Learning Dynamic Structure from Undersampled Data

John W. Cook
New College of Florida
johncookchicago@gmail.com

David Danks
Carnegie Mellon University
ddanks@cmu.edu

Sergey M. Plis
Mind Research Network
s.m.plis@gmail.com

Abstract

Most causal learning algorithms for time series data assume that the underlying generative process operates on approximately the same timescale as the measurement process (or that any differences do not impede learning). This assumption fails in many domains, and so we first show that undersampling creates learning challenges at the measurement timescale, even for simple generative processes. We then describe four algorithmic generalizations (some previously proposed, none previously tested)—two for continuous data, and two for either continuous or discrete data—and test them on simulated data. The results suggest that measurement timescale structure learning from undersampled time series data is feasible, but the appropriate model class needs to be used. Moreover, explicitly representing the possibility of undersampling can yield valuable regularization benefits.

1 INTRODUCTION

Time series data play a key role in many scientific problems. Standard methods for learning (causal) structure and parameters in dynamic time series assume that either the data *generation* timescale is approximately similar to the data *measurement* timescale, or any mismatch does not create novel learning challenges, even at the measurement level. However, many scientific problems involve significant differences between the generation and measurement timescales. For example, standard fMRI methods measure the brain's BOLD signal (believed to be a complex effect of underlying neural activity) roughly every two seconds, but neural activity almost certainly operates at a much faster timescale. The question thus arises: does timescale mismatch lead to distinctive learning challenges (beyond the "usual" statistical issues), even at the measurement level?[1]

More precisely, we focus on cases of *undersampling* in which the measurement timescale is slower than the generation timescale. We first show (in Section 2.1) that undersampling causes novel learning/parameter estimation problems for one of the most common models of data generation. In light of this result, we describe (Section 2.2) methods to learn the measurement timescale dependency structure, and test them in extensive simulations (Section 3.1). Finally, we conclude (Section 3.2) by examining what can be learned about the generative timescale structure from these measurement timescale data. Several algorithms have recently been developed to infer causal timescale structures from undersampled data [4, 6, 10, 11], but tests of those algorithms used (without evaluation) single methods for measurement timescale learning. We thus ask whether some of the present algorithms yield outputs that are superior for causal timescale estimation,[2] and whether use of such algorithms provides a "regularization" benefit that improves *measurement* timescale estimation.

2 MODELS AND ALGORITHMS

2.1 VAR MODELS AND UNDERSAMPLING

Let $\mathbf{X} = \langle X(1), \ldots, X(v) \rangle$ be a set of random variables. A standard framework for (discrete-time) dynamical systems is the *V*ector *A*uto*R*egression (VAR) model, whose simplest form is:

$$\mathbf{X}_t = \mathbf{A}_1 \mathbf{X}_{t-1} + \ldots + \mathbf{A}_l \mathbf{X}_{t-l} + \mathbf{e}_t \qquad (1)$$

[1] Of course, there are many additional challenges in causal learning from fMRI data.

[2] In theory, algorithm A might have more total errors than algorithm B at the measurement timescale, but A's errors might be less problematic for causal timescale estimation.

where subscripts denote timesteps; $\mathbf{A}_i$ is a matrix encoding the *direct* impact of $\mathbf{X}_{t-i}$ on $\mathbf{X}_t$; and $\mathbf{e}_t$ is the vector of serially uncorrelated noise factors with simultaneous covariance matrix $\mathbf{\Sigma}$.

Let $P_{\mathcal{M}}(\mathbf{X}_t|\mathbf{X}_{t-1},\ldots,\mathbf{X}_{t-l})$ be the conditional distribution induced by VAR model $\mathcal{M}$. We follow standard practice and assume only that $P_{\mathcal{M}}$ is stationary; $P(\mathbf{X}_t)$ need not be stationary over time. $\mathbf{\Sigma}$ is assumed to be diagonal; non-diagonal $\mathbf{\Sigma}$ correspond to structural vector autoregression models, which we address later. The maximum l (l_{max}) such that $\mathbf{A}_r = \mathbf{0}$ for all $r > l_{\text{max}}$ is the *order* of the VAR model. Provably, undersampling does not increase the order of a VAR model, and so the measurement timescale order is the same as at the causal timescale [2]. For simplicity, we focus here on first-order ($l_{\text{max}} = 1$) VAR models with $\mathbf{A}_1 = \mathbf{A}$.

$\mathbf{A}$ encodes the influence of the previous timestep on the current time, and can be represented as a directed acyclic graph $\mathcal{G}$ over nodes for $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ with $X(j)_{t-1} \rightarrow X(i)_t$ iff $A_{ij} \neq 0$. We use both matrix and graph language as appropriate. Define the *density* ρ of $\mathbf{A}$ (or $\mathcal{G}$) to be the fraction of non-zero elements (or present edges).

Let $\mathbf{D}^1 = \{\mathbf{X}_0, \mathbf{X}_1, \ldots\}$ be the data at the timescale of the underlying VAR model. These data are *undersampled at rate* u when $\mathbf{D}^u = \{\mathbf{X}_0, \mathbf{X}_u, \ldots, \mathbf{X}_{ku}, \ldots\}$ for $k \in \mathbb{Z}^+$. In general, superscripts will denote undersample rate. We also use superscripts to modify time indices; for example, $(t-1)^u$ denotes the previous time step in $\mathbf{D}^u$, which corresponds to $t-u$ in $\mathbf{D}^1$.

