a metric that is not directly related to directed paths nor to connected node pairs. We propose a simple modification to Reisach et al.'s algorithm that counts each pair of nodes exactly once (see Appendix B for python code), changing sortability values quantitatively, but not qualitatively (see Figure 2 last subplot for an example using R2-scores instead of variance). Reisach et al. (2021) observed that the static data generation techniques from Section 2.2.1 produce strongly varsortable datasets, and we confirm this computationally (Figure A2a and c). Lohse and Wahl (2024) showed that generated time series datasets such as the C4C data are also often highly varsortable even when they are stable over time. Varsortability can be removed from generated data post-hoc via standardization.

R2-sortability Reisach et al. (2023) found that the joint distributions for SCM parameters that produce varsortability in simulated datasets leave another artifact in the data that cannot be removed after the fact: increasing fractional cause-explained variance along the topological order. Though ground truth causal structure is needed to calculate cause-explained variance, they show that the coefficient of determination (R^2, Glantz et al., 2017)—which is measurable in the absence of causal knowledge—gives an upper-bound for the fraction of cause-explained variance and is an effective proxy for recovering the topological order from simulated static datasets (see Figure A2b and c).

The extension of R2-sortability to the time series setting is less obvious than it is for varsortability; Lohse and Wahl (2024) extend the definition by focusing on the fraction of cause-explained variance due to *distinct processes*, calculating R^2 for $X_i(t)$ based on $\{X_j(t-\tau)|j \neq i, \tau \in 0...\tau_{\max}\}$, and excluding the past of X_i. We refer to this definition as R2*-sortability. Strong varsortability appears to be associated with strong reverse-R2*-sortability in both real and simulated time series

datasets (Lohse and Wahl, 2024). The reversal may be because variables with large variance are likely to be affected more by their own past than by distinct processes when all causal edge weights are drawn from the same distribution. We assert that a better extension of R2-sortability to the time series setting should include the past of the entire system, thus approximating the fraction of cause-explained variance in the time series graph. Figure A1 shows that this modified definition can revert the association so that varsortability is associated with R2-sortability, as in the static case.

Real Systems While simulated static and time series datasets are characterized by strong var- and R2-sortability patterns, Lohse and Wahl (2024) found that two real-world time series datasets—a river flow dataset where the causal relationships can reasonably be given by the locations of data collection relative to the river flow (Tran et al., 2024), and data from a physical Causal Chamber in which the researcher controls the causal relationships (Gamella et al., 2024)—demonstrate opposite extremes in var- and R2*-sortability, suggesting that simulated sortability tendencies are unrealistic.

2.2.3. AVOIDING SORTABILITY

There have been a few attempts in the literature to alleviate var- and R2-sortability in generated static benchmark data. Mooij et al. (2020) assign unit noise, then divide it and the causal coefficients $\boldsymbol{a_i}$ for each node X_i by $(1 + \sum_j a_{ji}^2)^{1/2}$: the standard deviation X_i would have if its parents were independent (we call this 'IPA' for 'Independent Parents Assumption'). Datasets generated by IPA have weaker expected varsortability; but, in combination with the authors' choice of distribution for drawing a_{ji}, IPA produces reverse var- and R2-sortable datasets on average (Figure A2d-f). Squires et al. (2022) account for covariance of the parents but enforce an even split between explainable variance and noise by generating data without noise, dividing the data and $\boldsymbol{a_i}$ by $\sqrt{2}$ times the sample standard deviation, and then adding noise with $s_i = \frac{\sqrt{2}}{2}$ (we call this '50-50'). 50-50 lacks any tendency for varsortability (Figure A2g and i), but produces a tendency toward reverse-R2-sortability that is just as strong as the original tendency toward R2-sortability in standard data generation techniques (Figure A2h and i). Ormaniec et al. (2024) also standardize during data generation, but they scale $\boldsymbol{a_i}$ and s_i by the full sample standard deviation $\sqrt{\frac{1}{P-1}\sum_p^P \left[U_j^{(p)2} + \left(\sum_j a_{ji} X_j^{(p)}\right)^2\right]}$, thus completely removing varsortability (Figure A2j). They name this hybrid SCM-standardization approach the 'iSCM', but it can be viewed as a different sampling of the distribution of SCMs, like all approaches discussed in Section 2.2.1 and here. The iSCM reduces R2-sortability enough that it is not noticeable on the sparse graphs featured in the main body of their paper, but a tendency toward reverse R2-sortability is still visible for denser graphs (Figure A2k and l). Finally, Andrews and Kummerfeld (2024) introduce the "DAG Onion" method (DaO), which uniformly samples joint probability distributions consistent with a graph, then samples SCMs that could produce it. Contrary to the motivation given in the paper, DaO produces data that is strongly reverse varsortable (Figure A2m and o) and mildly (reverse) R2-sortable for sparse (dense) graphs (Figure A2n and o).

3. Thought Experiments

With so many different ways to sample SCMs, the question should not be how to remove any exploitable artifact, but rather, what artifacts do we expect to see, and how do we want to sample the space of possible SCMs? Do we expect sortability tendencies? Should all variables be equally noisy? Does it make sense to draw causal coefficients uniformly or to limit how large or small the coefficients can be? The following thought experiments explore these questions.

3.1. Toy Model: Sortability in Standard SCM Generation Techniques

To gain intuition for the standard generation techniques described in Section 2.2, we inspect a toy SCM generation process that always assigns unit noise and causal coefficients. Though not random, it retains the features that lead to sortability in commonly-used random SCM generation techniques. We begin by examining chains of different lengths. For a chain of length 2, such as $X \to Y$, our toy SCM generation method would produce X with variance 1 – all of it noise, and Y with variance 2 – half of which is noise and half of which is causally explainable. If we added a third variable to the chain, it would have a variance of 3, one third of which is noise and two thirds of which is explainable. Thus, this toy generation process yields SCMs where the total variance and fraction of causally-explainable variance for a variable reflects the number of *ancestors* it has. This results in var- and R2-sortability because the maximum number of ancestors a variable can have in an acyclic SCM is limited by its place in the topological order. Although this relationship is not deterministic when coefficients are assigned randomly, it holds on average for SCM generation techniques using unit noise because all root nodes must have a variance of 1 (all of it noise), while any variable with ancestors must have a variance that is larger than 1 (because the noise is independent from the cause-explained variance) by an amount that is, on average, proportional to the variance of its parents. Thus, we note that causal discovery algorithms relying on the equal variance assumption (Ng et al., 2024; Peters and Bühlmann, 2013; Chen et al., 2019) implicitly assume sortability.

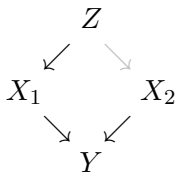
3.2. Modeling Choices and Marginalization: Number of non-Parent Ancestors

When appending variables to a chain, increasing variance along the topological order may at first appear to be a reasonable physical pattern, but it becomes clearly inappropriate when we recognize that the number of ancestors a variable has in an SCM is a *modeling choice*, not a fundamental physical property of the system. Given an SCM, one can always define a smaller—but equally-valid—SCM over any subset of the original variables through *marginalization*, which absorbs each excluded variable into the exogenous noise terms of its children. If we begin with a causally sufficient SCM (the noise terms are jointly independent) and every marginalized variable has at most one child, then the marginalized SCM will also be causally sufficient. Likewise, it is often possible to expand an SCM representing a particular physical system by including additional parents and ancestors. Is there a limit to how many predecessors one can add to a physically-meaningful SCM? This would mean reaching an indisputable "root cause", which is not influenced by anything at all. Whether such a "root cause" exists is a theological question rather than a scientific one, but personal beliefs cannot affect the conclusion that, in practice, there is no "root cause" that can be included in an SCM, because those who believe in an absolute power also believe that it is unknowable.

We want the outputs of our SCM generation process to represent physical properties of real systems, not arbitrary modeling choices. Let's assume our example chain from Section 3.1 represents some real physical system. How does our representation of the system change when we apply marginalization transformations? If we marginalize out X, we are left with the causal chain $Y \to Z$, where the variance of Y is still 2 and the variance of Z is still 3. In this example and in general, the total variance of a variable has no relationship to the number of parents or ancestors included in the SCM. But while the fraction of causally explainable variance of Z is still $2/3$, removing X renders Y completely unexplainable. The difference is that X was a *parent* of Y, but only an *ancestor* of Z, making it clear that the fraction of causally explainable variance in an SCM representing a real system is related to the number of *parents* included in the scientist's model, rather than the number of *ancestors*. A realistic SCM generation process should not produce variances that reflect the struc-

ture of the causal graph, and so the variables may as well be standardized; but it is acceptable—and expected—for probability distribution of the fraction of variance due to noise to be related to the number of *parents* (but not ancestors) a variable has in the causal graph used to model the system.

3.3. Remote Changes: Independence of Cause and Mechanism



Say some physical system can be represented by the causal graph given by the black arrows in the diagram to the right and the SCM with unit coefficients and noise. The variances of Z and X_2 would be 1, the variance of X_1 would be 2, and the variance of Y would be 4, with three fourths of it explainable.

