Discovering Temporal Causal Relations from Subsampled Data

Mingming Gong[*1] MINGMING.GONG@STUDENT.UTS.EDU.AU
Kun Zhang[*2,3] KZHANG@TUEBINGEN.MPG.DE
Bernhard Schölkopf[2] BS@TUEBINGEN.MPG.DE
Dacheng Tao[1] DACHENG.TAO@UTS.EDU.AU
Philipp Geiger[2] PGEIGER@TUEBINGEN.MPG.DE

[1] Centre for Quantum Computation and Intelligent Systems, FEIT, University of Technology, Sydney, NSW, Australia
[2] Max Plank Institute for Intelligent Systems, Tübingen 72076, Germany
[3] Information Sciences Institute, University of Southern California

Abstract

Granger causal analysis has been an important tool for causal analysis for time series in various fields, including neuroscience and economics, and recently it has been extended to include instantaneous effects between the time series to explain the contemporaneous dependence in the residuals. In this paper, we assume that the time series at the true causal frequency follow the vector autoregressive model. We show that when the data resolution becomes lower due to subsampling, neither the original Granger causal analysis nor the extended one is able to discover the underlying causal relations. We then aim to answer the following question: can we estimate the temporal causal relations at the right causal frequency from the subsampled data? Traditionally this suffers from the identifiability problems: under the Gaussianity assumption of the data, the solutions are generally not unique. We prove that, however, if the noise terms are non-Gaussian, the underlying model for the high-frequency data is identifiable from subsampled data under mild conditions. We then propose an Expectation-Maximization (EM) approach and a variational inference approach to recover temporal causal relations from such subsampled data. Experimental results on both simulated and real data are reported to illustrate the performance of the proposed approaches.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. * Equal contribution.

1. Introduction

Granger causal analysis (Granger, 1980) has been widely used to find the temporal causal relations from time series. Time series x_1 is said to cause times series x_2 in the Granger's sense, if and only if the past and current values of x_1 contain useful information to predict the future values of x_2 that are not contained elsewhere.[1] In practice, although its nonlinear or nonparametric extensions exist, Granger causal analysis usually assumes a linear model, and consequently, the Granger causal relations can be seen by fitting the vector autoregressive (VAR) regression model (Sims, 1980). When using VAR to estimate temporal causal relations, one assumes that the data are obtained at the right causal frequency, i.e., the VAR model serves as an approximator to the true data-generating process. However, in practice the causal frequency is usually unknown, and the data are available at some fixed frequency such as daily, weekly, or monthly. As a consequence, the sampling frequency of the data is usually different from the true causal frequency.

There are two typical aggregation schemes to generate low-resolution or low-frequency data from high frequency ones. One is by subsampling or systematic sampling: for every k consecutive observations, one is kept, the rest being skipped. We call k the *subsampling factor*. The other is to take the local averages of k consecutive, non-overlapping observations as the new observations. See Silvestrini & Veredas (2008) for a survey on aggregation of univariate and multivariate time series models. Subsampling is a common phenomenon in time series, and is our main focus in

[1] In physics, it might be more mathematically tractable to construct theoretical models in continuous time, and often an exact description requires the use of continuous time. However, we would like to note that some time series are inherently discrete; an example is the dividend paid by a company to shareholders in successive years. Furthermore, even for continuous processes, their causal interactions may take place at discrete points.

this paper. As observations are temporally aggregated, the observed "causal structure" may be different from the original true one. As claimed in (Weiss, 1984), "some care needs to be taken in causality testing, as causality is defined for the true processes and not for the equation on the (temporally) aggregated or sampled data." Various contributions have been made on how temporal aggregation changes the Granger causal relations in the data (Rajaguru & Abeysinghe, 2008; Breitung & Swanson, 2002); for instance, temporal aggregation could cause spurious instantaneous correlations in the time series. However, little work has been done to recover the temporal causal relations at the proper causal frequency from the aggregated (or subsampled) data. In this paper, we are concerned with whether it is possible to recover the original causal relations at the causal frequency from the subsampled data, and if it is, how to do so.

Even if the original time series are generated by a VAR, as the time resolution becomes lower, one can see that the residuals are no longer contemporaneously independent (Wei, 2006, Chapter 20). To account for that, in addition to the time delayed causal relations, it was proposed to incorporate instantaneous effects between the variables (Hyvärinen et al., 2010). This extension has received considerable interest in neuroscience and economics. However, it is not clear how the discovered causal relations are related to those at the original causal frequency. In particular, as stated in (Granger, 1988), it was advocated that there is no true instantaneous causality;[2] spurious instantaneous causality may be found whenever the interval at which data are collected is lower than the causal frequency. In this paper, our results indicate that the instantaneous causal relations estimated by those methods are usually different from the true ones at the causal frequency.

We aim to recover the linear temporal causal relations from the subsampled data. We assume that the original time series at the causal frequency are stationary. The difficulty comes from the information loss in the missing observations caused by subsampling. It has been shown in (Palm & Nijman, 1984; Harvey, 1989) that with only the second-order information of the low-resolution data, usually the temporal causal relations are not identifiable. We assume that the error or noise terms are non-Gaussian, and under some additional mild conditions on the temporal causal relations, we show that interestingly, they can be uniquely recovered from subsampled data. To this end, we adopt the mixture of Gaussians for the distributions of the noise terms, and propose two estimation approaches. One is based on the Expectation-Maximization (EM) algorithm; however, its computational complexity increases very rapidly along with the dimension of the time series and the subsampling factor k. The other resorts to the variational inference framework, making the estimation procedure computationally efficient.

[2]Instantaneous causality might happen, say, in quantum physics. Here we focus on temporal causality.

There has been plenty of work in economics for temporal disaggregation of the low-resolution time series, with or without the side information from related indicators observed at the desired high frequency (Harvey & Chung, 2000; Moauro & Savio, 2005; Proietti, 2006). However, temporal disaggregation does not imply that the temporal causal relations in the high frequency data can be correctly recovered. The autocovariance structure of the low-resolution time series usually does not contain enough information to identify all parameters in the high-frequency model (Palm & Nijman, 1984; Harvey, 1989), and little attention has been paid to find further conditions to ensure that such parameters are identifiable. The work by Danks & Plis (2014) aims to infer the causal structure at the correct causal frequency directly from the causal structure learned from the subsampled data; they do not assume any specific form for the causal relations and their method is completely nonparametric, but on the other hand, an MCMC search is needed, which involves high computational load, and their method cannot estimate the causal strength.

This paper is organized as follows. In Section 2 we review Granger causal analysis with instantaneous effects, which was recently proposed for finding causal relations in time series when the VAR residuals are contemporaneously dependent. In Section 3 we study the effect of decreasing the temporal resolution of the time series by subsampling; in particular, it is found that unfortunately, both the VAR model and Granger causal analysis with instantaneous effects fail to recover the temporal causal relations underlying the data at the causal frequency. We then investigate whether it is possible to recover the original temporal causal relations from subsampled data. Interestingly, under the non-Gaussianity assumption of the data as well as other mild assumptions, we prove that the temporal causal relations at the causal frequency can be recovered from subsampled data. Next, in Section 4 we propose practical methods, including the EM algorithm and variational inference procedure, to achieve so. In Section 5 we report experimental results on both simulated and real data. Finally, Section 6 concludes the paper.

