SANDIA REPORT
SAND2023-05141
Printed June, 2023

Sandia National Laboratories

Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems

J. Jake Nichol, Michael Weylandt, Mark Smith, Laura P. Swiler

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods

ABSTRACT

Causal discovery algorithms construct hypothesized *causal graphs* that depict *causal dependencies* among variables in observational data. While powerful, the accuracy of these algorithms is highly sensitive to the underlying dynamics of the system in ways that have not been fully characterized in the literature. In this report, we benchmark the PCMCI causal discovery algorithm in its application to gridded spatiotemporal systems. Effectively computing grid-level causal graphs on large grids will enable analysis of the causal impacts of transient and mobile spatial phenomena in large systems, such as the Earth's climate. We evaluate the performance of PCMCI with a set of structural causal models, using simulated spatial vector autoregressive processes in one- and two-dimensions. We develop computational and analytical tools for characterizing these processes and their associated causal graphs.

Our findings suggest that direct application of PCMCI is not suitable for the analysis of dynamical spatiotemporal gridded systems, such as climatological data, without significant preprocessing and down-scaling of the data. PCMCI requires unrealistic sample sizes to achieve acceptable performance on even modestly sized problems and suffers from a notable curse of dimensionality. This work suggests that, even under generous structural assumptions, significant additional algorithmic improvements are needed before causal discovery algorithms can be reliably applied to grid-level outputs of earth system models.

ACKNOWLEDGMENTS

We thank members of the CLDERA Grand Challenge LDRD project team for helpful discussions and comments on an early draft of this manuscript. JJN also acknowledges support from his Ph. D. advisors, Dr. Matthew Fricke and Dr. Melanie Moses of the Department of Computer Science at the University of New Mexico.

JJN and MS developed the 1D model and performed and analyzed relevant simulations. JJN and MW developed the 2D model, characterized its VAR dynamics, and performed and analyzed relevant simulations. JJN and MW wrote and edited the manuscript. LPS supervised all research and edited the manuscript.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Abbreviation	Definition
ANM	Additive Noise Model
CI	Conditional Independence
DAG	Directed Acyclic Graph
DOE	Department of Energy
ENSO	El Niño Southern Oscillation
FCI	Fast Causal Inference [algorithm]
LiNGAM	Linear Non-Gaussian Acyclic Model
LPCMCI	Latent-PCMCI
MCC	Matthews Correlation Coefficient
NDD	Neighborhood Dependency Density
NDM	Neighborhood Dynamics Matrix
OCE	Optimal Causal Entropy
PC	Peter-Clark [algorithm]
PCA	Principle Component Analysis
PCMCI	PC -Momentary Conditional Independence [algorithm]
PDAG	Partially Directed Acyclic Graph
PDE	Partial Differential Equation
SCM	Structural Causal Model
SEM	Structural Equation Model
VAR	Vector Autoregressive [model]

1. INTRODUCTION

Automated causal structure discovery is an exciting frontier of data-driven science and domain-informed machine learning, but techniques for causal discovery are still rather untested in complex domains. As part of a larger investigation of causal discovery and attribution in climate systems, we investigate the performance of a state-of-the-art algorithm for causal discovery from climate data. The algorithm returns a *causal graphical model* of the given variables. Causal graphical models are usually directed acyclic graphs (DAGs) that relate the causal dependence (graph edges) between variables (graph nodes). Due to the scientific, computational, and statistical difficulties of characterizing climate systems, we instead draw upon well-established techniques for the *benchmarking* of machine learning algorithms for the evaluation of causal discovery. Our results highlight the limitations of modern causal discovery approaches and demonstrate the unreliable performance of these algorithms, even in the most amenable scenarios.

To create the benchmark test cases and perform the various studies we show in this report, we rely on the ideas of benchmarking. According to Olson et al. [1], "the term benchmarking is used in machine learning to refer to the evaluation and comparison of ML methods regarding their ability to learn patterns in 'benchmark' datasets that have been applied as 'standards'. Benchmarking could be thought of simply as a sanity check to confirm that a new method successfully runs as expected and can reliably find simple patterns that existing methods are known to identify." There are many benchmark datasets available: readers may be familiar with the ImageNet database which is commonly used for image classification test problems [2]. Recently, there has been a growth in scientific machine learning benchmarks as well, see Thiyagalingam et al. [3, 4]. The benchmarking approach typically involves a few main steps: identification of training datasets which provide the benchmark data or "gold standard" data, identification of the algorithm or method being tested and associated algorithm choices that might be examined (*e.g.* number of layers in a neural network, activation function used, optimization algorithm to determine hyperparameters, etc.), and a set of performance metrics with which to evaluate the algorithm. Depending on the extent and focus of the benchmark exercise, the ML algorithm can be run with many algorithm choices and the "best" choices can be identified, according to the performance metrics which typically involve "goodness of fit" with respect to predicting the benchmark data but which also may include time to train, time to make a prediction or inference, amount of computing power needed, etc.

We note that causal discovery does not necessarily fall into the machine learning category: it involves aspects of statistical modeling and network inference. However, we feel the benchmark terminology as defined above represents the goal of our efforts well. We also have leveraged verification and validation concepts from the computational science community which focuses on PDE solutions for physical systems, with the goal of improving the credibility of computational models and assessing their predictive capability [5–8]. There are some aspects of verification, specifically solution verification, in the work presented in this report. In the subsequent sections, however, we use the benchmarking terminology.

Benchmarking becomes more challenging for structure-learning algorithms (such as causal discovery), because they require a complete ground-truth graph to evaluate correctness, rather than additional obser-

vations as traditional machine learning requires. This typically limits structure-learning benchmarking to high-fidelity simulation output or hypothesized ground-truth, developed from randomized control trials. While there are a number of metrics that measure the performance of a machine learning model (such as cross-validation error, leave-one-out error, *etc.*), they typically only apply to models predicting additional data points from observational probability distributions, rather than intervention distributions[1], because they capture the ability of the model to represent the training and/or testing data. They do not address other questions such as the correct implementation of the algorithm or the properties and performance it exhibits on various classes of problems. For causal modeling and causal discovery algorithms, there has been limited work specifically seeking to address the issue of "is the inferred graph or causal structure that the algorithm produces correct?" though the works of Runge [10], Runge et al. [11] provide limited, but promising, initial results in this space. In this work, we seek to partially address this important lacuna.

In this work, we report the results of an extensive benchmarking exercise for the PCMCI algorithm of Runge et al. [11]. We specifically focus on the performance of this algorithm as applied to data with spatial and temporal dependence. Our results rely upon a simulation framework inspired by statistical models for time series and by the spatial dynamics of cellular automata. While limited benchmarking of PCMCI has previously been performed, ours is distinguished by a thorough analysis of the effect of spatial structure on performance.

1.1. Background and Related Work

The philosophical and statistical aspects of causal inference and causal discovery are subtle but powerful and our discussion here is necessarily informal. For a further discussion of these issues, we refer the reader to the books by Peters et al. [9] and by Pearl and Mackenzie [12], as well as the many references therein.

1.1.1. Structural Causal Modelling

Causal network discovery, or causal structure learning, is the process of estimating a causal graph[2] of an underlying *structural causal model* (SCM) from observational data[3] and subject matter expertise[4]. An SCM is a semi-mechanistic model, which augments a classical statistical model with a notion of causal structure.[5] While exact estimation of the SCM is typically impossible, it is often possible to accurately estimate the causal network associated with that SCM. A causal network is a DAG representation of the SCM, where variables represent different aspects of the data and directed edges connect "cause" to "effect."

[1] Intervention distributions are what causal graphs predict. We omit discussion of that topic and refer the reader to Peters et al. [9, p. 120-121]

[2] Also known as a *causal network*.

[3] Observational data is characterized as non-experimental data; it contains no planned interventions or controls.

[4] Subject matter expertise is represented by critical causal assumptions, which causal discovery algorithms leverage to reason about the statistical properties found in observational data.

[5] Classical probabilistic statistical models do not naturally incorporate causal structure, instead representing data as a simultaneous draw from an underlying probability distribution. Any temporal object, such as the sample path of Brownian motion, is a draw of a single time-indexed object from an underlying space, rather than a system obeying causal laws.

Many SCMs imply the same causal network, but, under reasonable assumptions,[6] there is a unique DAG for any SCM. When considering SCMs of temporal data, there exist multiple ways of depicting the causal network; see the works by Eichler [13] and Peters et al. [9, p. 198] for details.

As a simple example, consider the following SCM:

$$\begin{aligned} W_t &:= 0.9W_{t-1} + \eta_t^W \\ X_t &:= 0.8X_{t-1} + 0.4W_{t-1} + 0.2Z_{t-3} + \eta_t^X \\ Y_t &:= 0.5Y_{t-1} + 0.2X_{t-2} + \eta_t^Y \\ Z_t &:= 0.6Z_{t-1} + 0.3Y_{t-1} + \eta_t^Z \end{aligned} \tag{1.1}$$

where each $\eta \sim \mathcal{N}(0,1)$ is IID Gaussian noise. These relations form a SCM for simulated realizations of this process.

Figure 1-1 is a causal graph for the SCM in Equation (1.1). Specifically, it is a *time series graph* [10], which captures the temporal dependencies of each node. Each node is a temporally lagged instantiation of each variable. Notice that each variable is autocorrelated in Equation (1.1), with a link between itself and its past self, over 1 lag. The 2 and 3 lag dependencies in $X \to Y$ and $Z \to X$, respectively, are also depicted passing over their respective lag lengths. Without the lagged representation, time-delayed feedbacks[7] would be illustrated as cycles, which violates an important assumption of causal graphs: acyclicity.

While an SCM maps to a DAG, causal network discovery algorithms often output partially-directed acyclic graphs (PDAGs) [14], in which some edges are undirected. Undirected edges indicate a dependence was identified, but not the direction of dependence. Edges sometimes fail to be oriented because of violated assumptions or too little data, but most causal discovery algorithms can only estimate up to the correct Markov equivalence class of graphs, even when assumptions are met and sampling is sufficient. See Peters et al. [9, p. 102] for more on the Markov equivalence of graphs.

Estimated graphs can be annotated with more information indicating the strength of dependence between nodes, causal effect size, causal susceptibility, *etc.* [15, 16], but in this work, we are only concerned with estimating the topology (edge structure) of the time series causal network.

Algorithms for reconstructing causal networks from data generated by an SCM are discussed in the next section.

1.1.2. Causal Discovery & the PCMCI Algorithm

Many algorithms for causal discovery have been proposed in the previous 30 years, most notably the PC algorithm [17], named for its authors Peter Spirtes and Clark Glymour, the Fast Causal Inference (FCI) [17], and the Linear Non-Gaussian Acyclic Model (LiNGAM) [18]. While these general-purpose

[6] These assumptions include causal faithfulness, the causal Markov condition, and causal sufficiency. Put simply, the faithfulness assumption states that separation of two nodes in the causal network is implied by independence, the causal Markov condition states that separation in the graph implies independence in the data, and causal sufficiency states that we have included all common causes of two or more variables in the analysis. Again, for a more detailed discussion, see the books of Peters et al. [9] and Pearl and Mackenzie [12].

[7] Such as that from $X \to Y \to Z \to X$ over 2, 1, and 3 lags, respectively.

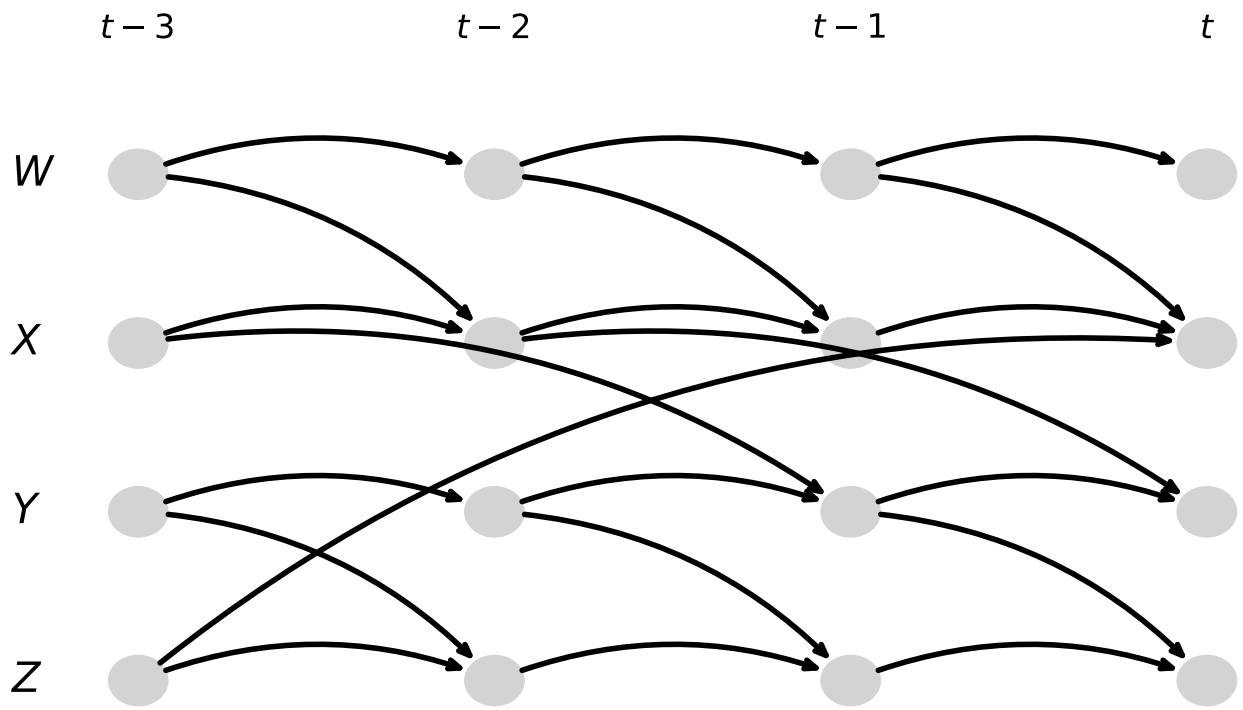


Figure 1-1. A time series graph representation of the SCM in Equation (1.1). By associating each variable with a node for each time lag, it is possible to fully capture relationship between variables and their temporal ancestors.

algorithms are primarly designed for non-temporal data, temporally-aware variants of these algorithms exist [16] as well as novel approaches specific to time series, such as the Optimal Causal Entropy (OCE) algorithm [19, 20]. In this work, we consider the PC-Momentary Conditional Independence (PCMCI) algorithm of Runge et al. [16]. We focus on PCMCI because it was specifically designed to deal with the complex temporal structure of climate data and it has found wide use among the causal climate community [15, 21–25].

PCMCI modifies the classical PC algorithm [17] by adding so-called "Momentary Conditional Independence" tests. These tests take advantage of the temporal structure of the data to greatly reduce the number of potential causal effects, thereby decreasing the space of possible causal networks and improving inferential performance. Like the PC algorithm, the output of PCMCI is a PDAG, however, the time-order of lagged dependencies helps PCMCI orient more edges than it would without temporal information.

The standard variant of PCMCI assumes all causal relationships work on a lag and that there are no contemporaneous dependencies in the data. While we focus on the standard PCMCI algorithm, our simulation study could easily be applied to PCMCI variants, including the Latent-PCMCI of Gerhardus and Runge [26], which allows for unobserved confounders, and PCMCI+ of Runge [27], which allows for contemporaneous dependencies.

Runge et al. [11] detail PCMCI thoroughly and provide an open-source implementation of the approach[8]. PCMCI is a two-phase algorithm: the first phase uses a modified version of the PC algorithm to construct a sparse causal PDAG; this modified algorithm, which they call PC_1, performs a series of iterative conditional

[8] https://jakobrunge.github.io/tigramite/

independence (CI) tests in a search for the causal parents of each variable. PC_1 modifies this search to only condition on the potential confounders with the largest correlations to the variables in question. While this significantly increases computational performance, the full impact of this heuristic modification has not yet been fully characterized.

The second phase of the PCMCI algorithm uses MCI tests to prune this graph in an attempt to eliminate temporally-induced spurious causality. MCI tests extend traditional conditional independence tests by conditioning on lagged (time-shifted) observations of variables. In doing so, they specifically examine whether apparent causal dependencies are artifacts of autocorrelation and prune these spurious graph edges and reduce the false positive rate of PC_1.

As with the original PC algorithm, both the PC_1 and MCI steps of PCMCI can be used with arbitrary conditional independence tests. Test with the conditional Pearson correlation, the *partial correlation*, are easily implemented and widely used, but their performance is only guaranteed for (jointly) Gaussian data. Peters et al. [9] discuss alternative independence tests; see also the discussion by Runge [10].

