# Causal Discovery from Temporally Aggregated Time Series

**Mingming Gong**[*†], **Kun Zhang**[†], **Bernhard Schölkopf**[‡], **Clark Glymour**[†], **Dacheng Tao**[♯]
[*]Centre for Artificial Intelligence, FEIT, University of Technology Sydney, NSW, Australia
[†]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
[‡]Max Plank Institute for Intelligent Systems, Tübingen, Germany
[♯]School of Information Technologies, FEIT, University of Sydney, NSW, Australia

## Abstract

Discovering causal structure of a dynamical system from observed time series is a traditional and important problem. In many practical applications, observed data are obtained by applying subsampling or temporally aggregation to the original causal processes, making it difficult to discover the underlying causal relations. Subsampling refers to the procedure that for every $k$ consecutive observations, one is kept, the rest being skipped, and recently some advances have been made in causal discovery from such data. With temporal aggregation, the local averages or sums of $k$ consecutive, non-overlapping observations in the causal process are computed as new observations, and causal discovery from such data is even harder. In this paper, we investigate how to recover causal relations at the original causal frequency from temporally aggregated data when $k$ is known. Assuming the time series at the causal frequency follows a vector autoregressive (VAR) model, we show that the causal structure at the causal frequency is identifiable from aggregated time series if the noise terms are independent and non-Gaussian and some other technical conditions hold. We then present an estimation method based on non-Gaussian state-space modeling and evaluate its performance on both synthetic and real data.

## 1 INTRODUCTION

Causal modeling (Spirtes et al., 2001; Pearl, 2000) of time series data has been widely applied in many fields such as econometrics (Ghysels et al., 2016), neuroscience (Zhou et al., 2014), and climate science (Van Nes et al., 2015). Classical causal discovery approaches, e.g., Granger causality test (Granger, 1969), usually assume that the data measurement frequency matches the true causal frequency of the underlying physical process. However, since the true causal frequency is usually unknown, the time series data are often measured at the frequency lower than the causal frequency. For example, some econometric indicators such as GDP and non-farm payroll are usually recorded at quarterly and monthly scales. Causal interactions between the processes, however, may take place at the weekly or fortnightly scales (Ghysels et al., 2016). In neuroscience, imaging technologies have relatively low temporal resolutions, while many high frequency neuronal interactions are important for understanding neuronal dynamics (Zhou et al., 2014). In these situations, the available observations have a lower resolution than the underlying causal process.

There are two typical schemes to generate low-resolution or low-frequency data from high-frequency ones (Silvestrini & Veredas, 2008; Marcellino, 1999). One is by subsampling: for every $k$ consecutive observations, one is kept, the rest being skipped. The other is temporally aggregation, i.e., taking the local averages or sums of $k$ consecutive, non-overlapping observations from the underlying causal process as new observations. For instance, the time series of interest, money supply, and temperature are usually obtained by subsampling; in contrast, the U.S. nominal GDP was obtained by aggregation – it refers to a total number of dollars spent over a time period.

Numerous contributions have been made on analyzing the effects of the above two schemes to generate low-resolution data on the properties of the time series such as estimated causal relations and exogeneity (Tiao, 1972; Weiss, 1984; Granger, 1987; Marcellino, 1999; Rajaguru & Abeysinghe, 2008). These studies found that temporal aggregation can lead to errors in the estimated causal relations if not properly addressed. For example, Breitung & Swanson (2002) examined the impact of temporal aggregation on Granger causality in vector autoregressive (VAR) models and found that the results of Granger causal analysis heavily depend on temporal aggregation.

Recovering the high frequency causal relations from temporally aggregated data is a very hard problem due to information loss in the aggregation process. A classical way to

discover high frequency causal relations from temporally aggregated data is to first disaggregate the low frequency time series to high frequency ones and then apply standard causal discovery methods on the disaggregated data. Temporal disaggregation of low resolution time series has been extensively studied in the econometric and statistical literature (Boot et al., 1967; Stram & Wei, 1986; Harvey & Chung, 2000; Moauro & Savio, 2005; Proietti, 2006), which is clearly an even harder problem than discovering causal relations.

Recently, a set of methods have been proposed to estimate the causal structure at the causal frequency from *subsampled* data without resorting to disaggregation techniques (Hyttinen et al., 2016; Gong et al., 2015; Plis et al., 2015a; Danks & Plis, 2013). Plis et al. (2015a,b) first inferred the causal structure from the subsampled data, and then searched for the causal structure at the causal frequency from the causal structure inferred in the first step. Based on this framework, Hyttinen et al. (2016) proposed a much faster inference method using a general purpose Boolean constraint solver. Gong et al. (2015) proposed a model-based approach and examined the identifiability of the underlying vector autoregressive model (VAR) at the causal frequency from subsampled time series. They showed that the causal transition matrix is identifiable if the noise process is non-Gaussian. This work was recently extended to mixed frequency data by structural VAR modeling (Tank et al., 2017) . However, how to estimate causal relations from aggregated data still remains unknown.

Compared to subsampling, *temporal aggregation* is perhaps more widely used to produce low-resolution time series, especially in economics and finance. However, the effect of temporal aggregation is more complex, and accordingly it is technically more difficult to recover the underlying causal relations from such data. Specifically, because the noise terms are generated by a larger number of independent components and thus the mixing matrix contains a more complicated structure, the estimation is both statistically and computationally harder.

The objective of this paper is to seek a possible solution to this problem, by studying the theoretical identifiability of the underlying causal relations and developing a practical causal discovery algorithm. Following (Gong et al., 2015), we assume that the high-frequency data follow a VAR model, the error terms are non-Gaussian, and there are no confounders (Geiger et al., 2015). We show that the original causal relation can be estimated from the aggregated data with known $k$, under a set of technical conditions.

Moreover, we propose an estimation method based on non-Gaussian state-space modeling of the aggregated data. Since the exact inference in the non-Gaussian state-space model is intractable, we estimate the model parameters using the particle stochastic approximation EM (PSAEM) algorithm (Lindsten, 2013; Svensson et al., 2014), which combines the efficient conditional particle filter with ancestor sampling (CPF-AS) (Lindsten et al., 2014) with the stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999). Interestingly, in the extreme case where the aggregation factor $k$ becomes larger and larger, we show that the observed time series will become independent and identically distributed (*i.i.d.*), and we study to what extent the underlying time-delayed causal relations can be recovered from the instantaneous dependence in the observed data.

## 2 EFFECT OF TEMPORAL AGGREGATION

In the linear case, Granger causal analysis Granger (1969) can be done by fitting the following first-order VAR model (Sims, 1980):

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{e}_t, \tag{1}$$

where $\mathbf{x}_t = (x_{t,1}, x_{t,2}, ..., x_{t,n})^\intercal$ is the vector of the observed data, $\mathbf{e}_t = (e_{t,1}, ..., e_{t,n})^\intercal$ is the temporally and contemporaneously independent noise process, and $\mathbf{A}$ is the causal transition matrix containing temporal causal relations.

### 2.1 WITH A FINITE $k$

Gong et al. (2015) studied causal discovery from subsampled data. With subsampling, the observations $\tilde{\mathbf{x}}^s_{1:T} \triangleq (\tilde{\mathbf{x}}^s_1, \tilde{\mathbf{x}}^s_2, ..., \tilde{\mathbf{x}}^s_T) = (\mathbf{x}_1, \mathbf{x}_{1+k}, ..., \mathbf{x}_{1+(T-1)k})$ follow

$$\begin{aligned}\tilde{\mathbf{x}}^s_t &= \mathbf{A}\mathbf{x}_{1+(t-1)k-1} + \mathbf{e}_{1+(t-1)k} \\ &= \mathbf{A}(\mathbf{A}\mathbf{x}_{1+(t-1)k-2} + \mathbf{e}_{1+(t-1)k-1}) + \mathbf{e}_{1+(t-1)k} \\ &= ... \\ &= \mathbf{A}^k\tilde{\mathbf{x}}^s_{t-1} + \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{1+(t-1)k-l},\end{aligned} \tag{2}$$

which turns out to be a VAR model with temporally independent and contemporaneously dependent noise process. They demonstrated that it is possible to identify the high-resolution causal relation $\mathbf{A}$ from the low-resolution observations $\tilde{\mathbf{x}}^s_{1:T}$ if the noise terms are non-Gaussian.