2. Granger Causality and Its Extension with Instantaneous Effects

For Granger causal analysis in the linear case (Granger, 1980), one fits the following VAR model (Sims, 1980) on the data:

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

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

Now let us assume that $\mathbf{x}_t$ also contains instantaneous effects. Let $\mathbf{B}$ contains the instantaneous causal relations between $\mathbf{x}_t$. Equation (1) changes to

$$\begin{aligned} &\mathbf{x}_t = \mathbf{B}\mathbf{x}_t + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{e}_t, \\ \Rightarrow &(\mathbf{I} - \mathbf{B})\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{e}_t, \\ \Rightarrow &\mathbf{x}_t = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{A}\mathbf{x}_{t-1} + (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}_t. \end{aligned} \tag{2}$$

To estimate all involved parameters in Granger causality with instantaneous effects, wo estimation procedures have been proposed. The two-step method first estimates the errors in the above VAR model and then apply independent component analysis (ICA) (Hyvärinen et al., 2001) on the estimated errors Hyvärinen et al. (2008). The other is based on multichannel blind deconvolution, which is statistically more efficient (Zhang & Hyvärinen, 2009).

3. Identifiability of the Causal Relations from Subsampled Data

Suppose the original high-resolution data were generated by (1). We consider low-resolution data generated by subsampling (or systematic sampling) with the subsampling factor k. Here we are interested in finding the causal transition matrix $\mathbf{A}$ which generated the original data from the subsampled data. Traditionally, if one uses only the second-order information, this suffers from parameter identification issues (Palm & Nijman, 1984), i.e., the same subsampled (low-frequency) model may disaggregate to several high frequency models, which are observationally equivalent at the low frequency.

3.1. Effect of Subsampling (Systematic Sampling)

Suppose that due to the low resolution of the data, there is an observation every k time steps. That is, the low-resolution observations $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T)$ are $(\mathbf{x}_1, \mathbf{x}_{1+k}, ..., \mathbf{x}_{1+(T-1)k})$; here we have assumed that the first sampled point is $\mathbf{x}_1$. We then have

$$\begin{aligned} \tilde{\mathbf{x}}_{t+1} &= \mathbf{x}_{1+tk} = \mathbf{A}\mathbf{x}_{1+tk-1} + \mathbf{e}_{1+tk} \\ &= \mathbf{A}(\mathbf{A}\mathbf{x}_{1+tk-2} + \mathbf{e}_{1+tk-1}) + \mathbf{e}_{1+tk} \\ &= ... \\ &= \mathbf{A}^k\tilde{\mathbf{x}}_t + \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{1+tk-l}. \end{aligned} \tag{3}$$

We denote by $\vec{\mathbf{e}}_{t+1}$ the noise term, i.e., $\vec{\mathbf{e}}_{t+1} = \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{1+tk-l}$. We call $(\mathbf{A}, \mathbf{e}, k)$ the representation of the kth order subsampled time series $\tilde{\mathbf{x}}_t$.

Equation (3) follows the vector autoregression (VAR) model, and then the following result directly follows.

Theorem 1. *If one fits a VAR model on the subsampled data $\tilde{\mathbf{x}}_t$ generated according to (3), as done by the traditional Granger causal analysis (Granger, 1980), the discovery temporal causal relations are given by A^k as the sample size $T \to \infty$.*

It has been pointed out (Marcellino, 1999) that the estimated time-delayed causal relation is not a time series property invariant to temporal aggregation.[3] Let us give an illustration on this.

Misleading Granger causal relations in subsampled data: An illustration Suppose $A = \begin{bmatrix} 0.8 & 0.5 \\ 0 & -0.8 \end{bmatrix}$. Consider the case where $k = 2$. The corresponding VAR model would be

$$\tilde{\mathbf{x}}_t = \mathbf{A}^2\tilde{\mathbf{x}}_{t-1} + \vec{\mathbf{e}}_t = \begin{bmatrix} 0.64 & 0 \\ 0 & 0.64 \end{bmatrix}\tilde{\mathbf{x}}_{t-1} + \vec{\mathbf{e}}_t.$$

That is, the causal influence from $\mathbf{x}_{2,t-1}$ to $\mathbf{x}_{1t}$ is missing in the corresponding subsampled data (with $k = 2$).

Suppose $A = \begin{bmatrix} 0.6 & 0.6 \\ 0.6 & -0.6 \end{bmatrix}$. Then the VAR model on the subsampled data is

$$\tilde{\mathbf{x}}_t = \mathbf{A}^2\tilde{\mathbf{x}}_{t-1} + \vec{\mathbf{e}}_t,$$

where $\mathbf{A}^2 = \begin{bmatrix} 0.72 & 0 \\ 0 & 0.72 \end{bmatrix}$, $\vec{\mathbf{e}}_t = \mathbf{e}'_t + \mathbf{A}\mathbf{e}'_{t-1} = \begin{bmatrix} e'_{1t} \\ e'_{2t} \end{bmatrix} + \begin{bmatrix} 0.6 & 0.6 \\ 0.6 & -0.6 \end{bmatrix} \cdot \begin{bmatrix} e'_{1,t-1} \\ e'_{2,t-1} \end{bmatrix}$, and $\mathbf{e}'_{t-l} = \mathbf{e}_{1+(t-1)k-l}$. Clearly the delayed causal relations between x_{1t} and x_{2t} are missing. Furthermore, one can see that $\mathbb{C}\text{ov}(\vec{e}_{1t}, \vec{e}_{2t}) = 0$. If e'_{it} are Gaussian, $\vec{e}_{1t}$ and $\vec{e}_{2t}$ are independent from each other, and thus there are no instantaneous causal effects. If they are non-Gaussian, $\vec{\mathbf{e}}_t$ is a linear mixture of four independent components, which are e'_{1t}, e'_{2t}, $e'_{1,t-1}$, and $e'_{2,t-1}$, and it is not possible to decompose it into two independent components; that is, the Granger causal model with instantaneous effects does not hold for the subsampled data.

3.2. Identifiability of the Causal Relations at the Causal Frequency

Suppose the system (1) is stable. Then all eigenvalues of $\mathbf{A}$ have modulus smaller than one (Lütkepohl, 2005). As

[3] More precisely, it gives a comprehensive study on the effects of temporal aggregation on exogeneity, causality, cointegration, unit roots, seasonal unit roots, impulse response functions, and trend-cycles decompositions; it finds that cointegration and unit roots are invariant to temporal aggregation, whereas the other properties are not (Marcellino, 1999).

a consequence, the eigenvalues of $\mathbf{A}^k$ become smaller and smaller as k increases, and the estimate of $\mathbf{A}^k$ by fitting the VAR model on $\tilde{\mathbf{X}}$ involves large estimation errors on finite samples. Moreover, even if we can estimate $\mathbf{A}^k$ perfectly, given the value of $\mathbf{A}^k$, there are usually a large number of possible solutions to $\mathbf{A}$ (Mitchell, 2003), which is different from the case where $\mathbf{A}$ is a scalar.[4]

An important issue is the identifiability of $\mathbf{A}$, i.e., whether it is possible to identify the original temporal causal relations, as implied by $\mathbf{A}$, from the low-resolution subsampled data $\tilde{\mathbf{x}}_t$. In other words, suppose $\tilde{\mathbf{x}}_t$ also admits another representation $(\mathbf{A}', \mathbf{e}', k)$, and we aim to see the relationship between $\mathbf{A}'$ and $\mathbf{A}$; in particular, if we always have $\mathbf{A}' = \mathbf{A}$, then as $n \to \infty$, the causal relationship at the correct resolution, $\mathbf{A}$, can be uniquely recovered from the low-resolution data. In fact, it has been demonstrated in (Palm & Nijman, 1984) that with only the second-order information, usually $\mathbf{A}$ is not identifiable. That is, the same low-frequency model may disaggregate to several high frequency models, which are observationally equivalent at the low frequency (according to the second-order statistics). However, we shall see when non-Gaussianity of the data is considered, the identifiability of $\mathbf{A}$ is achievable.