Finally, we note that while PCMCI is commonly used for climate data, it does not take advantage of the spatial structure typically present in such data. Rather than dealing with spatial structure explicitly, common practice is to summarize data into non-spatial components before applying PCMCI. This summarization is typically done with a statistical technique such as Principal Components Analysis (PCA) or variants thereof or by using external climate knowledge to divide spatial data into pre-defined regions or modes, which are assumed to have no further spatial dependencies [15, 22, 25, 28, 29]. While powerful, these approaches have several drawbacks: PCA-type approaches construct features that are composed of all of the features of the underlying data, so the implied causal relationships are often of an "all-to-all" nature; *a priori* knowledge is useful for well-studied climate phenomena but is difficult to apply to novel studies. In this work, we consider working with unaggregated spatial data observed on a regular grid, such as the output of a large-scale earth system model or geo-referenced observational data. As we will see below, this approach poses novel difficulties in simulation and estimation.

1.2. Contributions

In this paper, we perform an extensive simulation study to benchmark the performance of PCMCI on a set of spatially-inspired SCMs. By using data generated from a known SCM, we are able to accurately quantify the performance of PCMCI on a variety of metrics. In addition to the analysis of PCMCI, our data simulation procedures may be of independent interest. Our findings inform the feasibility of causal discovery from real and simulated climate data and identify several challenges that must be addressed before applying these algorithms at scale.

Section 2 introduces the mathematical framework used to generate spatiotemporal data generation studies, while section 3 describes the specific parameter values used in our simulations. The results of our simulation studies are shown in Section 4, along with a detailed discussion of their implications for causal discovery practice. Finally, Section 5 summarizes our results and discusses potential directions of future research.

2. METHODS

2.1. Spatiotemporal Data Generation Models

Causal dependencies in multivariate data are often expressed as SCMs, *e.g.*, SCM (1.1). If there exists a direct causal dependence from X to Y, which we denote $X \to Y$, then we posit a relationship of the form:

$$Y := f_Y(X) + \eta_Y \qquad (\eta_Y \perp\!\!\!\perp X) \tag{2.1}$$

where f_Y is a (measurable) function relating the cause variable X to the effect variable Y and η_Y is additive noise. If X is random, then we assume X and η are independent ($\eta_Y \perp\!\!\!\perp X$), though this assumption may be relaxed in some circumstances.[1] In the common case where $f_Y(\cdot)$ is a linear function of X, we recover the well studied class of linear structural equation models (SEM) [12]. As Peters et al. [9] discuss, the assumption of additive noise in Equation (2.1) is not essential, but it is standard in the field and we will use it throughout our analysis.

The SCM (2.1) is an additive noise model[2] (ANM) [9, p. 50], a restriction on the class of SCMs that is also useful for identifying variables which do not exhibit a causal effect on Y. Suppose that

$$Y = f(X,Z) + \eta_Y$$

for some function f. It can be shown that Z is not a parent of Y if there exists some function $g(X)$ such that $f(X,Z) = g(X)$ for all (X,Z) or equivalently $Y = g(X) + \eta_Y$.

When modeling temporal data, the ANM (2.1) must be modified to allow for a variable to depend on its previous values. Let $\boldsymbol{X}_t$ be the state of a system of interest at time t; we make two standard assumptions on the behavior of $\boldsymbol{X}_t$:

T1) Lagged dependence: $X_{i,t} \not\to X_{j,t-\tau}$[3] for any (i,j) and any $\tau \geq 0$.

T2) Temporal Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over time.

These assumptions are essentially unavoidable in causal analysis of temporal data: Assumption T1 states that causal dependencies follow the "arrow of time" while Assumption T2 implies that there is a fixed causal structure that we are seeking to estimate. If T2 did not hold, then it is unclear what our target of

[1] We also assume that $Y \notin X$, *i.e.*, that Y does not appear on both sides of equation (2.1): this is essentially equivalent to the common assumption that the causal graph of the system is a DAG.

[2] Following the notation in Peters et al. [9], we will hereafter use assignment (:=) when describing SCM definitions, and equivalence (=) when specifying ANMs and, later, autoregressive models. In this work, the ANMs and autoregressive models are generative models, so they are no less causal.

[3] For our purposes, $\not\to$ indicates no direct dependence between variables.

estimation actually is.[4] Under these assumptions, the ANM for a system with only a single temporal lag[5] becomes:

$$\boldsymbol{X}_t = f(\boldsymbol{X}_{t-1}) + \boldsymbol{\eta} \tag{2.2}$$

where, as before, η is an independent noise variable. In the temporal context, where the effect of the randomly sampled $\eta_{i,t}$ terms persists over time, we will typically refer to the $\eta_{i,t}$ terms as *innovations* rather than *error* or *noise* to emphasize that they are not measurement error, but rather are the fundamental driving element of the system.

In simulation settings, $f(\cdot)$ often represents one step of a (explicit) PDE solver [9]. If $f(\cdot)$ is a linear function, then Equation (2.2) is a Vector Autoregressive (VAR) model [10, 11] and can be written as

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}$$

where $\boldsymbol{A}$ is a fixed matrix encoding the causal dynamics of the system. Specifically, we note that the sparsity pattern of $\boldsymbol{A}$ exactly captures the causal structure of the system:

$$X_{i,t-1} \to X_{j,t} \Leftrightarrow A_{ij} \neq 0$$

As we will observe in the sequel, this property of VARs is particularly useful when simulating from and estimating causal structure in temporal data.

So far, our development has not posited any spatial structure to $\boldsymbol{X}_t$, only the temporal lagged-dependence structure of Equation (2.2). We next introduce two spatial causal assumptions that parallel our temporal assumptions:

S1) Neighborhood dependence: if (i, j) are not neighbors (in a problem specific sense) then $X_i \not\to X_j$.

S2) Spatial Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over space.

Assumption S1 attempts to capture a sense of "locality" and to disallow "action at a distance." When applying this assumption to physical systems, this implies a certain relationship between the temporal and spatial discretizations used: at sufficiently low observation rates, it is possible for a causal effect to exist beyond immediate neighbors.[6] We do not explore the details of that relationship here, but we do note that similar concerns are well-studied in the design of numerical differential equation solvers where spatial and temporal discretizations must be chosen in a suitably consistent manner. Like Assumption T2, Assumption S2 ensures that PCMCI is learning the same causal structure throughout the space. Assumption S2 is not essential in this application and can be easily relaxed. These dynamics are similar to rule-based cellular automata (CA), where the state of each cell is dependent on its immediate neighbors and the update rules are fixed across all cells and time steps.

Under these assumptions, we obtain the single-lag spatiotemporal ANM:

$$X_{i,t} = f\left(X_{i,t-1}, \{X_{j,t-1}\}_{j \in \mathcal{N}(i)}\right) + \eta_{i,t}$$

[4] Assumption T2 can be weakened to only require the *causal* structure of the dynamics, and not the full dynamics, to remain constant over time, but we do not pursue this relaxation.

[5] For higher order lags, we have $\boldsymbol{X}_t = \sum_{\tau=1}^{T} f_\tau(\boldsymbol{X}_{t-\tau}) + \boldsymbol{\eta}$, but we omit higher lags for simplicity of exposition unless noted otherwise. See Peters et al. [9, p. 208] for additional discussion.

[6] For example, consider a simple system in which $X_{i+1,t+1} = X_{i,t} + \eta_{i,t}$ for all (i,t). If i is interpreted as a spatial coordinate in a single dimension, this system satisfies S1. If we reduce our sampling and can only observe $\boldsymbol{Y}_t = \boldsymbol{X}_t, \boldsymbol{Y}_{t+1} = \boldsymbol{X}_{t+2}, \ldots, \boldsymbol{Y}_{t+\tau} = \boldsymbol{X}_{t+2\tau}$, we instead have the causal relationship $Y_{i+2,t} = Y_{i,t}$ which appears to violate S1.

where $\mathcal{N}(i)$ denotes the neighborhood of i, . If f is further assumed to be linear, then we have

$$X_{i,t} = \alpha X_{i,t-1} + \sum_{j \in \mathcal{N}(i)} \beta_j X_{j,t-1} + \eta_{i,t} \tag{2.3}$$

The sparsity of the α and β_j coefficients dictates the causal structure of $\boldsymbol{X}_t$. We will occasionally refer to α as a temporal autocorrelation coefficient and β_j as a cross-dependence coefficient, though they are not numerically equal to the actual autocorrelation function of the process $\boldsymbol{X}_t$.

It is clear that Equation (2.3) can be again expressed as a linear VAR system, with the spatial assumptions S1 and S2 posing additional constraints on the structure of the dynamics (coefficient) matrix. In the next two sections, we characterize these constraints for one- and two-dimensional systems, leaving higher-dimensional systems to the reader.

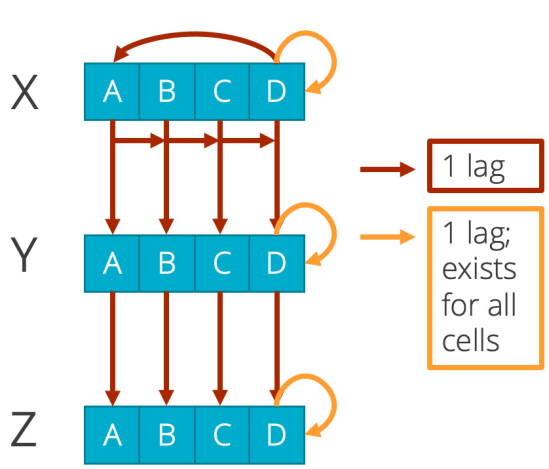
Specifically, we consider two spatial cases to evaluate different kinds of spatiotemporal dynamics. In Section 2.1.1, we consider a multivariate, multi-lagged model supported on a one-(spatial)-dimensional array. In Section 2.1.2, we consider a univariate single-lag model supported on a two-(spatial)-dimensional array. For both models, we assume the underlying space has a toroidal topology, with the leftmost and rightmost elements of the one-dimensional space being neighbors, and similarly for the topmost and bottommost elements in the the two-dimensional case.[7] In one-dimension, the torus is a circle, while the two-dimensional torus is a "donut" shape. We note that this topology differs from that of the surface of a sphere, in that moving far north does not have the same effect as moving far to the west and that there is no analogue of a pole where all cells coincide, but our results can be extended to that setting. Under these two settings, we design an extensive simulation study to characterize the performance of causal discovery algorithms on spatial data.

[7] More informally, we simulate dynamics in a world which "wraps" like the classic arcade game PacMan.

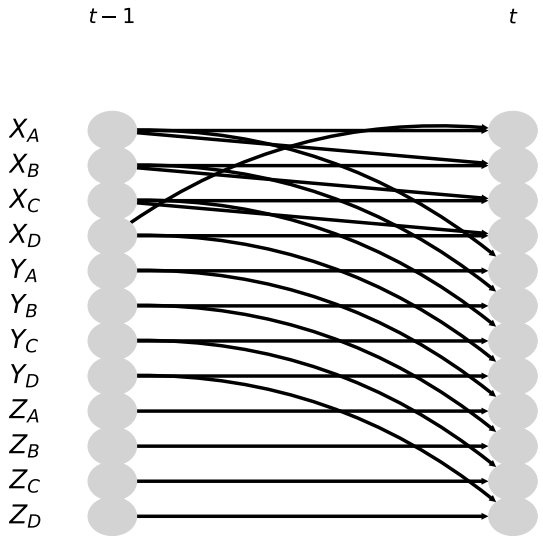
2.1.1. Model Definition: One Spatial Dimension

We first consider simulating causal dynamics on a one-dimensional spatial lattice of size N. Under our assumption S2, we note that each cell can only causally depend on itself and its immediate left and right neighbors, suitably lagged. We further consider a "multivariate" setting in which multiple variables are observed for each cell, and where the causal structure for different variables may not coincide.

We describe the structure of our one-dimensional model in some detail, noting that most of the intuition transfers to the two-dimensional case we consider in the following section. On a lattice of size $N = 4$, we observe three variables, X, Y, Z. Within a single variable, only X exhibits spatial dependencies, such that each cell depends on the neighbor to its left. The causal structure between variables is $X \to Y \to Z$. This sort of model is suitable for simplified modeling of atmospheric aerosol advection and their interaction with radiation and atmospheric temperatures: for some aerosol species, wind can advect aerosols to spatially neighboring regions, while the causal structure $X \to Y \to Z$ reflects the aerosol particles' radiation absorption and subsequent temperature impact, *e.g.*, $H_2SO_4 \to$ radiative flux $\to$ atmospheric temperature. See Figure 2-1a for a spatial illustration of this structure, and Figure 2-1b for a time series graph of the same example. Figure 2-1a is an example of a *summary graph* [9, p. 199].



(a) Spatial representation of the SCM. Auto-correlation orange links exist for all variables, but most are omitted for readability.



(b) The time series graph of the SCM.

Figure 2-1. Causal graphs of variables X, Y, Z at grid cells A, B, C, D, for the SCM defined by Equation (2.4). Here, each variable exhibits temporal autocorrelation at each grid cell (orange arrows), while we observe spatial dependence among X and cross-variable dependence $X \to Y \to Z$. All dependencies occur after a single lag.

If we assume linear dynamics for this system, we obtain the SCM:

$$
\begin{aligned}
X_{A,t} &:= \alpha_{X,A} X_{A,t-1} + \beta_{X,A} X_{D,t-1} + \eta_{X,A,t} \\
X_{B,t} &:= \alpha_{X,B} X_{B,t-1} + \beta_{X,B} X_{A,t-1} + \eta_{X,B,t} \\
X_{C,t} &:= \alpha_{X,C} X_{C,t-1} + \beta_{X,C} X_{B,t-1} + \eta_{X,C,t} \\
X_{D,t} &:= \alpha_{X,D} X_{D,t-1} + \beta_{X,D} X_{C,t-1} + \eta_{X,D,t} \\
\\
Y_{A,t} &:= \alpha_{Y,A} Y_{A,t-1} + \gamma_{X\to Y,A} X_{A,t-1} + \eta_{Y,A,t} \\
Y_{B,t} &:= \alpha_{Y,B} Y_{B,t-1} + \gamma_{X\to Y,B} X_{B,t-1} + \eta_{Y,B,t} \\
Y_{C,t} &:= \alpha_{Y,C} Y_{C,t-1} + \gamma_{X\to Y,C} X_{C,t-1} + \eta_{Y,C,t} \\
Y_{D,t} &:= \alpha_{Y,D} Y_{D,t-1} + \gamma_{X\to Y,D} X_{D,t-1} + \eta_{Y,D,t} \\
\\
Z_{A,t} &:= \alpha_{Z,A} Z_{A,t-1} + \gamma_{Y\to Z,A} Y_{A,t-1} + \eta_{Z,A,t} \\
Z_{B,t} &:= \alpha_{Z,B} Z_{B,t-1} + \gamma_{Y\to Z,B} Y_{B,t-1} + \eta_{Z,B,t} \\
Z_{C,t} &:= \alpha_{Z,C} Z_{C,t-1} + \gamma_{Y\to Z,C} Y_{C,t-1} + \eta_{Z,C,t} \\
Z_{D,t} &:= \alpha_{Z,D} Z_{D,t-1} + \gamma_{Y\to Z,D} Y_{D,t-1} + \eta_{Z,D,t}
\end{aligned}
\tag{2.4}
$$

Because this system is linear, we have an equivalent vector autoregressive (VAR) process representation, $\boldsymbol{\chi} = \boldsymbol{\Gamma} + \boldsymbol{\eta}$:

$$
\begin{bmatrix} X_{A,t} \\ X_{B,t} \\ X_{C,t} \\ X_{D,t} \\ Y_{A,t} \\ Y_{B,t} \\ Y_{C,t} \\ Y_{D,t} \\ Z_{A,t} \\ Z_{B,t} \\ Z_{C,t} \\ Z_{D,t} \end{bmatrix} =
\begin{bmatrix}
\alpha_{X,A} & 0 & 0 & \beta_{X,A} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\beta_{X,B} & \alpha_{X,B} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \beta_{X,C} & \alpha_{X,C} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \beta_{X,D} & \alpha_{X,D} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_{X\to Y,A} & 0 & 0 & 0 & \alpha_{Y,A} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \gamma_{X\to Y,B} & 0 & 0 & 0 & \alpha_{Y,B} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \gamma_{X\to Y,C} & 0 & 0 & 0 & \alpha_{Y,C} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \gamma_{X\to Y,D} & 0 & 0 & 0 & \alpha_{Y,D} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \gamma_{Y\to Z,A} & 0 & 0 & 0 & \alpha_{Z,A} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z,B} & 0 & 0 & 0 & \alpha_{Z,B} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z,C} & 0 & 0 & 0 & \alpha_{Z,C} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z,D} & 0 & 0 & 0 & 0\alpha_{Z,D}
\end{bmatrix}
\begin{bmatrix} X_{A,t-1} \\ X_{B,t-1} \\ X_{C,t-1} \\ X_{D,t-1} \\ Y_{A,t-1} \\ Y_{B,t-1} \\ Y_{C,t-1} \\ Y_{D,t-1} \\ Z_{A,t-1} \\ Z_{B,t-1} \\ Z_{C,t-1} \\ Z_{D,t-1} \end{bmatrix} +
\begin{bmatrix} \eta_{X,A,t-1} \\ \eta_{X,B,t-1} \\ \eta_{X,C,t-1} \\ \eta_{X,D,t-1} \\ \eta_{Y,A,t-1} \\ \eta_{Y,B,t-1} \\ \eta_{Y,C,t-1} \\ \eta_{Y,D,t-1} \\ \eta_{Z,A,t-1} \\ \eta_{Z,B,t-1} \\ \eta_{Z,C,t-1} \\ \eta_{Z,D,t-1} \end{bmatrix}
$$