Suppose $\mathbf{D}^1$ is generated from $P_{\mathcal{M}}(\mathbf{X}_t|\mathbf{X}_{t-1})$ for VAR model $\mathcal{M}$. One key question for measurement timescale learning is whether there is always a VAR model $\mathcal{M}^u$ such that $P_{\mathcal{M}^u}$ can fit $\mathbf{D}^u$ (in the large sample limit). Theorem 2.1 provides a negative answer to this question: frequently (though not always), there is no VAR model for undersampled data. That is, VAR models are not generally "closed" under the operation of undersampling.

Theorem 2.1. *Let $\mathcal{M}$ be a first-order VAR with $P_{\mathcal{M}}(\mathbf{X}_t|\mathbf{X}_{t-1})$. For $u > 1$, there is a first-order VAR $\mathcal{M}^u$ such that $P_{\mathcal{M}^u}(\mathbf{X}_t|\mathbf{X}_{(t-1)^u}) = P_{\mathcal{M}}(\mathbf{X}_t|\mathbf{X}_{t-u})$ if and (almost always) only if there is no c with $A_{ic}, A_{jc} \neq 0$ for $i \neq j$ (i.e., $\mathcal{G}$ has no $X(i)_t \leftarrow X(c)_{t-1} \rightarrow X(j)_t$ structures).*

Proof. Let $\mathcal{M}$ be an arbitrary first-order VAR (so multivariate Gaussian). After algebra, $\mathcal{M}$ undersampled by u yields (see also [4]):

$$\mathbf{X}_t = (\mathbf{A})^u \mathbf{X}_{t-u} + \sum_{i=0}^{u-1} (\mathbf{A})^i \mathbf{e}_{t-i} \qquad (2)$$

By assumption, $\mathbf{X}_{t-u}$ is independent of $\mathbf{e}_{t-i}$ for $i \leq u$, and so there is a suitable VAR model $\mathcal{M}^u$ iff $\sum_{i=0}^{u-1}(\mathbf{A})^i \mathbf{e}_{t-i} = \mathbf{f}_t$ has the correct noise properties. $\mathbf{f}_t$ must be serially uncorrelated since the $\mathbf{e}_t$ are. Thus, we must determine if $\mathbf{\Sigma}_{\mathbf{f}}$ is diagonal, which will hold iff each $e(j)_{t-i}$ occurs in the expansion for (at most) one $f(k)_t$.

($\Leftarrow$) Assume there is no appropriate c. Expansion of the summed $\mathbf{e}_{t-i}$ shows that each $e(c)_{t-i}$ occurs in at most one $f(j)_t$, and so $\mathbf{\Sigma}_{\mathbf{f}}$ is diagonal.

($\Rightarrow$) Assume there is such a c. Thus, at least one $e(c)_{t-1}$ will occur in multiple $f(k)_t$ expansions. Those terms will cancel out of all-but-one $f(k)_t$ expansion only if the relevant $\mathbf{A}$ entries (perhaps exponentiated) exactly balance; such exact parameter balancing happens for only Lebesgue measure zero of $\mathbf{A}$-parameter space [9]. Hence, there is almost always no such c. □

Although VAR models are not closed under undersampling, they might nonetheless be able to approximate undersampled time series arbitrarily closely. To test for this possibility, we randomly produced 1000 stable VAR models with 20 variables and edge/matrix density $\rho = 0.2$; generated 4000 samples; undersampled that datastream at $u \in \{1, 2, 3, 4\}$; and then used only the first 1000 datapoints of each data series (regular and undersampled) to estimate an optimal VAR model. Figure 1 shows BIC scores of the final models for each u,[3] where the variability of those scores encodes model selection uncertainty [12]. Notably, the score distributions for each u are significantly different from each other, with BIC increasing as u increases. These large differences in BIC score distributions vividly demonstrate that undersampling results in data that are outside of the VAR model class, and sometimes very far outside.

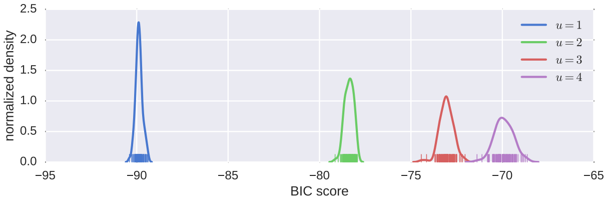


Figure 1: Model selection uncertainty (distribution of BIC scores) for VAR models given undersampled data.

Estimation of a VAR model is insufficient for modeling undersampled time series, even when the data were truly generated by a VAR model, so we must consider alternative models. Recall that the qualitative structure of VAR $\mathcal{M}$ can be represented by DAG $\mathcal{G}^1$ over $2\mathbf{X}$ (nodes for every $X(i)_{t-1}$ and $X(i)_t$), and edges corresponding to

[3] All models have the same number of parameters, so data likelihoods would make the same point.

non-zero $\mathbf{A}$ entries. In contrast, the relevant graphical model class for $\mathcal{G}^u$ has $2\mathbf{X}$ nodes for $\mathbf{X}_t$ and $\mathbf{X}_{(t-1)^u}$, which need not be $\mathbf{X}_{t-1}$. More importantly, this graph can have both (1) $X(i)_{(t-1)^u} \rightarrow X(j)_t$ iff there is a directed path $X(i)_{t-u} \rightarrow \ldots \rightarrow X(j)_t$ in $\mathcal{G}^1$; and (2) $X(i)_t \leftrightarrow X(j)_t$ iff there is $X(i)_t \leftarrow \ldots X(c)_{t-k} \ldots \rightarrow X(j)_t$ in $\mathcal{G}^1$ for $k < u$ [2]. The bidirected edges capture the non-diagonal $\mathbf{e}_t$ correlation structure described in Theorem 2.1.