In this paper, we are concerned with the temporally aggregated data $\tilde{\mathbf{x}}_{1:T} \triangleq (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T)$, which are obtained by taking the average (or sum) of every non-overlapping $k$ points, i.e., $\tilde{\mathbf{x}}_t = \frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_{i+(t-1)k}$, where

$$\mathbf{x}_{i+(t-1)k} = \mathbf{A}^k\mathbf{x}_{i+(t-2)k} + \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{i+(t-1)k-l}.$$

Taking the average of the above equation over $i = 1, 2, \ldots, k$, we have

$$\tilde{\mathbf{x}}_t = \frac{\mathbf{A}^k}{k}\sum_{i=1}^{k}\mathbf{x}_{i+(t-2)k} + \frac{1}{k}\sum_{i=1}^{k}\Big(\sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{i+(t-1)k-l}\Big)$$

$$= \mathbf{A}^k \tilde{\mathbf{x}}_{t-1} + \frac{1}{k}\Big( \sum_{m=0}^{k-1} (\sum_{n=0}^{m} \mathbf{A}^n) \mathbf{e}_{tk-m} + \sum_{m=1}^{k-1} (\sum_{n=m}^{k-1} \mathbf{A}^n) \mathbf{e}_{(t-1)k-m+1} \Big), \quad (3)$$

which is a vector autoregressive-moving-average (VARMA) model with one autoregressive term and two moving-average terms:

$$\begin{aligned} \tilde{\mathbf{x}}_t &= \mathbf{A}^k \tilde{\mathbf{x}}_{t-1} + \vec{\mathbf{e}}_t \\ &= \mathbf{A}^k \tilde{\mathbf{x}}_{t-1} + \mathbf{M}_0 \boldsymbol{\epsilon}_t + \mathbf{M}_1 \boldsymbol{\epsilon}_{t-1}, \end{aligned} \quad (4)$$

where $\boldsymbol{\epsilon}_t = \frac{1}{k}[\mathbf{e}_{tk}, \mathbf{e}_{tk-1}, \ldots, \mathbf{e}_{(t-1)k+1}]^\intercal$, $\mathbf{M}_0 = [\mathbf{I}, \mathbf{I} + \mathbf{A}, \ldots, \sum_{n=0}^{k-1} \mathbf{A}^n]$, and $\mathbf{M}_1 = [\sum_{n=1}^{k-1} \mathbf{A}^n, \ldots, \mathbf{A}^{k-1}, \mathbf{0}]$. Here, $\mathbf{I}$ represents the $n \times n$ identity matrix, and $\mathbf{0}$ represents the $n \times n$ zero matrix. We call $(\mathbf{A}, \mathbf{e}, k)$ the representation of the $k$-th order aggregated time series $\tilde{\mathbf{x}}$. Clearly $\mathbf{A}$ cannot be recovered by simply fitting a VAR model on $\tilde{\mathbf{x}}_t$, as done by Granger causal analysis. Even if we using VARMA modeling, we are only guaranteed to identify $\mathbf{A}^k$ instead of the original $\mathbf{A}$. In Section 3, we will show under what conditions can we identify the causal relation $\mathbf{A}$ at the causal frequency from the aggregated time series $\tilde{\mathbf{x}}_{1:T}$.

### 2.2 WHEN $k \to \infty$

Interestingly, causal discovery from aggregated data with a large aggregation factor $k$ seems to have a wide range of applications. For instance, in the stock market, the causal influences between stocks take place very quickly (as indicated by the efficient market hypothesis), but we usually work with low-frequency data such as daily returns. The daily return is the sum of high-frequency returns within the same day. Discovering the causal interactions between stocks from their daily returns then become a problem of causal discovery from aggregated data with a large $k$.

When the aggregation factor $k$ is very large, $\vec{\mathbf{e}}_t$ becomes a mixture of numerous independent components. Fortunately, we can use a simple model to approximate the generating process of the aggregated data. From (1), we have

$$\sum_{i=1}^{k} \mathbf{x}_{i+(t-1)k+1} = \mathbf{A} \sum_{i=1}^{k} \mathbf{x}_{i+(t-1)k} + \sum_{i=1}^{k} \mathbf{e}_{i+(t-1)k+1},$$

that is,

$$\begin{aligned} \tilde{\mathbf{x}}_t &= \frac{1}{k} \sum_{i=1}^{k} \mathbf{x}_{i+(t-1)k+1} \\ &= \mathbf{A}\Big[\frac{1}{k} \sum_{i=1}^{k} \mathbf{x}_{i+(t-1)k+1} - \frac{1}{k}(\mathbf{x}_{1+tk} - \mathbf{x}_{1+(t-1)k})\Big] + \\ &\qquad \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{i+(t-1)k+1}. \end{aligned}$$

Denote by $\bar{\mathbf{e}}_t$ the error term above, i.e., $\bar{\mathbf{e}}_t = \frac{1}{k} \sum_{i=1}^{k} \mathbf{e}_{i+(t-1)k+1}$. Note that $\bar{\mathbf{e}}_t$ has contemporaneous independent components. Since

$$\frac{1}{k}(\mathbf{x}_{1+tk} - \mathbf{x}_{1+(t-1)k}) \to 0, \text{ as } k \to \infty,$$

we have

$$\tilde{\mathbf{x}}_t = \mathbf{A}\tilde{\mathbf{x}}_t + \bar{\mathbf{e}}_t, \quad (5)$$

as $k \to \infty$. This is a linear instantaneous causal model for the components of $\tilde{\mathbf{x}}_t$ because the components of the total error term, $\bar{\mathbf{e}}_t$, are still contemporaneously independent. When the error terms are non-Gaussian, it has the same form as the Linear, Non-Gaussian Model (LiNG) (Lacerda et al., 2008); when the causal relations are further assumed to be acyclic, it follows the form of the Linear, Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006). The difference is that in LiNG or LiNGAM, the self-loop influences, $A_{ii}$, are assumed to be zero. We will also investigate the identifiably of $\mathbf{A}$ in this case in Section 3.

## 3 IDENTIFIABILITY OF CAUSAL RELATIONS IN A

We investigate the identifiability of the high-resolution causal transition matrix $\mathbf{A}$ from the aggregated time time series $\tilde{\mathbf{x}}_{1:T}$. In other words, suppose $\tilde{\mathbf{x}}$ also admits another representation $(\mathbf{A}', \mathbf{e}', k)$, we aim to see whether it is always the case that $\mathbf{A} = \mathbf{A}'$ as the sample size $T \to \infty$. If the noise terms follow the Gaussian distribution, $\mathbf{A}$ is usually not identifiable (Palm & Nijman, 1984). Recently, it has been shown that $\mathbf{A}$ is identifiable from subsampled time series if the noise terms are non-Gaussian (Gong et al., 2015). However, this does not give rise to the identifiability of $\mathbf{A}$ from aggregated time series – the latter is much more difficult to see as the aggregated model described in (3) has a more complicated structure. Here, we show that, in the exact model (3), $\mathbf{A}$ is identifiable from the aggregated data under appropriate conditions; furthermore, as $k \to \infty$, the approximate model (5) holds, and $\mathbf{A}$ is partially identifiable from the aggregated data, but the identification procedure is computationally much more efficient.

First, we will show that $\mathbf{A}^k$ can be identified by fitting the VARMA model (4). We make the following assumption.

*A*1. At least one of the $\tau$-step ($\tau \geq 2$) delayed cross covariance matrices of $\tilde{\mathbf{x}}_t$, $\mathbb{E}[\tilde{\mathbf{x}}_{t-1}\tilde{\mathbf{x}}_{t-\tau}^\intercal]$, is invertible.

Since $\boldsymbol{\epsilon}_t$ is both temporally and contemporaneously independent, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\epsilon}_{t-1}$ are independent of $\tilde{\mathbf{x}}_{t-\tau}$, which implies that

$$\mathbb{E}[\boldsymbol{\epsilon}_t \tilde{\mathbf{x}}_{t-\tau}^\intercal] = 0, \ \mathbb{E}[\boldsymbol{\epsilon}_{t-1} \tilde{\mathbf{x}}_{t-\tau}^\intercal] = 0.$$

Multiplying both sides of (4) from the right side by $\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}$ and taking the expectation, we have

$$\begin{aligned}&\mathbb{E}[\tilde{\mathbf{x}}_t\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}]\\&= \mathbf{A}^k\mathbb{E}[\tilde{\mathbf{x}}_{t-1}\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}] + \mathbf{M}_0\mathbb{E}[\boldsymbol{\epsilon}_t\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}] + \mathbf{M}_1\mathbb{E}[\boldsymbol{\epsilon}_{t-1}\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}]\\&= \mathbf{A}^k\mathbb{E}[\tilde{\mathbf{x}}_{t-1}\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}]. \qquad (6)\end{aligned}$$

Under the assumption $A_1$, we can first see that $\mathbf{A}^k$ is identifiable:

$$\mathbf{A}^k = \mathbb{E}[\tilde{\mathbf{x}}_t\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}] \cdot \mathbb{E}[\tilde{\mathbf{x}}_{t-1}\tilde{\mathbf{x}}_{t-\tau}^{\mathsf{T}}]^{-1}. \qquad (7)$$

### 3.1 IDENTIFIABILITY WITH FINITE $k$

Substituting the above equation into (3), one can then find $\vec{\mathbf{e}}_t$, which is defined to be linear mixtures of $(2k-1)$ noise terms, i.e., $\mathbf{e}_{tk}$, $\mathbf{e}_{tk-1}$, ..., and $\mathbf{e}_{(t-2)k+2}$. In the following, we will concentrate on the identifiablity of $\mathbf{A}$ from $\vec{\mathbf{e}}$.