Here, the α parameters control the temporal autocorrelation of each cell-variable series with itself, the β parameters control the spatial dependence within a variable, and the γ parameters capture cross-variable dependencies. In this scenario, we assume only variable X has spatial dependencies within the same variable, while variables Y and Z exhibit only autocorrelation and the cross-variable structure $X \to Y \to Z$. If we further assume causal stationarity for this model (Assumption S2), these dynamics simplify further to $\tilde{\boldsymbol{\chi}} = \tilde{\boldsymbol{\Gamma}} + \boldsymbol{\eta}$:

$$
\begin{bmatrix} X_{A,t} \\ X_{B,t} \\ X_{C,t} \\ X_{D,t} \\ Y_{A,t} \\ Y_{B,t} \\ Y_{C,t} \\ Y_{D,t} \\ Z_{A,t} \\ Z_{B,t} \\ Z_{C,t} \\ Z_{D,t} \end{bmatrix} =
\begin{bmatrix}
\alpha_X & 0 & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\beta & \alpha_X & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \beta & \alpha_X & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \beta & \alpha_X & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_{X\to Y} & 0 & 0 & 0 & \alpha_Y & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \gamma_{X\to Y} & 0 & 0 & 0 & \alpha_Y & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \gamma_{X\to Y} & 0 & 0 & 0 & \alpha_Y & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \gamma_{X\to Y} & 0 & 0 & 0 & \alpha_Y & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \gamma_{Y\to Z} & 0 & 0 & 0 & \alpha_Z & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z} & 0 & 0 & 0 & \alpha_Z & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z} & 0 & 0 & 0 & \alpha_Z & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{Y\to Z} & 0 & 0 & 0 & \alpha_Z
\end{bmatrix}
\begin{bmatrix} X_{A,t-1} \\ X_{B,t-1} \\ X_{C,t-1} \\ X_{D,t-1} \\ Y_{A,t-1} \\ Y_{B,t-1} \\ Y_{C,t-1} \\ Y_{D,t-1} \\ Z_{A,t-1} \\ Z_{B,t-1} \\ Z_{C,t-1} \\ Z_{D,t-1} \end{bmatrix} +
\begin{bmatrix} \eta_{X,A,t-1} \\ \eta_{X,B,t-1} \\ \eta_{X,C,t-1} \\ \eta_{X,D,t-1} \\ \eta_{Y,A,t-1} \\ \eta_{Y,B,t-1} \\ \eta_{Y,C,t-1} \\ \eta_{Y,D,t-1} \\ \eta_{Z,A,t-1} \\ \eta_{Z,B,t-1} \\ \eta_{Z,C,t-1} \\ \eta_{Z,D,t-1} \end{bmatrix}
$$

That is:

- $\alpha_v = \alpha_{v,\ell}$ for all variables v and spatial locations ℓ;
- $\beta = \beta_{X,\ell}$ for all spatial locations ℓ;
- $\gamma_{v\to w} = \gamma_{v,w,\ell}$ for all variables v, w and all spatial locations ℓ

Further examination of this matrix reveals several sub-blocks with circulant structure, including an α_X, β block, a $\gamma_{X\to Y}$ block, a $\gamma_{Y\to Z}$ block, and α_Y and α_Z blocks: we will return to this observation in the next section.

The specific values of α_v, β, and $\gamma_{v\to w}$ determine whether the resulting stochastic process has spatiotemporal statistical stationarity, which we will call "stability" for brevity. PCMCI assumes the given time series are statistically stationary, so we need to filter the coefficients that constitute a stable process. To do that, we constructed a *companion matrix* [30, p. 259], which is of the general form:

$$\boldsymbol{F} = \begin{bmatrix} \tilde{\boldsymbol{\Gamma}}_{t-1} & \tilde{\boldsymbol{\Gamma}}_{t-2} & \dots & \tilde{\boldsymbol{\Gamma}}_{t-\tau} \\ \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix}$$

for τ lags in the model. The companion matrix is a matrix composed of the $\tilde{\Gamma}$ coefficient matrices (defined above), and the identity matrices and zero matrices that match the size of $\tilde{\Gamma}$. If all eigenvalues of the companion matrix are less than one, then the chosen coefficients will constitute a stable system [30, p. 259]. In Section 3.1, we describe a two-lag system used for experiments, and the companion matrix we used for determining stability is given by:

$$\boldsymbol{F_1} = \begin{bmatrix} \tilde{\boldsymbol{\Gamma}}_{t-1} & \tilde{\boldsymbol{\Gamma}}_{t-2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$$

In Section 3.1 we give specifics of the various model parameters used in our simulations. Because our spatiotemporal model thus reduces to a standard VAR process, for which the PCMCI causal discovery algorithm has previously been found to be effective, we note that our results complement and extend what has previously been shown for the PCMCI algorithm [11].

2.1.2. Model Definition: Two Spatial Dimensions

We next consider simulating causal dynamics on a two-dimensional finite lattice of dimension N. As before, we require that the simulated system has VAR-type dynamics and satisfies assumptions S1-2 and T1-2.

In two spatial dimensions, Assumption S2 implies that each cell has eight neighbors in its so-called "Moore neighborhood"[8], yielding a total of nine potential causal parents (eight neighboring cells and the dependent cell's own previous value). As such, the causal dynamics of the system are dictated by a 3-by-3 matrix, which we term the *neighborhood dependence matrix* (NDM). To simulate dynamics from the NDM, we

[8] In the study of cellular automata, the Moore neighborhood of a cell includes both orthogonal and diagonal neighbors, while the von Neumann neighborhood includes only orthogonal (up, down, left, right) neighbors.

update each element of $\boldsymbol{X}_t$ by taking the inner product of the NDM and the immediate neighborhood of a grid cell: that is,

$$X_{ij,t} = \langle X_{\mathscr{N}(ij),t-1}, \mathbf{NDM} \rangle + \eta_{ij,t} = \mathrm{Trace}(X_{\mathscr{N}(ij),t-1}^{\top} \mathbf{NDM}) + \eta_{ij,t}$$

where $X_{\mathscr{N}(ij)}$ is the submatrix of X consisting of the $(i,j)^{\text{th}}$ element and its immediate neighbors. The NDM defines an invariant "update kernel" which is applied separately to each grid cell in order to simulate its expected value at the next time step. As such, the NDM update dynamics are a sliding dot product[9] of the NDM and the spatial grid, defined by $\boldsymbol{X}_t$:

$$\boldsymbol{X}_t = \mathbf{NDM} \star \boldsymbol{X}_{t-1} + \boldsymbol{\eta}_t \tag{2.5}$$

For two matrices $\mathbf{A} \in \mathbf{R}^{n \times n}$ and $\mathbf{B} \in \mathbf{R}^{N \times N}$, we define their sliding dot product $\mathbf{C} \in \mathbf{R}^{N \times N}$ to be the matrix with $(k,l)^{\text{th}}$ element given by

$$C_{kl} = \sum_{i=-\lceil n/2 \rceil}^{\lceil n/2 \rceil} \sum_{j=-\lceil n/2 \rceil}^{\lceil n/2 \rceil} A(k+i \bmod N, l+j \bmod N) B(2i+1, 2j+j). \tag{2.6}$$

where the mod operator is used to enforce wrapping at the boundaries of our lattice. In our context, the dimension of the sliding dot product kernel $\boldsymbol{A} = \mathbf{NDM}$ is fixed as $n = 3$, reflecting the size of the local neighborhood of each cell; the dimension of the state variable $\mathbf{B} = \boldsymbol{X}_t$ varies with the size of the lattice.

While it is possible to simulate dynamics according to Equation (2.5) for any NDM, the resulting multivariate time series is not statistically stationary without additional assumptions on **NDM**. In order to guarantee stationarity, we seek to represent Equation (2.5) as a (linear) VAR model and apply standard stationarity requirements [30]. In particular, we know that if we have VAR dynamics of the form

$$\boldsymbol{Y}_t = \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\eta}_t$$

the time series $\{\boldsymbol{Y}_t\}$ is stationary if $\|\boldsymbol{A}\|_{\text{op}} < 1$, where $\|\cdot\|_{\text{op}}$ denotes the operator or spectral norm of a matrix, *i.e.*, the magnitude of its largest (possibly complex) eigenvalue. Hence, for a given NDM $\boldsymbol{A}$, it suffices to find a matrix $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{N^2 \times N^2}$ such that

$$\mathrm{vec}(\boldsymbol{X}_t) = \widetilde{\boldsymbol{A}}\mathrm{vec}(\boldsymbol{X}_{t-1}) + \mathrm{vec}(\eta_t) \tag{2.7}$$

Figure 2-2 demonstrates how the NDM, $\boldsymbol{A}$, can be used to form an equivalent VAR coefficient matrix, $\widetilde{\boldsymbol{A}}$. For each grid cell, a suitably padded and shifted version of the NDM is constructed and then multiplied with the previous length N^2 state vector, $\mathrm{vec}(\boldsymbol{X}_{t-1})$. Repeating this process for all N^2 grid cells creates the N^2-by-N^2 coefficient matrix for the VAR representation. We do not seek to fully characterize the algebraic properties of this matrix here, but we do note that it exhibits a block convolutional structure, as shown in Figure 2-3; that is, it has the form of a N-by-N circulant matrix where each element is itself an N-by-N (sub)block matrix. Because the sliding dot product is closely related to a convolution, this circulant block structure is not unexpected.

[9] Denoted by $\star$; also known as a cross-correlation in signal processing, or a flipped convolution à la convolutional neural networks.

With this representation in hand, we are now able to characterize NDMs that give statistically stationary spatiotemporal data (which for brevity we will call "stable NDMs"): a 3-by-3 NDM, $\boldsymbol{A}$ yields stable dynamics if its equivalent N-by-N VAR coefficient matrix $\widetilde{\boldsymbol{A}}$ satisfies $\|\widetilde{\boldsymbol{A}}\|_{\text{op}} < 1$.

In our simulations below, we leverage this characterization as the basis of an Accept-Reject sampling scheme for statistically stationary NDM matrices from the asymmetric Gaussian ensemble. See Algorithm 1. While the efficiency of Algorithm 1 was more than sufficient for this study, more work is needed to efficiently sample stationary NDMs on larger grids. We note that, though natural, this characterization of stationary NDMs does not appear to have been previously considered in the literature and the VAR representation appears to be novel. Previous simulation studies of PCMCI, such as that of Runge [10] and Runge et al. [11], do not sample from the space of stable NDMs and instead explicitly construct a selection of SCMs with small coefficients whose stationarity is then verified empirically through simulation.

Algorithm 1 Sampling Stable Gaussian NDMs: Accept/Reject Algorithm

- **Output:** $\boldsymbol{A}$ sampled from $\boldsymbol{A} \sim \mathcal{N}(\mathbb{R}^{3\times 3})|\boldsymbol{A}$ is stationary

---

- **Repeat:**
  1. Sample $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$ from the 9-dimensional standard Gaussian distribution
  2. Construct $\widetilde{\boldsymbol{A}}$ according to the process of Figure 2-2
  3. **If** $\|\widetilde{\boldsymbol{A}}\|_{\text{op}} < 1$ **return** $\boldsymbol{A}$

In our two-dimensional simulation studies below, we only consider the single-lag single-variable VAR defined by Equations (2.5) and (2.7). Extensions to more complex models are straight-forward. For our model, the multilag extension of Equation 2.5 is given by

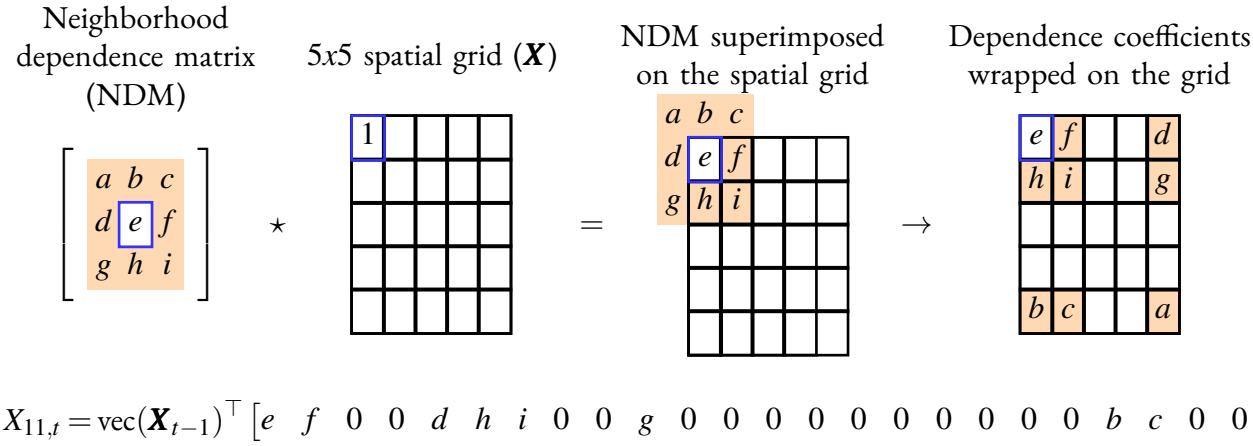
$$\boldsymbol{X_t} = \sum_{\ell=1}^{L} \boldsymbol{A}_\ell \star \boldsymbol{X}_{t-\ell} + \boldsymbol{\eta}_t \tag{2.8}$$

for L lags, while the multilag, the *multivariate* extension of Equation 2.5 is given by

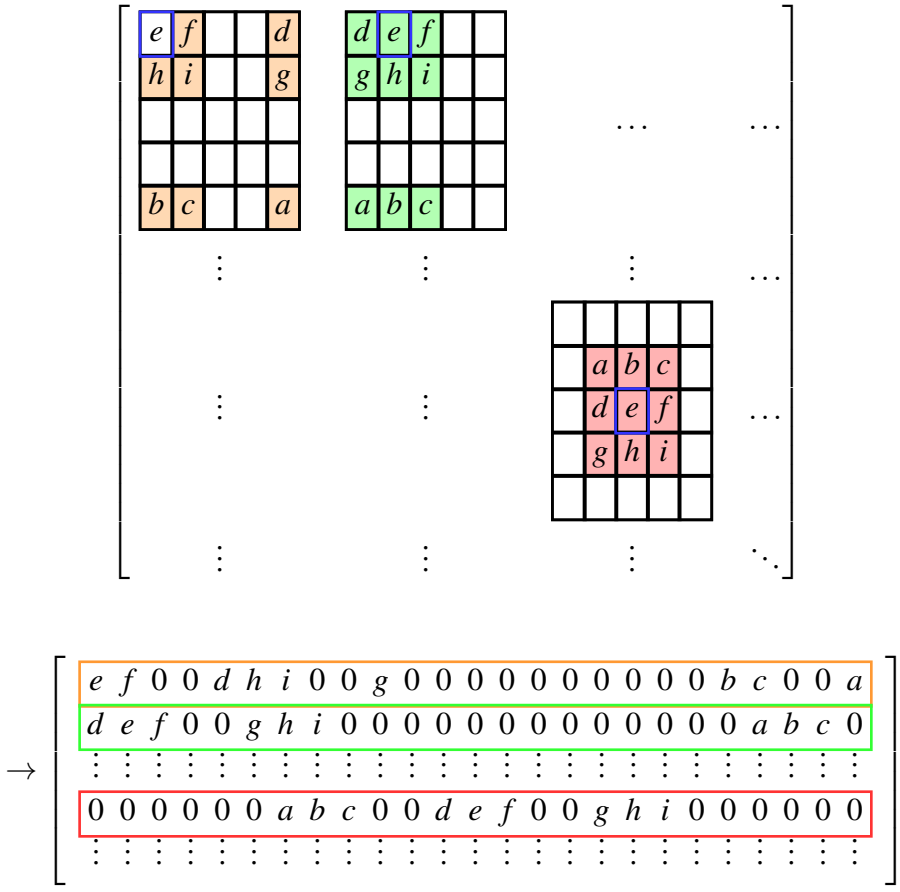
$$\boldsymbol{X}_{\boldsymbol{t}}^{(J)} = \sum_{\ell=1}^{L} \sum_{j=1}^{\mathfrak{J}} \boldsymbol{A}_{\boldsymbol{\ell}}^{(j\to J)} \star \boldsymbol{X}_{t-\ell}^{(j)} + \boldsymbol{\eta}_t \qquad \text{for } J = 1,\ldots,\mathfrak{J} \tag{2.9}$$

for L lags and $\mathfrak{J}$ variables. Here $\boldsymbol{A}_\ell$ denotes the lag-ℓ NDM while $\boldsymbol{A}_\ell^{(j\to J)}$ denotes the multivariate dependence NDM of J on j at lag ℓ.