Now, say we intervene on X_2, activating the gray edge by changing the structural assignment to $X_2 := -\frac{\sqrt{2}}{2}(Z + U_2)$ with $U_2 \sim \mathcal{N}(0, 1)$. The variance of X_2 is still 1, but now it has a covariance with X_1 of $\mathbb{E}[(Z)(-\frac{\sqrt{2}}{2}Z)] = -\frac{\sqrt{2}}{2}$. The physical process determining Y should not change, but Y's variance is now $\mathbb{E}[X_1^2] + 2\mathbb{E}[X_1X_2] + \mathbb{E}[X_2^2] + 1 = (2) + 2(-\frac{\sqrt{2}}{2}) + (1) + 1 = 4 - \sqrt{2}$, and its fraction of cause-explainable variance is $\frac{3-\sqrt{2}}{4-\sqrt{2}} = \frac{10-\sqrt{2}}{14}$. Though the actual functional parameters for Y do not respond to the correlation of Y's parents, because the total variance has changed, the fraction of explainable variance changes, and all *standardized* functional parameters must be scaled by the old standard deviation divided by the new standard deviation (see Section 2.1.1). Since the scalar is the same for all parameters, the *relative* magnitudes and signs of the parameters (the "mechanisms") will not change, and so we would not want the distribution of these relationships in randomly sampled standardized parameters for Y to reflect the correlation between X_1 and X_2 (the "causes"), consistent with the principle of separation of cause and mechanism.

3.4. Desired SCM-Generation Characteristics

In Section 3.2, we find that the variance of a variable is independent of modeled causal structure. Since its value depends on the chosen units and units are not specified, it is arbitrary. In Section 3.2, we find that the fraction of unexplained variance is independent of the number of non-parent ancestors, and in Section 3.3, we find that the functional parameters are independent up to uniform scaling of the influence of remote structural features and functional parameters on the correlation of parents. We conclude that the distribution of functional parameters produced by an SCM generation process should depend only on the number of parents (local structure). This can be seen as a generalization of the local Markov property (which relates data to the SCM that generated it) to the SCM generation process itself. Finally, clearly, no physical limit on the ratio between causal coefficients and noise variance exists.

Definition 1 *An* **SCM generation process** *over nodes V and some function space with parameterization $\mathbf{\Theta}$ defines the distribution $\mathcal{P}(\mathbf{\Theta}|E) = \Pi_i \mathcal{P}(\mathbf{\Theta}_i|E, \mathbf{\Theta} \setminus \mathbf{\Theta}_i)$ for every (random) input graph $\mathcal{G} = (V, E)$, where $\mathbf{\Theta}_i$ is a vector of the parameters of the structural assignment for $X_i \in V$. It is...*

Definition 2 **unitless** *if $\forall E, X_i \in V$, $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1$.*

Definition 3 **Markov-consistent** *if for any adjacency matrices E, E' and nodes $X_i, X_j \in V$ with the same number of parents, $\exists c \in \mathbb{R} | c \neq 0 \wedge \mathcal{P}(\tilde{\mathbf{\Theta}}_i) = \mathcal{P}(c \odot \tilde{\mathbf{\Theta}}_j)$ where $\odot$ is the* scaling product *for the parameterization $\mathbf{\Theta}$ and $\tilde{\mathbf{\Theta}}_i$ are elements of $\mathbf{\Theta}_i$ that are random given E_i (see Appendix A).*

Definition 4 **unrestricted** *if the distribution of the ratio of variance due to a single parent and to noise for each node has full support on $(0, \infty)$.*

4. Method

We propose Algorithm 1. If a variable $\hat{X}_i$ in a standardized system has d_i independent parents, we may imagine its variance as a unit d_i-ball where each Cartesian coordinate represents the causal coefficient and standard deviation from one of the parents, and $1 - r^2 < 1$ is the variance due to noise. We draw coefficients randomly and independently by sampling uniformly from this d_i-ball, which can be accomplished by drawing $a''_{ji} \sim \mathcal{N}(0,1)$ iid from the standard normal distribution and scaling $\boldsymbol{a_i}'' = (a''_{ji})_j$ by $r_i/||\boldsymbol{a_i}''||$, where $r_i \sim \mathcal{U}^{1/d_i}(0,1)$ is drawn from the d_ith root of the uniform distribution from 0 to 1 (Harman and Lacko, 2010). The fraction of the volume located near the outside of a d_i-ball increases with d_i, so the expected fraction of variance due to noise is inversely related to the number of parents (average noisiness can be adjusted via the distribution for $r_i^{d_i}$), but recalling Section 3.2, this is not concerning. To ensure unit variance with correlated parents, we standardize (because $a'_{ji} = a''_{ji} = 0$ when $\hat{X}_j \notin pa(\hat{X}_i) \subseteq \hat{X}_{<i}$, we only need $\hat{X}_{<i}$ to do this).

Algorithm 1 UUMC SCM Generation

Require: a graph $\mathcal{G} = \left(\{\hat{X}_i\}, E\right)$ with topological order $i \in 1...N$ and adjacency matrix $E = (e_{ji}) \in \{0,1\}^{N\times N}$ with $N \in \mathbb{N}$
Ensure: $j \geq i \Rightarrow e_{ji} = 0$
$R = (\rho_{ji}) \leftarrow \mathbb{I}_N$
$\boldsymbol{s} = (\hat{s}_i) \leftarrow \mathbf{1}_N$
$\hat{A} = (\hat{a}_{ji}) \leftarrow E$
for $i \in 1...N$ **do**
  $d_i \leftarrow \sum_j \boldsymbol{e}_i = \#pa(\hat{X}_i)$
  if $d_i > 0$ **then**
    draw $\boldsymbol{a}''_i \sim \mathcal{N}(0,1)^N$
    draw $r_i \sim \mathcal{U}^{1/d_i}(0,1)$
    $a''_{ji} \leftarrow a''_{ji} e_{ji}\ \forall j \in 1...N$
    $\boldsymbol{a}'_i \leftarrow \frac{r_i}{\sum_j a''^2_{ji}} \boldsymbol{a}''_i;\ s'_i \leftarrow \sqrt{1 - r_i^2}$
    $\sigma'_i \leftarrow \sqrt{\boldsymbol{a'_i}^T R \boldsymbol{a'_i} + s'^2_i}$
    $\hat{a}_{ji} \leftarrow \frac{a'_{ji}}{\sigma'_i}\ \forall j \in 1...N;\ \hat{s}_i \leftarrow \frac{s'_i}{\sigma'_i}$
    $\rho_{ij} = \rho_{ji} \leftarrow \sum_k \hat{a}_{ki}\rho_{jk}\ \forall j < i$
  end
end
return $\hat{A}$

Theorem 1 *Algorithm 1 is unitless, unrestricted, and Markov-consistent.*

Proof Assume that $\hat{X}_{<i}$ are standardized. Then we have $\hat{\mu}_i = \sum_j \hat{a}_{ji}(0) + (0) = 0$ and $\hat{\sigma}_i^2 = \frac{1}{\sigma'^2_i}\left(\mathbb{E}\left[\left(\sum_j a'_{ji}\hat{X}_j\right)^2\right] + (s'_i)^2\right) = \frac{\sum_{jk} a'_{ji}a'_{ki}\rho_{jk} + s'^2_i}{\boldsymbol{a'_i}^T R \boldsymbol{a'_i} + s_i^2} = 1$, so Algorithm 1 is unitless. It is Markov-consistent because $\boldsymbol{a''_i}$ are drawn identically and independently, the distribution for r depends only on d_i (the number of parents), and the final modification is scaling all parameters by $\sigma'_i||\boldsymbol{a_i}''||$. Finally, Algorithm 1 is unrestricted because the ratio of the variance due to an individual parent to that due to noise is proportional to $\frac{r_i^2}{1-r_i^2}$, which approaches 0 as $r \to 0$ and approaches ∞ as $r \to 1$. ■

Remark 1 *While the iSCM is also Markov-consistent, it is not unitless because it generates data that is standardized for a finite sample rather than in the infinite sample limit, and it is not unrestricted because it produces SCMs where the ratio of the magnitude of a causal coefficient $|\hat{a}_{ji}|$ to the noise standard deviation $\hat{s}_i$ for $\hat{X}_i$ is uniformly distributed on the interval $[0.5, 2]$.*

5. Experiments

5.1. Sortability Properties

As intended, our SCM generation method results in an expectation of neutral varsortability for ER graphs (Figure 1a and c). Though the SCM is constructed to be standardized in the infinite sample

limit, finite sample size still leads to a wide range of simulated varsortability values. Likewise, this data-generation method produces SCMs with a realistically-wide range of R2-sortability (Figure 1b). This is because, in the absence of varsortability, R2 scores are highest for the node with the most neighbors, regardless of their causal order. We confirm this by examining the R2 scores of nodes in the three types of unshielded triples in Figure 2a-c. The nodes are labeled by topological order; the nodes with 2 neighbors include node 2 for the collider, node 1 for the chain, and node 0 for the confounder. The expected R2 scores for nodes with only one neighbor are low in each case, while the expected R2 score for the node with 2 neighbors is higher, favoring nodes at different positions in the topological order depending on the connectivity of the graph.