Let

$$\mathbf{H} \triangleq [\mathbf{I}, \mathbf{I}+\mathbf{A}, \ldots, \sum_{l=0}^{k-1}\mathbf{A}^l, \sum_{l=1}^{k-1}\mathbf{A}^l, \ldots, \sum_{l=k-2}^{k-1}\mathbf{A}^l, \mathbf{A}^{k-1}]. \qquad (8)$$

The error terms in (3) correspond to the following mixing procedure of random vectors:

$$\begin{aligned}&\vec{\mathbf{e}} = \mathbf{H}\tilde{\mathbf{e}}, \text{ where} \qquad (9)\\&\tilde{\mathbf{e}} = (e_1^{(0)}, ..., e_n^{(0)}, e_1^{(1)}, ..., e_n^{(1)}, ..., e_1^{(2k-2)}, ..., e_n^{(2k-2)})^{\mathsf{T}}.\end{aligned}$$

Here, $\mathbf{e}^l = (e_1^{(l)}, ..., e_n^{(l)})^{\mathsf{T}}$ together with the time index $t$ represent $\mathbf{e}_{tk-l}$. The components of $\tilde{\mathbf{e}}$ are independent, and for each $i$, $e_i^{(l)}$, $l = 0, ..., 2k-2$, have the same distribution $p_{e_i}$. Under the condition that $p_{e_i}$ is non-Gaussian for each $i$, $\mathbf{H}$ can be estimated up to the permutation and scaling indeterminacies (including the sign indeterminacy) of the columns, as given in the following proposition.

**Proposition 1.** *Suppose that all $p_{e_i}$ are non-Gaussian. Given $k$ and $\tilde{\mathbf{x}}_{1:T}$ generated according to (3), $\mathbf{H}$ can be determined up to permutation and scaling of columns.*

For the proof of Proposition 1, please refer to (Gong et al., 2015).

We make the following assumptions on the underlying dynamic process (1) and the distributions $p_{e_i}$, and then we have the identifiability result for the causal transition matrix $\mathbf{A}$.

A2. The system is stable, in that all eigenvalues of $\mathbf{A}$ have modulus smaller than one.

A3. The distributions $p_{e_i}$ are different for different $i$ after re-scaling by any non-zero scale factor, their characteristic functions are all analytic (or they are all non-vanishing), and none of them has an exponent factor with a polynomial of degree at least 2.

The following identifiability result on $\mathbf{A}$ states that in various situations, $\mathbf{A}$ for the original high frequency data is fully identifiable.

**Theorem 1.** *Suppose all of $e_{it}$ are non-Gaussian, and that the data $\tilde{\mathbf{x}}_t$ are generated by (3) and that it also admits another $k$th order aggregation representation $(\mathbf{A}', \mathbf{e}', k)$. Let assumptions A1 and A2 hold. When the number of observed data points $T \to \infty$, the following statements are true.*

*(i) $\mathbf{A}'$ can be represented as $\mathbf{A}' - \mathbf{I} = (\mathbf{A} - \mathbf{I})\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with $1$ or $-1$ on its diagonal. If we constrain all the self influences, represented by diagonal entries of $\mathbf{A}$ and $\mathbf{A}'$, to be no greater than 1, then $\mathbf{A}' = \mathbf{A}$.*

*(ii) If each $p_{e_i}$ is asymmetric, we have $\mathbf{A}' = \mathbf{A}$.*

A complete proof of Theorem 1 can be found in Section 6.

### 3.2 IDENTIFIABILITY AS $k \to \infty$

We have shown that $\mathbf{A}$ is identifiable from aggregated data (3) when $k$ is finite. However, when $k$ becomes larger, estimating $\mathbf{A}$ will encounter more difficulty because more independent components in (9) are involved. When $k = \infty$, it is not necessary for Proposition 1, as well as Theorem 1, to hold, because $\vec{\mathbf{e}}$ in (9) is the mixture of an infinite number of independent components.

Interestingly, as $k \to \infty$, $\tilde{\mathbf{x}}_t$ follows an instantaneous causal model in the i.i.d. case, as shown in (5). We will then answer the following two questions. In this case, can we still estimate $\mathbf{A}$ from aggregated data? If we can, is there an efficient procedure to do so?

Equation (5) implies $(\mathbf{I} - \mathbf{A})\tilde{\mathbf{x}}_t = \bar{\mathbf{e}}_t$. That is, applying the linear transformation $(\mathbf{I} - \mathbf{A})$ on $\tilde{\mathbf{x}}_t$ produces independent components, as components of $\bar{\mathbf{e}}_t$. This can be achieved by the independent component analysis (ICA) procedure (Hyvärinen et al., 2001), and $(\mathbf{I} - \mathbf{A})$ can be estimated up to row scaling and permutation indeterminacies. We then have the following observations.

First, the diagonal entries of $\mathbf{A}$, $A_{ii}$, which represent the self influences or "self-loops" of the time-delayed causal relations, cannot be determined (Lacerda et al., 2008). (Here we have assumed $A_{ii} \neq 1$.) This is because the scale of each row of $(\mathbf{I} - \mathbf{A})$ is unknown, and so is $(1 - A_{ii})$.

Let $\mathbf{D_A}$ be the diagonal matrix with $A_{11}, A_{22}, ..., A_{nn}$ on its diagonal. Equation (5) is equivalent to

$$\begin{aligned}&(\mathbf{I} - \mathbf{D_A})\tilde{\mathbf{x}}_t = (\mathbf{A} - \mathbf{D_A})\tilde{\mathbf{x}}_t + \bar{\mathbf{e}}_t\\\Rightarrow &\tilde{\mathbf{x}}_t = \underbrace{(\mathbf{I} - \mathbf{D_A})^{-1}(\mathbf{A} - \mathbf{D_A})}_{\triangleq \mathbf{A}^{NoSelfLoop}}\tilde{\mathbf{x}}_t + (\mathbf{I} - \mathbf{D_A})^{-1}\bar{\mathbf{e}}_t. \qquad (10)\end{aligned}$$

Secondly, suppose there is no feedback loop between the processes after removing the self-loops, meaning that $(\mathbf{A} - \mathbf{D_A})$ can be permuted to a strictly lower-triangular matrix by equal row and column permutations. According to the LiNGAM model, which assumes these is no self-loop, $(\mathbf{I} - \mathbf{D_A})^{-1}(\mathbf{A} - \mathbf{D_A})$ in (10) can be uniquely estimated (Shimizu et al., 2006). In other words, if one applies LiNGAM analysis on $\tilde{\mathbf{x}}_t$, the estimated causal coefficients from th $i$th variable to the $j$th variable is actually $(1 - A_{jj})^{-1}A_{ji}$. From this we can see whether $A_{ji}$ is zero or not; furthermore, if the self-loops $A_{jj}$ are given by prior knowledge, then $\mathbf{A}$ is fully identifiable.

Thirdly, suppose there exist feedback loops between the processes after removing the self-loops. In this case, $(\mathbf{A} - \mathbf{D_A})$ cannot be permuted to a strictly lower-triangular matrix by equal row and column permutations. The identifiability of $\mathbf{A}$ in (5) has been studied by Lacerda et al. (2008): suppose the feedback loops are disjoint, although in theory there are multiple solutions to $\mathbf{A}^{NoSelfLoop}$, the most stable solution (the product of the coefficients in $\mathbf{A}^{NoSelfLoop}$ along each loop is minimized) is unique.

# 4 ESTIMATING THE CAUSAL RELATIONS FROM AGGREGATED DATA

In this section, we present the algorithm to estimate $\mathbf{A}$ from aggregated data with finite $k$. Clearly, the larger $k$, the more difficult it is to estimate $\mathbf{A}$ from aggregated data. Therefore, when $k$ is relatively large (say, larger than 6), we advocate the methods given in Section 3.2 to partially estimate $\mathbf{A}$.

Since the identifiability of $\mathbf{A}$ from aggregated data relies on the non-Gaussianity of the error terms, we use Gaussian mixtures to represent their distributions. It is natural to do parameter estimation with the Expectation-Maximization algorithm, which, unfortunately, involves a large number of Gaussian components. To avoid this issue, we propose to use the Stochastic Approximation EM (SAEM) algorithm, as a variant of EM, and further resort to conditional particle filtering with ancestor sampling (CPF-AS) to achieve computational efficiency.

## 4.1 STATE-SPACE MODELING

We can consider (3) as a special *state-space* model:

$$\begin{aligned}\tilde{\mathbf{x}}_t &= \mathbf{A}^k\tilde{\mathbf{x}}_{t-1} + \mathbf{H}\tilde{\mathbf{e}}_t \\ &= \mathbf{A}^k\tilde{\mathbf{x}}_{t-1} + \begin{bmatrix}\mathbf{I} & \sum_{l=0}^{1}\mathbf{A}^l & \dots & \sum_{l=0}^{k-1}\mathbf{A}^l\end{bmatrix} \\ &\quad \begin{bmatrix}\sum_{l=1}^{k-1}\mathbf{A}^l & \dots & \sum_{l=k-2}^{k-1}\mathbf{A}^l & \mathbf{A}^{k-1}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{e}}_t^{(0)} \\ \tilde{\mathbf{e}}_t^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_t^{(2k-2)}\end{bmatrix},\end{aligned} \tag{11}$$

where

$$\begin{bmatrix}\tilde{\mathbf{e}}_t^{(0)} \\ \tilde{\mathbf{e}}_t^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_t^{(2k-2)}\end{bmatrix} = \mathbf{F}\begin{bmatrix}\tilde{\mathbf{e}}_{t-1}^{(0)} \\ \tilde{\mathbf{e}}_{t-1}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{t-1}^{(2k-2)}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{tk} \\ \mathbf{e}_{tk-1} \\ \vdots \\ \mathbf{e}_{(t-1)k+1} \\ \mathbf{0}_{(nk-n)\times 1}\end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix}\mathbf{0}_{nk\times(nk-n)} & \mathbf{0}_{nk\times(nk)} \\ \mathbf{I}_{(nk-n)\times(nk-n)} & \mathbf{0}_{(nk-n)\times(nk)}\end{bmatrix},$$

$\tilde{\mathbf{e}}_t^{(l)} = \mathbf{e}_{tk-l}$, and the noise terms $\mathbf{e}_{tk}, \mathbf{e}_{tk-1}, ..., \mathbf{e}_{(t-1)k+1}$ share the same distribution for the same channel and are mutually independent. Since the non-Gaussianity is essential to the identifiability of $\mathbf{A}$, we use a Gaussian mixture model to represent each channel of noise term $\mathbf{e}$, i.e., $p_{e_i} = \sum_{c=1}^{m} w_{i,c}\mathcal{N}(e_i|\mu_{i,c}, \sigma_{i,c}^2)$, where $w_{i,c} \geq 0$, and $\sum_{c=1}^{m}\pi_{i,c} = 1$, for $i = 1, ..., n$. Correspondingly, each channel of $\tilde{\mathbf{e}}$ is also represented by a Gaussian mixture model.