Finally we note that the single variable VAR(1) here represents the easiest case for causal discovery algorithms. The introduction of more lags, more variables, or non-linear dependencies would only increase the difficulty of causal discovery. As such, the experiments we show below represent an *upper bound* on the performance of PCMCI as applied in more realistic scenarios.



(a) Mapping the action of a neighborhood dependence matrix (NDM) on a single grid cell to a matrix representation. As the NDM is applied to the top left grid cell of the 5×5 spatial grid, the update incorporates all 8 neighbors, which wrap both vertically and horizontally around the edge of our 2D torus. The action of the NDM on a particular grid cell is represented by the top right matrix, which can easily be seen to be equivalent to the vector-matrix product formulation shown below.



(b) Constructing the matrix representation of the NDM action for the entire grid. The process described in Figure is repeated for each grid cell in the 5×5 lattice, which produces a 5×5 matrix, each element of which is a 5×5 matrix reflecting the NDM on a particular cell. Vectorizing these matrices yields the full 25×25-update matrix shown in the final row.

Figure 2-2. Spatial Updates in the Two-Dimensional Model (Section 2.1.2). The 3×3 NDM is expanded to a $N^2 \times N^2$ matrix which fully characterizes the action of the NDM and can be used to analyze the behavior of the resulting system. The sparsity pattern of this matrix is reflected in the time series causal network for this process.

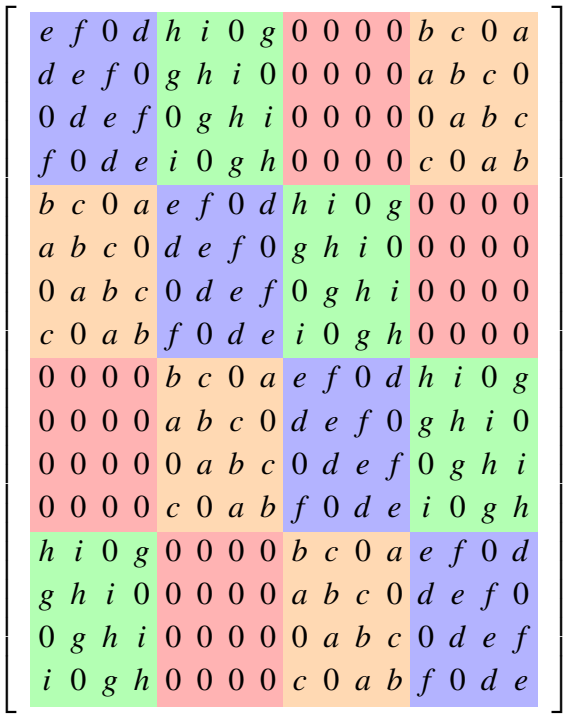


Figure 2-3. Dynamics matrix for the 3×3 NDM $\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ as applied on a 4×4 lattice. Note the "nested circulant" structure of this matrix, where each colored block has a circulant structure, as well as the block circulant structure of the dynamics matrix as a whole.

2.2. PCMCI Algorithm: Tuning Parameters

The PCMCI algorithm has two tuning parameters which must be set by the analyst:

- $\tau_{\max}$, the maximum dependence lag
- α_{PC}, the significance threshold used for each conditional independence test

$\tau_{\max}$ can be chosen based on expert knowledge of the system to determine the maximum hypothetical time for causality to propagate. In general, setting $\tau_{\max}$ too low will significantly distort the estimated causal structure, while setting $\tau_{\max}$ too high will slightly increase the runtime and the false positive rate of PCMCI; as such, users should err on the high side of possible values of $\tau_{\max}$ when the optimal value is unknown.

The PCMCI algorithm uses the α_{PC} parameter for pruning links in the PC Condition Selection phase of the algorithm. During this phase, the (classical) PC Condition Selection algorithm is used for Markov blanket discovery, where it proceeds by running a series of conditional independence tests and removes the link between two variables if the associated test has a p-value less than α_{PC}. As Runge et al. [11] notes, PCMCI does not account for dependencies among the various independence tests or for multiple testing and α_{PC} is better interpreted as a regularization parameter than a statistical significance level, as the false

positive rate of the PCMCI algorithm is not controlled. *Ceteris paribus*, decreasing α_{PC} will result in a sparser estimated causal graph.

Other free parameters include a *minimum* lag $\tau_{\min}$, autocorrelation control parameters p_X and p_Y, and a final threshold level α_G which is applied as a heuristic multiplicity correction. The roles of these parameters are described in more detail by Runge et al. [11] and we do not vary them in our analysis.

3. SIMULATION DESIGN

3.1. Simulation Design: One Spatial Dimension

In order to assess the performance of PCMCI on our one-dimensional model, we fixed a grid size of $N = 10$ and considered five variables observed at each grid cell, V, W, X, Y, Z. Only variables V and Y exhibited spatial dependence: with a left-to-right dependence at one lag and a right-to-left dependence at two lags ($V_{i-1,t-1} \to V_{i,t}$ and $V_{i+2,t-2} \to V_{i,t}$ and similarly for Y). Our simulation design is depicted in Figure 3-1.

Runge et al. [16] note that temporal autocorrelation is typically a severe difficulty for causal discovery algorithms. The PCMCI algorithm was developed specifically to abate these difficulties [16]. To assess the performance of PCMCI, we sampled autocorrelation, which we call coefficient a, from the range $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, with a common autocorrelation used for all variables and grid cells. We consider many of these high degrees of autocorrelation, as autocorrelation is a notable aspect of the climate science questions motivating this study.

We sampled both within-variable spatial and between-variable dependence coefficients, which we call coefficient c, from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, to assess the performance of PCMCI under a range of dependence structures. These dependence coefficients were held constant at all grid cells. Innovations ($\eta_{i,t}$) were sampled from the standard normal distribution. We generated time series with T time samples ranging from $\{50, 150, 250, 350, 475, 575, 675, 775, 900, 1000\}$.

Parameter combinations that failed to exhibit stable dynamics were excluded from our analysis. We ran 30 replicate simulation runs for each stable parameter combination. The number of possible simulation runs is $30,000$, however, because most coefficient combinations were not stable, the number of runs completed was $4,500$. The specific coefficients used are detailed in Section 4.2.

3.2. Simulation Design: Two Spatial Dimensions

In order to characterize the performance of PCMCI in a variety of regimes, we considered the following simulation parameters for our two-dimensional model:

- Number of Time Samples (T): $\{50, 150, 250, 350, 475, 575, 675, 775, 900, 1000\}$
- Grid Size (N): $\{4x4, 5x5, 6x6, 7x7, 8x8, 9x9, 10x10\}$
- Innovation Scale ($\sigma = \mathrm{sd}(\eta_{i,t})$): $\{0.1, 0.5, 1.0, 2.0, 4.0\}$
- Neighborhood Dependence Density (NDD): $\frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{4}{9}, \frac{5}{9}, \frac{6}{9}, \frac{7}{9}, \frac{8}{9}, \frac{9}{9}$

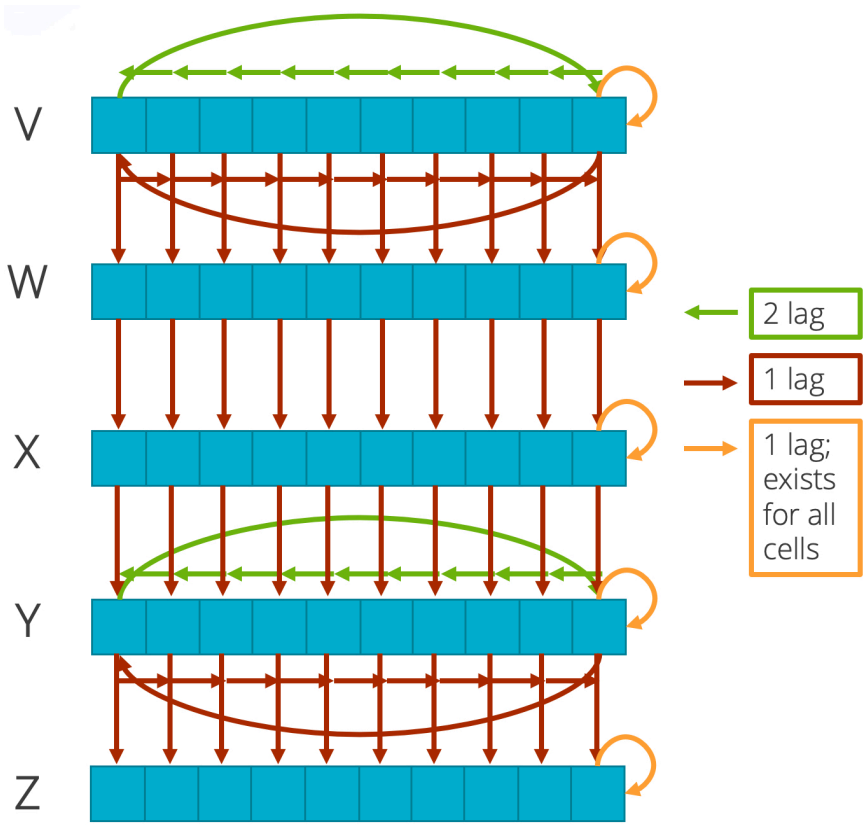


Figure 3-1. A causal graph for the one-dimensional simulation model. The five variables V, W, X, Y, Z are each observed on 10 grid cells. Each variable exhibits temporal autocorrelation (orange), while only V and Y exhibit spatial (left/right) dependencies. Cross-variable dependencies exist at every grid cell according to the causal structure $V \to W \to X \to Y \to Z$. Both the cross-variable and left-to-right dependencies occur at one lag (red), while the right-to-left dependencies occur at two lags (green). This graph has $50 = 5 \times 10$ nodes and 130 edges: the time series causal graph would have $100 = 50 \times$ (max lag = 2) nodes.

Here, σ controls the scale of Gaussian innovations added to each element of $\boldsymbol{X}_t$, and the NDD measures the number of causal parents implied by the NDM. When NDD $= \frac{1}{9}$, there is only one dependence between neighboring grid cells[1]; increasing NDD adds more dependencies; NDD $= 1 = \frac{9}{9}$ implies a fully connected (local) causal system. For each of these 3,150 parameter combinations, we generated 30 random stationary NDMs, yielding a total of 94,500 NDMs, from which we generated 94,500 time series.

In order to simulate these dynamics, statistically stationary NDMs are sampled using Algorithm 1. In order to avoid causal signals that are too small to be detected, we additionally only considered NDMs whose non-zero elements had magnitude at least 0.1. Because the NDMs selected were guaranteed to be stable, we encountered no numerical difficulties in our data generation process.

We generated the innovations $\eta_{i,t}$ from a suitable mean-zero normal distribution and used a Gaussian condition independence test in PCMCI. If a specific distribution for $\eta_{i,t}$ is not assumed, non-parametric independence tests can be used, though these have a higher sample complexity and require longer observational series (greater T).

[1]Sometimes dependence is between a grid cell and itself, such that nodes are autocorrelated and there is no cross-dependence.

4. RESULTS

4.1. Performance Measures

To compare the PCMCI-estimated causal graphs with the underlying SCM-implied causal graphs, we report discovery performance using several measures of classification accuracy; in particular we show the F_1-score and the Matthews Correlation Coefficient. Additional accuracy measures appear in the Appendix to this report.

The F_1 score is a popular measure of classification accuracy, which attempts to balance the precision and recall of a classifier. Specifically, the F_1 score is defined as [31]:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \text{Harmonic Mean}(\text{Precision}, \text{Recall}) \tag{4.1}$$

where precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

and TP, FP, and FN are the counts of true positives, false positives, and false negatives, respectively.[1] F_1 ranges from 0.0 to 1.0, with 0.0 indicating perfect disagreement, that is the estimated graph is the complement of the true graph, and 1.0 indicating exact graph recovery.

We note that the F_1 score is undefined when $TP = 0$, as both Precision and Recall are 0, which would occur if there are no links in the true graph (i.e. all varaibles are independent). We note that the F_1 score can equivalently be expressed as [32]:

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{4.4}$$

As such we, define F_1 to be 1.0 if $FP, FN = 0$ as the estimated graph is correctly fully sparse and 0.0 if $FP > 0$ or $FN > 0$.

We additionally report the Matthews Correlation Coefficient (MCC), also called the ϕ coefficient. Unlike F_1, MCC depends on true negatives and is symmetric in the positive and negative labels: that is, if we

[1] In our context, positives refer to the existence of a link while negatives refer to absence of a causal link. In other contexts, it may be more natural to refer to the absence of a causal link as a scientific finding, as the baseline assumption is that dependencies exist among all measured variables. The MCC measurement we report is invariant to this switch of labels.

compare the *complements* of the true graph and the estimated graph, representing causal independence, we get the same MCC. Chicco [32] defined MCC as follows[2]:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4.5}$$

which ranges from $[-1,1]$. $MCC = -1$ implies the model is perfectly incorrect, $MCC = 0$ indicates a level of accuracy consistent with random guessing, and $MCC = 1$ indicates perfect graph recovery.

As before, we take care to define MCC for the case of sparse graphs (causal independence). MCC is undefined in any of these four cases:

1. if $TP = 0$ **AND** $FP = 0$
2. if $TP = 0$ **AND** $FN = 0$
3. if $TN = 0$ **AND** $FP = 0$
4. if $TN = 0$ **AND** $FN = 0$

We handle these cases separately, assigning values of $\{-1, 0, +1\}$ as appropriate to the causal discovery problem.

1. If $TP = 0$ **AND** $FP = 0$, then the estimated graph is fully sparse:
   a) if $FN = 0$, then the true graph is also fully sparse and we take $MCC = 1$;
   b) if $FN \neq 0$ **AND** $TN = 0$, then the true graph is fully connected, the estimated graph missed all causal relationships, and we take $MCC = -1$;
   c) if $FN \neq 0$ **AND** $TN \neq 0$, then some, but not all, of the causal independence relationships of the estimated graph are false and we take $MCC = 0$.
2. If $TP = 0$ **AND** $FN = 0$, then the true graph is fully sparse:
   a) if $FP = 0$, then the estimated graph is also fully sparse and we take $MCC = 1$;
   b) if $FP \neq 0$ **AND** $TN = 0$, then the estimated graph is fully connected, which is exactly wrong, and we take $MCC = -1$;
   c) if $FP \neq 0$ **AND** $TN \neq 0$, then the estimated graph has implies some spurious causal dependencies and we take $MCC = 0$.
3. If $TN = 0$ **AND** $FP = 0$, then the true graph is fully connected:
   a) if $TP = 0$, then the estimated graph is fully sparse, which is exactly wrong, and we take $MCC = -1$;
   b) if $TP \neq 0$ **AND** $FN = 0$, then estimated graph is fully connected and we take $MCC = 1$;
   c) if $TP \neq 0$ **AND** $FN \neq 0$, then the estimated graph omitted some, but not all, causal relationships and we take $MCC = 0$.

[2] Derived from an earlier definition by Matthews [33].

4. If $TN = 0$ **AND** $FN = 0$, then estimated graph is fully connected:

    a) if $TP = 0$, then the true graph is fully sparse, so the algorithm is perfectly incorrect, and we take $MCC = -1$;

    b) if $TP \neq 0$ **AND** $FP = 0$, then the true graph is also fully connected and we take $MCC = 1$.

    c) if $TP \neq 0$ **AND** $FP \neq 0$, some, but not all, of the estimated causal dependencies are spurious and we take $MCC = 0$.

		c			
		0.1	0.2	0.3	0.4
	0.1	X	X	X	X
	0.2	X	X	X	
	0.3	X	X	X	
a	0.4	X	X		
	0.5	X	X		
	0.6	X			
	0.7	X			

Table 4-1. The stable autocorrelation (a) and cross-correlation (c) dependence coefficients identified for the one-dimensional model.

4.2. One-dimensional model

The results of the simulation study described in Section 3.1 are shown in Figure 4-1. For each simulation, we provided PCMCI with the correct maximum lag ($\tau_{\max} = 2$) and set the threshold parameters to relatively stringent values ($\alpha_{PC} = 0.01$, $\alpha_G = 0.01$). Internal to PCMCI, we used a Gaussian partial correlation test for independence testing, as our data was generated from a linear-Gaussian VAR.

Recall from Section 3.1 that autocorrelated dependence is labeled coefficient a, and within-variable and between-variable cross-correlation dependence is labeled coefficient c. As Figure 4-1 shows, only a minority of a and c dependence coefficients were found to be stable. a was able to reach as high as 0.7, while c was only able to reach as high as 0.4. Table 4-1 shows the specific a and c combinations that were identified as stable in this model. The specific stable coefficient combinations would likely change with a different model formulation, *e.g.*, different spatial dependence structures.

Figure 4-1 shows that PCMCI performed better with more time samples, but performance was limited by the particular a and c coefficients. The algorithm performed better where either coefficient was larger, but particularly when c was larger. For example, when $c = 0.1$, more time samples made little to no difference in performance beyond 250 samples.