When confounders and colliders occur with equal frequency, one might expect the R2-sortability tendencies for these different structures to exactly cancel out. Instead, we see a positive skew (Figures 2f and 1b), meaning our data is *reverse* R2-sortable on average, similar to the iSCM and IPA approaches (Figure A2e and k). This holds for random graphs of all densities (Figure 1b-c). Should we expect such an artifact? Analytically, yes. The expected R2 score is not the same for the middle node of different types of triples (proof in Appendix D). For a collider such as $\hat{A} \to \hat{B} \leftarrow \hat{C}$, the R2 score for $\hat{B}$ is upper-bounded by $1 - \hat{s}_b^2$, because there is no information in $\hat{A}$ or $\hat{C}$ that could possibly help predict $\hat{U}_B$. However, for a confounded triple such as $\hat{A} \leftarrow \hat{B} \to \hat{C}$, the R2 score is actually lower-bounded by $1 - \min(\hat{s}_A^2, \hat{s}_C^2)$, and can get arbitrarily close to 1. This is because all information from the parent influences every child. While that information is obscured by the noise terms of the children, these noise terms are mutually independent, so they partially cancel each other during the attempted reconstruction of the parent node.

We confirm this computationally: while R2-scores are equally distributed for nodes with one neighbor in all types of unshielded triples (Figure 2d), the R2 score for the middle, or *hub*, node of the collider (green) is on average slightly lower than that of the other structures (Figure 2e). This pattern holds for larger structures as well. Naturally by the law of large numbers, as we increase the number of children for a confounding hub node, its expected R2 score increases (Figure 3, blue) – a pattern which should hold regardless of the data generation strategy. On the other hand, the behavior of the R2 score of a 'colliding' hub node as we increase the number of parents (red) reflects our choice to reduce the fraction of variance due to noise for variables with more parents. The similarity in the relationship of R2 scores to the number of neighbors for hub nodes from these different structures gives us confidence that this was a realistic choice. Despite the similarity, the upper-bound for the collider means that it always lags the R2-score for the hub of a confounder.

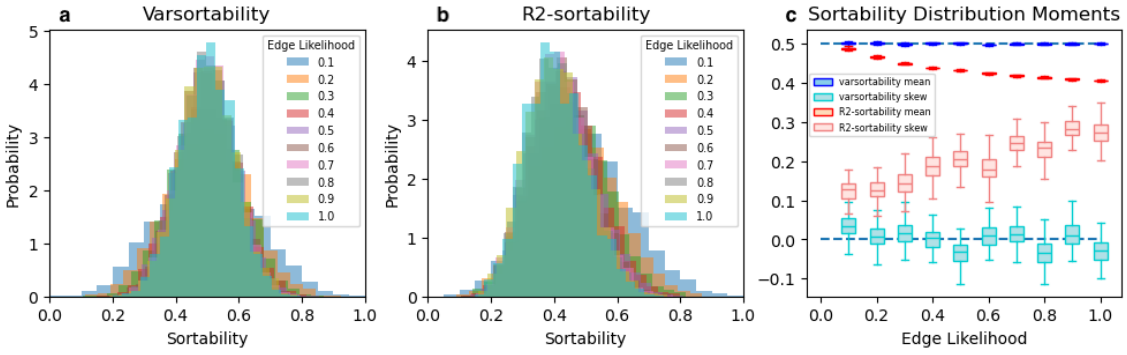


Figure 1: Sortability probability distribution functions for static ER graphs with 20 nodes and varying edge probabilities, based on 5000 randomly generated SCMs and 100 data samples.

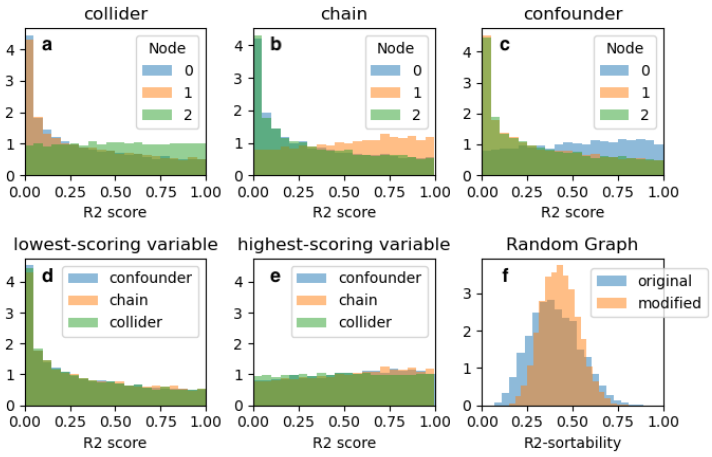


Figure 2: **a-c** R2 scores for nodes in the three kinds of unshielded triples, labeled by topological order, based on 10000 random SCMs and 500 samples. **d** Highest and **e** lowest scoring node from each triple. **f** R2-sortability for ER graphs with 20 nodes and edge probability 0.5, based on 5000 random SCMs and 100 samples, according to the original (blue) and modified (orange) sortability metrics (see Section 2.2.2).

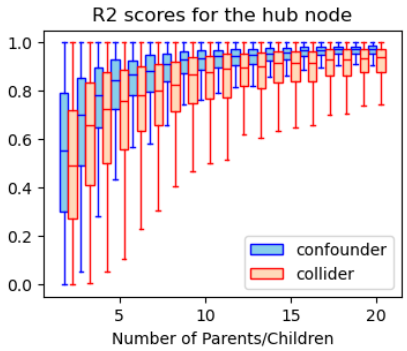


Figure 3: R2 scores for the hub node of colliders and confounders with different numbers of parents/children, based on 500 randomly-generated static colliders and confounders of each size and 500 simulated data points.

5.2. Causal Discovery Evaluation

In Figure 4, we examine the impact of data generation strategy on the apparent performance of various static causal discovery algorithms (mostly as implemented in gCastle (Zhang et al., 2021)). PC(-stable) are constraint-based methods relying on conditional independence testing (Kalisch and Buehlmann, 2005) and the *faithfulness assumption* (that connections in the causal graph manifest as statistical relationships in the data) to discover partially-oriented equivalence classes. The other methods learn fully-oriented graphs; Greedy Equivalence Search (GES) is score-based (Chickering, 2002), NOTEARS is gradient-based (Zheng et al., 2018) like the winner of the C4C competition (Weichwald et al., 2020), and Best Order Score Search (BOSS) is a permutation-based algorithm that uses a score based on conditional independence constraints and relaxes the faithfulness assumption by relying more strongly on causal sufficiency.

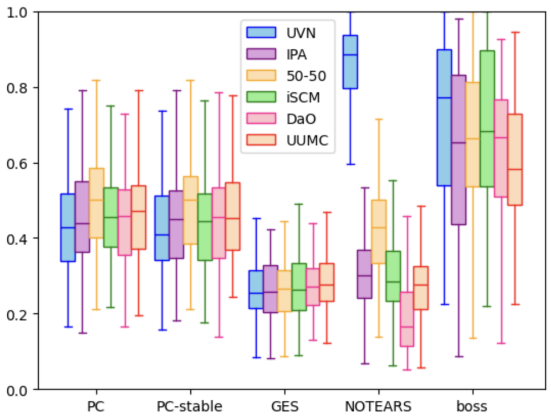


Figure 4: F1-score comparison of static causal discovery algorithms on benchmark data generated in different ways (see Section 2.2) based on 100 SCMs with 10 variables and 500 data samples. "UUMC" is the method proposed here.

In blue, we show the performance of these algorithms on data generated in the standard way (denoted UVN for unit-variance-noise, see Section 2.2.1) according to gcastle's directed F1 score. NOTEARS, which was introduced accompanied by an evaluation on UVN data, far outperforms the other algorithms on this dataset, followed closely only by BOSS. Note that the apparent performance of PC(-stable) is limited because the F1 score is not implemented for equivalence classes.

In purple, yellow, green, and pink, we present the performance of these algorithms on data generated according to the four proposals for addressing sortability artifacts discussed in Section 2.2.3. By far the most dramatic change is the reduced performance of NOTEARS for any of these approaches, all of which reduce varsortability, and especially for DaO (pink), which exhibits reverse varsortability (Figure A2). Though not nearly as dramatic, the performance of BOSS is also markedly reduced for all modified data generation approaches. Within the modified approaches, the performances of score-search algorithms like GES and BOSS remain relatively constant, but the constraint-based algorithms perform noticeably better on the 50-50 data (yellow), likely because this approach ensures that causal effects are not overshadowed by noise while avoiding computational faithfulness violations due to near-determinism (see Runge, 2018, for a discussion of the relevant assumptions). Interestingly, NOTEARS also performs better on the 50-50 data, suggesting that it is more effective when the true fraction of explainable variance is similar for all variables.

Finally, the performance of these algorithms on data from UUMC SCMs generated using Algorithm 1 is shown in red. Compared to the iSCM, our approach yields a marginally higher performance for the PC algorithms and a marginally lower performance for NOTEARS, slightly increasing the apparent marginal performance of constraint-based relative to gradient-based algorithms. Furthermore, while BOSS has the highest performance of all the algorithms on data that is not varsortable, it performs notably worse on the UUMC data than any of the other modified approaches, materially reducing the apparent performance improvement provided by BOSS.