These graphical models correspond to Structural VAR (SVAR) models, and there are efficient algorithms for parameter estimation given the graphical structure. However, there has been almost no research on SVAR structure learning algorithms. We now turn to exploring multiple such methods, and also testing their measurement timescale performance on simulated data (Section 3.1), where we are able to determine the "ground truth" for undersampled structure using the forward inference algorithm of Danks and Plis [2].

2.2 GENERALIZED ALGORITHMS

Prior structure learning research involving undersampled data has focused on algorithms for inferring causal timescale structure from measurement timescale inputs [4, 6, 10, 11]. As such, those papers used measurement timescale structure learning algorithms, though none of them tested the performance of those methods *at the measurement timescale*. We describe four generalizations of existing time series structure learning algorithms—three that have previously been mentioned—that accomodate the possibility of undersampled data.

As noted above, the key graphical impact of undersampling is to produce bidirected edges, and so the generalized algorithms all search for not only directed between-time edges, but also bidirected within-time edges. For the purposes of this paper, we have assumed that the generating structure is a VAR model (though we relax that assumption in Section 3.1), and so all four algorithms can be used on continuous-valued data. Two algorithms can also be applied to discrete-valued data, and we explain the necessary adjustments in the appropriate sections.

2.2.1 SVAR Estimation

For linear Gaussian data, undersampled data can be represented as a first-order SVAR model [3, 7]:

$$\mathbf{X}_t = \mathbf{B}\mathbf{X}_t + \mathbf{A}\mathbf{X}_{(t-1)^u} + \epsilon_t \tag{3}$$

where the diagonal elements of $\mathbf{B}$ are normalized to 1 and elements of ϵ_t are independent. In general, this model is underdetermined. When the SVAR results from undersampling, however, $\mathcal{G}^u$ will have only bidirected within-time edges which are symmetric, and the corresponding within-time matrix $\mathbf{B}$ must also be symmetric (non-zero $\mathbf{B}$ entries for bidirected edges). Non-zero $\mathbf{A}$ entries encode between-timestep directed edges.

Given $\mathbf{D}^u$, we can directly estimate the SVAR model structure as done in [10] by finding the $\mathbf{A}, \mathbf{B}$ that optimize the log-likelihood of the data, subject to two constraints: symmetry of $\mathbf{B}$, and small matrix entries made into (structural) zeroes. Precise mathematical formulations are provided in Eqs. (4)-(6), where $\mathbf{X}_{-1}$ denotes the values of $\mathbf{X}$ shifted one step back relative to $\mathbf{X}$.

$$\ln \mathcal{L}_c(\mathbf{A}, \mathbf{B}) \propto T \ln |\mathbf{B}| - \frac{1}{2}\,\mathrm{trace}(\Sigma_{\mathbf{X}} \mathbf{B}^T \mathbf{B}) \tag{4}$$

$$\Sigma_{\mathbf{X}} = \mathbf{Y}\mathbf{Y}^T \tag{5}$$

$$\mathbf{Y} = \mathbf{X} - \mathbf{A}\mathbf{X}_{-1}, \tag{6}$$

2.2.2 Score-based Graph Search

We also examined existing graphical structure search algorithms, though adapted for potentially undersampled data. Score-based search procedures find the graph that maximizes some score, typically likelihood-based. We adapted the FGS algorithm[4]—a computationally efficient version of Greedy Equivalence Search (GES)—that searches through the space of (graph) equivalence classes in a greedy fashion based on BIC score. Despite being a greedy search, FGS/GES is correct in the large sample limit [1].

In general, the true measurement timescale graph $\mathcal{G}^u$ can have bidirected edges, but FGS cannot output such edges. Thus, no simple adaptation can be provably correct for all possible data. We instead considered more heuristic adaptations of FGS that might nonetheless be successful on smaller sample sizes.

The most straightforward way to adapt FGS is to not search over graphs that posit impossible connections (e.g., $X(i)_t \rightarrow X(j)_{t-1}$), and then adjust any within-time edges. In preliminary investigations, however, we found that this adjustment led to a fractured search space, and so the algorithm was frequently trapped at a local maximum, typically a very sparse graph.

Instead, we adapted FGS by post-processing the output. We first ran normal FGS for graphs over $2\mathbf{X}$, without any constraints encoding temporal information. We then transformed the FGS output graph $\mathcal{G}_{\mathcal{FGS}}$ into $\mathcal{G}$ by edge-wise adjustments, as shown in Algorithm 1. The resulting algorithm provided the best overall error rates.

Appropriate scores have also been developed for discrete-valued data, and Algorithm 1 can be easily

[4] We used the python-wrapped version of FGS from Tetrad.