In Figure 4-2, we show PCMCI performance as a function of autocorrelation. Figures 4-2a, 4-2b, and 4-2c depict this when $c = 0.1$, $c = 0.2$, and $c = 0.3$, respectively. Again we see that F_1 score increases as the a coefficient increases. Note the differently scaled Y-axes between the panels; the F_1 score reaches higher magnitudes when c is larger. This suggests that within the confines of a stable system, larger autocorrelation increase the signal-to-noise ratio, making the dynamics more easily identifiable. It does not appear that autocorrelation specifically is a detriment to structure identification.

In Figure 4-3, we show PCMCI performance as a function of cross-correlation. Figures 4-3a, 4-3b, and 4-3c depict this when $a = 0.1$, $a = 0.2$, and $a = 0.3$, respectively. We more clearly see that F_1 score increases as the c coefficient increases. Note the differently scaled Y-axes between the panels; performance reaches higher magnitudes when a is larger. Like autocorrelation, larger cross-correlation increases performance, likely because of an improved signal-to-noise ratio. Larger autocorrelation and larger cross-correlation combined results in the best performance.

Finally, in Figure 4-4, we show PCMCI performance as a function of T time samples. Each data point includes all a and c values. We observe a clear pattern that PCMCI performance increases as a function of T, regardless of coefficient values.

4.3. Two-dimensional model

In this section, we present the results of the simulation study described in Sections 3.2. Recall that, for the two-dimensional simulations, we had only a single variable and that the complexity of the problem was controlled by the 3-by-3 neighborhood dynamics matrix, suitably expanded for the larger grid. For each simulation, we provided PCMCI with the correct maximum lag ($\tau_{\max} = 1$) and set the threshold parameters to relatively stringent values ($\alpha_{PC} = 0.01$, $\alpha_G = 0.01$). Internal to PCMCI, we used a Gaussian partial correlation test for independence testing, as our data was generated from a linear-Gaussian VAR.

In Figure 4-5, we examine the effect of grid size (N) on both the F_1 and MCC scores, with other parameters fixed to $\sigma = 1.0$, $NDD = \frac{3}{9}$, and $T = 1000$. While we observe a high degree of variance in this plot, it is clear that performance degrades on larger grid sizes, though at a relatively slow rate if we recall that the problem dimensionality increases *quadratically* in N. As F_1 and MCC are highly correlated, we only depict MCC in subsequent figures. Appendix A features alternate performance measures.

Figure 4-6 depicts the effect of varying the sample length (T). We clearly observe a sub-linear growth in accuracy, as would be expected from the decreasing marginal information of additional samples.[3] Figure 4-7 further depicts the effect of T for various values of grid size, N, and connectivity (NDD). Here we observe that neither grid size nor connectivity have significant impact on PCMCI performance, but that, as expected, there is a small decrease in performance as the grid size increases.

Figure 4-8 highlights the effect of graph density on PCMCI performance. From this plot, it is clear that PCMCI performance is marginally impacted by number of causal relationships increases, and increasing T removes these minimal effects. Comparing results columnwise, we again observe a relatively limited effect of grid size on our results. While Figure 4-8, clearly indicates that PCMCI is able to recover the true graph in the large sample limit, this provides limited guidance for analysts considering the use of causal discovery from data of limited sample size.

In Figure 4-9, we attempt to answer the question "how many samples will I need to expect success"? Because the threshold for "success" is problem dependent, we instead estimate the probability of MCC $> m$ for various values of m. For moderately stringent thresholds ($m \approx 0.7$), we see that $T = 500$ samples appear sufficient for even large grid sizes, while even $T = 1000$ samples may be insufficient at highly stringent thresholds ($m = 0.9$). From these plots it is clear that, while average MCC performance may not vary significantly in grid size, the *dependability* of PCMCI clearly decreases rapidly in N.

Finally, Figure 4-10 investigates the effect of the innovation scale ($\sigma = \mathrm{sd}(\eta_{i,t})$) on PCMCI performance. Empirically, we observe no systematic effect of σ on performance: we hypothesize that this is because σ controls the magnitude of both the additive Gaussian innovations and the signal component $\tilde{\boldsymbol{A}}\mathrm{vec}(\boldsymbol{X}_t)$, leaving the effective signal-to-noise ratio of the problem unchanged. While we do not show this analytically

[3] Via general statistical principles, we expect MSE $\propto T^{-1/2}$, and note that MCC is a non-linear, but monotonic, function of estimation accuracy.

for the causal discovery problem, we do note that a similar phenomenon occurs in the estimation of VAR coefficients.[4]

Additional results, including analysis of the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) appear in Appendix A. Those plots indicate few false positives across different simulation regimes and that decreases in MCC are primarily driven by false negatives, indicating large numbers of samples are necessary to correctly identify causal effects. While varying the PCMCI thresholding parameters α_{PC} and α_G may adjust the balance of false negatives and false positives, we do not explore the effect of those parameters in this work.

[4] Briefly, let $\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}$ for $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2 I)$. Then

$$\begin{aligned} \text{Cov}(\boldsymbol{X}_t) &= \text{Cov}(\boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\eta}) \\ &= \boldsymbol{A}\text{Cov}(\boldsymbol{X}_t)\boldsymbol{A}^T + \sigma^2 \boldsymbol{I} \\ \implies \text{vec}(\text{Cov}(\boldsymbol{X}_t)) &= \sigma^2 (\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1}. \end{aligned}$$

Additionally recalling that the variance of the OLS estimator is given by $\text{Cov}(\text{vec}(\hat{\beta})) = (\text{vec}(\boldsymbol{X})\text{vec}(\boldsymbol{X})^T)^{-1} \otimes \sigma^2 \boldsymbol{I}$, we have $\text{Cov}(\text{vec}(\hat{\beta})) \approx [\sigma^2(\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1}]^{-1} \otimes \sigma^2 \boldsymbol{I} = (\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1} \otimes \boldsymbol{I}$ which does not depend on σ.

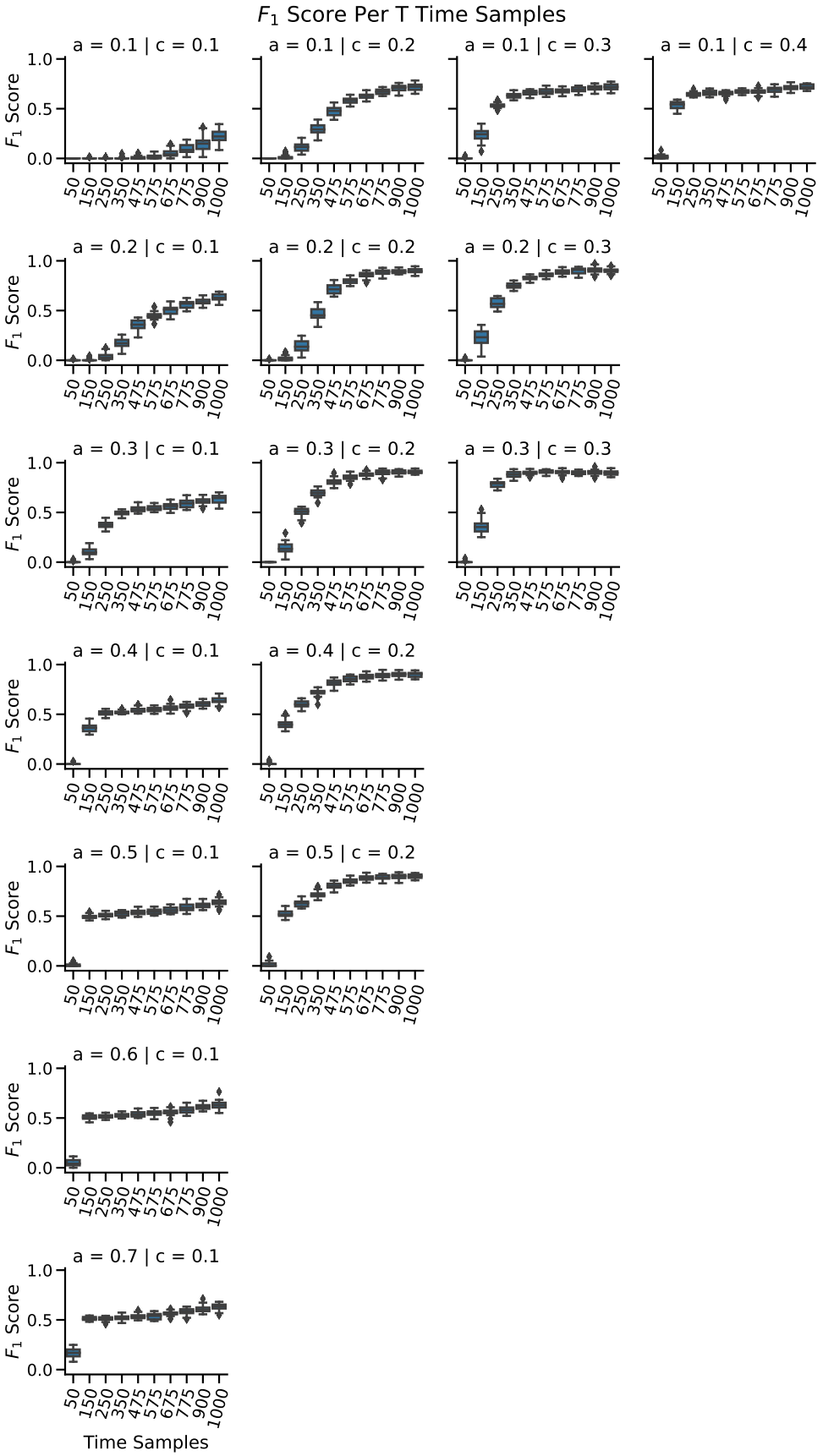
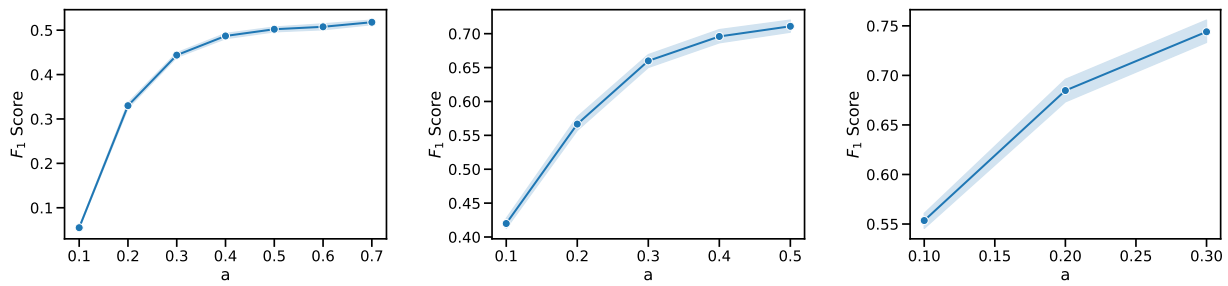
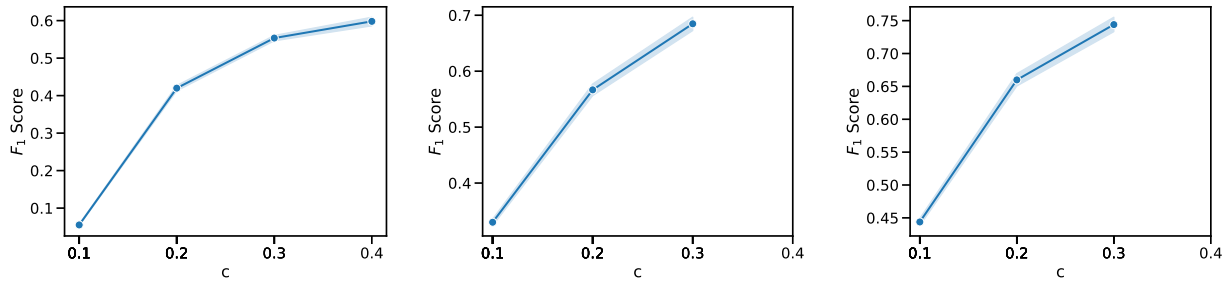


Figure 4-1. F_1 **scores for the One-Dimensional model of Sections 3.1 and 3.1. Coefficients** a **and** c **represent autocorrelation and cross-correlation dependence coefficients, respectively, where cross-correlation relates to both variable-to-variable and cell-to-cell dependencies. Only stable coefficient combinations are shown. PCMCI performs better with more time samples, larger** a **values, and larger** c **values. When** a **or** c **are sufficiently large, the system becomes unstable.**



(a) $c = 0.1$ **(b)** $c = 0.2$ **(c)** $c = 0.3$

Figure 4-2. F_1 **score results from the one-dimensional spatial example with varying autocorrelation,** a**, and constant cross-correlation,** c**. Each data point includes all possible** T **time samples. Note the different Y axes.**



(a) $a = 0.1$ **(b)** $a = 0.2$ **(c)** $a = 0.3$

Figure 4-3. F_1 **score results from the one-dimensional spatial example with varying cross-correlation,** c**, and constant autocorrelation,** a**. Each data point includes all possible** T **time samples. Note the different Y axes.**

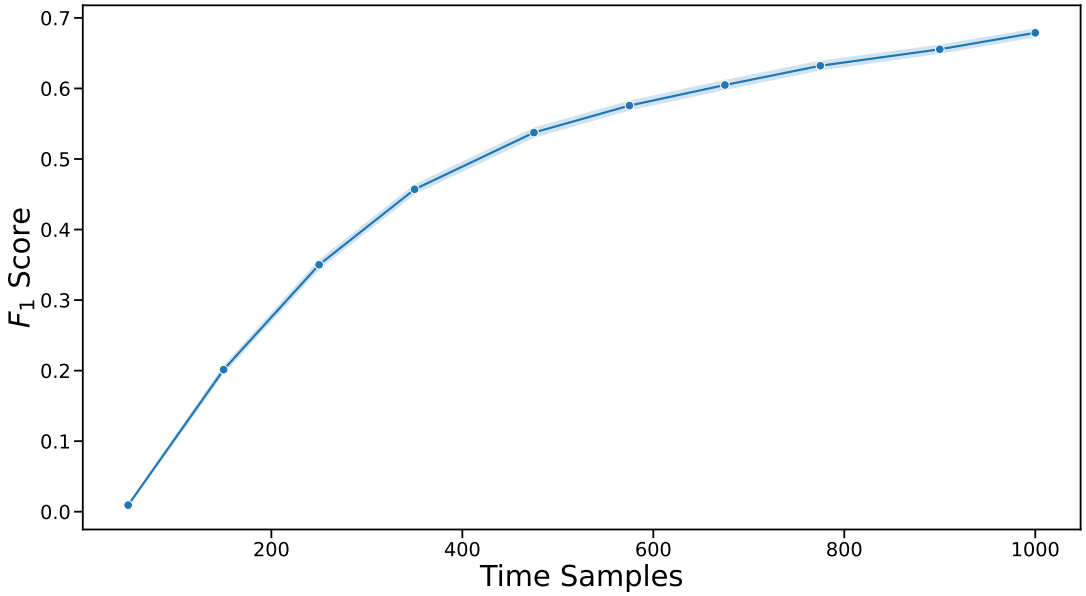


Figure 4-4. F_1 **score results from the one-dimensional spatial example with varying** T **time samples. Each data point includes all possible** a **and** c **dependence coefficients.**

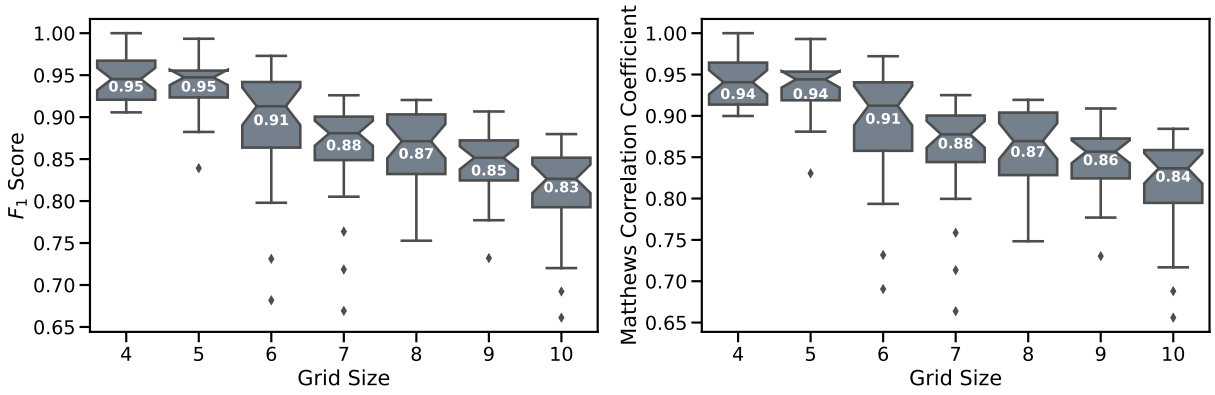


Figure 4-5. Effect of grid size (N) on PCMCI F_1 and MCC scores. Both metrics decrease relatively slowly in N. Other simulation parameters are fixed to $\sigma = 1.0$, $NDD = \frac{3}{9}$, and $T = 1000$.