6. Extension to Time Series Data

Extending the method to the time series setting is non-trivial. Even if we insist that our method reduces to the static case when $\tau_{\max} = 0$ thus forcing all dependencies to be contemporaneous, the extension to lagged (auto-)dependencies is not obvious. Note, for instance, that the discrete auto-dependencies in an SVAR model (equation 3) are naturally unitless—$a_{ii}(\tau) = \hat{a}_{ii}(\tau)$, where $\hat{a}_{ii}(\tau)$ are the auto-dependencies in the unitless SVAR model defined by eq. 4—while other lagged dependencies ($a_{ji}(\tau),\ j \neq i, 0 < \tau \leq \tau_{\max}$) are unit-dependent. Furthermore, we can no longer simply use coefficients from the causal model to represent the contribution of a parent process to its child, since the parent may affect the child process via multiple lags.

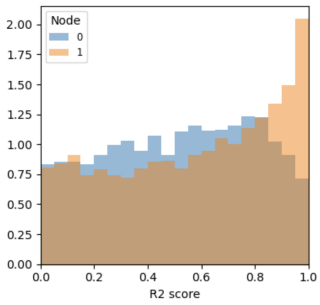


Figure 5: R2-scores for random SCMs corresponding to the time series graph given by $\hat{X}_0(t-1) \to \hat{X}_0(t) \to \hat{X}_1(t) \to \hat{X}_1(t+1)$.

In Appendix E, we propose a theoretical foundation for drawing causal parameters based on a comparison to continuous processes. We approximate the contribution of a parent process with the sum of the causal coefficients at all lags, inspired by the direct transfer function at frequency 0 (see Appendix E.2 and Reiter et al., 2023). Auto-dependence is drawn from $\mathcal{U}(0, 1)$, the fraction of variance due to noise is drawn relative to the part of the total variance not due to auto-dependence.

Both auto- and cross-dependencies are adjusted to respond to a random sampling rate, inspired by discrete sampling of Ornstein-Uhlenbeck processes (see Appendix E.1), and we show that hidden confounding from subsampling can be bounded based on the sampling frequency. Finally, we solve a system of equations yielding the (auto-)covariances and scaling parameters for noise and the cross-coefficients to achieve standardized time series. The full method is given in Algorithm A1. It is designed to produce datasets with unit variance, but R2-sortability tendencies may emerge.

Figure 5 shows that for a simple 2-variable SVAR process with lag-1 auto-dependencies and $\hat{X}_0(t) \to \hat{X}_1(t)$, the distribution of R2 scores for the source node (blue) differs from that of the target node (orange). On the whole, time series datasets generated this way tend toward R2-sortability.

7. Discussion

We examine data generation techniques that aim to avoid var- and R2-sortability, and evaluate their success in removing these signatures of the underlying causal structure. Notably, the Dag-Onion method—which is meant to naturally avoid nonphysical artifacts and ensure simulation of data with all possible correlation structures—results in strong reverse varsortability and different R2-sortability tendencies for graphs of different densities. Also, while the iSCM eliminates varsortability, it demonstrates mild reverse R2-sortability that is stronger for denser graphs. Is this a problem?

We propose a method for sampling Structural Causal Models (SCMs) that is based on theoretical arguments that appropriately distinguish between physical properties of a system and arbitrary modeling choices such as the number of included ancestors or the choice of units. We find that its R2-sortability properties are similar to those of the iSCM. For static graphs, we demonstrate both theoretically and computationally that a node's R2-scores on average reflect its connectivity, but that they are a poor indicator for the topological order. Thus, we assert that while a tendency for varsortability is a nonphysical artifact of standard data generation processes, R2-scores do reflect aspects of the causal structure in any system that can be represented by an SCM. To the extent that R2-scores are related to the topological order, *reverse* R2-sortability is more likely than R2-sortability, because the fact that SCMs are not deterministic limits how well one can predict a child, but multiple children can serve as independent observations of a parent, producing arbitrarily good reconstructions of the parent in the limit of many children.

Despite having similar sortability properties, when deployed for evaluation of causal discovery algorithms, our approach results in mild improvements for constraint-based over gradient-based algorithms and a dramatic drop in the performance of the permutation-based algorithm BOSS when compared to evaluations using the iSCM. This provides evidence that artifacts other than var- and R2-sortability, such as restrictions on the ratios of functional parameters, can affect the absolute and relative performances of causal discovery algorithms even when they do not explicitly exploit these artifacts, and highlights the need for intentional SCM sampling methods like that proposed here, rather than attempting to remove one artifact at a time from standard sampling methods.

Finally, we take a first step toward intentional sampling of SVAR models. Surprisingly, we find that our method produces time series datasets that are slightly R2-sortable on average. Our method accounts for additional noise due to subsampling, but neglects the covariance of the noise terms in multivariate systems. Though we bound the magnitude of hidden confounding in terms of the sampling rate, there is a need to rigorously address the issue of pervasive hidden-confounding that is a necessary consequence of subsampling continuous systems.

Acknowledgments

RH, JW, UN and JR received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). JW received support from the German Federal Ministry of Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007).

References

Bryan Andrews and Erich Kummerfeld. Better simulations for validating causal discovery with the dag-adaptation of the onion method, 2024.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of dags using the best order score search and grow shrink trees. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63945–63956. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c9cde817d04811ba28e44071bd9f76a5-Paper-Conference.pdf.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL https://www.science.org/doi/abs/10.1126/science.286.5439.509.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad P Kording, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. *arXiv preprint arXiv:2412.01953*, 2024.

Gustau Camps-Valls, Andreas Gerhardus, Urmi Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2023.10.005. URL https://www.sciencedirect.com/science/article/pii/S0370157323003411. Discovering causal relations and equations from data.

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 09 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz049. URL https://doi.org/10.1093/biomet/asz049.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery, 2023. URL https://arxiv.org/abs/2310.01753.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Paul Erdös and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. The causal chambers: Real physical systems as a testbed for ai methodology, 2024.

Stanton A. Glantz, Bryan K. Slinker, and Torsten B. Neilands. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill Education, New York, NY, 3 edition, 2017. URL accessbiomedicalscience.mhmedical.com/content.aspx?aid=1141896746.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Martin Roth, and Steffen Sonntag. causalAssembly: Generating realistic production data for benchmarking causal discovery. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/gobler24a.html.

Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. *Cause effect pairs in machine learning*. Springer, 2019.

Radoslav Harman and Vladimír Lacko. On decompositional algorithms for uniform sampling from n-spheres and n-balls. *Journal of Multivariate Analysis*, 101(10):2297–2304, 2010. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2010.06.002. URL https://www.sciencedirect.com/science/article/pii/S0047259X10001211.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Christoph Käding and Jakob Runge. Distinguishing cause and effect in bivariate structural causal models: A systematic investigation. *Journal of Machine Learning Research*, 24(278):1–144, 2023. URL http://jmlr.org/papers/v24/22-0151.html.

Markus Kalisch and Peter Buehlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm, 2005. URL https://arxiv.org/abs/math/0510436.

Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data, 2021. URL https://arxiv.org/abs/2104.08043.

Christopher Lohse and Jonas Wahl. Sortability of time series data. *arXiv preprint arXiv:2407.13313*, 2024.

Helmut Lütkepohl. *New introduction to multiple time series analysis*, chapter 9.1.1 The A-Model. Springer Science & Business Media, 2005a.

Helmut Lütkepohl. *New introduction to multiple time series analysis*, chapter 2.1.1 Stable VAR(p) Processes. Springer Science & Business Media, 2005b.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

Arnold Neumaier and Tapio Schneider. Multivariate autoregressive and ornstein-uhlenbeck processes: estimates for order, parameters, spectral information, and confidence regions. *ACM Transactions in Mathematical Software*, 1998.

Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 71–105. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/ng24a.html.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, 2022. doi: https://doi.org/10.1002/widm.1449. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1449.

Weronika Ormaniec, Scott Sussex, Lars Lorch, Bernhard Schölkopf, and Andreas Krause. Standardizing structural causal models, 2024. URL https://arxiv.org/abs/2406.11601.

Judea Pearl. *A Theory of Inferred Causation*, page 41–64. Cambridge University Press, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 11 2013. ISSN 0006-3444. doi: 10.1093/biomet/ast043. URL https://doi.org/10.1093/biomet/ast043.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27772–27784. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e987eff4a7c7b7e580d659feb6f60c1a-Paper.pdf.

Alexander Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. A scale-invariant sorting criterion to find a causal order in additive noise models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 785–807. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/027e86facfe7c1ea52ca1fca7bc1402b-Paper-Conference.pdf.

Nicolas-Domenic Reiter, Andreas Gerhardus, Jonas Wahl, and Jakob Runge. Causal inference on process graphs part i. *arXiv preprint arXiv:2305.11561*, 2023.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

Jakob Runge, Jobst Heitzig, Norbert Marwan, and Jürgen Kurths. Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Phys. Rev. E*, 86:061121, Dec 2012. doi: 10.1103/PhysRevE.86.061121. URL https://link.aps.org/doi/10.1103/PhysRevE.86.061121.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz Marí, and Gustau Camps-Valls. The causality for climate competition. In Hugo Jair Escalante and Raia Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 110–120. PMLR, 08–14 Dec 2020. URL https://proceedings.mlr.press/v123/runge20a.html.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

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

Chandler Squires, Annie Yun, Eshaan Nichani, Raj Agrawal, and Caroline Uhler. Causal structure discovery between clusters of nodes induced by latent factors. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 669–687. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/squires22a.html.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022. doi: 10.1017/eds.2022.11.