We aim to estimate the parameters $\mathbf{A}$ and the noise terms (if necessary) in the above state-space model. We introduce the additional latent variable $\tilde{\mathbf{z}}_t = (\tilde{z}_{t,1}, ..., \tilde{z}_{t,nk})^\intercal$, in which $\tilde{z}_{t,j} \in \{1, \dots, m\}$, to model the distribution of noise terms $\tilde{\mathbf{e}}_t$ by Gaussian mixture models. The joint distribution of the state-space model (11) over both observed and unobserved variables is given by

$$\begin{aligned}&p(\tilde{\mathbf{x}}_{1:T}, \tilde{\mathbf{e}}_{1:T}, \tilde{\mathbf{z}}_{1:T}) = p(\tilde{\mathbf{z}}_1)p(\tilde{\mathbf{e}}_1|\tilde{\mathbf{z}}_1)p(\tilde{\mathbf{x}}_1|\tilde{\mathbf{e}}_1) \\ &\prod_{t=2}^{T} p(\tilde{\mathbf{z}}_t)p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{e}}_{t-1}, \tilde{\mathbf{z}}_t)p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{e}}_t).\end{aligned} \tag{12}$$

The distributions in (12) are specified as follows:

$$p(\tilde{\mathbf{z}}_t) = \prod_{j=1}^{nk} p(\tilde{z}_{t,j}) = \prod_{j=1}^{nk}\tilde{\pi}_{j,\tilde{z}_{t,j}}, \tag{13a}$$

$$p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{e}}_{t-1}, \tilde{\mathbf{z}}_t) = \mathcal{N}(\tilde{\mathbf{e}}_t|\mathbf{F}\tilde{\mathbf{e}}_{t-1} + \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t), \tag{13b}$$

$$p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{e}}_t, \tilde{\mathbf{x}}_{t-1}) = \mathcal{N}(\tilde{\mathbf{x}}_t|\mathbf{A}^k\tilde{\mathbf{x}}_{t-1} + \mathbf{H}\tilde{\mathbf{e}}_t, \boldsymbol{\Lambda}). \tag{13c}$$

Since there are no additional additive noise terms in the model, we fix $\boldsymbol{\Lambda}$ to a small value in our estimation algorithm for regularization. $\tilde{\boldsymbol{\mu}}_t$ is the conditional mean of $\tilde{\mathbf{e}}_t$, i.e., $\tilde{\boldsymbol{\mu}}_t = [\tilde{\mu}_{1,\tilde{z}_{t,1}}, \dots, \tilde{\mu}_{nk,\tilde{z}_{t,nk}}, \mathbf{0}_{1\times n(k-1)}]^\intercal$. $\tilde{\boldsymbol{\Sigma}}_t$ is a diagonal matrix containing the conditional variance parameters of $\tilde{\mathbf{e}}_t$, i.e., $\tilde{\boldsymbol{\Sigma}}_t = \text{diag}([\tilde{\sigma}^2_{1,\tilde{z}_{t,1}}, \dots, \tilde{\sigma}^2_{nk,\tilde{z}_{t,nk}}, \mathbf{0}_{1\times n(k-1)}])$. According to the structure of $\tilde{\mathbf{e}}$, the parameters $\tilde{\pi}_{j,\tilde{z}_{t,j}}$, $\tilde{\mu}_{j,\tilde{z}_{t,j}}$, and $\tilde{\sigma}_{j,\tilde{z}_{t,j}}$ are controlled by the parameters of $\mathbf{e}$, i.e., $\tilde{\pi}_{i+nl,c} = \pi_{i,c}$, $\tilde{\mu}_{i+nl,c} = \mu_{i,c}$, and $\tilde{\sigma}_{i+nl,c} = \sigma_{i,c}$, for $i = 1, ..., n,\ l = 0, ..., k-1$, and $c = 1, ..., m$.

### 4.2 STOCHASTIC APPROXIMATION EM

The expectation-maximization (EM) algorithm is usually adopted to find the maximum likelihood estimation of the parameters in a probabilistic model with unobserved variables. We can estimate the parameters $\theta = (\mathbf{A}, w_{i,c}, \mu_{i,c}, \sigma_{i,c})$ in (12) using the EM algorithm that iteratively maximizes the lower bound of the marginal log-likelihood $p_\theta(\tilde{\mathbf{x}}_{1:T}) = \log \sum_{\tilde{\mathbf{z}}_{1:T}} \int p_\theta(\tilde{\mathbf{x}}_{1:T}, \tilde{\mathbf{e}}_{1:T}, \tilde{\mathbf{z}}_{1:T}) d_{\tilde{\mathbf{e}}_{1:T}}$. In the E-step, at the $k$-th iteration, given the parameters $\theta_{k-1}$ estimated from the $(k-1)$-th iteration, the EM algorithm firstly computes the posterior distribution $p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{x}}_{1:T})$ and then computes the lower bound $\mathcal{Q}(\theta, \theta_{k-1}) = \sum_{\tilde{\mathbf{z}}_{1:T}} \int p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{x}}_{1:T}) \log p_\theta(\tilde{\mathbf{x}}_{1:T}, \tilde{\mathbf{e}}_{1:T}, \tilde{\mathbf{z}}_{1:T}) d_{\tilde{\mathbf{e}}_{1:T}}$. In the M-step, the parameters are updated as $\theta_k = \arg\max_\theta \mathcal{Q}(\theta, \theta_{k-1})$.

However, we note that the number of Gaussian mixtures in the posterior distribution grows exponentially with the dimension of the time series, $n$, the number of aggregation factor, $k$, and the duration of time series $T$. Therefore, computing the exact posterior $p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{x}}_{1:T})$ and $\mathcal{Q}(\theta, \theta_{k-1})$ is intractable in this situation. A possible solution is to adopt the monte carlo EM (MCEM) algorithm (Wei & Tanner, 1990), which approximately calculates $\mathcal{Q}(\theta, \theta_{k-1})$ using samples drawn from the posterior distribution $p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{x}}_{1:T})$. However, MCEM makes inefficient use of generated samples, as it discards samples generated in the previous EM iterations. Therefore, a large number of sample points are required for each iteration, which is computationally expensive when the sampling method is complex.

To reduce the number of simulated sample points, we propose to use the stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999), which only requires of a single realization of the unobserved variables at each iteration. At the $k$-th iteration, the E-step and M-step are replaced by the following:

**E-step:** Generate a single sample point $\tilde{\mathbf{z}}_{1:T}[k]$ from the posterior $p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T} | \tilde{\mathbf{x}}_{1:T})$, and compute

$$\begin{aligned}\hat{\mathcal{Q}}_k(\theta) = (1-\gamma_k)\hat{\mathcal{Q}}_{k-1}(\theta) \\ + \gamma_k \int p_{\theta_{k-1}}(\tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{z}}_{1:T}[k], \tilde{\mathbf{x}}_{1:T}) \\ \log p_\theta(\tilde{\mathbf{x}}_{1:T}, \tilde{\mathbf{e}}_{1:T}, \tilde{\mathbf{z}}_{1:T}[k]) d_{\tilde{\mathbf{e}}_{1:T}}. \quad (14)\end{aligned}$$

**M-step:** Update parameters by $\theta_k = \arg\max_\theta \hat{\mathcal{Q}}_k(\theta)$.

In (14), $\{\gamma_k\}_{k=1}^\infty$ is a sequence of decreasing step sizes satisfying $\sum_{k=1}^\infty \gamma_k = \infty$ and $\sum_{k=1}^\infty \gamma_k^2 < \infty$. Here we use Rao-Blackwellization (Svensson et al., 2014) to avoid sampling $\tilde{\mathbf{e}}_{1:T}$ because it is analytically integrable. It has been shown in (Delyon et al., 1999) that the resulting sequence $\{\theta_k\}_{k \geq 1}$ will converge to a stationary point of $p_\theta(\tilde{\mathbf{x}}_{1:T})$ under weak assumptions.