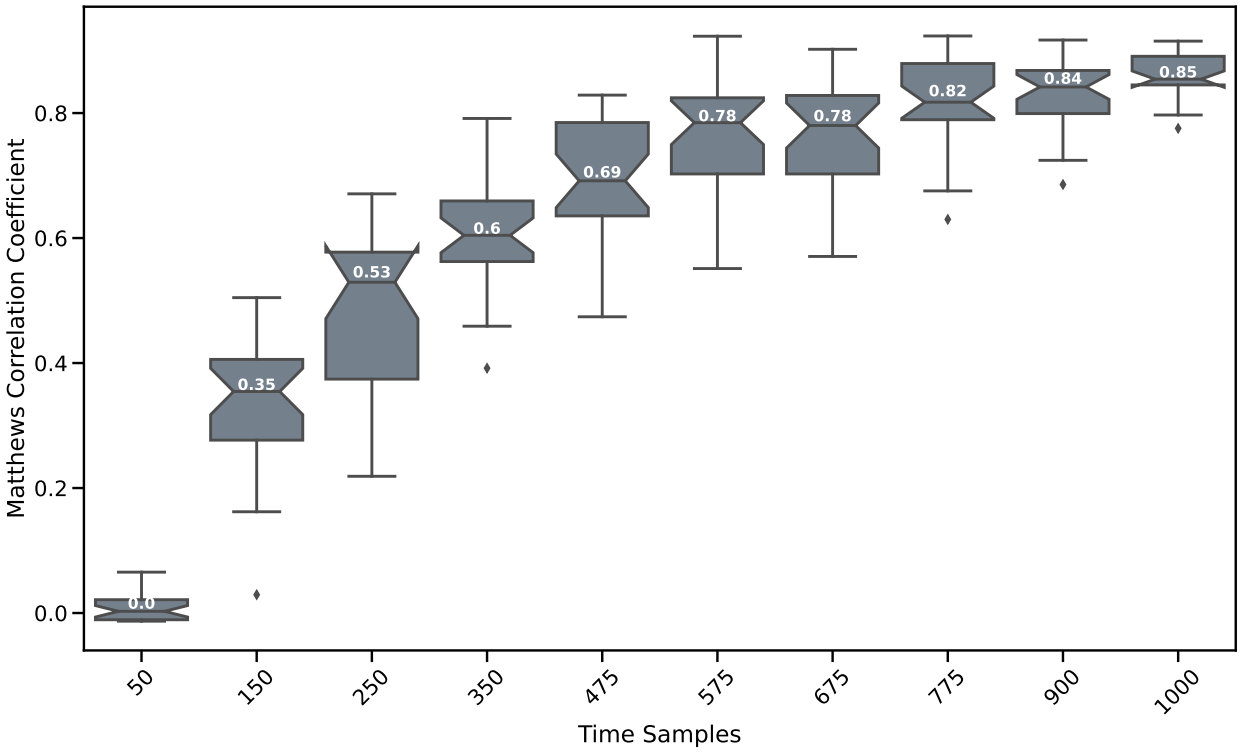


Figure 4-6. Effect of increasing sample size (T) on PCMCI performance (MCC). Performance increases sublinearly in T, with $T > 575$ being necessary to obtain acceptable performance (MCC > 0.7). Box labels report median MCC across replicates. Other simulation parameters fixed as $N = 10$, $\sigma = 1.0$, and NDD $= \frac{6}{9}$.

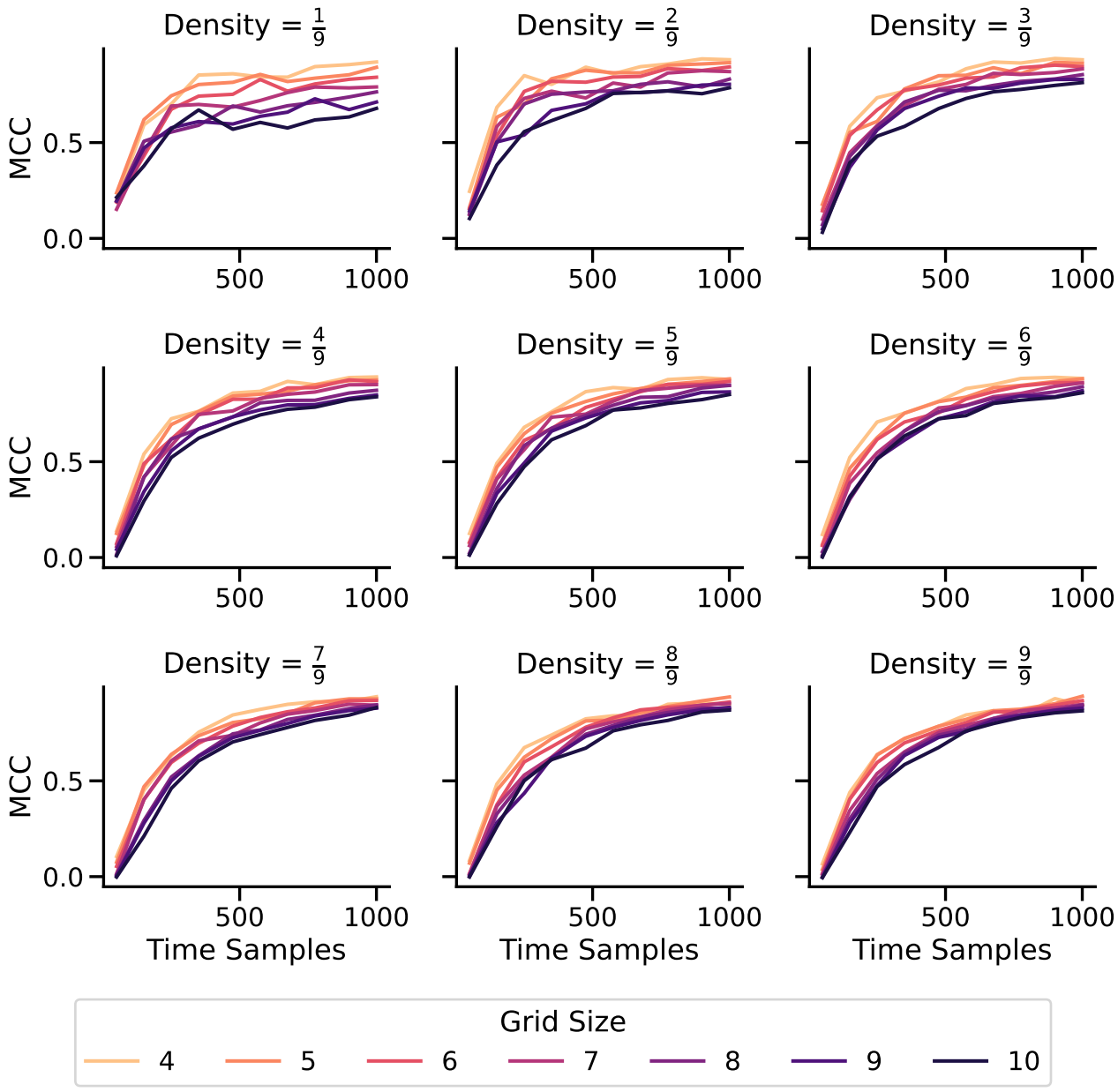


Figure 4-7. Effect of sample size, (T) grid size, (N), and neighborhood dependence density on PCMCI performance (MCC). For sufficiently large sample sizes, PCMCI is able to consistently recover the true graph structure; the effect of grid size and NDD are less pronounced than T. Values shown are mean performance over 30 replicates.

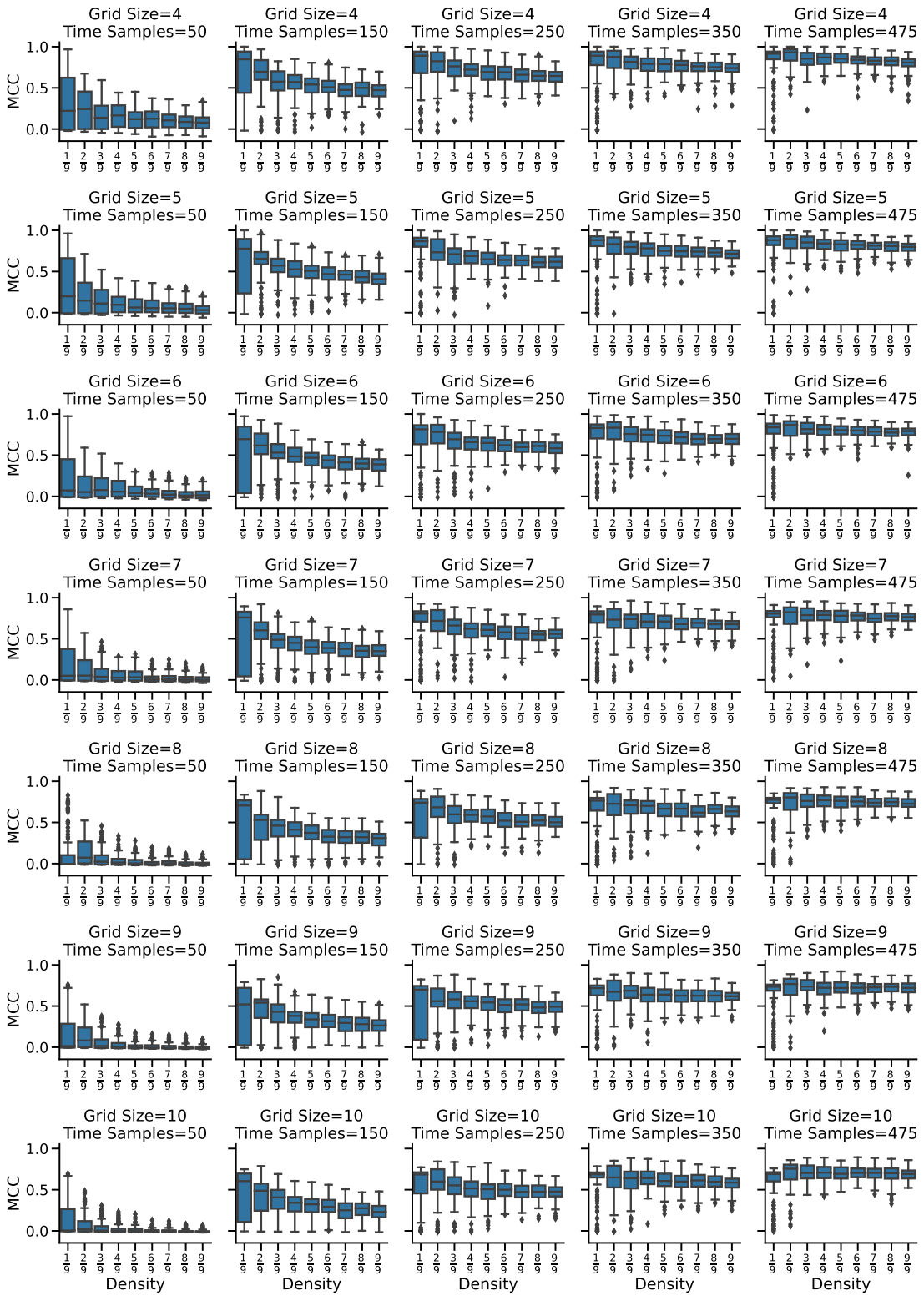


Figure 4-8. Effect of sample size, (T) grid size, (N), and neighborhood dependence density on PCMCI performance (MCC). For sufficiently large sample sizes, PCMCI is able to consistently recover the true graph structure; the effect of grid size and NDD are limited. Values shown are mean performance over 30 replicates. $\sigma = 1$ for all simulations.

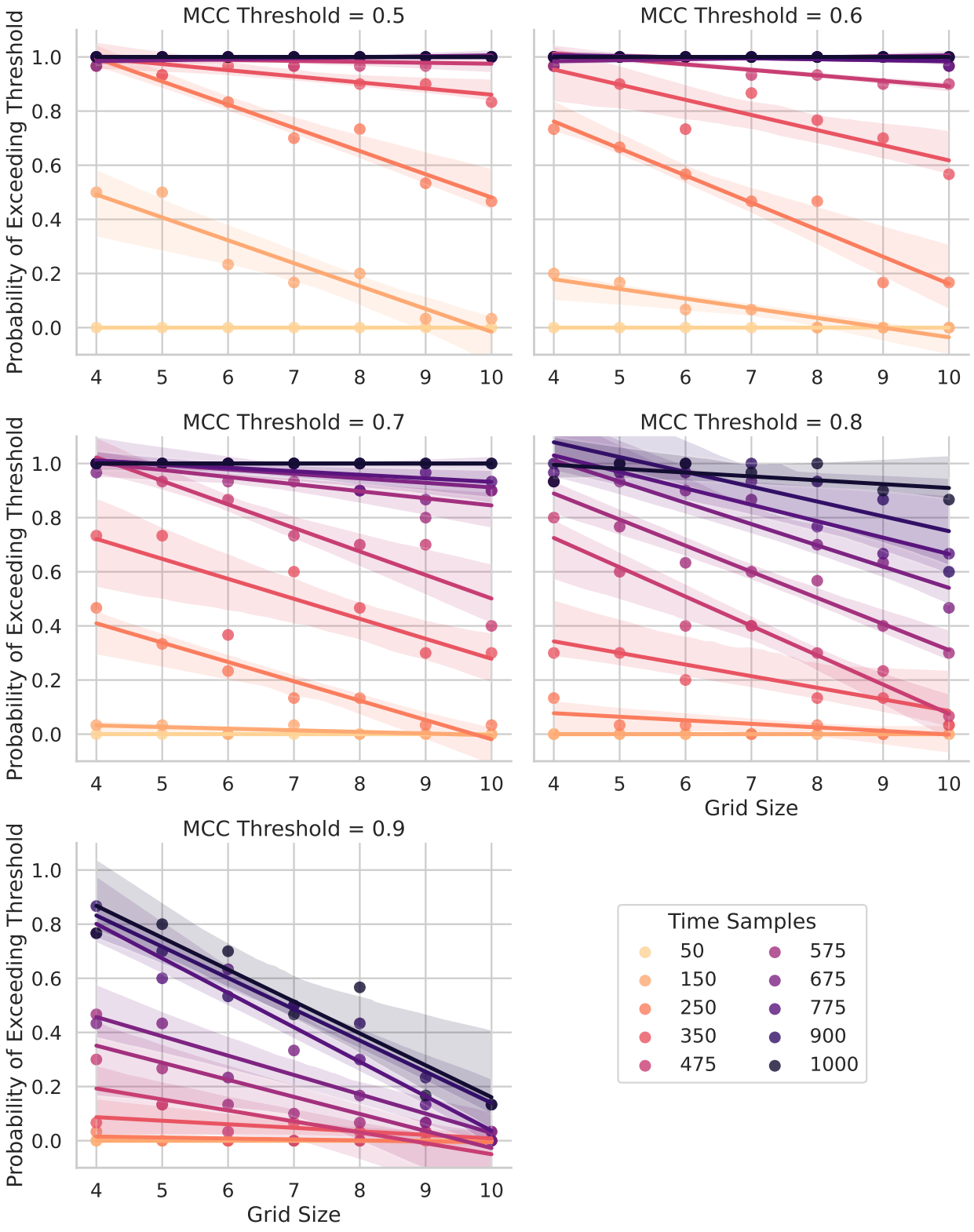


Figure 4-9. Probability of PCMCI Success as a function of grid size N and sample size T, with success defined as MCC above a user-defined threshold. Results are empirical probabilities over 30 replicates: σ and neighborhood density are fixed to 1.0 and $\frac{6}{9}$ respectively. Lines depict a simple linear model of grid size on success probability, with shaded regions depicting (non-multiplicity adjusted) confidence intervals.