Jacobo Torán. On the hardness of graph isomorphism. *SIAM Journal on Computing*, 33(5):1093–1108, 2004. doi: 10.1137/S009753970241096X. URL https://doi.org/10.1137/S009753970241096X.

Ngoc Mai Tran, Johannes Buck, and Claudia Klüppelberg. Estimating a directed tree for extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad165, 02 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad165. URL https://doi.org/10.1093/jrsssb/qkad165.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7:1–12, 2006.

Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In Hugo Jair Escalante and Raia

Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36. PMLR, 08–14 Dec 2020. URL https://proceedings.mlr.press/v123/weichwald20a.html.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery, 2021.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Appendix A. The Scaling Product

Definition 5 $\boldsymbol{\Theta}$ *is a* **parameterization** *of some stochastic functional space* $\mathcal{F}$ *over variables* V *if any stochastic structural assignment* $X_i := f(V)$ *with* $f \in \mathcal{F}$ *can be notated by a vector of parameters* $\boldsymbol{\Theta}_i$.

The space of linear additive Gaussian noise functions defined by Equation 1 can be parameterized by a vector consisting of causal coefficients followed by the noise mean and standard deviation. For a node X_i in a graph with N nodes, $\boldsymbol{\Theta}_i = (a_{1i}, \ldots, a_{Ni}, m_i, s_i)$. To extract the parameters that are random variables, we have the shorter vector $\tilde{\boldsymbol{\Theta}}_i = (a_{ji} | X_j \in pa(X_i), m_i, s_i)$.

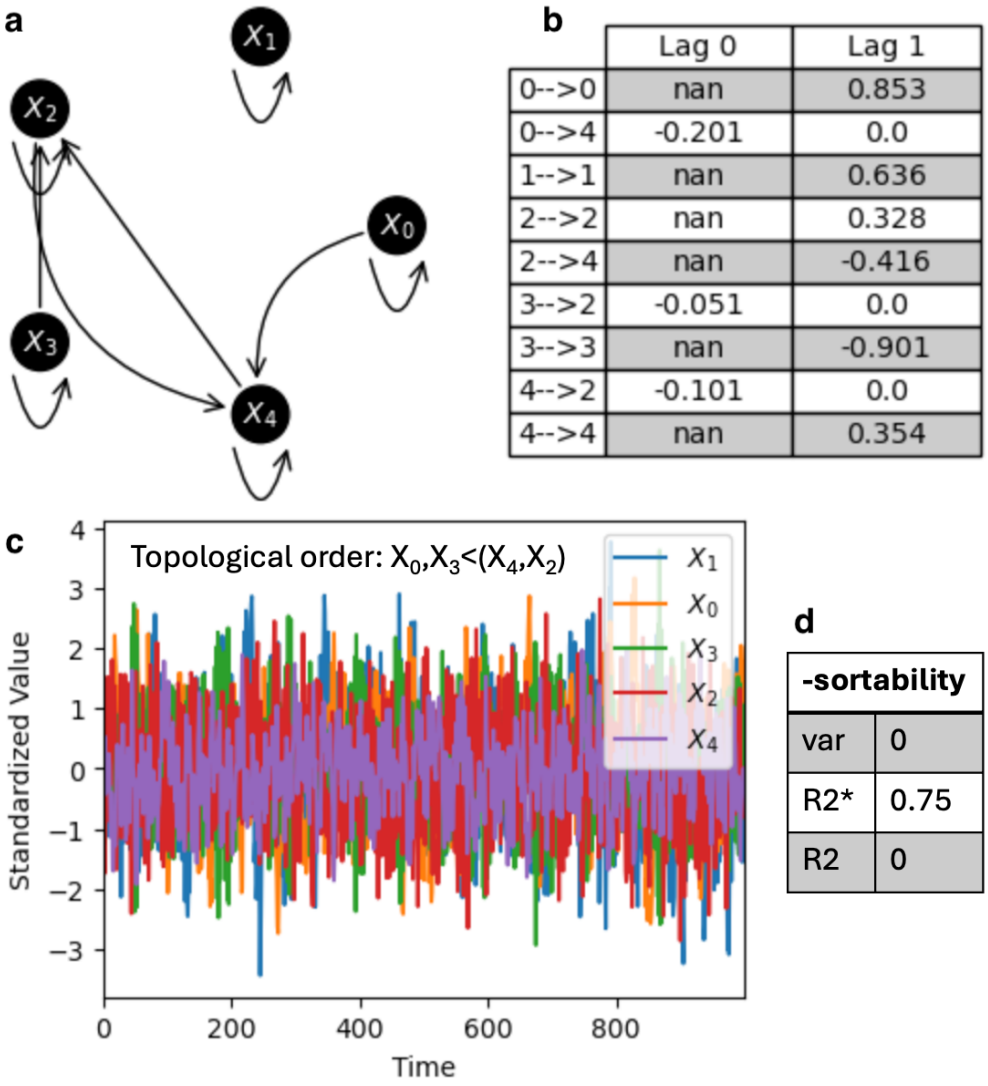
Definition 6 *The* **scaling product** $\odot$ *associated with a parameterization* $\boldsymbol{\Theta}$ *of a functional space* $\mathcal{F}$ *is defined such that* $f_i \in \mathcal{F} \Rightarrow cf_i \in \mathcal{F}$ *can be expressed* $c \odot \boldsymbol{\Theta}_i \; \forall c \in \mathbb{R}$.

The scaling product for the linear additive Gaussian noise functional parameterization discussed above is simply element-wise multiplication because $f_i(V) = \sum_{j=1}^{\#V} a_{ji} X_j + U_i,\ U_i \sim \mathcal{N}(m_i, s_i) \Rightarrow cf_i = c(\sum_{j=1}^{\#V} a_{ji} X_j + U_i) = \sum_{j=1}^{\#V} (ca_{ji}) X_j + U_i',\ U_i' \sim \mathcal{N}(cm_i, cs_i)$ if $\mathcal{N}(m_i, s_i)$ is defined as the normal distribution centered at m_i with standard deviation s_i. However, if $\mathcal{N}(m_i, s_i)$ is defined as the normal distribution with *variance* s_i, then $\odot$ is defined such that $c \odot \boldsymbol{\Theta} = (ca_{1j}, \ldots, ca_{Nj}, cm_i, \sqrt{c}s_i)$. Any parameterization will induce such a scaling product.

Appendix B. Modified Sortability Definitions

Here, we present code for calculating sortability which is modified from Reisach et al. (2021) [https://github.com/CausalDisco/CausalDisco/blob/main/LICENSE] such that it (1) accepts cyclic graphs and avoids comparing nodes within the same cyclic component (lines 13-17, 29, and 38-39), and (2) compares each pair of nodes connected by a causal path exactly once (lines 24, 28, and 35 are added, and `check_now` replaces `Ek` in lines 30-34). Then, in Figure A1, we demonstrate the difference between the time series extensions of R2-sortability proposed by Lohse and Wahl (2024), denoted R2*-sortability, and that proposed here, denote R2-sortability.

```
1  def sortability(M, W, tol=1e-9):
2      '''Takes a 1 x d metric M, such as variance or R2, and an N x N adjacency
3      matrix W, where the j,i-th entry corresponds to the edge weight for j->i,
4      and returns a value indicating how well M reflects the causal order.
5      Modified from <https://github.com/Scriddie/Varsortability> to avoid double-
6      counting node pairs connected by causal paths of multiple lengths, and to
7      accept graphs with cycles in the manner described by Christopher Lohse and
8      Jonas Wahl in "Sortability of Time Series Data" (Submitted to the Causal
9      Inference for Time Series Data Workshop at the 40th Conference on
10     Uncertainty in Artificial Intelligence). '''
11     E = W != 0
12     #Find ancestral relationships to avoid comparison within cycles
13     Ek = E.copy()
14     anc = Ek.copy()*False
15     for path_len in range(self.N):
16         anc = anc | Ek
17         Ek = Ek.dot(E)
18     #reset Ek to keep track of paths of various lengths
19     Ek = E.copy()
20
21     n_paths = 0
22     n_correctly_ordered_paths = 0
23
24     checked_paths = Ek.copy()*False
25
26     for path_len in range(E.shape[0] - 1):
27         check_now = (Ek
28                      & ~ checked_paths # to avoid double counting
29                      & ~ anc.T) #to avoid comparison within a cycle
30         n_paths += check_now.sum()
31         n_correctly_ordered_paths += (check_now * M / M.T > 1 + tol).sum()
32         n_correctly_ordered_paths += 1/2*(
33             (check_now * M / M.T <= 1 + tol) *
34             (check_now * M / M.T >=  1 - tol)).sum()
35         checked_paths = checked_paths | check_now
36         Ek = Ek.dot(E) #examine paths of length path_len+=1
37
38     if n_paths == 0: #in case all nodes are in the same cycle
39         return 0.5
40     return n_correctly_ordered_paths / n_paths
```



	Lag 0	Lag 1
0-->0	nan	0.853
0-->4	-0.201	0.0
1-->1	nan	0.636
2-->2	nan	0.328
2-->4	nan	-0.416
3-->2	-0.051	0.0
3-->3	nan	-0.901
4-->2	-0.101	0.0
4-->4	nan	0.354

-sortability	
var	0
R2*	0.75
R2	0

Figure A1: **An example SVAR model where reverse-varsortability is associated with R2*-sortability, but with reverse-R2-sortability.** **a** The summary graph for the underlying SVAR model with $N = 5$ and $\tau_{\max} = 1$. **b** The coefficients from the underlying SVAR model that are associated with each edge from the summary graph, where different lags appear in different columns. In this table, 0.0 and `nan` appear when there is an edge that exists at one lag but not the other. `nan` is used instead of 0.0 if a 0-lag edge in that orientation would run counter to the topological order of the variables in the SVAR model. **c** Simulated data for this SVAR process, where the time series are ordered by decreasing variance for best visibility. The topological order of the variables according to the summary graph is written at the top of this plot. **d** Time series var-, R2*-, and R2-sortability for this dataset.