Algorithm 1: Modified FGS Algorithm

```
Data: $\mathbf{D}^u = \{\mathbf{X}_0, \mathbf{X}_u, \ldots, \mathbf{X}_{ku}, \ldots\}$ for unknown $u$
Output: $\mathcal{G}^u$
// run FGS
$\mathcal{G}^u_{FGS} \leftarrow FGS(\mathbf{D})$;
// create output $\mathcal{G}^u$
$\mathcal{G}^u \leftarrow$ empty graph over nodes for $2\mathbf{X}$;
// adjust FGS output
forall edges $E \in \mathcal{G}^u_{FGS}$ do
    if $E = X(i)_{t-1} \rightarrow X(j)_t$ then
        add $X(i)_{t-1} \rightarrow X(j)_t$ to $\mathcal{G}^u$
    else if $E = X(i)_{t-1} \leftarrow X(j)_t$ then
        add $X(i)_{t-1} \rightarrow X(j)_t$ to $\mathcal{G}^u$
    else if $E = X(i)_t \rightarrow X(j)_t$ then
        add $X(i)_t \leftrightarrow X(j)_t$ to $\mathcal{G}^u$
return $\mathcal{G}^u$
```

Algorithm 2: Modified PC Algorithm

```
Data: $\mathbf{D}^u = \{\mathbf{X}_0, \mathbf{X}_u, \ldots, \mathbf{X}_{ku}, \ldots\}$ for unknown $u$
Output: $\mathcal{G}^u$
// create initial, complete $\mathcal{G}^u$
$\mathcal{G}^u \leftarrow$ empty graph over nodes for $2\mathbf{X}$;
forall $i, j \in \{1, \ldots, |\mathbf{X}|\}$ do
    add $X(i)_{t-1} \rightarrow X(j)_t$ to $\mathcal{G}^u$;
    if $i \neq j$ then
        add $X(i)_t \leftrightarrow X(j)_t$ to $\mathcal{G}^u$
// remove directed edges
for $N \leftarrow 0$ to $|\mathbf{X}| - 2$ do
    forall $i, j$ s.t. $X(i)_{t-1} \rightarrow X(j)_t$ in $\mathcal{G}^u$ do
        forall $\mathbf{S} \subseteq \mathbf{pa}(X(j)_t)$ s.t. $|\mathbf{S}| = N$ do
            if $X(i)_{t-1} \perp X(j)_t | \mathbf{S}$ then
                remove $X(i)_{t-1} \rightarrow X(j)_t$ from $\mathcal{G}^u$;
// remove bidirected edges
forall $i \neq j \in \{1, \ldots, |\mathbf{X}|\}$ do
    if $X(i)_t \perp X(j)_t | \mathbf{pa}(X(i)_t) \cup \mathbf{pa}(X(j)_t)$ then
        remove $X(i)_t \leftrightarrow X(j)_t$ from $\mathcal{G}^u$;
return $\mathcal{G}^u$
```

modified to use a different score in the first step. We adapted GOBNILP, which uses local scores to find optimal graphs, as this adaptation performed best among a range of potential adjustments that we considered.

2.2.3 Constraint-based Graph Search

Constraint-based search methods find the equivalence class of graphs that predicts the pattern of independencies and associations found in the data [14]. For computational and statistical reasons, constraint-based search algorithms do not compute every possible independence/association, but rather a dynamically determined set based on earlier results in the search algorithm.

The PC algorithm [14] has previously been adapted for time series data [8], though that version assumed that the measurement and causal/generative timescales were approximately equal. Thus, it will not necessarily work for learning measurement timescale structure given undersampled data.

Instead, we used a version of the PC algorithm that (a) starts with a graph containing only possible edges (rather than the usual complete graph); and then (b) sequentially attempts to remove directed then bidirected edges, in the usual PC manner. Algorithm 2 provides more specific details about the resulting algorithm.

For continuous data, we tested for (conditional) independence using OLS regression, and judged independence if the resulting coefficient was not significantly different from zero. This version was previously used by [11], though without any exploration of its performance on measurement timescale data. For discrete-valued data, we used a conditional χ^2 test insted of OLS. These independence tests are the same as those used in standard implementations of the PC algorithm; our adjustment was only in which independence tests were performed, not the tests used.

2.2.4 Information-theoretic Search

Finally, we consider graphical model search algorithms based on information-theoretic measures. Granger Causality (GC) [5] is one of the most widely-used "causal" search algorithms for time series data.[5] Prior work has shown that GC provides unreliable information about the *causal* timescale given undersampled data [13], but its performance on measurement timescale data is unknown, though a similar algorithm was used by [6].

The key intuition underlying GC-based search is that $X(i)_{t-1}$ Granger-causes $X(j)_t$ just when $X(i)_{t-1}$ provides information about $X(j)_t$, even conditioning on all other variables in the past. More specifically, let $M_X(\mathbf{S})$ be some class of models that predict X given $\mathbf{S}$ as input (e.g., density estimator, mutual information calculation, etc.). We add $X(i)_{t-1} \rightarrow X(j)_t$ only if $M_{X(j)_t}(\mathbf{X}_{t-1} \setminus X(i)_{t-1}) \neq M_{X(j)_t}(\mathbf{X}_{t-1})$. For bidirected edges, as shown in Algorithm 3, we use a second round of tests to determine whether to include bidirected edges in $\mathcal{G}^u$. We tested the modified GC algorithm only on continuous-valued data. Notably, the modified GC algorithm is much simpler and faster than the other generalized algorithms.