Let

$$\mathbf{L} \triangleq [\mathbf{I}\ \mathbf{A}\ \mathbf{A}^2 \cdots\ \mathbf{A}^{k-1}]. \tag{4}$$

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

$$\vec{\mathbf{e}} = \mathbf{L}\tilde{\mathbf{e}}, \text{ where} \tag{5}$$
$$\tilde{\mathbf{e}} = (e_1^{(0)}, ..., e_n^{(0)}, e_1^{(1)}, ..., e_n^{(1)}, ..., e_1^{(k-1)}, ..., e_n^{(k-1)})^\intercal.$$

The components of $\tilde{\mathbf{e}}$ are independent, and for each i, $e_i^{(l)}$, $l = 0, ..., k-1$, have the same distribution p_{e_i}.

First, we note that under the condition that p_{e_i} are non-Gaussian, $\mathbf{L}$ can be estimated up to the permutation and scaling indeterminacies (including the sign indeterminacy) of the columns, as given in the following lemma.

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

Proof. For the proof, let us introduce the following lemma. It was proven in Kagan et al. (1973, Theorem 10.3.1).

Lemma 1. *Let $\vec{\mathbf{e}} = \mathbf{J}\mathbf{r}$ and $\vec{\mathbf{e}} = \mathbf{M}\mathbf{s}$ be two representations of the n-dimensioal random vector $\vec{\mathbf{e}}$, where $\mathbf{J}$ and $\mathbf{M}$ are constant matrices of orders $n \times l$ and $n \times m$, respectively, and $\mathbf{r} = (r_1, ..., r_l)^\intercal$ and $\mathbf{s} = (s_1, ..., s_m)^\intercal$ are random vectors with independent components. Then the following assertions hold.*

(i) If the ith column of $\mathbf{J}$ is not proportional to any column of $\mathbf{M}$, then r_i is Gaussian.

(ii) If the ith column of $\mathbf{J}$ is proportional to the jth column of $\mathbf{M}$, then the logarithms of the characteristic functions of r_i and s_j differ by a polynomial in a neighborhood of the origin.

Equation (3) is a VAR model, and by making use of the second-order statistical information (i.e., autocovariances), we can estimate $\mathbf{A}^k$ and get rid of the contribution of the first term in (3). Then we focus on the noise part, which is given in (5). Since all p_{e_i} are non-Gaussian, according to (i) of Lemma 1 or Theorem 1 in (Eriksson & Koivunen, 2004), we know that $\mathbf{L}$ can be determined up to the permutation and scaling of columns. □

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 $\mathbf{A}$.

A1. The system is stable, in that all eigenvalues of $\mathbf{A}$ have modulus smaller than one.

A2. 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-resolution data is fully identifiable.

Theorem 2. *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 kth order subsampling 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 $\mathbf{A}' = \mathbf{A}\mathbf{D}_1$, where $\mathbf{D}_1$ is a diagonal matrix with 1 or -1 on its diagonal. If we constrain the self influences, represented by diagonal entries of $\mathbf{A}$ and $\mathbf{A}'$, to be positive,[5] then $\mathbf{A}' = \mathbf{A}$.

[4] For instance, the 2×2 identity matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ has infinite symmetric rational square roots given by $\frac{1}{a}\begin{bmatrix} b & c \\ c & -b \end{bmatrix}$, $\frac{1}{a}\begin{bmatrix} b & -c \\ -c & -b \end{bmatrix}$, $\frac{1}{a}\begin{bmatrix} -b & c \\ c & b \end{bmatrix}$, and $\frac{1}{a}\begin{bmatrix} -b & -c \\ -c & b \end{bmatrix}$, where b is a arbitrary nonnegative integer and c and a are arbitrary positive integers such that $b^2 + c^2 = a^2$ (Mitchell, 2003).

[5] We note that this is usually the case in neuroscience and economics.

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

(iii) If $\mathbf{A}$ is of full rank, all its diagonal entries are non-zero, and the graph implied by $\mathbf{A}$ is weakly connected,[6] then we have that $\mathbf{A}' = \mathbf{A}$ for odd k and that $\mathbf{A}'$ must be $\mathbf{A}$ or $-\mathbf{A}$ for even k.

A complete proof of Theorem 2 can be found in Section 1 of the Supplementary Material.

3.3. Relation to Granger Causality with Instantaneous Effects

In general, the estimated error terms in the subsampled times series are not spatially independent any more. The contemporaneous dependence in the noise terms inspired the model of Granger causality with instantaneous effects (Reale & Tunnicliffe Wilson, 2001; Hyvärinen et al., 2010); see (2). This model might provide an approximation to the underlying causal relations; however, in principle it does not hold for the low-resolution data obtained by subsampling, as one can see from the following theorem.[7]

Theorem 3. *Suppose the subsampled data $\tilde{\mathbf{x}}_t$ were generated by (3) and that all of e_{it} are non-Gaussian. Further assume that $\mathbf{A}$ is not diagonal, such that there exist causal relations between different time series. As $T \to \infty$, for the subsampled data $\tilde{\mathbf{x}}_t$, the model of Granger causality with instantaneous effects, represented by (2), does not hold, in that the error terms estimated with the VAR model are not linear mixtures of only n independent components.*

A complete proof of Theorem 3 can be found in Section 2 of the Supplementary Material.

4. Estimating the Temporal Causal Relations from Subsampled data

As stated in the previous section, to recover the temporal causal relations from systematically subsampled data, we have to make use of the non-Gaussianity of the data. Therefore, we use a Gaussian mixture model to represent each noise term p_{e_i}, i.e., $p_{e_i} = \sum_{c=1}^{m} w_{i,c}\mathcal{N}(e_i|\mu_{i,c}, \sigma_{i,c}^2)$, where $w_{i,c} \geq 0$, $\sum_{c=1}^{m} w_{i,c} = 1$, and $\sum_{c=1}^{m} w_{i,c}\mu_{i,c} = 0, i = 1, ..., n$. The VAR model on the low resolution data (3) can be simplified as

$$\tilde{\mathbf{x}}_t = \mathbf{A}^k\tilde{\mathbf{x}}_{t-1} + \mathbf{L}\tilde{\mathbf{e}}_t, \tag{6}$$

where $\tilde{\mathbf{e}}_t = (\mathbf{e}_{1+(t-1)k}^\intercal, \mathbf{e}_{1+(t-1)k-1}^\intercal, ..., \mathbf{e}_{1+(t-1)k-(k-1)}^\intercal)^\intercal$. It can be seen that each component of $\tilde{\mathbf{e}}$ can also be represented using a Gaussian mixture model $p_{\tilde{e}_i} = \sum_{z_i=1}^{m} \tilde{w}_{i,z_i}\mathcal{N}(\tilde{e}_i|\tilde{\mu}_{i,z_i}, \tilde{\sigma}_{i,z_i}^2)$, $i = 1, 2, ..., nk$. According to the structure of $\tilde{\mathbf{e}}$, some components of $\tilde{\mathbf{e}}$ share the same Gaussian mixture parameters, i.e., $\tilde{w}_{j+nl,c} = w_{j,c}$, $\tilde{\mu}_{j+nl,c} = \mu_{j,c}$, $\tilde{\sigma}_{j+nl,c} = \sigma_{j,c}$, $j = 1, ..., n$, $l = 0, ..., k-1$, $c = 1, ..., m$.