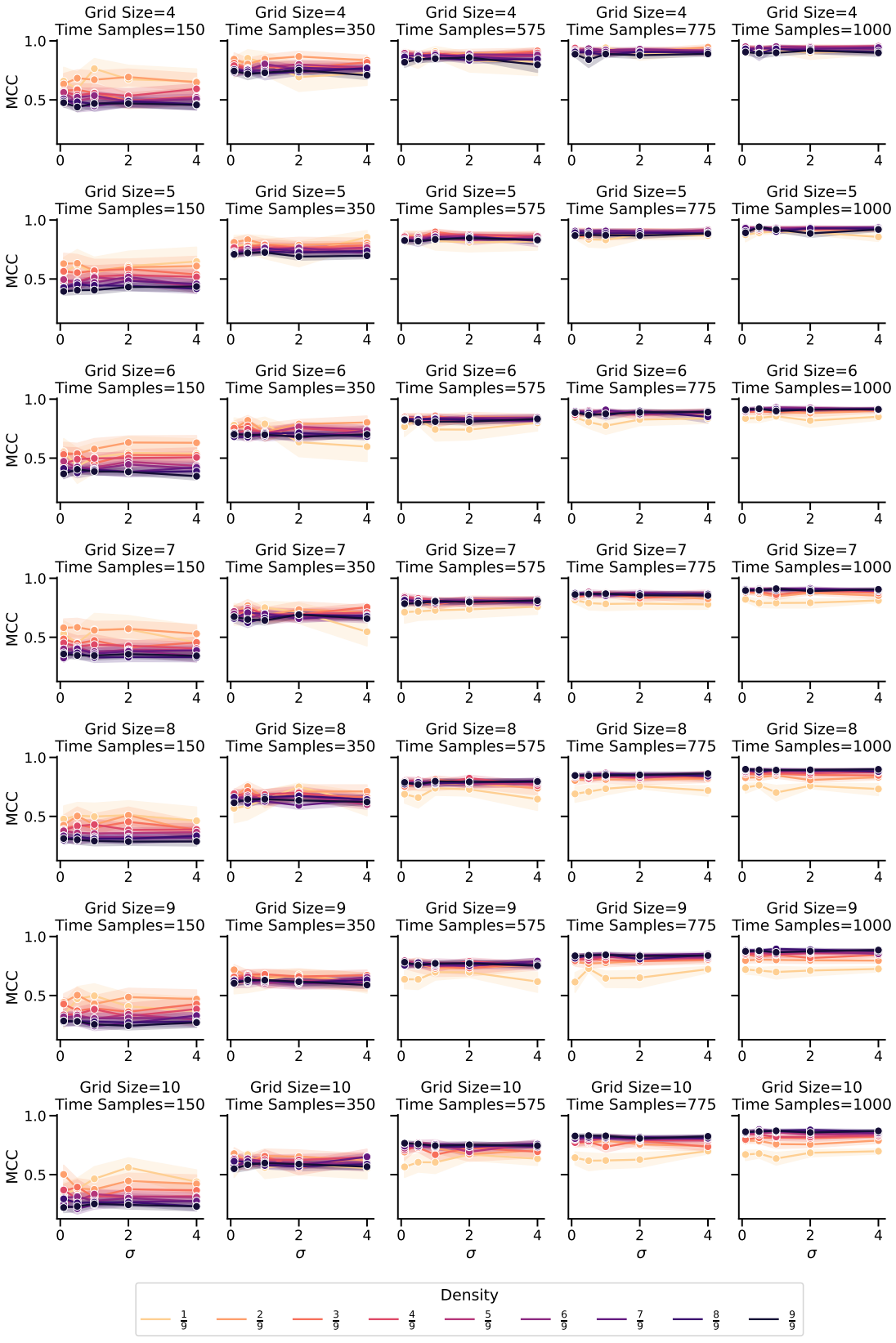


Figure 4-10. Effect of Innovation Magnitude (σ) on PCMCI performance (MCC). Changing σ appears to have no systematic effect on PCMCI performance.

5. DISCUSSION

In this work, we investigated the performance of the PCMCI causal network discovery algorithm for linear-VAR systems in one- and two-dimensional space. We varied the length of the observed time series, the size of the underlying grid, and the density of the underlying causal graph, and found significant effects of each. Our results provide a robust characterization of PCMCI performance on spatiotemporal systems and highlight several avenues of future inquiry.

Most notably, we found that $T \approx 1000$ samples were necessary to for consistent high-accuracy causal discovery across the various scenarios we considered (see Figure 4-4 and Figure 4-9). While this is consistent with the asymptotic consistency of PCMCI, these extreme sample sizes are unrealistic for the climate data analytics motivating this study. Note that we restricted our analysis to linear-Gaussian systems, which enables PCMCI to reduce the difficult problem of testing conditional independence to the relatively easier problem of estimating partial correlations. While it is possible to use PCMCI with more general conditional independence tests, these tests have a far higher sample complexity, and would require a *far greater sample size* to achieve consistent performance.

By contrast, the effect of the grid size was relatively minor, suggesting that performance gains may attainable through clever use of this spatial structure. Changing the number of true causal effects had notable impacts on certain performance measures, but further work is needed to determine whether this scenario is inherently more difficult for causal discovery or whether it is an artifact of the specific accuracy measures we used, *e.g.*, the number of true positives for an empty graph.

We note that in our study of the one-dimensional model, we found that PCMCI tolerated high autocorrelation well. This result is somewhat unexpected, given previous work showing that causal discovery algorithms tend to handle autocorrelation poorly. However, PCMCI was developed to be robust to autocorrelation [10]. The clearest conclusion, apart from the aforementioned benefits of more time samples, was that larger causal dependence coefficients were beneficial, regardless of whether they were autocorrelational or cross-correlational coefficients.

Finally, our study of the two-dimensional model also provided several computational advances that may be of independent interest, including characterization of the sliding dot product and VAR representations of our model, an easy-to-implement check for stability of the resulting VAR process, and an effective algorithm for sampling from the space of stable dynamics.

As shown in Figure 4-9, the probability of "successful" graph recovery is highly sensitive to both the sample size and the grid size. As the number of potential causal parents for a single grid cell increases quadratically in N, this is perhaps unavoidable. More generally, causal discovery algorithms are known to suffer from the curse of dimensionality, particularly when applied on the grid-level in spatiotemporal systems as the both the potential causal parents and the number of grid cells studied increase rapidly in the grid size [10, 15, 22, 25].

In the climate context, the underlying grids are far larger than those considered in this study, necessitating extremely large sample sizes. Unfortunately, our causal stationarity assumptions (Assumption T2 and S2) are less likely to hold over these extended time frames. To avoid this problem, some works have artificially reduced the problem dimensionality by replacing grid cells with pre-defined regions of climatological interest [15, 21, 22, 25]. They made attempts to benchmark their results with either simulated or theoretical expectations. However, their simulations were not of grid-cell-level causal dynamics, as ours are, and their studies on natural climate data could not be benchmarked rigorously. Finally, we note that these approaches are only appropriate for long-term climate analyses in which well-defined spatially-stable statistically-regular modes are the objects of study. We do not expect these approaches to perform well when studying "one-off" climate events, in which relevant regions are rarely known *a priori*, making dimension reduction a far more challenging task.

Finally, we note that our study only considered samples from the stationary distribution of a linear system driven by Gaussian innovations. As a result, our simulated data is itself Gaussian and does not reflect structures that may be found in climate data, *e.g.*, the El Niño Southern Oscillation (ENSO) or, on shorter scales, major storms. It is unclear how PCMCI would perform when applied to these stable structures, as they have complex spatiotemporal dynamics.

Causal discovery is an important aspect of modern climate research and there is a need for algorithms that can scalably and accurately determine causal structure from grid-level data. While PCMCI is quite data-hungry on large grids and observational climate data are quite limited, additional insights can be gleaned from the analysis of large simulation ensembles. Currently, PC-family algorithms do not incorporate spatial structure: in future work, we hope to investigate the use of spatial structure to reduce the dimensionality of the causal discovery problem.

Causal discovery remains a challenging task, particularly in the climate domain. As simulation and observational data continues to grow in size and scope, there is a pressing need for approaches that can perform robustly at a range of time- and spatial-scales, ranging from storm tracking to diffusion of volcanic aerosols to long-term natural and anthropogenic climate changes. The benchmarking techniques and simulations of this paper give insight into the weaknesses of current approaches and suggest new avenues of causal discovery research.

REFERENCES

[1] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10:1–13, 2017.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[3] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scientific machine learning benchmarks. *Nature Reviews Physics*, 4(6):413–420, 2022. doi:10.1038/s42254-022-00441-7.

[4] Jeyan Thiyagalingam, Kuangdai Leng, Samuel Jackson, Juri Papay, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scimlbench: A benchmarking suite for ai for science, 2021. URL https://github.com/stfc-sciml/sciml-bench.

[5] Osman Balci. Verification, Validation, and Certification of Modeling and Simulation Applications. *Proceedings of the 2003 Winter Simulation Conference, 2003*, 1:150–158, 2003. doi:10.1109/wsc.2003.1261418.

[6] William L. Oberkampf and Christopher J. Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, 2010. ISBN 9780511760396. doi:10.1017/cbo9780511760396.016.

[7] National Research Council. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. The National Academies Press, Washington, DC, 2012. ISBN 978-0-309-25634-6. doi:10.17226/13395.

[8] R G Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1):12–24, 2013. ISSN 1747-7778. doi:10.1057/jos.2012.20.

[9] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 9780262037310.

[10] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018. ISSN 1054-1500. doi:10.1063/1.5025050.

[11] 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): 4996—5023, 2019. doi:https://doi.org/10.1126/sciadv.aau4996. URL http://advances.sciencemag.org/.

[12] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

[13] Michael Eichler. In *AISTATS 2010: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 193–200, 2010. URL https://proceedings.mlr.press/v9/eichler10a.html.

[14] Judea Pearl. Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688, 1995. ISSN 0006-3444. doi:10.2307/2337329.

[15] Jakob Runge, Vladimir Petoukhov, Jonathan F. Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6(1):8502, 2015. doi:10.1038/ncomms9502.

[16] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Munoz-Mari, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Scholkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1), 2019. ISSN 20411723. doi:10.1038/s41467-019-10105-3.

[17] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer, 1993. doi:10.1007/978-1-4612-2748-9.

[18] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvarinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL https://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf.

[19] Jie Sun and Erik M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014. ISSN 0167-2789. doi:10.1016/j.physd.2013.07.001.

[20] Jie Sun, Dane Taylor, and Erik M Bollt. Causal Network Inference by Optimal Causation Entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015. doi:10.1137/140956166.

[21] Marlene Kretschmer, Dim Coumou, Jonathan F. Donges, and Jakob Runge. Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation. *Journal of Climate*, 29(11):4069–4081, 2016. ISSN 0894-8755. doi:10.1175/jcli-d-15-0654.1.

[22] Peer Nowack, Jakob Runge, Veronika Eyring, and Joanna D. Haigh. Causal networks for climate model evaluation and constrained projections. *Nature Communications 2020 11:1*, 11(1):1—11, 2020. ISSN 2041-1723. doi:10.1038/s41467-020-15195-y. URL http://www.nature.com/articles/s41467-020-15195-y.

[23] Zachary S. Kaufman, Nicole Feldl, Wilbert Weijer, and Milena Veneziani. Causal Interactions Between Southern Ocean Polynyas and High-Latitude Atmosphere-Ocean Variability. *Journal of Climate*, 33(11):4891–4905, 2020. ISSN 0894-8755. doi:10.1175/jcli-d-19-0525.1.

[24] Christopher Krich, Jakob Runge, Diego G. Miralles, Mirco Migliavacca, Oscar Perez-Priego, Tarek El-Madany, Arnaud Carrara, and Miguel D. Mahecha. Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach. *Biogeosciences*, 17(4):1033–1061, 2020. doi:10.5194/bg-17-1033-2020.

[25] 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, 2022. doi:10.1017/eds.2022.11.

[26] Andreas Gerhardus and Jakob Runge. LPCMCI: Causal Discovery in Time Series with Latent Confounders. In *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc., 2020. doi:10.5194/egusphere-egu21-8259. URL https://proceedings.neurips.cc/paper/2020/file/94e70705efae423efda1088614128d0b-Paper.pdf.

[27] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 1388–1397. PMLR, 2020. URL https://proceedings.mlr.press/v124/runge20a.html.

[28] Yi Deng and Imme Ebert-Uphoff. Weakening of atmospheric information flow in a warming climate in the Community Climate System Model. *Geophysical Research Letters*, 41(1):193–200, 2014. ISSN 1944-8007. doi:10.1002/2013gl058646.

[29] Imme Ebert-Uphoff and Yi Deng. Causal Discovery from Spatio-Temporal Data with Applications to Climate Science. *2014 13th International Conference on Machine Learning and Applications*, pages 606–613, 2014. doi:10.1109/icmla.2014.96.

[30] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994. ISBN 9780691042893.

[31] Nancy Chinchor. MUC-4 Evaluation Metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, page 22–29, USA, 1992. Association for Computational Linguistics. ISBN 1558602739. doi:10.3115/1072064.1072067. URL https://doi.org/10.3115/1072064.1072067.

[32] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, 2017. ISSN 1756-0381. doi:10.1186/s13040-017-0155-3.

[33] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. ISSN 0005-2795. doi:https://doi.org/10.1016/0005-2795(75)90109-9. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

APPENDIX A. Additional Simulation Results: Two-Dimensional Model

In this section, we depict various performance rates of PCMCI in our two-dimensional simulation study (Section 3.2). Here we report:

$$FDR = \frac{FP}{TP+FP} \quad \text{(False Discovery Rate, Figure A-1)}$$

$$TPR = \frac{TP}{TP+FN} = \frac{TP}{P} \quad \text{(True Positive Rate, Figure A-2)}$$

$$FNR = \frac{FN}{TP+FN} = \frac{FN}{P} \quad \text{(False Negative Rate, A-3)}$$

$$TNR = \frac{TN}{TN+FP} = \frac{TN}{N} \quad \text{(True Negative Rate, A-4)}$$

$$FPR = \frac{FP}{TN+FP} = \frac{FP}{N} \quad \text{(False Positive Rate, A-5)}$$

where FDR is the false discovery rate; TP, FP, TN, FN are the number of true positives, false positives, true negatives, and false negatives, respectively; and P, N are the number of edges and non-edges in the true graph.

As with the one-dimensional model, PCMCI exhibits a bias towards non-discovery, with low true and false positive rates across scenarios. The FPR is almost always kept near 0, indicating that we can have a high degree of confidence in the causal effects identified by PCMCI, but that it has limited statistical power at moderate sample sizes.

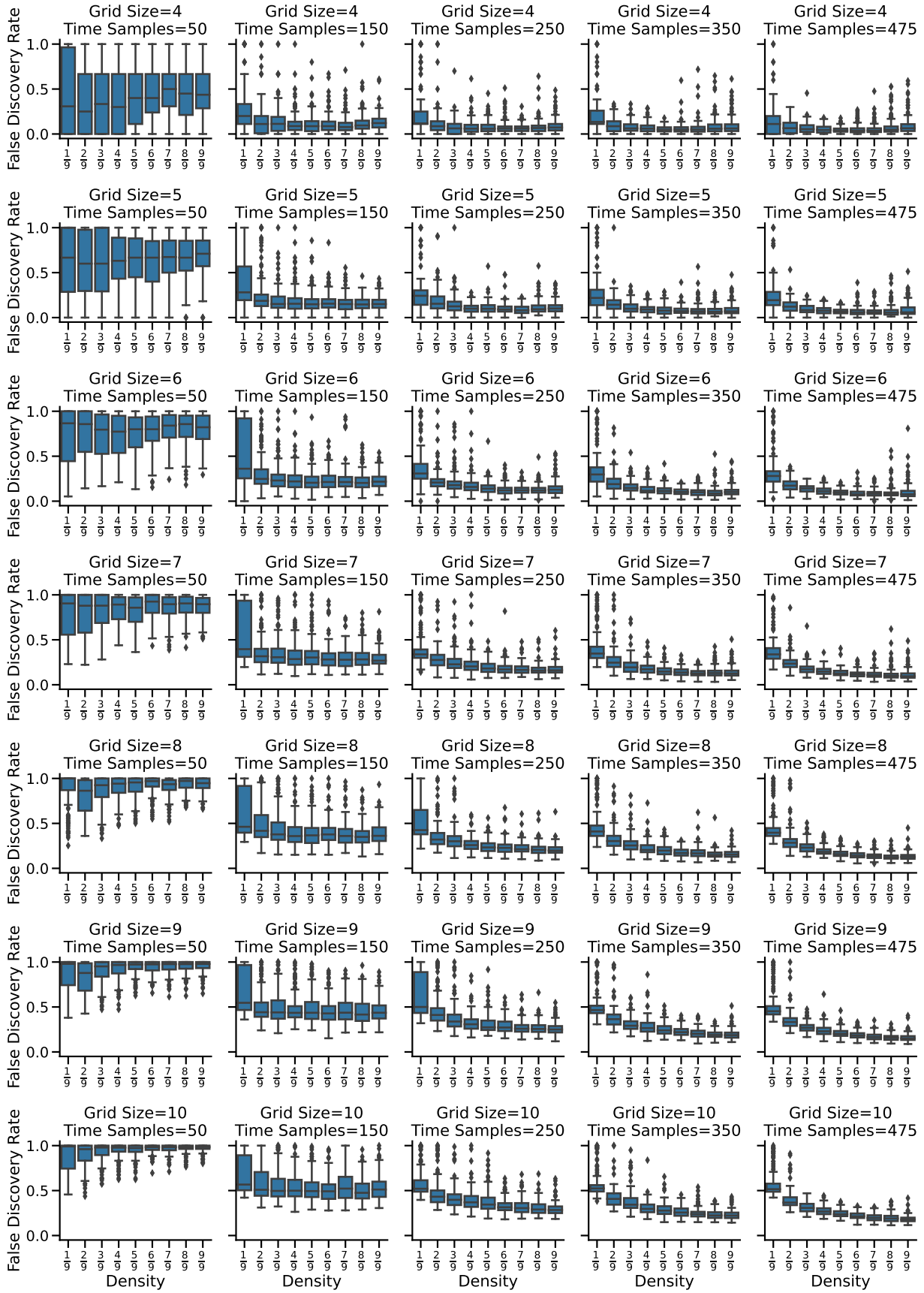


Figure A-1. False Discovery Rate of PCMCI under the scenarios described in Section 3.2. PCMCI consistently exhibits low FDR for $T > 50$. FDR decreases with the number of causal effects (density) and with increasing time samples.