[5] We use scare quotes as GC provides causal information only under very specific conditions.

```
Algorithm 3: Modified GC Algorithm
Data: $\mathbf{D}^u = \{\mathbf{X}_0, \mathbf{X}_u, \ldots, \mathbf{X}_{ku}, \ldots\}$ for unknown $u$
Output: $\mathcal{G}^u$
// create initial empty $\mathcal{G}^u$
$\mathcal{G}^u \leftarrow$ empty graph over nodes for $2\mathbf{X}$;
// add directed edges
forall $X(i)_{t-1}, X(j)_t \in \mathcal{G}^u$ do
    if $M_{X(j)_t}(\mathbf{X}_{t-1} \setminus X(i)_{t-1}) \neq M_{X(j)_t}(\mathbf{X}_{t-1})$ then
        add $X(i)_{t-1} \rightarrow X(j)_t$ to $\mathcal{G}^u$;
// add bidirected edges
forall $X(i)_t, X(j)_t \in \mathcal{G}^u$ do
    if $M_{X(j)_t}(\mathbf{X}_{t-1}) \neq M_{X(j)_t}(\mathbf{X}_{t-1} \cup X(i)_t)$ then
        add $X(i)_t \leftrightarrow X(j)_t$ to $\mathcal{G}^u$;
return $\mathcal{G}^u$
```

2.2.5 Validating the Generalizations

A generalized algorithm should perform approximately the same as the original algorithm for any data that satisfy the original algorithm's assumptions. To validate these generalizations, we compared outputs for each pair of continuous-data search algorithms (original vs. generalized) for 100 randomly generated VAR models with $|\mathbf{X}| = V \in \{10, 15, \ldots, 30\}$, $\rho = 0.2$, and $N = 1000$. For all comparisons, we used $u = 1$, as that satisfies the original algorithms' assumption that the measurement and causal timescales are the same. Since our interest is simply whether the outputs are the same, we calculated the symmetric difference of the edge sets for the outputs, which is also the Hamming distance between binary representations of output graphs. Notice that the output graphs of the generalized algorithms can include bidirected edges, but those of the original algorithms cannot. Thus, any bidirected edge is automatically an error (on this measure). Figure 2 shows that the generalized algorithms performed almost identically to the original algorithms, thereby validating the generalizations (and implementations).

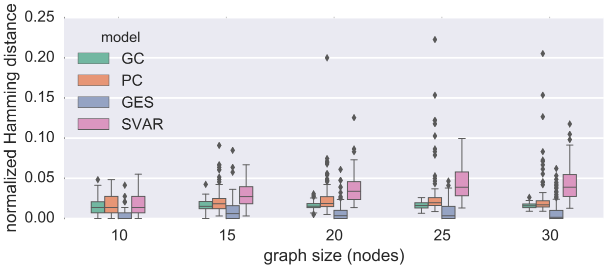


Figure 2: Hamming distance between outputs of original and generalized algorithms on $\mathbf{D}^1$ data.

3 RESULTS

3.1 SIMULATION TESTS

We first examine the performance of these generalized algorithms on simulated data for which we can compute the ground truth undersampled structure, and so algorithm error rates. For all simulation tests, we did the following for each algorithm $\mathcal{A}$:

1. Generate a random VAR model $\mathcal{M}$ with random graph $\mathcal{G}$ and $\mathbf{A}$ values (normalized to ensure that the time series does not diverge)[6]
2. Sample (non-equilibrium) time series $\mathbf{D}^1$ from $\mathcal{M}$
3. Undersample by u to produce $\mathbf{D}^u$
4. Use $\mathcal{A}$ to determine $\mathcal{G}_{out}$ given $\mathbf{D}^u$
5. Compute errors of commission (i.e., false edge positives) and omission (false edge negatives) in $\mathcal{G}_{out}$, using $\mathcal{G}^u$ (the theoretically predicted graph when undersampling by u)

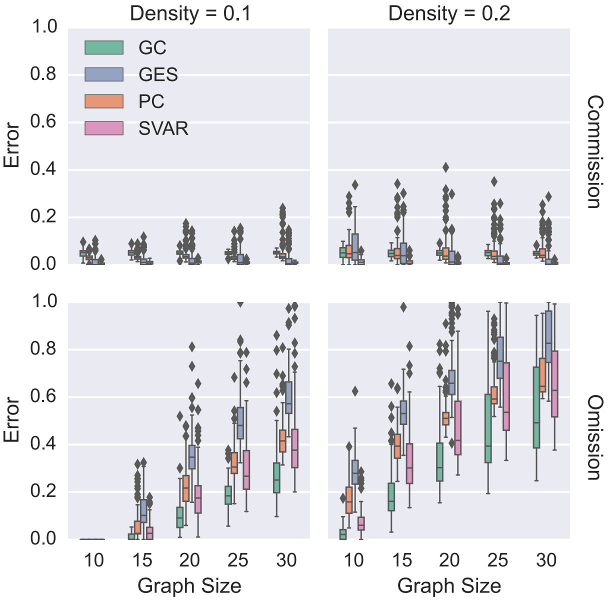
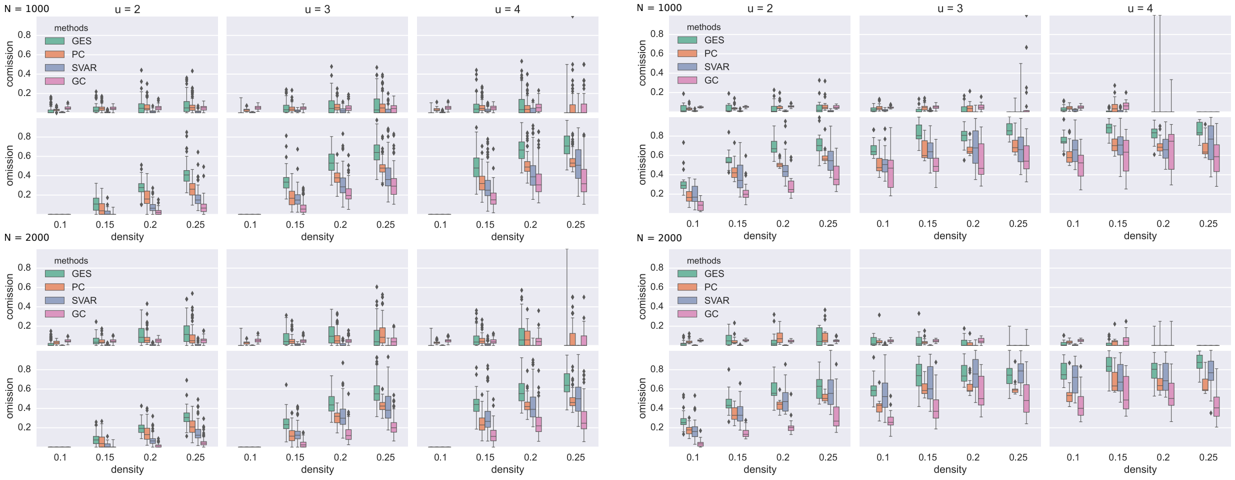