Consequently, we can write down the conditional distribution $p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1})$ as

$$p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}) = \sum_{\mathbf{z}_t} p(\mathbf{z}_t) \int p(\tilde{\mathbf{e}}_t|\mathbf{z}_t)p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{e}}_t, \tilde{\mathbf{x}}_{t-1})d\tilde{\mathbf{e}}_t,$$

where $\mathbf{z}_t = (z_{t,1}, ..., z_{t,nk})^\intercal$, $p(\mathbf{z}_t) = \prod_{i=1}^{nk} p(z_{t,i}) = \prod_{i=1}^{nk} \tilde{w}_{i,z_{t,i}}$, $p(\tilde{\mathbf{e}}_t|\mathbf{z}_t) = \prod_{i=1}^{nk} p(\tilde{e}_{t,i}|z_{t,i}) = \prod_{i=1}^{nk} \mathcal{N}(\tilde{e}_{t,i}|\tilde{\mu}_{i,z_{t,i}}, \tilde{\sigma}_{i,z_{t,i}}^2)$, and $p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{e}}_t, \tilde{\mathbf{x}}_{t-1}) = \mathcal{N}(\tilde{\mathbf{x}}_t|\mathbf{A}^k\tilde{\mathbf{x}}_{t-1} + \mathbf{L}\tilde{\mathbf{e}}_t, \Lambda)$. Here we assume a fixed and small Λ for regularization, because there is no additional additive noise term in (6). The model can be seen as an extension of the Independent Factor Analysis (IFA) (Attias, 1999) with additional constraints on the model parameters.

4.1. Parameter Estimation via EM algorithm

Given the subsampling factor k, we use the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimation of the model parameters $\Theta = (\mathbf{A}, w_{i,c}, \mu_{i,c}, \sigma_{i,c})$. Considering $\mathbf{z}_t$ and $\tilde{\mathbf{e}}_t$ as latent variables, we maximize the EM lower bound $\mathcal{L}(q, \Theta)$ of the data log-likelihood $\sum_t \ln p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \Theta)$ with respect to parameters Θ (M step) and find the distribution $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t)$ over the latent variables (E step) alternately until convergence.

In the E step, given the parameters Θ' from the previous iteration, the lower bound $\mathcal{L}(q, \Theta')$ is maximized with respect to q, resulting in $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\Theta') = p(\mathbf{z}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')p(\tilde{\mathbf{e}}_t|\mathbf{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')$, which is the posterior distribution of the latent variables.

In the M step, given the posterior distribution $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\Theta')$, the lower bound is maximized with respect to the parameters Θ. Because the EM lower bound can be decomposed into several terms which only depend on subsets of the parameters, the parameters can be updated independently. However, $w_{i,c}$ and $\mu_{i,c}$ must be updated jointly due to the constraints $\sum_{c=1}^{m} w_{i,c} = 1, \sum_{c=1}^{m} w_{i,c}\mu_{i,c} = 0, i = 1, ..., n$. This is a constrained nonlinear programming problem and we solve it by interior point methods (Byrd et al., 1999). After updating $w_{i,c}$ and $\mu_{i,c}$, we update $\sigma_{i,c}$ which has a closed form solution. Because the lower bound involves $\mathbf{A}^l$, $l = 1, ...k$, $\mathbf{A}$ has no analytic solutions. Thus

[6] In an undirected graph, two vertices x_i and x_j are called connected if it contains a path from x_i to x_j. A undirected graph is said to be connected if every pair of vertices in the graph is connected, and furthermore, a directed graph is called weakly connected if replacing all of its directed edges with undirected edges produces a connected undirected graph (Diestel, 1997).

[7] In this paper we assume causal sufficiency, that is, there is no hidden time series which causes more than one observed time series. However, we note that in the confounded case, it is still the case that in principle, the model of Granger causality with instantaneous effects does not hold.

we update $\mathbf{A}$ via the conjugate gradient descent algorithm. In practice, the convergence of EM algorithm is very slow when the the noise variance Λ approches zero. We adopt the *adaptive overrelaxed* EM (Salakhutdinov & Roweis, 2003) algorithm to obtain a faster rate of convergence. Details of the EM algorithm can be found in Section 3 of the Supplementary Material.

4.2. Mean Field Approximation

One problem with the EM algorithm is that the number of Gaussian mixture components will increase exponentially in nk. Thus, in the E step, the posterior marginals $p(\tilde{e}_{t,i}, z_{t,i}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1})$ would involve m^{nk} sums at each iteration. To make the algorithm computationally more efficient, we make the mean field assumption and approximate the true posterior $p(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1})$ with the factorized distribution $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t) = q(\mathbf{z}_t)q(\tilde{\mathbf{e}}_t)$. Using the factorized posterior distribution, we can obtain the posterior of $\tilde{\mathbf{e}}_t$ and $z_{t,i}$ independently. Therefore, the computational load is linear in nk. The variational EM lower bound is

$$\mathcal{L} = \sum_t \sum_{\mathbf{z}_t} q(\mathbf{z}_t) \int q(\tilde{\mathbf{e}}_t) \ln p(\tilde{\mathbf{x}}_t, \tilde{\mathbf{e}}_t, \mathbf{z}_t|\tilde{\mathbf{x}}_{t-1}, \Theta)\, d\tilde{\mathbf{e}}_t - \sum_t \sum_{\mathbf{z}_t} q(\mathbf{z}_t) \ln q(\mathbf{z}_t) - \sum_t \int q(\tilde{\mathbf{e}}_t) \ln q(\tilde{\mathbf{e}}_t)\, d\tilde{\mathbf{e}}_t.$$

The variational M step is similar to the M step in the original EM algorithm. In the E step, given Θ' from the previous M step, $q(\mathbf{z}_t|\Theta')$ and $q(\tilde{\mathbf{e}}_t|\Theta')$ are updated alternately by maximizing the lower bound:

$$q(\mathbf{z}_t|\Theta') \propto \exp \langle \ln p(\tilde{\mathbf{x}}_t, \tilde{\mathbf{e}}_t, \mathbf{z}_t|\tilde{\mathbf{x}}_{t-1}, \Theta') \rangle_{q(\tilde{\mathbf{e}}_t|\Theta')}, \quad (7)$$

$$q(\tilde{\mathbf{e}}_t|\Theta') \propto \exp \langle \ln p(\tilde{\mathbf{x}}_t, \tilde{\mathbf{e}}_t, \mathbf{z}_t|\tilde{\mathbf{x}}_{t-1}, \Theta') \rangle_{q(\mathbf{z}_t|\Theta')}. \quad (8)$$

In (7), the expectation of the log-likelihood with respect to $q(\tilde{\mathbf{e}}_t|\Theta')$ is calculated as

$$\langle \ln p(\tilde{\mathbf{x}}_t, \tilde{\mathbf{e}}_t, \mathbf{z}_t|\tilde{\mathbf{x}}_{t-1}, \Theta') \rangle_{q(\tilde{\mathbf{e}}_t|\Theta')} = \sum_{i=1}^{nk} \ln p(z_{t,i}) + \sum_{i=1}^{nk} \ln p(L_{t,i}|z_{t,i}) + \text{const},$$

where

$$\ln p(L_{t,i}|z_{t,i}) = -\frac{\left\langle (\tilde{e}_{t,i} - \tilde{\mu}'_{i,z_{t,i}})^2 \right\rangle_{q(\tilde{e}_{t,i}|\Theta')}}{2\tilde{\sigma}'^2_{i,z_{t,i}}} - \ln \tilde{\sigma}'_{i,z_{t,i}}.$$

Thus, the posterior $q(\mathbf{z}_t|\Theta')$ can be obtained as

$$q(z_{t,i}|\Theta') = \frac{p(L_{t,i}|z_{t,i})p(z_{t,i})}{\sum_{z'_{t,i}=1}^{m} p(L_{t,i}|z'_{t,i})p(z'_{t,i})}. \quad (9)$$

It can be seen that the computational complexity of the posteriors $q(z_{t,i}|\Theta')$ is linear in nk. In (8), the expectation of the log-likelihood with respect to $q(\mathbf{z}_t|\Theta')$ is in the form of a log-likelihood of joint Gaussian distribution and $q(\tilde{\mathbf{e}}_t|\Theta')$ thus can be efficiently obtained from the Gaussian posterior distribution.