### 4.3 CONDITIONAL PARTICLE FILTER WITH ANCESTOR SAMPLING

In our model, sampling from the posterior $p_{\theta_{k-1}}(\tilde{\mathbf{z}}_{1:T} | \tilde{\mathbf{x}}_{1:T})$ is usually performed using forward filter/backward simulator particle smoother, which typically requires a large number of particles to generate a smooth backward trajectory $\tilde{\mathbf{z}}_{1:T}[k]$. To reduce the number of required particles, we use the Markovian version of SAEM (Kuhn & Lavielle, 2004), which samples from a Markov kernel $\mathcal{M}_{\theta_{t-1}}$, leaving the posterior distribution invariant. Specifically, let $\tilde{\mathbf{z}}_{1:T}[k-1]$ be the previous draw from the Markov kernel, the current state is sampled by $\tilde{\mathbf{z}}_{1:T}[k] \sim \mathcal{M}_{\theta_{k-1}}(\cdot | \tilde{\mathbf{z}}_{1:T}[k-1])$. Following (Lindsten, 2013; Svensson et al., 2014), we construct the Markov kernel using Rao-Blackwellized conditional particle filter with ancestor sampling (RB-CPF-AS) (Lindsten et al., 2014), which was originally proposed for Gibbs sampling.

The machinery inside RB-CPF-AS resembles a standard particle filter, with two main differences: one particle trajectory is deterministically set to a reference trajectory $\tilde{\mathbf{z}}'_{1:T}$, and the ancestors of the reference trajectory are randomly chosen and stored during the algorithm execution. Algorithm 1 gives a brief description of the RB-CPF-AS algorithm. Let $\{\tilde{\mathbf{z}}^i_{1:t-1}, w^i_{t-1}\}_{i=1}^N$ be the approximation of $p_\theta(\tilde{\mathbf{z}}_{1:t-1} | \tilde{\mathbf{x}}_{1:t-1})$, RB-CPF-AS propagates this sample to time $t$ by introducing the auxiliary variables $\{a^i_t\}_{i=1}^N$, referred to as *ancestor indices*. To generate $\tilde{\mathbf{z}}^i_t$ for the first $N-1$ particle trajectories, we first sample the ancestor index according to $P(a^i_t = j) \propto w^j_{t-1}$, and then sample $\tilde{\mathbf{z}}^i_{t,j}$ according to $p_\theta(\tilde{\mathbf{z}}_t | \tilde{\mathbf{z}}^{a^i_t}_{t-1}) = \prod_{j=1}^{nk} p_\theta(\tilde{z}_{t,j})$. The first $N-1$ trajectories are then augmented as $\tilde{\mathbf{z}}^i_{1:t} = \{\tilde{\mathbf{z}}^{a^i_t}_{1:t-1}, \tilde{\mathbf{z}}^i_t\}$. The $N$-th particle is set to the reference particle, $\tilde{\mathbf{z}}^N_t = \tilde{\mathbf{z}}'_t$, and the ancestor index $a^N_t$ is sampled according to

$$\begin{aligned}&P(a^N_t = i) \\ &\propto p_\theta(\tilde{\mathbf{z}}^i_{1:t-1} | \tilde{\mathbf{z}}'_{t:T}, \tilde{\mathbf{x}}_{1:T}) \\ &\propto p_\theta(\tilde{\mathbf{x}}_{t:T}, \tilde{\mathbf{z}}'_{t:T} | \tilde{\mathbf{x}}_{1:t-1}, \tilde{\mathbf{z}}^i_{1:t-1}) p_\theta(\tilde{\mathbf{z}}^i_{1:t-1} | \tilde{\mathbf{x}}_{1:t-1}), \quad (15)\end{aligned}$$

where $p_\theta(\tilde{\mathbf{z}}^i_{1:t-1} | \tilde{\mathbf{x}}_{1:t-1}) = w^i_{t-1}$ and

$$\begin{aligned}p_\theta(\tilde{\mathbf{x}}_{t:T}, \tilde{\mathbf{z}}'_{t:T} | \tilde{\mathbf{x}}_{1:t-1}, \tilde{\mathbf{z}}^i_{1:t-1}) \\ \propto |\mathbf{M}^i_{t-1}|^{-1/2} \exp(-\frac{1}{2}\eta^i_{t-1}). \quad (16)\end{aligned}$$

Conditioned on $\tilde{\mathbf{z}}^i_{1:T}$, we can calculate $p(\tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{z}}^i_{1:T}, \tilde{\mathbf{x}}_{1:T})$ using Kalman filter and Rauch-Tung-Striebel (RTS) smoother. The filtering, prediction, and smoothing PDFs are

$$p_\theta(\tilde{\mathbf{e}}_t | \tilde{\mathbf{z}}^i_{1:t}, \tilde{\mathbf{x}}_{1:t}) = \mathcal{N}(\tilde{\mathbf{e}}_t | \hat{\boldsymbol{\mu}}^i_{f,t}, \hat{\boldsymbol{\Sigma}}^i_{f,t}), \quad (17a)$$

$$p_\theta(\tilde{\mathbf{e}}_{t+1} | \tilde{\mathbf{z}}^i_{1:t}, \tilde{\mathbf{x}}_{1:t}) = \mathcal{N}(\tilde{\mathbf{e}}_{t+1} | \hat{\boldsymbol{\mu}}^i_{p,t}, \hat{\boldsymbol{\Sigma}}^i_{p,t}), \quad (17b)$$

$$p_\theta(\tilde{\mathbf{e}}_t | \tilde{\mathbf{z}}^i_{1:T}, \tilde{\mathbf{x}}_{1:T}) = \mathcal{N}(\tilde{\mathbf{e}}_t | \hat{\boldsymbol{\mu}}^i_{s,t}, \hat{\boldsymbol{\Sigma}}^i_{s,t}), \quad (17c)$$

respectively. Let $\|\mathbf{e}\|^2_{\mathbf{\Omega}} = \mathbf{e}^\intercal \mathbf{\Omega} \mathbf{e}$ and $\hat{\mathbf{\Sigma}}^i_{f,t} = \mathbf{\Gamma}^i_{f,t} \mathbf{\Gamma}^{i,\intercal}_{f,t}$. In (16),

$$\mathbf{M}^i_t = \mathbf{\Gamma}^{i,\intercal}_{f,t} \mathbf{\Omega}_t \mathbf{\Gamma}^i_{f,t} + \mathbf{I}, \tag{18a}$$

$$\eta^i_t = \|\hat{\boldsymbol{\mu}}^i_{f,t}\|^2_{\mathbf{\Omega}_t} - 2\boldsymbol{\lambda}^\intercal_t \hat{\boldsymbol{\mu}}^i_{f,t} - \|\mathbf{\Gamma}^{i,\intercal}_{f,t}(\boldsymbol{\lambda}_t - \mathbf{\Omega}_t \hat{\boldsymbol{\mu}}^i_{f,t})\|^2_{\mathbf{M}^{-1}_t}, \tag{18b}$$

where

$$\mathbf{\Omega}_t = \mathbf{F}^\intercal (\mathbf{I} - \hat{\mathbf{\Omega}}_{t+1} \tilde{\mathbf{\Sigma}}^{'1/2}_{t+1} \mathbf{\Upsilon}^{-1}_{t+1} \tilde{\mathbf{\Sigma}}^{'1/2}_{t+1}) \hat{\mathbf{\Omega}}_{t+1} \mathbf{F}, \tag{19a}$$

$$\boldsymbol{\lambda}_t = \mathbf{F}^\intercal (\mathbf{I} - \hat{\mathbf{\Omega}}_{t+1} \tilde{\mathbf{\Sigma}}^{'1/2}_{t+1} \mathbf{\Upsilon}^{-1}_{t+1} \tilde{\mathbf{\Sigma}}^{'1/2}_{t+1}) (\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\mathbf{\Omega}}_{t+1} \tilde{\boldsymbol{\mu}}^{'}_{t+1}), \tag{19b}$$

$$\hat{\mathbf{\Omega}}_t = \mathbf{\Omega}_t + \mathbf{H}^\intercal \mathbf{\Lambda}^{-1} \mathbf{H}, \tag{19c}$$

$$\hat{\boldsymbol{\lambda}}_t = \boldsymbol{\lambda}_t + \mathbf{H}^\intercal \mathbf{\Lambda}^{-1} (\tilde{\mathbf{x}}_t - \mathbf{A}^k \tilde{\mathbf{x}}_{t-1}), \tag{19d}$$

$$\mathbf{\Upsilon}_t = \tilde{\mathbf{\Sigma}}^{'1/2}_t \hat{\mathbf{\Omega}}_t \tilde{\mathbf{\Sigma}}^{'1/2}_t + \mathbf{I}, \tag{19e}$$

$$\tilde{\boldsymbol{\mu}}^{'}_t = [\tilde{\mu}_{1,\tilde{z}^{'}_{t,1}}, \ldots, \tilde{\mu}_{nk,\tilde{z}^{'}_{t,nk}}, \mathbf{0}_{1\times n(k-1)}]^\intercal, \tag{19f}$$

$$\tilde{\mathbf{\Sigma}}^{'}_t = \mathrm{diag}([\tilde{\sigma}^2_{1,\tilde{z}^{'}_{t,1}}, \ldots, \tilde{\sigma}^2_{nk,\tilde{z}^{'}_{t,nk}}, \mathbf{0}_{1\times n(k-1)}]). \tag{19g}$$

With $\mathbf{\Omega}_T = 0$ and $\boldsymbol{\lambda}_T = 0$, $\{\mathbf{\Omega}_t, \boldsymbol{\lambda}_t\}^T_{t=1}$ can be computed recursively for $\tilde{\mathbf{z}}^{'}_{1:T}$ using (19a)-(19g). Once all the ancestors $\{a^i_t\}^N_{i=1}$ have been sampled, we can calculate the new particle weights as follows

$$\begin{aligned} w^i_t &\propto p_\theta(\tilde{\mathbf{x}}_t | \tilde{\mathbf{z}}^i_{1:t}, \tilde{\mathbf{x}}_{1:t-1}) \\ &= \mathcal{N}(\tilde{\mathbf{x}}_t | \mathbf{F} \hat{\boldsymbol{\mu}}^i_{p,t+1} + \mathbf{A}^k \tilde{\mathbf{x}}_{t-1}, \mathbf{F} \hat{\mathbf{\Sigma}}^i_{p,t+1} \mathbf{F}^\intercal + \mathbf{\Lambda}). \end{aligned} \tag{20}$$

After all the particle trajectories have been generated, we obtain $\tilde{\mathbf{z}}_{1:T}[k]$ by sampling from these trajectories according to the weights $\{W^i_T\}^N_{i=1}$ at time $T$.