Figure 3: Estimation error as a function of graph size for $u = 2$ & $N = 1000$.

We first tested algorithm performance as a function of both graph size and density for continuous data. Figure 3

[6] For discrete variables, we need transition probabilities—$P(X(i)_t|\mathbf{pa}(i)_{t-1})$—that are generated as follows (all variables have m possible values): For the base case of $X_{t-1} \rightarrow Y_t$, we construct a random $1-1$ map $f : X \mapsto Y$, and set $P(Y = f(x)|X = x) = A$ for constant $0 < A < 1$, and $P(Y \neq f(x)|X = x) = \frac{1-A}{m-1}$. If there are multiple parents, then we first construct parent-specific conditional distributions as above, and then set $P(X(i)_t|\mathbf{pa}(i)_{t-1})$ to be the renormalized product of those parent-specific conditionals. This method ensures that each parent has a non-neglible impact on the child.



(a) $V = 10$

(b) $V = 20$

Figure 4: Estimation error as a function of ρ for different u, N for continuous-data algorithms applied to either 100 10-node (Figure 4a) or 20 20-node (Figure 4b) random graphs.

plots commission and omission errors for 100 random graphs, $N = 1000$ post-undersampling datapoints, and $u = 2$. As expected, algorithm performance worsened as both V and ρ increased. Unsurprisingly, edge commission error rates were lower than edge ommission error rates: undersampling generally leads to weaker associations at the measurement timescale (compared to the causal generative timescale), so false positives should be less likely than false negatives. Interestingly, though, the absolute magnitude of the commission error rates was quite small for all V, ρ. In contrast, omission error rates grew rapidly as a function of V, particularly for $\rho = 0.2$. For example, for $N = 30, \rho = 0.2$, the generalized (for undersampling) PC algorithm typically outputs an empty graph. The GC algorithm outperforms the other generalized algorithms in omission error rates, while SVAR is the best for edge comission.

Given this basic understanding of the algorithms' performances, we then turned to a more general analysis. We focused on $V \in \{10, 20\}$, as the results in Figure 3 indicated that those were sufficiently different in complexity and performance. Figure 4 plots the commission and omission error rates for all four continuous-data algorithms across multiple values of N, u, ρ, V, with 100 random graphs per simulation setting.

Omission error rates were again higher than commission error rates, and all of the algorithms exhibited very low false positive rates (alternately, high specificity). Interestingly, both error *rates* increased—commission more slowly than omission—as the total *number* of edges in $\mathcal{G}^1$ increased, whether due to increases in V or ρ.

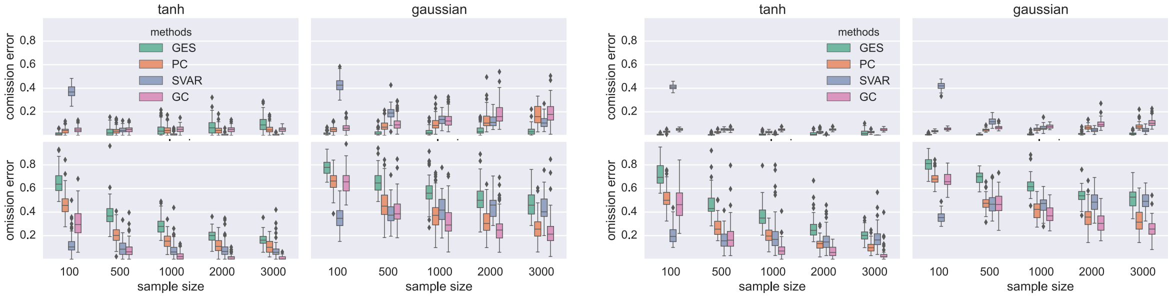
Unsurprisingly, omission error rates also increased with u. An edge in $\mathcal{G}^u$ corresponds to a directed path of length u in $\mathcal{G}^1$. In general, the association between endpoints of a directed path will be smaller than between adjacent variables on that path. All of these algorithms use associations to posit edges, so as the lengths of to-be-detected paths increase (i.e., as u increases), the estimation problem should become progressively more difficult.

Overall, we find that GC is the best-performing algorithm for these conditions, as its measurement timescale success occurs across a wide range of simulation parameter settings. SVAR is the next best performer for omission errors and is the best method with respect to comission errors.