4.3. Determination of the subsampling factor k

One practical issue is that the subsampling factor k is usually unknown. Therefore we need a principled way to choose the best k for our algorithms. In this paper, we used cross-validation on the log-likelihood of the models to choose the optimal k; specifically, we consider the value of k which gives the highest cross-validated log-likelihood as the optimal one. In our experiments, we used 5-fold cross validation.

5. Experimental Results

In this section we present experimental results on both simulated and read data to show the effectiveness of the proposed method to estimate the temporal causal relations from subsampled data. The objective function to be maximized by the proposed estimation methods is not convex. To avoid possible local optima, we used the transition matrix estimated by fitting VAR on the subsampled data to initialize the causal transition matrix $\mathbf{A}$, and use random initializations for the remaining parameters. With such an initialization scheme, we did not find any case where the proposed methods converge to unwanted solutions.

5.1. Simulated Data

To investigate the effectiveness of the proposed estimation methods, we conducted a series of simulations. We first generated the data at casual frequency by the VAR model (1) with randomly generated matrix $\mathbf{A}$ and independent Gaussian mixture noises $\mathbf{e}_t$. The elements in $\mathbf{A}$ are uniformly distributed between -0.5 and 0.5. The Gaussian mixture model contains two components for each dimension. We used both super-Gaussian and sub-Gaussian distributions for the noise terms. The parameters were $w_{i,1} = 0.8$, $w_{i,2} = 0.2$, $\mu_{i,1} = 0$, $\mu_{i,2} = 0$, $\sigma_{i,1} = 0.05$, $\sigma_{i,2} = 1$ for super-Gaussian noise and $w_{i,1} = 0.5$, $w_{i,2} = 0.5$, $\mu_{i,1} = -2$, $\sigma_{i,2} = 2$, $\sigma_{i,1} = 0.5$, $\sigma_{i,2} = 0.5$ for sub-Gaussian noise. Low-resolution observations were obtained by subsampling the high-resolution data by subsampling factor k. We tested data with dimension $n = 2$, subsampling factor $k = 2$ and 3, and sample size $T = 100$ and 300, respectively. We denote the proposed EM algorithm by Non-Gaussian EM (NG-EM) and the mean-field approximated algorithm by Non-Gaussian Mean-Field (NG-MF).

To our best knowledge, the problem considered in this paper has not been well studied, and we have not found

any existing method aiming at recovering the causal transition matrix from subsampled data. We compared our methods to two classical time series disaggregation methods: the Boot-Feibes-Lisma (BFL) method (Boot et al., 1967) and Stram-Wei (SW) method (Stram & Wei, 1986). These two methods try to recover the high resolution data using interpolation-based methods. To show the advantages of using non-Gaussianity of the data, we also compared our NG-EM and NG-MF with the method assuming Gaussian noise, denoted as G-EM, obtained by setting the noise distribution in NG-EM to a single Gaussian one. We repeated the experiments for 20 replications.

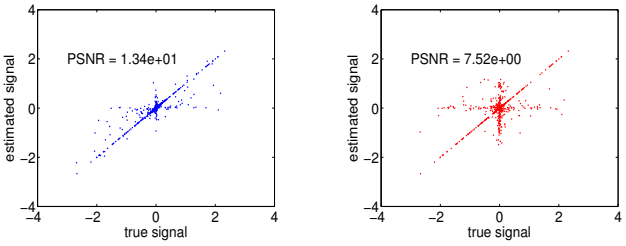
Table 1 shows the mean square error (MSE) of the estimated parameters $\mathbf{A}$. One can see that as the sample sizes T increases, our methods obtain better results. Furthermore, the estimation error increases with the subsampling factor k. Compared to other methods, our method achieves the lowest estimation error in the estimated $\mathbf{A}$. The method assuming Gaussian noise produces higher error because the solution is not unique and the algorithm may converge to an local optimal solution which is far away from the true $\mathbf{A}$. BFL and SW do not perform well because they are based on interpolation and thus lose some high frequency information. Our methods can also recover the causal-frequency data based on the estimated noise terms $\hat{\mathbf{e}}_t$. We used the posterior mean of noise terms as the estimate. Given the estimated noise, we can reconstruct the causal-frequency data based on the VAR model. Figure 1 gives the scatter plot of the estimated causal-frequency data against the true ones; one can see that NG-EM has a much better recovery performance than BFL, as indicated by a higher Peak Signal-to-Noise Ratio (PSNR). Moreover, as noted above, the causal-frequency data recovered by BFL cannot give a reliable estimation of $\mathbf{A}$.

To further illustrate the limitations of Gaussian noise models, we plot the contour of the log-likelihood function with respect to the two off-diagonal elements of $\hat{\mathbf{A}}$. Given a pair of off-diagonal elements, we optimized the log-likelihood over the diagonal elements. The off-diagonal elements were sampled from -0.8 to 0.8 at an interval of 0.01. The true causal matrix A is $\begin{bmatrix}0.65, & -0.16; & 0.15, & 0.65\end{bmatrix}$. Figure 2 shows the negative maximum log-likelihood function of non-Gaussian and Gaussian models computed from the subsampled data, with both super-Gaussian and sub-Gaussian noise terms. We used the same noise parameters as in the first simulation. It can be seen that, in both super-Gaussian (a & b) and sub-Gaussian case (c & d), the log-likelihood functions of Gaussian models have multiple solutions with the same likelihood value, while the log-likelihood functions of non-Gaussian models have only one global solution, which is around the true values. This is consistent with the theoretical results that the causal relations might not be uniquely determined using Gaussian noise models.

	super-Gaussian noise				sub-Gaussian noise			
	k=2		k=3		k=2		k=3	
	T=100	T=300	T=100	T=300	T=100	T=300	T=100	T=300
NG-EM	**7.27e-4**	**3.24e-4**	**1.70e-3**	**6.57e-4**	**5.76e-3**	**2.36e-3**	**1.31e-2**	**5.33e-3**
NG-MF	5.09e-3	2.62e-3	6.98e-3	5.22e-3	6.72e-3	3.31e-3	1.80e-2	6.17e-3
G-EM	1.33e-2	7.23e-3	1.63e-2	8.66e-3	3.56e-2	7.71e-3	2.64e-2	8.06e-3
BFL	3.89e-1	3.93e-1	4.87e-1	4.82e-1	3.61e-1	3.73e-1	4.80e-1	4.76e-1
SW	8.76e-2	8.51e-2	8.67e-2	8.47e-2	8.81e-2	8.73e-2	9.01e-2	8.57e-2

Table 1. Comparison of different methods on simulated super-Gaussian and sub-Gaussian data using Mean Square Error (MSE) between the true $\mathbf{A}$ and the estimated $\mathbf{A}$. The results are shown for different subsampling factors ($k = 2, 3$) and different length of data ($T = 100, 300$).