Appendix C. Benchmark Sortability Properties

Figure A2 compares the sortability properties of five data generation methods discussed in Sections 2.2.1 and 2.2.3. The standard unit-variance-noise approach (Zheng et al., 2018, first row) produces highly var- and R2-sortable datasets. The IPA approach (Mooij et al., 2020, second row) reduces and reverses the sortability tendencies of the data, while the 50-50 approach (Squires et al., 2022, third row) practically eliminates trends in varsortability but produces strongly reverse-R2-sortable data. The iSCM (Ormaniec et al., 2024, fourth row) eliminates varsortability and produces data that is gently, but noticeably, reverse-R2-sortable. Finally, DaO (Andrews and Kummerfeld, 2024, last row) produces data that is reverse varsortable and whose R2-sortability tendencies change with density. All trends except R2-sortability for DaO become stronger for denser graphs (last column).

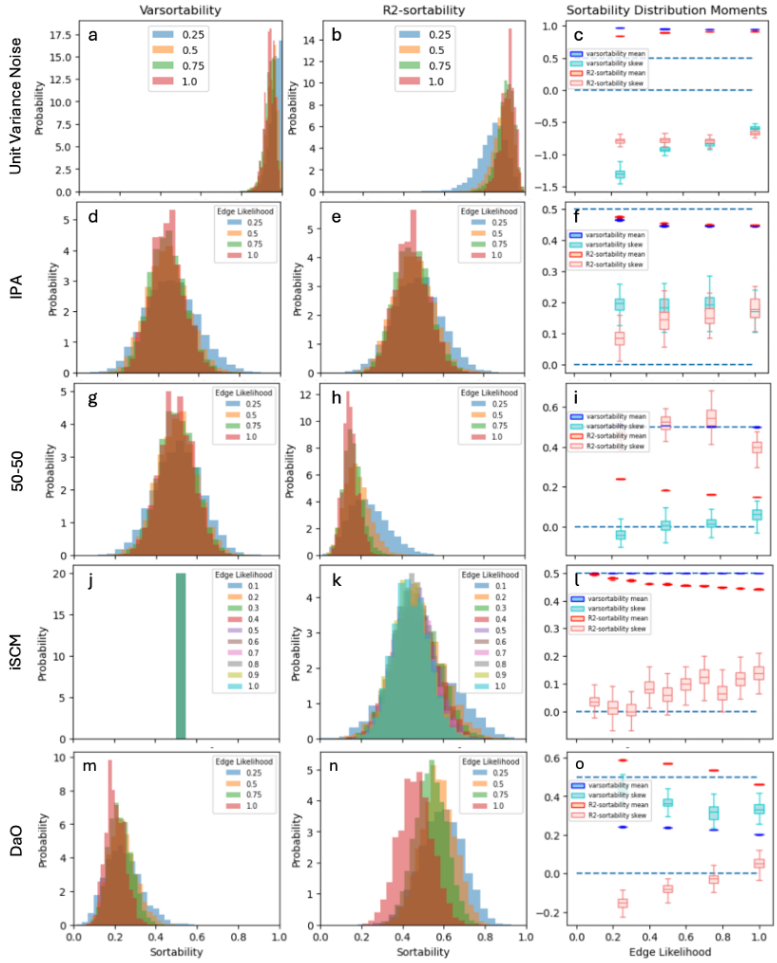


Figure A2: Var- and R2-sortability propbability distribution functions (pdfs, left and middle columns, respectively) for static ER SCMs with $N = 20$ and varying densities (colors) generated via different methods described in Sections 2.2.1 and 2.2.3 (different rows). The means and skews of the var- and R2-sortability pdfs are displayed in the right column as a function of graph density in blue and red, respectively, where the dashed horizontal lines at 0.5 and 0 mark neutral mean and skew, respectively. Pdfs are based on 5000 randomly generated SCMs with 100 observations.

Appendix D. R2 Asymmetry

D.1. R2-score

The R2-score for a variable $\hat{X}_i$ in a standardized dataset $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ with variables $\{\hat{X}_i | i \in 1...N\}$ and M samples is

$$R^2(\hat{X}_i) = 1 - \text{var}(\hat{X}_i - E[\hat{X}_i|\{\hat{X}_j|j \neq i\}])$$

We can estimate $E[\hat{X}_i|\{\hat{X}_j|j \neq i\}]$ using linear regression. The solution to the regression $\mathbf{y} = \mathbf{X}\beta + \epsilon$ is

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

yielding residuals

$$\epsilon = \mathbf{y} - \mathbf{X}\beta = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{I}_M - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$$

that estimate $\hat{X}_i - E[\hat{X}_i|\{\hat{X}_j|j \neq i\}]$.

D.2. SCMs for Unshielded Triples

We examine an SCM with standardized variables $\{\hat{A}, \hat{B}, \hat{C}\}$ and noise terms $\hat{U} = [\hat{U}_A, \hat{U}_B, \hat{U}_C]$, where $\mathbf{U} = [\mathbf{U_A}, \mathbf{U_B}, \mathbf{U_C}] \in \mathbb{R}^{M\times 3}$ is a matrix consisting of M samples of noise terms drawn iid from $\mathcal{N}(0, 1)$, and $[\mathbf{A}, \mathbf{B}, \mathbf{C}] \in \mathbb{R}^{M\times 3}$ is the resulting dataset.

D.2.1. COLLIDER

If $\hat{A} \rightarrow \hat{B} \leftarrow \hat{C}$ is a collider, it can be represented with the SCM:

$$\begin{cases} \hat{A} := \hat{U}_A \\ \hat{C} := \hat{U}_C \\ \hat{B} := a\hat{A} + c\hat{C} + b\hat{U}_B \end{cases}$$

where $a^2 + c^2 + b^2 = 1$. The sample values for an observational dataset can be expressed:

$$\begin{bmatrix}\mathbf{A} & \mathbf{B} & \mathbf{C}\end{bmatrix} := \mathbf{U}\begin{bmatrix}1 & a & 0\\ 0 & b & 0\\ 0 & c & 1\end{bmatrix}$$

and the covariance matrix for the dataset is

$$\begin{bmatrix}1 & a & 0\\ a & 1 & c\\ 0 & c & 1\end{bmatrix}$$

D.2.2. CONFOUNDER

If $\hat{A} \leftarrow \hat{B} \rightarrow \hat{C}$ is a confounded triple, it can be represented with the SCM:

$$\begin{cases} \hat{B} := \hat{U}_B \\ \hat{A} := a\hat{B} + \bar{a}\hat{U}_A \\ \hat{C} := c\hat{B} + \bar{c}\hat{U}_C \end{cases}$$

where $\bar{x} = \sqrt{1 - x^2}$. The sample values for an observational dataset can be expressed:

$$\begin{bmatrix}\mathbf{A} & \mathbf{B} & \mathbf{C}\end{bmatrix} := \mathbf{U}\begin{bmatrix}\bar{a} & 0 & 0\\ a & 1 & c\\ 0 & 0 & \bar{c}\end{bmatrix}$$

and the covariance matrix for the dataset is

$$\begin{bmatrix}1 & a & ac\\ a & 1 & c\\ ac & c & 1\end{bmatrix}$$

D.3. R2 Properties of Hub Nodes in Unshielded Triples

If we regress the middle node $\mathbf{B}$ on $\mathbf{A}$ and $\mathbf{C}$ in the collider, where our regression is consistent with the causal order, the regression will recover the causal model in the infinite sample limit. Thus, we will find $\epsilon \geq b\mathbf{U_B}$, yielding $R^2_B \leq 1 - b^2$ that is upper bounded by the cause-explained variance.