### 4.4 PARAMETER UPDATE

At the $k$-th M step, given the sample $\tilde{\mathbf{z}}_{1:T}[k]$ drawn by RB-CPF-AS, we can obtain $p_{\theta_{k-1}}(\tilde{\mathbf{e}}_t | \tilde{\mathbf{z}}_{1:T}[k], \tilde{\mathbf{x}}_{1:T}) = \mathcal{N}(\tilde{\mathbf{e}}_t | \hat{\boldsymbol{\mu}}_{s,t}, \hat{\mathbf{\Sigma}}_{s,t})$ using the RTS smoother. Then we have

$$\begin{aligned} &\int p_{\theta_{k-1}}(\tilde{\mathbf{e}}_{1:T} | \tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{x}}_{1:T}) \log p_\theta(\tilde{\mathbf{x}}_{1:T}, \tilde{\mathbf{e}}_{1:T}, \tilde{\mathbf{z}}_{1:T}) d_{\tilde{\mathbf{e}}_{1:T}} \\ &= \sum_{t=1}^T \log p(\tilde{\mathbf{z}}_t) \\ &\quad - \frac{1}{2} \sum_{t=1}^T \int q(\tilde{\mathbf{e}}^{'}_t)(\tilde{\mathbf{e}}^{'}_t - \tilde{\boldsymbol{\mu}}^{'}_t)^\intercal \tilde{\mathbf{\Sigma}}^{'-1}_t (\tilde{\mathbf{e}}^{'}_t - \tilde{\boldsymbol{\mu}}^{'}_t) d\tilde{\mathbf{e}}^{'}_t + \log |\tilde{\mathbf{\Sigma}}_t| \\ &\quad - \frac{1}{2} \sum_{t=1}^T \int q(\tilde{\mathbf{e}}_t)(\tilde{\mathbf{y}}_t - \mathbf{H}\tilde{\mathbf{e}}_t)^\intercal \mathbf{\Lambda}^{-1} (\tilde{\mathbf{y}}_t - \mathbf{H}\tilde{\mathbf{e}}_t) d\tilde{\mathbf{e}}_t \\ &\quad + \text{const}, \end{aligned} \tag{21}$$

where $\tilde{\mathbf{z}}_t = \tilde{\mathbf{z}}_t[k]$, $\tilde{\mathbf{e}}^{'}_t = [\mathbf{e}^\intercal_{tk}, \mathbf{e}^\intercal_{tk-1}, \ldots, \tilde{\mathbf{e}}^\intercal_{(t-1)k+1}]^\intercal$, $\tilde{\boldsymbol{\mu}}^{'}_t = [\tilde{\mu}_{1,\tilde{z}_{t,1}}, \ldots, \tilde{\mu}_{nk,\tilde{z}_{t,nk}}]^\intercal$, $\tilde{\mathbf{\Sigma}}^{'}_t = \mathrm{diag}([\tilde{\sigma}^2_{1,\tilde{z}_{t,1}}, \ldots, \tilde{\sigma}^2_{nk,\tilde{z}_{t,nk}}])$, $\tilde{\mathbf{y}}_t = \tilde{\mathbf{x}}_t - \mathbf{A}^k \tilde{\mathbf{x}}_{t-1}$, and $q(\tilde{\mathbf{e}}_t) = p_{\theta_{k-1}}(\tilde{\mathbf{e}}_t | \tilde{\mathbf{z}}_{1:T}, \tilde{\mathbf{x}}_{1:T})$.

**Algorithm 1:** RB-CPF-AS

**Input:** $\tilde{\mathbf{z}}^{'}_{1:T} = \tilde{\mathbf{z}}_{1:T}[k-1]$, $\theta = \theta_{k-1}$
**Output:** $\tilde{\mathbf{z}}_{1:T}[k] \sim \mathcal{M}_{\theta_{k-1}}(\cdot | \tilde{\mathbf{z}}_{1:T}[k-1])$
Compute $\{\mathbf{\Omega}_t, \boldsymbol{\lambda}_t\}^T_{t=1}$ according to (19a)-(19g)
Draw $\tilde{\mathbf{z}}^i_1$ with $\tilde{\mathbf{z}}^i_1 \sim p_\theta(\tilde{\mathbf{z}}_1)$ for $i = 1, \ldots, N$
Compute $\boldsymbol{\mu}^i_1$, $\mathbf{\Sigma}^i_1$, and $w^i_1$ for $i = 1, \ldots, N$
**for** *t=2 to T* **do**
  Draw $a^i_t$ with $P(a^i_t = j) \propto w^j_{t-1}$ for $i = 1, \ldots, N$
  // Resampling and ancestor sampling
  Draw $\tilde{\mathbf{z}}^i_t$ with $\tilde{\mathbf{z}}^i_t \sim p_\theta(\tilde{\mathbf{z}}_t)$ for $i = 1, \ldots, N$
  Compute $\{\mathbf{M}^i_{t-1}, \eta^i_{t-1}\}$ according to (18a) and (18b)
  Draw $a^i_N$ according to (15) for $i = 1, \ldots, N$
  // Particle propagation
  Set $\tilde{\mathbf{z}}^i_{1:t} = \{\tilde{\mathbf{z}}^{a^i_t}_{1:t-1}, \tilde{\mathbf{z}}^i_t\}$ for $i = 1, \ldots, N$
  Set $\hat{\boldsymbol{\mu}}^i_{f,1:t-1} = \hat{\boldsymbol{\mu}}^{a^i_t}_{f,1:t-1}$, $\hat{\boldsymbol{\mu}}^i_{p,1:t-1} = \hat{\boldsymbol{\mu}}^{a^i_t}_{p,1:t-1}$, $\hat{\mathbf{\Sigma}}^i_{f,1:t-1} = \hat{\mathbf{\Sigma}}^{a^i_t}_{f,t-1}$, $\hat{\mathbf{\Sigma}}^i_{p,1:t-1} = \hat{\mathbf{\Sigma}}^{a^i_t}_{p,1:t-1}$
  // Weighting
  Compute $\hat{\boldsymbol{\mu}}^i_{f,t}$, $\hat{\boldsymbol{\mu}}^i_{p,t}$, $\hat{\mathbf{\Sigma}}^i_{f,t}$, and $\hat{\mathbf{\Sigma}}^i_{p,t}$
  Compute weights $w^i_t$ according to (20)
Draw $J$ with $P(J = j) \propto W^j_T$ and set $\tilde{\mathbf{z}}_{1:T}[k] = \tilde{\mathbf{z}}^J_{1:T}$.

It can be seen that we only need sufficient statistics $\int \tilde{\mathbf{e}}_t q(\tilde{\mathbf{e}}_t) d\tilde{\mathbf{e}}_t$ and $\int \tilde{\mathbf{e}}_t \tilde{\mathbf{e}}^\intercal_t q(\tilde{\mathbf{e}}_t) d\tilde{\mathbf{e}}_t$ to maximize (21). Denoting a sufficient statistics at the $k$-th iteration as $S^k$, we use $\mathbb{S}^k = (1 - \gamma_k)\mathbb{S}^k + \gamma_k S^k$ to maximize $\hat{\mathcal{Q}}_k(\theta)$. To maximize $\hat{\mathcal{Q}}_k(\theta)$ with respect to $\mathbf{A}$, we compute the gradient of $\mathbf{A}$ in terms involving $\mathbf{A}^k$ and $\mathbf{H}$ and apply a conjugate gradient descent method as done in Gong et al. (2015).

## 5 EXPERIMENTS

In this section, we conduct empirical studies of the two estimation methods presented in Section 3.2 and Section 4 on both synthetic and real data to show their effectiveness.

### 5.1 SIMULATED DATA

We conduct a series of simulations to investigate the effectiveness of the proposed estimation methods. Following (Gong et al., 2015), we generated the data at the casual frequency using the VAR model (1) with randomly generated matrix $\mathbf{A}$ and independent Gaussian mixture noises $\mathbf{e}_t$. The elements in $\mathbf{A}$ were drawn from a uniform distribution $\mathcal{U}(-0.5, 0.5)$. The Gaussian mixture model contains two components for each channel. The parameters were $w_{1,1} = 0.2$, $w_{1,2} = 0.8$, $w_{2,1} = 0.3$, $w_{2,2} = 0.7$ ,$\mu_{i,1} = 0$, $\mu_{i,2} = 0$, $\sigma^2_{i,1} = 1e-4$, $\sigma^2_{1,2} = .1$, and $\sigma^2_{2,2} = 0.2$. Low-resolution observations were obtained by aggregating the high-resolution data using aggregation factor $k$. Similarly,

we also generated data with Gaussian noise (by setting $\sigma^2_{1,c} = 0.01$, $\sigma^2_{2,c} = 0.02$) for comparison of different methods. We tested data with dimension $n = 2$, aggregation factor $k = 2$ and 3, and sample size $T = 150$ and 300, respectively. For comparison,we replaced the Gaussian mixture models in our method with Gaussian noise models, leading to a method based on Gaussian noises. We denote the method proposed in Section 4 that performs causal discovery from temporally aggregated data as $\text{CDTA}_{\text{finite}}$ and the corresponding Gaussian counterpart as $\text{CDTA}_{\text{Gauss}}$. We also compare with the NG-EM method (Gong et al., 2015) on the aggregated data with non-Gaussian noises. The experiments are repeated for 10 replications .