Finally, to test the effectiveness of the subsampling factor determination scheme in Section 4.3, we applied cross-validation on 50 randomly generated subsampled time series of length $T = 100$ and found that this scheme always produces the correct value of k, no matter $k = 2$ or 3.



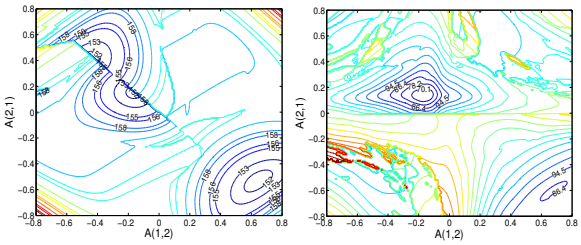
(a) Recovered by NG-EM (b) Recovered by BFL

Figure 1. Recovery of the causal-frequency data using the proposed EM method and the traditional methods: (a) The recovery results of the proposed NG-EM method ($PSNR = 13.4$); (b) The recovery results of the BFL method ($PSNR = 7.52$).

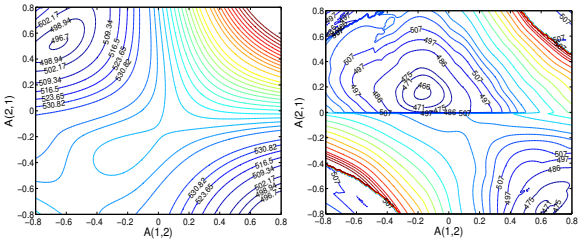
5.2. Real Data

We conducted experiments on the Temperature Ozone data and the Temperature in House data (Peters et al., 2013). We used the subsampling factor determination scheme in Section 4.3 to determine the optimal value of k as well as whether the frequency of the given data is lower than the causal frequency. For the data whose resolution is not lower than the "causal" one (which corresponds to the optimal sampling factor k determined by cross validation), we manually subsampled them to generate low-resolution data and then repeated the subsampling factor determination procedure to find the optimal causal frequency and the corresponding causal relations. Since the BFL and SW methods do not aim to estimate the causal relations at causal frequency, they are not suitable for comparison.

Temperature Ozone. The Temperature Ozone data is the 50th causal-effect pair from the website `https://webdav.tuebingen.mpg.de/cause-effect/`.



(a) Super-Gaussian, G-EM (b) Super-Gaussian, NG-EM



(c) Sub-Gaussian, G-EM (d) Sub-Gaussian, NG-EM

Figure 2. The contour plot of the negative log-likelihood function with repect to the two off-diagonal elmemnts of $\hat{\mathbf{A}}$: (a) negative log-likelihood function of the Gaussian model computed on super-Gaussian data, (b) negative log-likelihood function of the non-Gaussian model computed on super-Gaussian data, (c) negative log-likelihood function of the Gaussian model computed on sub-Gaussian data, (d) negative log-likelihood function of the non-Gaussian model computed on sub-Gaussian data.

The data have records of daily temperature X and ozone density Y. The ground truth is $Y \to X$. The cross-validated log-likelihood is $-89.638, -89.197$, and -90.246, respectively, as k ranges from 1 to 3. Therefore, we consider $k = 2$ as the best subsampling factor. The estimated transition matrix $\mathbf{A}$ for $k = 1, 2, 3$ is $\begin{bmatrix} 0.7285 & 0.1769 \\ -0.0378 & 0.9526 \end{bmatrix}$, $\begin{bmatrix} 0.8312 & 0.1370 \\ 0.0093 & 0.9537 \end{bmatrix}$, and $\begin{bmatrix} 0.8816 & 0.0989 \\ 0.0292 & 0.9462 \end{bmatrix}$, respectively. We can see from the results that the transition matrix $\mathbf{A}$ at $k = 2$ gives the weakest response from effect X to cause Y, which seems plausible.

Temperature in House. The Temperature in House dataset contains temperature recorded hourly in six rooms (1 - Shed, 2 - Outside, 3 - Kitchen Boiler, 4 - Living room, 5 - WC, 6 - Bathroom) of a house. We analyzed the causal relations $2 \to 3$ and $2 \to 4$ because they are relatively strong. For $2 \to 3$, the cross-validated log-likelihood is $183.596, 184.076, 184.139, 184.168$, and 184.183, respectively, as k ranges from 1 to 5. The estimated transition matrix $\mathbf{A}$ is $\begin{bmatrix} 0.9476 & -0.0024 \\ 0.0621 & 0.9394 \end{bmatrix}$, $\begin{bmatrix} 0.9735 & -0.0011 \\ 0.0329 & 0.9688 \end{bmatrix}$, $\begin{bmatrix} 0.9823 & -0.0007 \\ 0.0223 & 0.9790 \end{bmatrix}$, $\begin{bmatrix} 0.9867 & -0.0005 \\ 0.0169 & 0.9841 \end{bmatrix}$, and $\begin{bmatrix} 0.9894 & -0.0004 \\ 0.0136 & 0.9873 \end{bmatrix}$, respectively. It is interesting to note that the cross-validated likelihood always increases as k varies from 1 (corresponding to a 1-hour sampling interval) to 5 (12-minute sampling interval) . This indicates that the causal frequency is very high; in fact, note that 2 and 3 are adjacent, and it seems reasonable to consider the two processes as continuous ones. If we allow k to go to infinity, the VAR model provides an approximator to continuous processes.

For $2 \to 4$, the cross-validated log-likelihood is $273.533, 273.716, 322.347, 370.555$, and 370.547, respectively, as k ranges from 1 to 5. The estimated causal transition matrix is $\begin{bmatrix} 0.9416 & 0.0077 \\ 0.0638 & 0.9557 \end{bmatrix}$, $\begin{bmatrix} 0.9707 & 0.0037 \\ 0.0338 & 0.9764 \end{bmatrix}$, $\begin{bmatrix} 0.9804 & 0.0024 \\ 0.0228 & 0.9842 \end{bmatrix}$, $\begin{bmatrix} 0.9853 & 0.0018 \\ 0.0172 & 0.9880 \end{bmatrix}$, and $\begin{bmatrix} 0.9883 & 0.0014 \\ 0.0138 & 0.9905 \end{bmatrix}$. Here it seems that $k = 4$ (corresponding to a 15-minute sampling interval) should be preferred. Note that 2 and 4 are not adjacent. Causal influences between them take some time; in this case, a VAR model with a 15-minute sampling interval might provide a good approximation to the true processes.

6. Conclusion

Sometimes the observed time series were actually obtained by subsampling the true processes. We have considered the issue of recovering linear temporal causal relations at the true causal frequency from such time series. We were concerned with the situation, under certain mild conditions on the structure of the causal relations where the noise terms in the causal time series are non-Gaussian. We have shown that in this situation, the causal relations are identifiable. Two practical methods, one based on the EM algorithm and the other the variational inference framework, have been proposed to estimate the causal relations from low-resolution data. The method based on variational inference is computationally more efficient, and is recommended if the data have high dimensions or many points. As a line of our future research, we are trying to further improve the computational efficiency of the proposed methods, especially the one based on variational inference, to solve large-scale problems.