If we perform the same regression on the confounder instead, then substituting $\mathbf{X} = [\mathbf{A}, \mathbf{C}]$ and $\mathbf{y} = \mathbf{B}$ yields

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \begin{bmatrix}\mathbf{A}^T \\ \mathbf{C}^T\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{C}\end{bmatrix} = M\begin{bmatrix}1 & ac \\ ac & 1\end{bmatrix} \\
(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{M}\frac{1}{1-a^2c^2}\begin{bmatrix}1 & -ac \\ -ac & 1\end{bmatrix} \\
\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T &= \begin{bmatrix}\mathbf{A} & \mathbf{C}\end{bmatrix}\frac{1}{M}\frac{1}{1-a^2c^2}\begin{bmatrix}1 & -ac \\ -ac & 1\end{bmatrix}\begin{bmatrix}\mathbf{A}^T \\ \mathbf{C}^T\end{bmatrix} \\
&= \frac{1}{M}\frac{\mathbf{A}\mathbf{A}^T + \mathbf{C}\mathbf{C}^T - ac(\mathbf{A}\mathbf{C}^T + \mathbf{C}\mathbf{A}^T)}{1-a^2c^2} \\
\Rightarrow \epsilon &= \left(\sum_j \left(\delta_{ij} - \frac{1}{M}\frac{a_ia_j + c_ic_j - ac(a_ic_j + c_ia_j)}{(1-a^2c^2)}\right)b_j\right)_i \\
&= \mathbf{B} - \frac{\sigma_{AB}\mathbf{A} + \sigma_{BC}\mathbf{C} - ac(\sigma_{BC}\mathbf{A} + \sigma_{AB}\mathbf{C})}{1-a^2c^2} \\
&= \mathbf{B} - \frac{a\mathbf{A} + c\mathbf{C} - ac(c\mathbf{A} + a\mathbf{C})}{1-a^2c^2} = \mathbf{B} - \frac{a(1-c^2)\mathbf{A} + c(1-a^2)\mathbf{C}}{1-a^2c^2} \\
&= \mathbf{B} - \frac{a\bar{c}^2\mathbf{A} + c\bar{a}^2\mathbf{C}}{1-a^2c^2} = \mathbf{B} - \frac{a\bar{c}^2(a\mathbf{B} + \bar{a}\mathbf{U}_A) + c\bar{a}^2(c\mathbf{B} + \bar{c}\mathbf{U}_C)}{1-a^2c^2} \\
&= \frac{1-a^2c^2-a^2\bar{c}^2-\bar{a}^2c^2}{1-a^2c^2}\mathbf{B} - \frac{a\bar{a}\bar{c}^2}{1-a^2c^2}\mathbf{U}_A - \frac{c\bar{c}\bar{a}^2}{1-a^2c^2}\mathbf{U}_C \\
&= \frac{\bar{a}^2\bar{c}^2\mathbf{B} - a\bar{a}\bar{c}^2\mathbf{U}_A - c\bar{c}\bar{a}^2\mathbf{U}_C}{1-a^2c^2} \\
\Rightarrow \sigma_\epsilon^2 &= \frac{\bar{a}^4\bar{c}^4 + a^2\bar{a}^2\bar{c}^4 + c^2\bar{c}^2\bar{a}^4}{(1-a^2c^2)^2} = \frac{\bar{a}^4\bar{c}^4 + (1-\bar{a}^2)\bar{a}^2\bar{c}^4 + (1-\bar{c}^2)\bar{c}^2\bar{a}^4}{(1-(1-\bar{a}^2)(1-\bar{c}^2))^2} \\
&= \frac{\bar{a}^2\bar{c}^2(\bar{a}^2 + \bar{c}^2 - \bar{a}^2\bar{c}^2)}{(\bar{a}^2 + \bar{c}^2 - \bar{a}^2\bar{c}^2)^2} = \frac{\bar{a}^2\bar{c}^2}{\bar{a}^2 + \bar{c}^2 - \bar{a}^2\bar{c}^2}
\end{aligned}
$$

Without loss of generality, let $\bar{a}^2 = s^2 < 1$ be the larger noise term and $\bar{c}^2 = t^2s^2 < 1$ be the smaller noise term, with $t^2 < 1$. Then

$$\sigma_\epsilon^2 = \frac{s^2(s^2t^2)}{s^2 + s^2t^2 - s^2(s^2t^2)} = \frac{s^2t^2}{1 + t^2(1-s^2)} < s^2t^2 < s^2$$

and $R_B^2 = 1 - \sigma_\epsilon^2 > 1 - s^2t^2$ is lower-bounded by 1 minus the smaller of the two noise terms.

Appendix E. Time Series

Our formulation treats time discretely, as is common to several time series causal inference works, but many relevant systems we may wish to emulate exist in continuous time. Thus, we must understand the relationship between parameters from continuous differential equations and the causal coefficients in our discrete-time systems, and we begin by examining the continuous analog of autoregressive models: the Ornstein-Uhlenbeck process.

E.1. Continuous Analog: Ornstein-Uhlenbeck Processes

A univariate zero-mean Ornstein-Uhlenbeck process $X(t)$ is defined by the *Langevin equation*

$$dX = -\kappa X dt + \sigma dW$$

where W denotes a Wiener process with the property $W(t+\tau) - W(t) \sim \mathcal{N}(0, \tau)$ (Neumaier and Schneider, 1998). In this formulation, $\kappa \in \mathbb{R}^+$ is the rate of reversion to the mean and $\sigma \in \mathbb{R}^+$ is the noisiness of the process, and the variance of this system is $\sigma^2/2\kappa$ (note $\kappa \in \mathbb{R}^+$). As with our discrete formulations, κ is naturally unitless and would not be affected by standardizing X – instead, σ would be set to $\sqrt{2\kappa}$. A discrete sampling with time-step Δ after standardizing is

$$\hat{X}_\Delta(t+1) = e^{-\kappa\Delta}\hat{X}_\Delta(t) + \hat{U}(t),\ \hat{U}(t) \sim \mathcal{N}(0, \sqrt{1 - e^{-2\kappa\Delta}})$$

with causal coefficient $\hat{a}(\Delta) = (e^{-\kappa})^\Delta$ and noise standard deviation $\hat{s}(\Delta) = \sqrt{1 - \hat{a}(\Delta)^2}$. This univariate system has one degree of freedom, parameterized by Δ – the inverse of the sampling rate.

A multivariate Ornstein-Uhlenbeck process has more than one degree of freedom even when it has zero-mean and no interaction terms ($j \neq i \Rightarrow \kappa_{ji} = 0$), but there is still only one sampling rate. Thus, in addition to drawing a parameter related to Δ, we must draw parameters related to κ. As in the static case, it is acceptable if we draw more parameters than degrees of freedom as long as we obey the constraints of the unitless system. The parameters κ_{ii} must be drawn from $\mathbb{R}^+$, but $e^{-\kappa_{ii}} \in (0, 1)$ falls in the unit interval. Δ should also be drawn from $\mathbb{R}^+$, but a scientist employing time series causal methods should attempt to choose a sampling frequency that produces non-trivial auto-coefficients and noise terms. Thus, we choose to draw $e^{-\kappa} \sim \mathcal{U}(0, 1)$ uniformly and $\Delta \sim \mathcal{F}(100, 100)$ from Snedecor's F distribution with $d_1 = d_2 = 100$, which spans $\mathbb{R}^+$ and centers at 1. Since these auto-dependence parameters will not be affected by standardization, the cross-dependence terms and the standard deviation of the noise must adjust to fit the drawn auto-dependence. Given contemporaneous cross-dependencies, we can divide the remaining variance between the cross-dependence terms and noise using a d_i-ball in a manner similar to the static case.

How should we handle lagged cross-dependencies? Examine a bivariate Ornstein-Uhlenbeck process with a one-way causal dependency:

$$d\begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} = -\begin{bmatrix} \kappa_{11} & 0 \\ \kappa_{12} & \kappa_{22} \end{bmatrix}\begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} dt + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} d\begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix}$$

In its standardized form, $\hat{\kappa}_{ii} = \kappa_{ii}$, $\hat{\kappa}_{12} = \frac{\sigma_1}{\sigma_2}\sqrt{\frac{\kappa_{22}}{\kappa_{11}}}\kappa_{12}$, $\hat{\sigma}_1^2 = 2\kappa_{11}$, and $\hat{\sigma}_2^2 = 2\kappa_{22}\left(1 - \frac{2\hat{\kappa}_{12}^2}{\kappa_{22}(\kappa_{11}+\kappa_{22})}\right)$. While the auto-dependence terms $\kappa_{ii} \in \mathbb{R}^+$ must always be positive, the cross-dependence term $\hat{\kappa}_{12} \in \mathbb{R}$ may be negative. A discrete sampling of this standardized process with time-step Δ is

$$\begin{bmatrix} \hat{X}_1(t+1) \\ \hat{X}_2(t+1) \end{bmatrix}_\Delta = \begin{bmatrix} e^{-\kappa_{11}\Delta} & 0 \\ \hat{\kappa}_{12}\frac{e^{-\kappa_{11}\Delta} - e^{-\kappa_{22}\Delta}}{\kappa_{11}-\kappa_{22}} & e^{-\kappa_{22}\Delta} \end{bmatrix}\begin{bmatrix} \hat{X}_1(t) \\ \hat{X}_2(t) \end{bmatrix}_\Delta + \begin{bmatrix} \hat{U}_1(t) \\ \hat{U}_2(t) \end{bmatrix}_\Delta$$

The cross-term is a function of the auto-dependence terms κ_{ii} and Δ: $\hat{a}_{12}(\Delta) = \frac{\hat{\kappa}_{12}}{\kappa_{11}-\kappa_{22}}(\hat{a}_{11}(\Delta) - \hat{a}_{22}(\Delta))$. Since we draw $e^{-\kappa}$ from a uniform distribution and Δ from a distribution centered at 1, rather than drawing κ and Δ from $\mathbb{R}^+$, we will draw $\frac{\hat{\kappa}_{12}}{\kappa_{11}-\kappa_{22}}$ in a manner similar to the static case: starting with an initial draw from $\mathcal{N}(0, 1)$, dividing by the norm from all parents (in this case, canceling out that initial draw), and multiplying by $r_2 \sim \mathcal{U}(0, 1)$ (in this case, there is only one parent). Though we do not consider the auto-dependence to be a parent when making the initial draw for the cross-dependence, we must consider it when scaling to create a normalized time series. Since the auto-covariance depends on the cross-dependence, we must solve a system of equations.