One further question for the continuous-data algorithms is their robustness to nonlinear relationships. This question is particularly salient for SVAR estimation, as it assumes a linear model. Figure 5 shows commission and omission errors as a function of sample size for 100 randomly generated structures with $\langle V, \rho \rangle \in \{\langle 10, 0.2 \rangle, \langle 20, 0.1 \rangle\}$ and $u = 2$. We tested two different nonlinear transformations, each applied variable-wise after each time step:

- Hyperbolic tangent: $\tanh(X)$
- Gaussian radial basis function: $\phi(X) = e^{\frac{-\|X-\mu\|^2}{2\sigma^2}}$

The Gaussian RBF function significantly worsens performance compared to the $\tanh$ nonlinearity. Unsurprisingly, SVAR performance worsens the most, while the other algorithms are less affected. In particular, GCu still performs quite well. Commission error rates were higher for $\langle V = 20, \rho = 0.1 \rangle$ for both nonlinear functions, even though those graphs were less dense than $\langle V = 10, \rho = 0.2 \rangle$. Overall performance was, however,



(a) $V = 10, \rho = 0.2$ (b) $V = 20, \rho = 0.1$

Figure 5: Estimation error as function of sample size for nonlinear models.

not dramatically worse for certain algorithms, which suggests that nonlinear relationships do not present an insurmountable problem.

Finally, Figure 6 shows the performance of the two discrete data algorithmic variants. As with continuous data, the constraint-based search (DPC) systematically outperforms the score-based search (GOBNILP). The discrete data led to higher omission error rates, though with almost zero comission errors: both algorithms are far more likely to output almost-empty graphs.

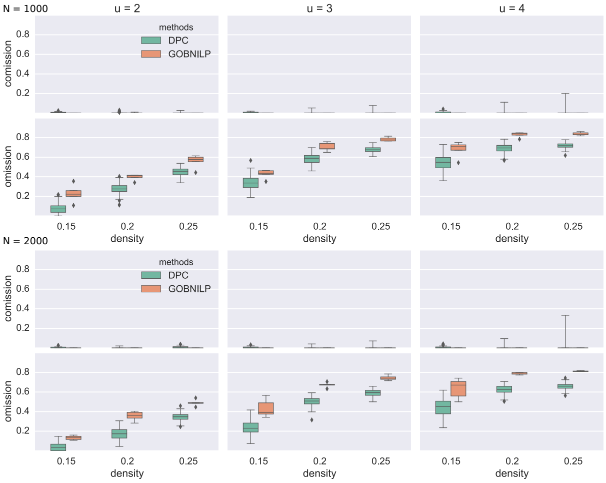


Figure 6: Estimation error of a function of ρ for different u, N for discrete data algorithms for $V = 10$.

3.2 BEYOND MEASUREMENT TIMESCALE

The different generalized algorithms exhibit substantial variation in estimation errors for $\mathcal{G}^u$, the measurement timescale structure. In many contexts, however, we are also interested in the causal timescale structure. Errors in measurement timescale estimation need not translate directly to causal timescale estimation: any particular measurement timescale estimation error could lead to many, or zero, errors in the causal timescale structure. Various algorithms have recently been developed to infer the space of possible $\mathcal{G}^1$ from $\mathcal{G}^u$ [6, 10, 11]. One challenge for all of these algorithms is that many potential $\mathcal{G}^u$ inputs have *no* corresponding $\mathcal{G}^1$; we refer to this as the "reachability" problem. Hence, it can be important to get "appropriately" close in the $\mathcal{G}^u$ learning. The standard responses to this problem are to either apply the search algorithm to neighbors of $\mathcal{G}^u$ until a reachable graph is found (as in [10, 11]); or use a constraint satisfaction-based approach [6]. We used the latter approach, as it is considerably faster.

We generated 100 random 8-node graphs for each $\rho \in \{0.17, 0.20, 0.25\}$, and 1000 undersampled datapoints ($u = 2$) for each graph. For each measurement timescale estimation algorithm $\mathcal{A}$, we first applied $\mathcal{A}$ to the data to obtain $\mathcal{G}^u$, and then passed that output to the causal timescale inference algorithm of Hyttinen et al. [6]. We computed three types of estimation errors: (i) $\mathcal{G}^u$ output by $\mathcal{A}$ vs. measurement timescale ground truth; (ii) inferred $\mathcal{G}^1$ vs. causal timescale ground truth; and (iii) $\mathcal{G}^u$ implied by inferred $\mathcal{G}^1$ vs. measurement timescale ground truth. We also measured execution clocktime, limited to one hour per graph.

Figure 7 shows the results of these simulations, with the three rows of the Figure corresponding to these three error calculations. As expected, the top row replicates the pattern of results from Section 3.1. The middle row demonstrates that not all measurement timescale estimation algorithms are the same: SVAR and GC provide notably better performance for $\mathcal{G}^1$ inference. Moreover, the bottom row shows that $\mathcal{G}^u$ estimation for SVAR and GC is improved, though not dramatically, by requiring there to be a $\mathcal{G}^1$ that could yield $\mathcal{G}^u$ given undersampling. That is, explicit modeling of undersampling provides a regularization benefit for $\mathcal{G}^u$ estimation. Note, for the already imprecise GES and PC the omission error increases. Overall, SVAR has comparably low error rates to GC, and also provides estimates, for which the $\mathcal{G}^1$ inference works much faster.