Acknowledgments

The authors thank Biwei Huang, Peter Spirtes, and Michel Besserve for helpful discussions. ZK was supported in part by DARPA grant No. W911NF-12-1-0034. GM and TD were supported by Australian Research Council Projects FT-130101457 and DP-140102164.

References

Attias, H. Independent factor analysis. *Neural Computation*, 11 (4):803–851, 1999.

Boot, J.C.G., Feibes, W., and Lisman, J.H.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.

Byrd, R.H., Hribar, M.E., and Nocedal, J. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

Danks, D. and Plis, S. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, 2014. To appear.

Diestel, R. *Graph Theory*. Springer-Verlag, 1997.

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

Granger, C. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 1980.

Granger, C. Some recent developments in a concept of causality. *Journal of Economietrics*, 39:199–211, 1988.

Harvey, A. C. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

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.

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

Hyvärinen, A., Shimizu, S., and Hoyer, P. O. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity. In *Proceedings of the 25th International Conference on Machine Learning (ICML2008)*, pp. 424–431, Helsinki, Finland, 2008.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, pp. 1709–1731, 2010.

Kagan, A. M., Linnik, Y. V., and Rao, C. R. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer, 2005.

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

Mitchell, D. W. Using pythagorean triples to generate square roots of I_2. *The Mathematical Gazette*, 87:499–500, 2003.

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

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

Peters, J., Janzing, D., and Schölkopf, B. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pp. 154–162, 2013.

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.

Reale, M. and Tunnicliffe Wilson, G. Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, 10(1-3):49–65, 2001.

Salakhutdinov, R. and Roweis, S. Adaptive overrelaxed bound optimization methods. In *Proceedings of the 20th International Conference on Machine Learning (ICML2003)*, pp. 664–671, 2003.

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.

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

Wei, W. S. *Time Series Analysis: Univariate and Multivariate Methods*. Pearson, 2006. 2nd Edition.

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

Zhang, K. and Hyvärinen, A. Acyclic causality discovery with additive noise: An information-theoretical perspective. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2009*, Bled, Slovenia, 2009.

Supplementary Document for "Discovering Temporal Causal Relations from Subsampled Data"

1. Proof of Theorem 2 in Section 3.2

Proof. Let us consider the limit when $T \to \infty$. According to (3), based on the second-order statistical information, one can uniquely determine $\mathbf{A}^k$ and $\mathbf{A}'^k$, that is,

$$\mathbf{A}^k = \mathbf{A}'^k. \tag{S1}$$

We can then determine the error term $\vec{\mathbf{e}}_t$. Then the corresponding random vector $\vec{\mathbf{e}}$ follows both the representation (5) and

$$\vec{\mathbf{e}} = \mathbf{L}'\tilde{\mathbf{e}}', \tag{S2}$$

where

$$\mathbf{L}' = [\mathbf{I}\ \mathbf{A}'\mathbf{A}'^2 \cdots \mathbf{A}'^{k-1}], \tag{S3}$$

and $\tilde{\mathbf{e}}' = ({e'}_1^{(0)}, ..., {e'}_n^{(0)}, {e'}_1^{(1)}, ..., {e'}_n^{(1)}, ..., {e'}_1^{(k-1)}, ..., {e'}_n^{(k-1)})^\intercal$ with ${e'}_i^{(l)}$, $l = 0, ..., k-1$, having the same distribution $p_{e'_i}$.

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

According to assumption A1, all eigenvalues of $\mathbf{A}$ have modulus smaller than one, and hence the eigenvalues of $\mathbf{A}\mathbf{A}^\intercal$ are smaller than 1. Then we know that for any n-dimensional vector v,

$$||\mathbf{A}v|| \leq ||\mathbf{A}|| \cdot ||v|| = \sqrt{||\mathbf{A}\mathbf{A}^\intercal||} \cdot ||v|| < ||v||.$$

According to the structure of $\mathbf{L}$, $L_{(l+1)n+i} = \mathbf{A}L_{ln+i}$. Considering L_{ln+i} as v in the above equation, one can see $||L_{(l+1)n+i}|| < ||L_{ln+i}||$, and similarly we have $||L'_{(l+1)n+j}|| < ||L'_{ln+j}||$. Hence, L'_{ln+j} is proportional to L_{ln+i}; more specifically, we have $L'_{ln+j} = \lambda_{li}L_{ln+i}$, where $\forall\ l$, λ_{li} have the same absolute value but possibly different signs. In particular, $L'_j = \lambda_{0i}L_i$. Bearing in mind that L_i and L'_j must be columns of $\mathbf{I}$, as implied by the structure of $\mathbf{L}$ and $\mathbf{L}'$, we can see that $\lambda_{0i} = 1$ and that $i = j$. Consequently, for $l > 0$, λ_{li} must be 1 or -1. Also considering the structures of $\mathbf{L}$ (4) and $\mathbf{L}'$ (S3), we see that $\forall l > 0$, $\mathbf{A}'^l = \mathbf{A}^l\mathbf{D}_l$, where $\mathbf{D}_l$ are diagonal matrices with 1 or -1 as their diagonal entries. If both $\mathbf{A}'$ and $\mathbf{A}$ have positive diagonal entries, $\mathbf{D}$ must be the identity matrix, i.e., $\mathbf{A}' = \mathbf{A}$. Therefore statement (i) is true.

We have shown that

$$L'_{ln+i} = \lambda_{li}L_{ln+i}, \tag{S4}$$

where $\lambda_{0i} = 1$ and for $l > 0$, λ_{li} are 1 or -1. We are now ready to prove (ii). If each p_{e_i} is asymmetric, e_i and $-e_i$ have different distributions. Consequently, the representation (S2) does not hold any more if one changes the signs of a subset of, but not all, non-zero elements of $\{L'_{ln+j} \mid l = 0, ..., k-1\}$. This implies that for non-zero L_{ln+i}, λ_{li}, including λ_{0i}, have the same sign, and they are therefore 1 since $\lambda_{0i} = 1$. Setting $l = 1$ in (S4) gives $\mathbf{A}' = \mathbf{A}$. That is, (ii) is true.

Let us now show that (iii) holds. If $k = 1$, this statement trivially holds. Now consider the case where $k > 1$. Because of (S1), we have

$$\mathbf{A}^{k-1}\mathbf{A} = \mathbf{A}'^{k-1}\mathbf{A}'. \tag{S5}$$

Since $\mathbf{A}$ is of full rank, $\mathbf{A}^{k-1}$ is also invertible. Recall $\mathbf{A}'^l = \mathbf{A}^l\mathbf{D}_l$. Denote by $d_{l,i}$ the (i,i)th entry of $\mathbf{D}_l$. Multiplying both sides of the above equation with $\mathbf{A}^{-(k-1)}$ from the left gives $\mathbf{A} = \mathbf{D}_{k-1}\mathbf{A}\mathbf{D}_1$, i.e., $\forall\ i\ \&\ j$, $a_{ij} = a_{ij}d_{k-1,i}d_{1,j}$.