With $v_\Delta(x) = \frac{1-e^{-x\Delta}}{x}$, the variances of the noise terms $\hat{\boldsymbol{U}}_\Delta(t)$ are:

$$\begin{aligned}\hat{s}_1^2 &= \hat{\sigma}_1^2 v_\Delta(2\kappa_{11}) = 2\kappa_{11} v_\Delta(2\kappa_{11}) = 1 - \hat{a}_{11}^2(\Delta) \\ \hat{s}_2^2 &= \big[\hat{\sigma}_1^2\hat{\kappa}_{12}^2\big(v_\Delta(2\kappa_{11}) - 2v_\Delta(\kappa_{11}+\kappa_{22}) + v_\Delta(2\kappa_{22})\big) \\ &\quad + \hat{\sigma}_2^2(\kappa_{11}-\kappa_{22})^2 v_\Delta(2\kappa_{22})\big]/(\kappa_{11}+\kappa_{22})^2\end{aligned}$$

The expression for $\hat{s}_1$ is fairly simple, echoing the expression from the static case. The expression for $\hat{s}_2$ is much more complicated because the total variance of $\hat{X}_2$ used for standardization depends on the system of equations defining the lagged co-variance terms. We will solve for this, rather than calculating it outright. There cross-covariance term is

$$\hat{s}_{12} = \frac{\hat{\sigma}_1^2\hat{\kappa}_{12}}{\kappa_{11}-\kappa_{22}}\left(v_\Delta(2\kappa_{11}) - v_\Delta(\kappa_{11}+\kappa_{22})\right) = \frac{2\kappa_{11}\hat{\kappa}_{12}}{\kappa_{11}-\kappa_{22}}\left(v_\Delta(2\kappa_{11}) - v_\Delta(\kappa_{11}+\kappa_{22})\right)$$

It is non-trivial whenever $\Delta \neq 0$, and positive when $\kappa_{12} < 0$. This implies that discrete sampling of a continuous system will *always* produce hidden confounding, but the magnitude of the confounding can be controlled by adjusting Δ. For simplicity, let's examine the situation when $\kappa_{22} = \kappa_{11} = \kappa$. The magnitude of the confounding $\frac{|\kappa_{12}|}{2\kappa}(1-(2\kappa\Delta+1)e^{-2\kappa\Delta}) < p$ is less than some value $p > 0$ as long as $\Delta < -\frac{1}{2}\left(W_{-1}\left(-\frac{1}{e}\left(1-\frac{2\kappa}{|\kappa_{12}|}p\right)\right)+1\right)$, where W_{-1} is the -1 branch of the Lambert W-Function. Thus, if Δ is chosen to be small enough, hidden confounding could be reasonably controlled. However, one must balance this goal with keeping Δ large enough that each variable has non-negligible noise at each time step.

E.2. Multiple Dependencies

If the discretization of a continuous process is an AR(p) process with $p > 1$, then the continuous process is described by a differential equation of order > 1. The number of auto-dependencies at discrete lags reflects the degree of non-linearity of the autodependence, rather than distinct parents that reduce the expected fraction of variance due to noise, as in the static case. A non-linear effect cannot be represented with a single parameter, but its effect on the mean of the child process can be quantified using the direct transfer function at frequency 0 (Reiter et al., 2023):

$$\hat{\mathfrak{h}}_{ji}(0) = \frac{\sum_{\tau=0}^{\tau_{\max}} \hat{a}_{ji}(\tau)}{1 - \sum_{\tau=0}^{\tau_{\max}} \hat{a}_{ii}(\tau)}$$

This measure includes the indirect effect of X_j on X_i via the past of X_i ($\hat{a}_{ii}(\tau)$), but, as discussed in Sections 6 and E.1, we wish to treat auto-dependence separately. Thus, we represent the contribution of each parent using only the numerator of this expression: $\sum_\tau \hat{a}_{ji}(\tau)$. These oscillatory processes have associated characteristic frequencies, and the coefficients in the discretization may vary widely based on the ratio of the sampling frequency to the characteristic frequency in a process called *aliasing*.

Sampling directly from frequency space is out of the scope of this paper, and since we are doing random generation, it does not matter if we preserve characteristic frequencies between our first draw and our final parameters. Therefore, we will do the initial draws of $a''_{ji}(\tau) \sim \mathcal{N}(0,1)$, but before moving on, we'll also draw contribution parameters b'_{ji} from $\mathcal{U}(0,1)$ if $i = j$ and from $\mathcal{N}(0,1)$ otherwise, and scale $a''_{ji}(\tau)$ by $\frac{b'_{ji}}{\sum_\tau a''_{ji}(\tau)}$. We will use $\{b'_{ji} \forall j \neq i\}$ as the cartesian coordinates in our d-ball. The full algorithm is given below.

Algorithm A1 Unitless SVAR Generation

Require: $\mathcal{G} = \left(\{\hat{X}_i(\tau)\}, E\right)$ a time series graph with topological order $i \in 1...N$ and adjacency matrix $E = (e_{ji}(\tau)) \in \{0,1\}^{N\times N\times(\tau_{\max}+1)}$ with $N \in \mathbb{N}$ and $\tau_{\max} \in \mathbb{Z}^+$

Ensure: $j \geq i \Rightarrow e_{ji}(0) = 0$

$R = (\rho_{ji}(\tau));\ \rho_{ii}(0) \leftarrow 1\ \forall i$

$\hat{\boldsymbol{s}} = (\hat{s}_i) \leftarrow \mathbf{1}_N$

$E_S = (h_{ji}) \leftarrow ||_\tau E$ is the adjacency matrix for the summary graph

$B' = (b'_{ji}) \leftarrow 0_{N,N}$

$\boldsymbol{C} = (C_i) \leftarrow \mathbf{1}_N$

draw $\Delta \sim \mathcal{F}(100, 100)$

for $i \in 1...N$ **do**

- $d_i \leftarrow \sum_{j\neq i} \boldsymbol{h}_i = \#pa_{E_S}(\hat{X}_i) \setminus \hat{X}_i$
- **if** $d_i > 0 || h_{ii} \neq 0$ **then**
  - draw $\boldsymbol{a''_i}(\tau) \sim \mathcal{N}(0,1)^{N\times(\tau_{\max}+1)}$
  - $a''_{ji}(\tau) \leftarrow a''_{ji}(\tau) e_{ji}(\tau)\ \forall j, \tau$
- **end**
- **if** $h_{ii} \neq 0$ **then**
  - draw $b'_{ii} \sim \mathcal{U}(0,1)$
  - $b_{ii} \leftarrow (b'_{ii})^\Delta$
  - $\hat{a}_{ii}(\tau) \leftarrow a'_{ii}(\tau) \leftarrow a''_{ii}(\tau) \frac{b_{ii}}{\sum_\tau a''_{ii}(\tau)}$
  - $s'_i \leftarrow \sqrt{1 - b_{ii}^2}$
- **end**
- **if** $d_i > 0$ **then**
  - draw $b''_{j\neq i,i} \sim \mathcal{N}(0,1)$
  - draw $r_i \sim \mathcal{U}^{1/d_i}(0,1)$
  - $b'_{j\neq i,i} \leftarrow \frac{r_i s'_i}{\sqrt{\sum_{j\neq i}(b'_{ji})^2}} b''_{ji}$
  - $s'_i \leftarrow s'_i \sqrt{1 - r_i^2}$
  - $b_{j\neq i,i} \leftarrow b'_{j\neq i,i} \frac{b_{jj}^\Delta - b_{ii}^\Delta}{b_{jj} - b_{ii}}$
  - $a'_{j\neq i,i}(\tau) \leftarrow a''_{j\neq i,i}(\tau) \frac{b_{j\neq i,i}}{\sum_\tau a''_{j\neq i,i}(\tau)}$
  - $C_i \leftarrow C_i;\ \hat{s}_i \leftarrow s'_i / C_i$
- **end**

end

$\hat{a}_{ji}(\tau) \leftarrow a'_{ji}(\tau)/C_i\ \forall i, \tau, j \neq i$

Solve for C_i and $\rho_{ji}(\tau)$ in

$$\begin{cases} 1 = \sum_{jk} \sum_{\tau\nu} a'_{ji}(\tau) a'_{ki}(\nu) \rho_{jk}(\tau - \nu) + (s'_i)^2\ \forall i \\ \rho_{ji}(\tau) = \sum_k \sum_\nu a'_{ki}(\nu) \rho_{jk}(\tau - \nu)\ \forall i, j, \tau \end{cases}$$

Check that $\det(\hat{A}(0)^{-1} + \mathbf{I}_N - \sum_{\tau=1}^{\tau_{\max}} \hat{A}(\tau) z^\tau) = 0$ yields roots z that lie in the unit circle.