Table 1 shows the mean squared error (MSE) of the estimated causal transition matrix $\mathbf{A}$. It can be seen that as the sample size $T$ increases, both the proposed $\text{CDTA}_{\text{finite}}$ and the baseline method $\text{CDTA}_{\text{Gauss}}$ obtain smaller estimation errors. On the non-Gaussian data, $\text{CDTA}_{\text{Gauss}}$ produces much higher errors than $\text{CDTA}_{\text{finite}}$. On the Gaussian data, neither $\text{CDTA}_{\text{finite}}$ nor $\text{CDTA}_{\text{Gauss}}$ can obtain accurate estimations. This is because the estimation algorithms can converge to many possible solutions that have the same marginal likelihood, if the data noises are Gaussian or the estimation algorithms assume a Gaussian noise model. The results are consistent with the theoretical results that the causal relations might not be uniquely determined using Gaussian noise models. It can also be seen that the NG-EM method fails on the aggregated data, because NG-EM is proposed for subsampled rather than aggregated data.

Further, we examined the performance of the method described in Section 3.2, denoted as $\text{CDTA}_{\text{infty}}$, with finite $k$ values. To achieve so, we generated aggregated data with $\mathbf{A} = \left[\begin{smallmatrix} 0.8 & 0.1 \\ 0.6 & 0.7 \end{smallmatrix}\right]$, aggregation factor $k = 2, 3, 4, 10$, and the same Gaussian mixture noise parameters described above. The true $\mathbf{A}^{\text{NoSelfLoop}}$ in this case can be calculated as $\mathbf{A}^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.5 \\ 2 & 0 \end{smallmatrix}\right]$. Using the linear instantaneous non-Gaussian model, we can obtain the estimations of $\mathbf{A}^{\text{NoSelfLoop}}$ on the aggregated data. The results for $k = 2, 3, 4, 10$ are given as follows:

$$\hat{\mathbf{A}}_2^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.52 \\ -13.5 & 0 \end{smallmatrix}\right],\ \hat{\mathbf{A}}_3^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.47 \\ 1.43 & 0 \end{smallmatrix}\right],$$
$$\hat{\mathbf{A}}_4^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.44 \\ 1.61 & 0 \end{smallmatrix}\right],\ \hat{\mathbf{A}}_{10}^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.51 \\ 2.03 & 0 \end{smallmatrix}\right]. \quad (22)$$

It seems that when $k \geq 4$, the linear instantaneous non-Gaussian causal model (10), which assumes that there is no self-loop, can estimate the corresponding $\mathbf{A}^{\text{NoSelfLoop}}$ accurately and very efficiently, at the cost of losing the self-loops in the original process. However, the self-loops can be estimated with $\text{CDTA}_{\text{finite}}$ when $k$ is reasonably large. As a cautionary notice, researchers should carefully interpret the estimated parameters produced by linear instantaneous causal models, which assume there is no self-loop; as a consequence, the linear instantaneous non-Gaussian causal model produces $(1 - A_{jj})^{-1}A_{ji}$, whose magnitude can be very different from the true causal parameter $A_{ji}$.

### 5.2 REAL DATA

We conducted experiments on the Temperature Ozone data (Mooij et al., 2016) and the macroeconomic data used in (Moneta, 2008). These two time series are collected by averaging the records during specified time intervals. For example, the Temperature Ozone data contain daily mean values of ozone and temperature of year 2009 in Chaumont and Switzerland. The macroeconomic data contain quarterly US macro variables for the period 1947:2 to 1994:1.

**Temperature Ozone.** The Temperature Ozone data is the 50th causal-effect pair from the website `https://webdav.tuebingen.mpg.de/cause-effect/`. The data have records of ozone density $X$ and daily mean temperature$Y$. The ground truth is causal relation is $Y \to X$. We first applied $\text{CDTA}_{\text{infty}}$ on the data, resulting in $\mathbf{A}^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.65 \\ 0.65 & 0 \end{smallmatrix}\right]$. From this result, we can find that instantaneous effects exist in both directions. This could possibly caused by aggregation with a small $k$ as the estimated $\mathbf{A}^{\text{NoSelfLoop}}$ is likely to be inaccurate. We then estimated the causal matrix by $\text{CDTA}_{\text{finite}}$. The estimated transition matrix $\mathbf{A}$ for $k = 1, 2, 3$ is $\left[\begin{smallmatrix} 0.8418 & 0.0945 \\ 0.0165 & 0.9785 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.8426 & 0.1151 \\ 0.0220 & 0.9702 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.8360 & 0.1150 \\ 0.0172 & 0.9729 \end{smallmatrix}\right]$, respectively. It seems that the estimated matrices sensibly captured the self-influences and cross-influences between the ozone and temperature processes.

**Macroeconomic Data** The data are quarterly U.S. observational on real aggregated macroeconomic variables. Here we consider the causal relations between two variables, including real balances $X$ and price inflation $Y$. $X$ denotes the logarithm of per captita M2 minus the logarithm of the implicit price deflator. $Y$ is the log of the implicit price deflator at the time $t$ minus log of the implicit price deflator at the time $t - 1$. Again, we first applied the $\text{CDTA}_{\text{infty}}$ to find the rough estimation of causal relations excluding self-loops. The estimated $\mathbf{A}^{\text{NoSelfLoop}}$ is $\left[\begin{smallmatrix} 0 & 0 \\ 0.3074 & 0 \end{smallmatrix}\right]$. This indicates that no influence from effect to cause can be estimated from the instantaneous dependencies, which is consistent with the ground truth. We also employed $\text{CDTA}_{\text{finite}}$ to obtain the estimation of a complete causal transition matrix $\mathbf{A}$. The estimated transition matrix $\mathbf{A}$ for $k = 1, 2, 3, 4$ is $\left[\begin{smallmatrix} 1.009 & -0.0013 \\ 0.1718 & 0.5796 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.9007 & -0.0011 \\ 0.2024 & 0.6420 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0.8503 & -0.0005 \\ 0.1378 & 0.7091 \end{smallmatrix}\right]$, respectively. We can see that $\mathbf{A}$ gives weaker responses from effect $Y$ to cause $X$ as $k$ increases. If we consider $k = 4$ as the aggregation factor, then we can calculate from the estimated $\mathbf{A}$ that $\mathbf{A}^{\text{NoSelfLoop}} = \left[\begin{smallmatrix} 0 & 0.033 \\ 0.4737 & 0 \end{smallmatrix}\right]$, which is close to the results estimated by $\text{CDTA}_{\text{infty}}$.

## 6 CONCLUSION

In this paper, we have investigate the problem of discovering high frequency causal relations from temporally aggregated time series. When the aggregation factor is finite, we proved

| Methods \ Data | non-Gaussian noise | | | | Gaussian noise | | | |
|---|---|---|---|---|---|---|---|---|
| | k=2 | | k=3 | | k=2 | | k=3 | |
| | T=150 | T=300 | T=150 | T=300 | T=150 | T=300 | T=150 | T=300 |
| $\text{CDTA}_{\text{finite}}$ | **2.10e-4** | **1.19e-4** | **8.17e-4** | **7.36e-4** | 1.42e-2 | 3.67e-3 | 7.63e-3 | 9.69e-3 |
| $\text{CDTA}_{\text{Gauss}}$ | 1.28e-2 | 4.49e-3 | 1.20e-2 | 7.22e-3 | 1.13e-2 | 3.08e-3 | 6.26e-2 | 9.07e-3 |
| NG-EM | 8.75e-2 | 8.51e-2 | 5.27e-1 | 1.88e-1 | - | - | - | - |

Table 1: MSE of different methods on simulated non-Gaussian and Gaussian data. The results are shown for subsampling factors ($k = 2, 3$) and lengths of data ($T = 150, 300$).

that the causal relations are fully identifiable if the underlying causal relations are linear and the noise process is non-Gaussian. We also show that the causal matrix that removes self-loops is identifiable from instantaneous dependencies, when the aggregation factor goes to infinity. Based on these results, we propose an algorithm to recover the complete causal matrix when the aggregation factor is relatively small an an very efficient algorithm to partially recover the matrix when the aggregation factor is relatively large. Future work will focus on automatically estimating the aggregation factor $k$ from data.