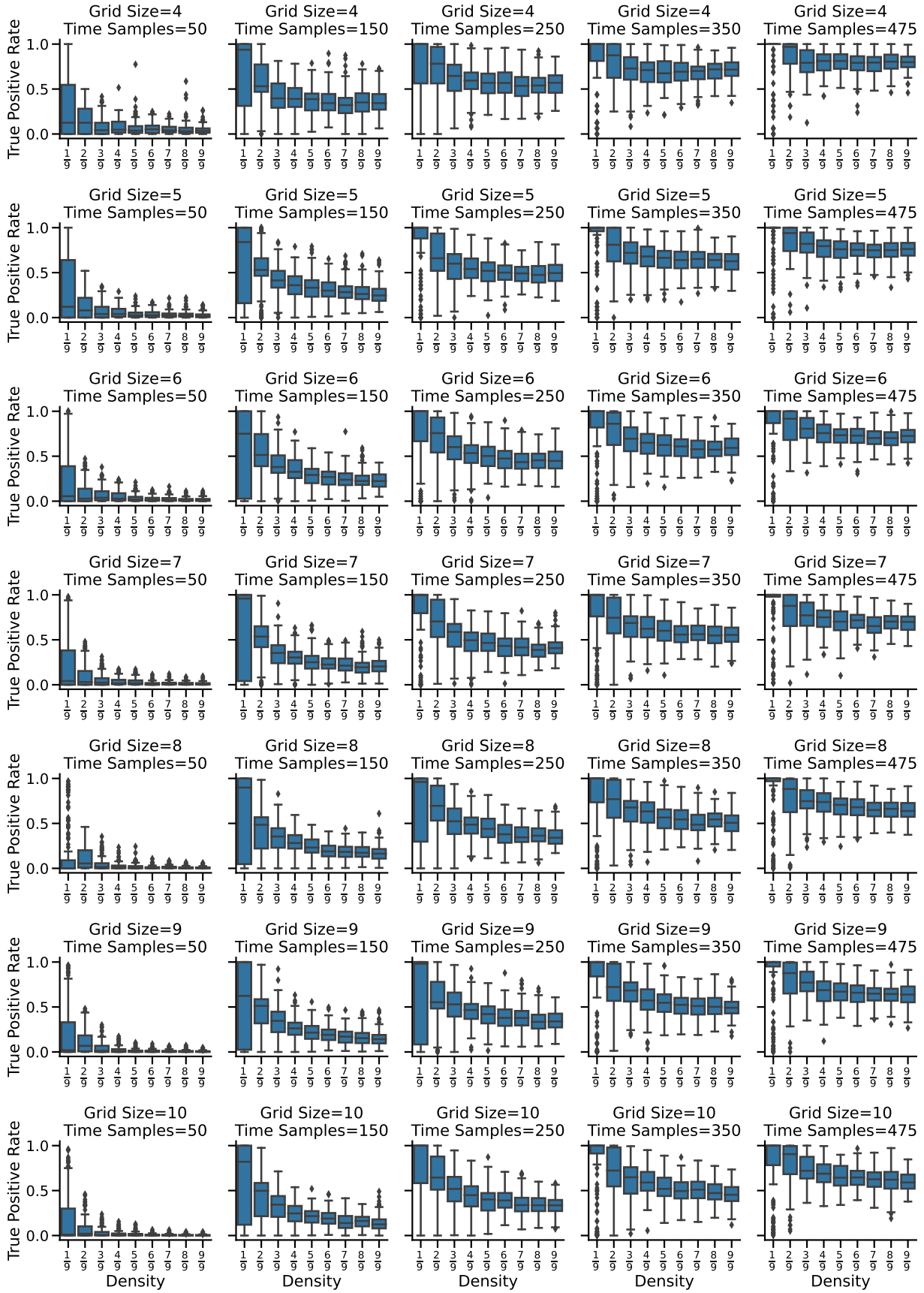


Figure A-2. True Positive Rate of PCMCI under the scenarios described in Section 3.2. PCMCI consistently exhibits low true positive rates for $T < 350$. TPR decreases with the number of causal effects and with increasing grid sizes.

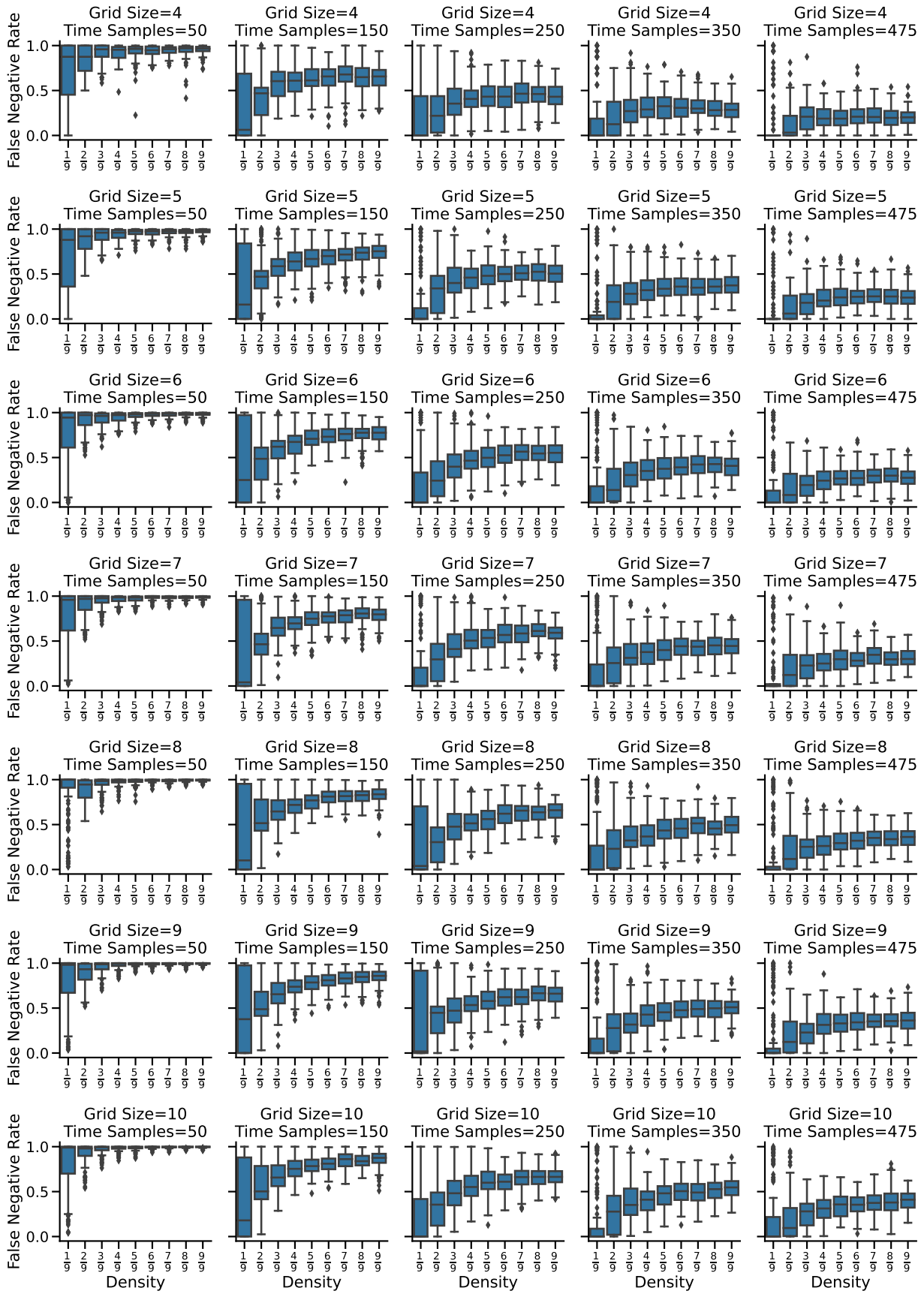


Figure A-3. False Negative Rate of PCMCI under the scenarios described in Section 3.2. PCMCI consistently exhibits relatively high false negative rates in all scenarios, indicating low statistical power. FNR generally increases with the number of causal effects and with increasing grid sizes.

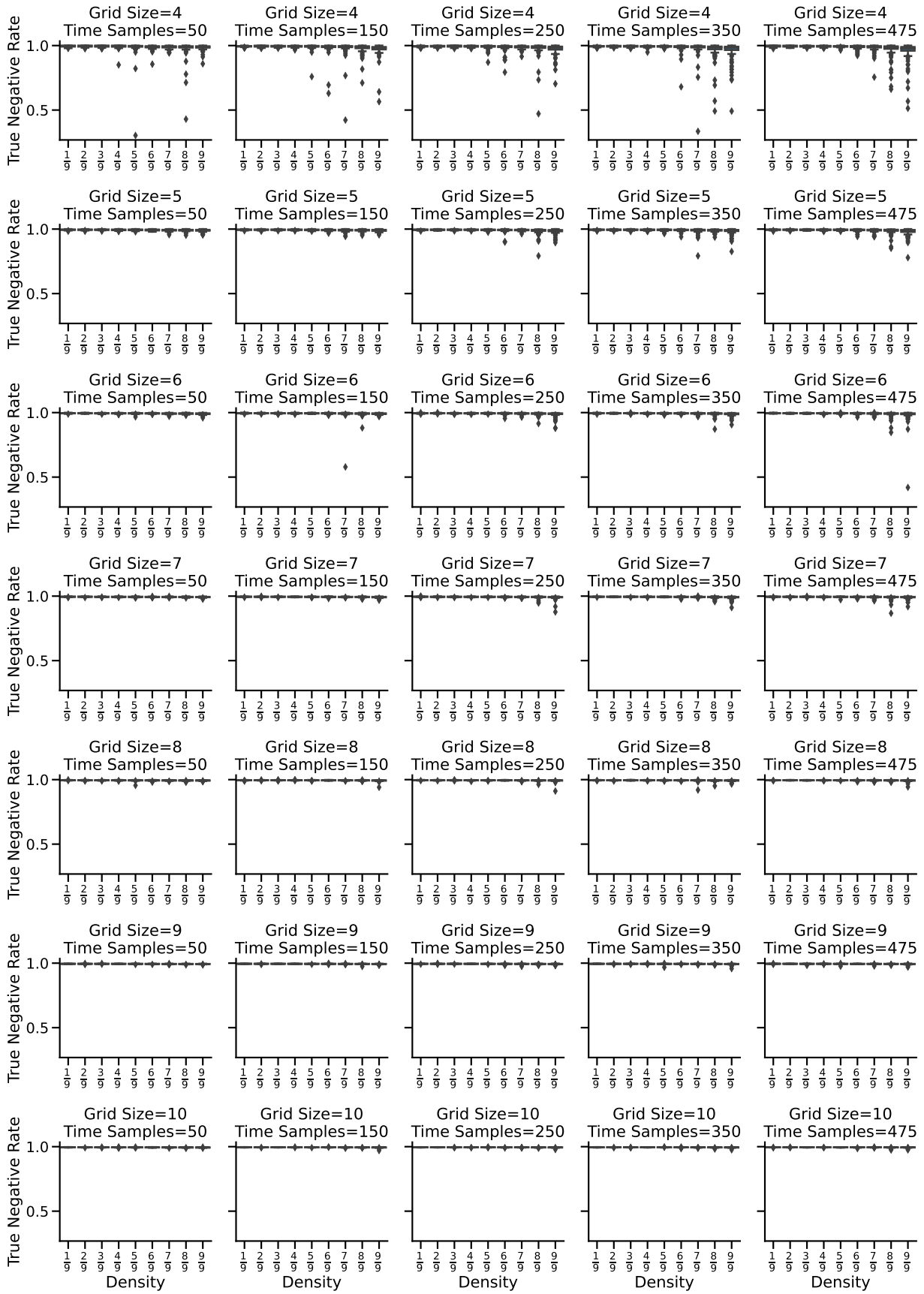


Figure A-4. True Negative Rate of PCMCI under the scenarios described in Section 3.2. PCMCI consistently exhibits near perfect true negative rates in all scenarios. To the extent it varies, TNR decreases with the number of causal effects and with decreasing grid sizes.

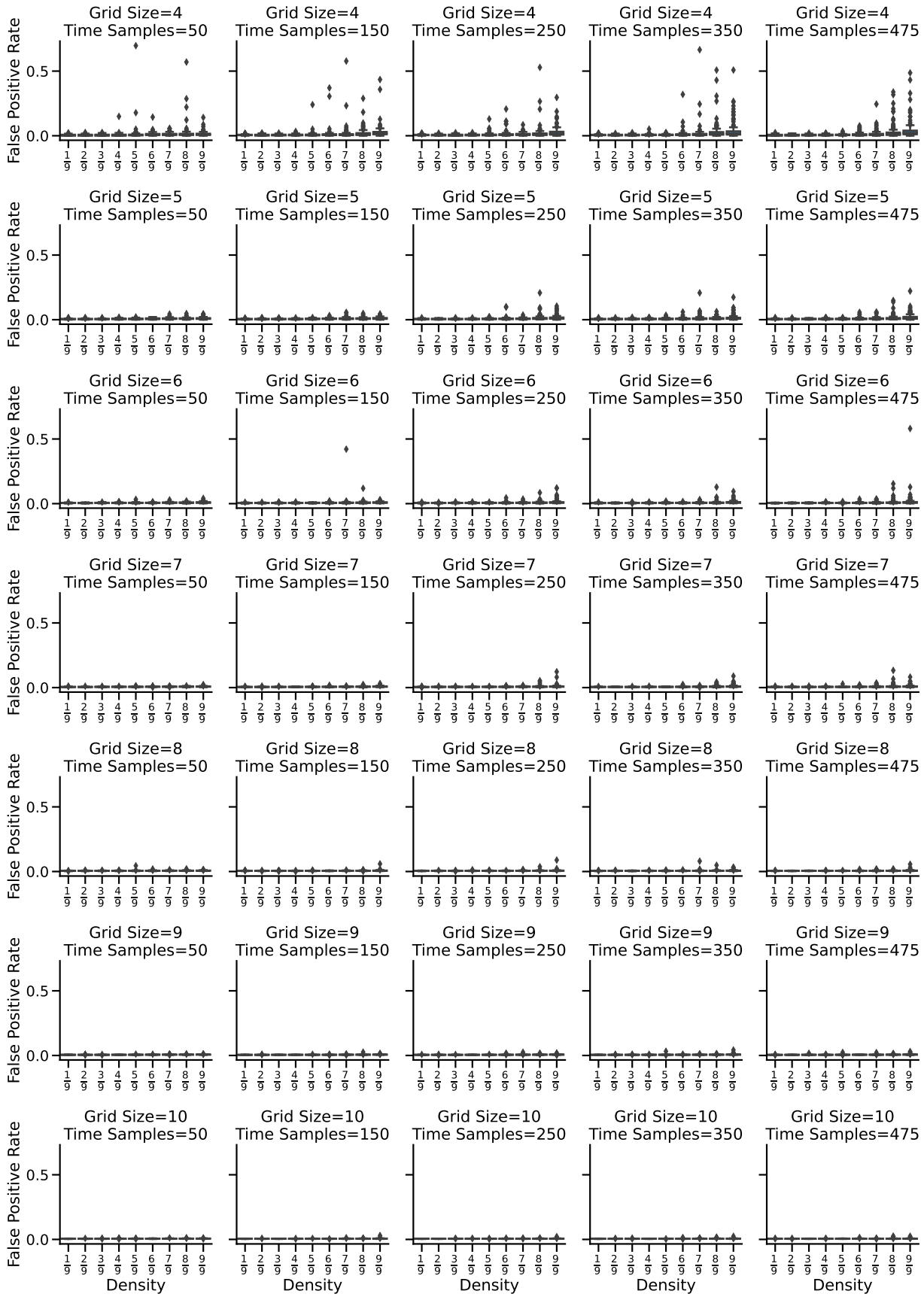


Figure A-5. False Positive Rate of PCMCI under the scenarios described in Section 3.2. PCMCI consistently exhibits near perfect false positive rates in all scenarios. To the extent it varies, FPR increases with the number of causal effects and with decreasing grid sizes.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Ronald Oldfield	1441	raoldfi@sandia.gov
Matt Peterson	1441	mgpeter@sandia.gov
Kara Peterson	1442	kjpeter@sandia.gov
Jay Brown	5493	jbrown2@sandia.gov
Aubrey Eckert	5573	acecker@sandia.gov
Lyndsay Shand	5573	lshand@sandia.gov
Irina Tezaur	8734	ikalash@sandia.gov
Meredith G.L. Brown	8931	merbrow@sandia.gov
Diana Bull	8931	dlbull@sandia.gov
Technical Library	1911	sanddocs@sandia.gov

Hardcopy—Internal

Number of Copies	Name	Org.	Mailstop
1	L. Martin, LDRD Office	1910	0359

Hardcopy—External

Number of Copies	Name(s)	Company Name and Company Mailing Address
1		

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.