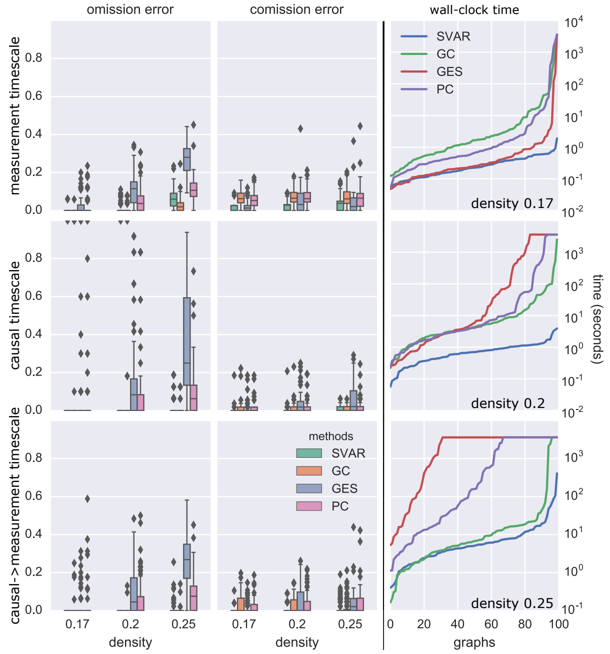


Figure 7: Error rates and clock-time plots as a function of ρ for 8-node graphs with $N = 1000$ at $u = 2$ for $\mathcal{G}^u$ estimation, $\mathcal{G}^1$ inference, and implied $\mathcal{G}^u$ for inferred $\mathcal{G}^1$.

4 CONCLUSION

Time series data are rapidly becoming almost-ubiquitous. In many of those domains, however, the relevant measurement processes are often much slower than the underlying generative or causal processes. As we showed in Section 2.1, this type of undersampling can create both theoretical and actual learning problems, as the undersampled data can have quite different distributional properties. We thus described and explored (with simulated data) generalizations of existing time series learning algorithms to discover measurement timescale structure from undersampled time series data.

For continuous-valued data, the generalization of Granger Causality clearly outperformed the other algorithms. The key to its success is almost certainly its focus on information gain, which is quite robust to the types of unusual distributions that can result from undersampling. Moreover, the generalized GC algorithm conducts fewer statistical tests, so is quite fast. At the same time, the statistical tests that it does perform can be very high-order, as they condition on $\mathcal{O}(|\mathbf{X}|)$ variables. Since high-order independence tests can be very unreliable for discrete-valued data, we expect that the generalized GC algorithm would not be the best choice for such data.

This paper provides the first benchmark results for structure learning algorithms at the measurement timescale applied to undersampled timeseries data. Some of these algorithms had previously been employed in other papers, but without careful examination of their measurement timescale performance. We obtained reasonably good results for some of the algorithms. Perhaps more importantly, the SVAR estimation and GC algorithms both learned measurement timescale structures that led to low error rates for *causal* timescale structure search. Moreover, the causal timescale structure search provided further regularization benefits for the measurement timescale structure search.

Various open problems remain, including the development of more generalized algorithms for discrete-valued data. The relatively high omission error rates are also cause for some concern, as the output graphs were almost always overly sparse. Some of these errors may be unavoidable, given that causal connections at the measurement timescale will almost always be weaker than those at the causal timescale. That is, $\mathcal{G}^u$ edges may just be harder to discover. Nonetheless, we are exploring algorithmic variations that allow the user to "tune" the algorithm for the desired trade-off between omission and comission errors.

Acknowledgements

Thanks to three anonymous reviewers for helpful comments. DD was supported by NSF IIS-1318815 & NIH U54HG008540 (from the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative). SMP was supported by NSF IIS-1318759 & NIH R01EB006841. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] David Maxwell Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2003.

[2] David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, volume 1, pages 1–10, 2013.

[3] S. Demiralp and K.D. Hoover. Searching for the causal structure of a vector autoregression*. *Oxford Bulletin of Economics and statistics*, 65(s1):745–767, 2003.

[4] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from sub-

sampled data. In *Proc. ICML*, pages 1898–1906, 2015.

[5] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[6] Antti Hyttinen, Sergey Plis, Matti Jrvisalo, Frederick Eberhardt, and David Danks. *Causal Discovery from Subsampled Time Series Data by Constraint Optimization*, volume 52. 8 2016.

[7] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2007.

[8] A. Moneta, N. Chlaß, D. Entner, and P. Hoyer. Causal search in structural vector autoregressive models. In *Journal of Machine Learning Research: Workshop and Conference Proceedings, Causality in Time Series (Proc. NIPS2009 Mini-Symposium on Causality in Time Series)*, volume 12, pages 95–114, 2011.

[9] M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Annals of Statistics*, 1:763765, 1973.

[10] Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems 28*, pages 1–9. Curran Associates, Inc., 2015.

[11] Sergey Plis, David Danks, and Jianyu Yang. Mesochronal structure learning. In *Proceedings of the Thirty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-15)*, Corvallis, Oregon, 2015. AUAI Press.

[12] Kristopher J Preacher and Edgar C Merkle. The problem of model selection uncertainty in structural equation modeling. *Psychological methods*, 17(1): 1, 2012.

[13] Anil K Seth, Paul Chorley, and Lionel C Barnett. Granger causality analysis of fmri bold signals is invariant to hemodynamic convolution but not downsampling. *Neuroimage*, 65:540–555, 2013.

[14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Springer, 1993.