## APPENDIX: PROOF OF THEOREM 1

*Proof.* Here we consider the limit when $T \to \infty$. According to the identifiability results of $\mathbf{A}^k$ (7), we have

$$\mathbf{A}^k = \mathbf{A}'^k. \tag{23}$$

We then consider the remaining error term $\vec{\mathbf{e}}_t$. The corresponding random vector $\vec{\mathbf{e}}$ follows both the representation (9) and

$$\mathbf{H}' \triangleq [\mathbf{I}, \ldots, \sum_{l=0}^{k-1} \mathbf{A}'^l, \sum_{l=1}^{k-1} \mathbf{A}'^l, \ldots, \sum_{l=k-2}^{k-1} \mathbf{A}'^l, \mathbf{A}'^{k-1}]. \tag{24}$$

$$\tilde{\mathbf{e}}' = ({e'}_1^{(0)}, ..., {e'}_n^{(0)}, {e'}_1^{(1)}, ..., {e'}_n^{(1)}, ..., {e'}_1^{(2k-2)}, ..., {e'}_n^{(2k-2)})^\intercal \tag{25}$$

with ${e'}_i^{(l)}$, $l = 0, ..., 2k-2$, having the same distribution $p_{e'_i}$.

According to Proposition 1, each column of $\mathbf{H}'$ is a scaled version of a column of $\mathbf{H}$. Denote by $H_{ln+i}$, $l = 0, ..., 2k-2$; $i = 1, ..., n$, the $(ln+i)$th column of $\mathbf{H}$, and similarly for $H'_{ln+i}$. According to the Uniqueness Theorem in Eriksson & Koivunen (2004), we know that under condition A2, for each $i$, there exists one and only one $j$ such that the distribution of $e_i^{(l)}$, $l = 0, ..., 2k-2$ (which have the same distribution), is the same as the distribution of ${e'}_j^{(l)}$, $l = 0, ..., 2k-2$, up to changes of location and scale. As a consequence, the columns $\{H'_{ln+j} \mid l = 0, ..., 2k-2\}$ correspond to $\{H_{ln+i} \mid l = 0, ..., 2k-2\}$ up to the permutation and scaling arbitrariness.

According to the structure of $\mathbf{H}$, $\forall\, m \leq k-1$, $H_{kn+i} = H_{m+i} + H_{m+k+i}$, and similarly we have $\forall\, m \leq k-1$, $H'_{kn+j} = H'_{m+j} + H'_{m+k+j}$. Hence, $H_{kn+i}$ is proportional to $H'_{kn+j}$, i.e., $H'_{kn+j} = \lambda_{ki} H_{kn+i}$. Assume that $\sum_{l=0}^{1} \mathbf{A}'^l$, ..., $\sum_{l=0}^{k-1} \mathbf{A}'^l$, $\sum_{l=1}^{k-1} \mathbf{A}'^l$, ..., $\sum_{l=k-2}^{k-1} \mathbf{A}'^l$, $\mathbf{A}'^{k-1}$ are non-diagonal matrices, we have $H'_j = \lambda_{0i} H_i$. Since $H_i$ and $H'_j$ must be columns of $\mathbf{I}$, as implied by the structure of $\mathbf{H}$ and $\mathbf{H}'$, we can see that $\lambda_{0i} = 1$ and that $i = j$. Consequently, $\lambda_{ki}$ must be 1 or -1. Let $\mathbf{B} = \mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{k-1}$ and $\mathbf{B}' = \mathbf{I} + \mathbf{A}' + \ldots + \mathbf{A}'^{k-1}$, we thus have $\mathbf{B} = \mathbf{B}'\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with 1 or -1 as its diagonal entries. Moreover, because $\mathbf{AB} - \mathbf{B} = \mathbf{A}^k - \mathbf{I}$ and $\mathbf{A}'\mathbf{B}' - \mathbf{B}' = \mathbf{A}'^k - \mathbf{I}$, and $\mathbf{A}^k = \mathbf{A}'^k$, we have

$$\mathbf{A}' - \mathbf{I} = (\mathbf{A} - \mathbf{I})\mathbf{D}. \tag{26}$$

If both $\mathbf{A}'$ and $\mathbf{A}$ have diagonal entries which are smaller than 1, $\mathbf{D}$ must be the identity matrix, i.e., $\mathbf{A}' = \mathbf{A}$. Therefore statement (i) is true.

If each $p_{e_i}$ is asymmetric, $e_i$ and $-e_i$ have different distributions. Consequently, the representation (24) does not hold any more if one changes the signs of a subset of, but not all, non-zero elements of $\{H'_{ln+j} \mid l = 0, ..., 2k-2\}$. This implies that for non-zero $H_{ln+i}$, $\lambda_{li}$, including $\lambda_{0i}$, have the same sign, and they are therefore 1 since $\lambda_{0i} = 1$. $\lambda_{ki} = 1$ leads to $\mathbf{D} = \mathbf{I}$ and thus gives $\mathbf{A}' = \mathbf{A}$. That is, (ii) is true.

□


### Acknowledgements

The authors would like to thank Dr. Tongliang Liu for helpful discussions. DT and MG would like to acknowledge the support from DP-140102164, FT-130101457, and LP-150100671. CG and KZ would like to acknowledge the support from NIH-1R01EB022858-01 FAIN-R01EB022858, NIH-1R01LM012087, and NIH-5U54HG008540-02 FAIN-U54HG008540.


## References


Boot, J.CG., Feibes, W., and Lisman, J. Hubertus C. Further methods of derivation of quarterly figures from annual data. *Applied Statistics*, pp. 65–75, 1967.

Breitung, J. and Swanson, N. R. Temporal aggregation and spurious instantaneous causality in multiple time series models. *Journal of Time Series Analysis*, 23:651–665, 2002.

Danks, D. and Plis, S. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, 2013.

Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pp. 94–128, 1999.

Eriksson, J. and Koivunen, V. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

Geiger, P., Zhang, K., Gong, M., Schölkopf, B., and Janzing, D. Causal inference by identification of vector autoregressive processes with hidden components. In *32nd International Conference on Machine Learning*, pp. 1917–1925. Microtome Publishing, 2015.

Ghysels, E., Hill, J.B., and Motegi, K. Testing for granger causality with mixed frequency data. *Journal of Econometrics*, 192(1): 207–230, 2016.

Gong, M., Zhang, K., Schölkopf, B., Tao, D., and Geiger, P. Discovering temporal causal relations from subsampled data. In *ICML*, pp. 1898–1906, 2015.

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

Granger, Clive WJ. Implications of aggregation with common factors. *Econometric Theory*, 3(02):208–222, 1987.

Harvey, A. C. and Chung, C. H. Estimating the underlying change in unemployment in the uk. *Journal of the Royal Statistics Society, Series A*, 163:303–309, 2000.

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., and Danks, D. Causal discovery from subsampled time series data by constraint optimization. In *International Conference on Probabilistic Graphical Models*, pp. 216–227, 2016.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

Kuhn, E. and Lavielle, M. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI2008)*, Helsinki, Finland, 2008.

Lindsten, F. An efficient stochastic approximation em algorithm using conditional particle filters. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6274–6278. IEEE, 2013.

Lindsten, F., Jordan, M. I, and Schön, T.B. Particle gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15 (1):2145–2184, 2014.

Marcellino, M. Some consequences of temporal aggregation in empirical analysis. *Journal of Business and Economic Statistics*, 17:129–136, 1999.

Moauro, F. and Savio, G. Temporal disaggregation using multivariate structural time series models. *Journal of Econometrics*, 8:210–234, 2005.

Moneta, A. Graphical causal models and vars: an empirical assessment of the real business cycles hypothesis. *Empirical Economics*, 35(2):275–300, 2008.

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

Palm, F. C. and Nijman, T. E. Missing observations in the dynamic regression model. *Econometrica*, 52:1415–1435, 1984.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Plis, S., Danks, D., Freeman, C., and Calhoun, V. Rate-agnostic (causal) structure learning. In *Advances in neural information processing systems*, pp. 3303–3311, 2015a.

Plis, S., Danks, D., and Yang, J. Mesochronal structure learning. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, volume 31. NIH Public Access, 2015b.

Proietti, T. Temporal disaggregation by state space methods: Dynamic regression methods revisited. *The Econometrics Journal*, 9:357–372, 2006.

Rajaguru, G. and Abeysinghe, T. Temporal aggregation, cointegration and causality inference. *Economics Letters*, 101:223–226, 2008.

Shimizu, S., Hoyer, P.O., Hyvärinen, A., and Kerminen, A.J. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Silvestrini, A. and Veredas, D. Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, 22:458–497, 2008.

Sims, C. A. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Stram, Daniel O and Wei, William WS. A methodological note on the disaggregation of time series totals. *Journal of Time Series Analysis*, 7(4):293–302, 1986.

Svensson, A., Schön, T.B., and Lindsten, F. Identification of jump markov linear models using particle filters. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pp. 6504–6509. IEEE, 2014.

Tank, A., Fox, E.B., and Shojaie, A. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed frequency time series. *arXiv preprint arXiv:1704.02519*, 2017.

Tiao, George C. Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, pp. 525–531, 1972.

Van Nes, E.H., Scheffer, M., Brovkin, V., Lenton, T.M., Ye, H., Deyle, E., and Sugihara, G. Causal feedbacks in climate change. *Nature Climate Change*, 5(5):445–448, 2015.

Wei, G.CG. and Tanner, M.A. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85 (411):699–704, 1990.

Weiss, A. Systematic sampling and temporal aggregation in time series models. *Journal of Econometrics*, 26:271–281, 1984.

Zhou, D., Zhang, Y., Xiao, Y., and Cai, D. Analysis of sampling artifacts on the granger causality analysis for topology extraction of neuronal dynamics. *Frontiers in computational neuroscience*, 8, 2014.