Thus, $\forall\ i\ \&\ j$ with $a_{ij} \neq 0$ we have $d_{k-1,i}d_{1,j} = 1$. Since a_{ii} are not zero, we have $d_{k-1,i} = d_{1,i}$. Consequently, $a_{ij} = a_{ij}d_{1,i}d_{1,j}$, and $\forall\ i\ \&\ j$ with $a_{ij} \neq 0$, $d_{1,i}d_{1,j} = 1$, or $d_{1,i} = d_{1,j}$. Furthermore, since the graph implied by $\mathbf{A}$ is weakly connected, for any two nodes i' and j', we know that there is a undirected path connecting them, such that $d_{1,i'} = d_{1,j'}$. In words, $\mathbf{D}_1$ is either $\mathbf{I}$ or $-\mathbf{I}$. Finally, if $k > 1$ is odd, $\mathbf{A'}^{k-1} = (\mathbf{A}\mathbf{D}_1)^{k-1} = \mathbf{A}^{k-1}$, and then (S5) implies that $\mathbf{A}' = \mathbf{A}$. (iii) then holds. □

2. Proof of Theorem 3 in Section 3.3

Proof. Suppose the model of Granger causality with instantaneous effects, (2), holds, the VAR error terms of $\tilde{\mathbf{x}}_t$ can be written as a linear transformation of n independent variables; denote by $\mathbf{W}$ this linear transformation.

On the other hand, the error terms $\vec{\mathbf{e}}_t$ admit the representation (5). Since $\mathbf{A}$ is not diagonal, $\mathbf{L}$ contains at least $(n+1)$ columns none of which is proportional to each other. Since all of e_{it} are non-Gaussian, Lemma 1 (i) implies that all columns in $\mathbf{L}$ are proportional to some columns in $\mathbf{W}$. This implies that $\mathbf{W}$ has at least $(n+1)$ columns none of which is proportional to each other; however, $\mathbf{W}$ has only n columns, resulting in a contradiction. Therefore the model of Granger causality with instantaneous effects does not hold. □

3. Details of the EM Algorithm in Section 4.1

Instead of directly maximizing the data log-likelihood $\sum_t \ln p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \Theta)$, the EM algorithm maximizes the lower bound of the data log-likelihood, i.e.,

$$\mathcal{L}(q, \Theta) = \sum_t \sum_{\mathbf{z}_t} \int q(\mathbf{z}_t, \tilde{\mathbf{e}}_t) \ln \frac{p(\tilde{\mathbf{x}}_t, \tilde{\mathbf{e}}_t, \mathbf{z}_t|\tilde{\mathbf{x}}_{t-1}, \Theta)}{q(\mathbf{z}_t, \tilde{\mathbf{e}}_t)}\, d\tilde{\mathbf{e}}_t, \tag{S6}$$

with respect to the distribution $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t)$ and the parameters Θ alternately until convergence.

E step In the E step, given the parameters Θ' from the previous M step, the lower bound is maximized with respect to $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t)$. The maximum lower bound is obtained when $q(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\Theta')$ equals the posterior distribution $p(\mathbf{z}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')p(\tilde{\mathbf{e}}_t|\mathbf{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')$. The posterior distribution is obtained as

$$p(\mathbf{z_t}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') = \frac{p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \mathbf{z}_t)p(\mathbf{z}_t)}{\sum_{\mathbf{z}'_t} p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \mathbf{z}'_t)p(\mathbf{z}'_t)}, \tag{S7}$$

$$\begin{aligned} p(\tilde{\mathbf{e}}_t|\mathbf{z}_t, \tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') = &\mathcal{N}(\tilde{\mathbf{e}}_t|\tilde{\boldsymbol{\mu}}_{\mathbf{z}_t} + \tilde{\Sigma}_{\mathbf{z}_t}^{\intercal}\mathbf{L}^{\intercal}(\mathbf{L}\tilde{\Sigma}_{\mathbf{z}_t}\mathbf{L}^{\intercal} + \Lambda)^{-1} \\ &(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1} - \mathbf{L}\tilde{\boldsymbol{\mu}}_{\mathbf{z}_t}), \tilde{\Sigma}_{\mathbf{z}_t} - \tilde{\Sigma}_{\mathbf{z}_t}^{\intercal} \\ &\mathbf{L}^{\intercal}(\mathbf{L}\tilde{\Sigma}_{\mathbf{z}_t}\mathbf{L}^{\intercal} + \Lambda)^{-1}\mathbf{L}\tilde{\Sigma}_{\mathbf{z}_t}), \end{aligned} \tag{S8}$$

where $\tilde{\boldsymbol{\mu}}_{\mathbf{z}_t} = (\tilde{\mu}_{1,z_{t,1}}, ..., \tilde{\mu}_{nk,z_{t,nk}})^{\intercal}$ and $\tilde{\Sigma}_{\mathbf{z}_t} = \text{diag}(\tilde{\sigma}^2_{1,z_{t,1}}, ..., \tilde{\sigma}^2_{nk,z_{t,nk}})$.

M step In the M step, given the posterior distributions (S7) (S8) from the E step, the parameters are updated by maximizing the lower bound with respect to Θ. The lower bound can be decompsed into four terms each of which only contains a subset of the parameters, i.e.,

$$\mathcal{L}(q, \Theta) = \mathcal{L}_1(q, w) + \mathcal{L}_2(q, \mu, \sigma) + \mathcal{L}_3(q, \mathbf{A}) + \mathcal{L}_4(q). \tag{S9}$$

The four terms are calculated as

$$\mathcal{L}_1 = \sum_t \sum_{i=1}^{nk} \sum_{z_{t,i}=1}^{m} p(z_{t,i}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(z_{t,i}) = \sum_t \sum_{i=1}^{nk} \sum_{z_{t,i}=1}^{p} p(z_{t,i}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln \tilde{w}_{i,z_{t,i}}, \tag{S10}$$

$$\begin{aligned} \mathcal{L}_2 &= \sum_t \sum_{i=1}^{nk} \sum_{z_{t,i}=1}^{m} \int p(\tilde{e}_{t,i}, z_{t,i}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(\tilde{e}_{t,i}|z_{t,i}) d\tilde{e}_{t,i} \\ &= -\frac{1}{2} \sum_t \sum_{i=1}^{nk} \sum_{z_{t,i}=1}^{m} \int p(\tilde{e}_{t,i}, z_{t,i}|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \left(\frac{(\tilde{e}_i - \tilde{\mu}_{i,z_{t,i}})^2}{\tilde{\sigma}^2_{i,z_{t,i}}} + \ln 2\pi + 2 \ln \tilde{\sigma}_{i,z_{t,i}} \right) d\tilde{e}_{t,i}, \end{aligned} \tag{S11}$$

$$\begin{aligned}\mathcal{L}_3 &= \sum_t \int p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{e}}_t) d\tilde{\mathbf{e}}_t, \\ &= -\frac{1}{2}\sum_t \Big\{ \left[(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})^\intercal \Lambda^{-1} (\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})\right] - 2(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})^\intercal \Lambda^{-1}\mathbf{L} \left\langle \tilde{\mathbf{e}}_t \right\rangle_{p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta'))} \\ &+ Tr\left(\mathbf{L}^\intercal\Lambda^{-1}\mathbf{L}\left\langle \tilde{\mathbf{e}}_t\tilde{\mathbf{e}}_t^\intercal \right\rangle_{p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')}\right) + \ln|\Lambda| + n\ln 2\pi \Big\}, \end{aligned} \tag{S12}$$

$$\mathcal{L}_4 = -\sum_t \sum_{\mathbf{z}_t} \int p(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') d\tilde{\mathbf{e}}_t, \tag{S13}$$

where $\langle f(e)\rangle_{p(e)} = \int p(e)f(e)de$.

Due to the zero mean constraints on the noises, $\mu_{i,c}$ and $w_{i,c}$ are updated by maximize $\mathcal{L}_1 + \mathcal{L}_2$ with the constraints $\sum_{c=1}^m w_{i,c} = 1, \sum_{c=1}^m w_{i,c}\mu_{i,c} = 0, i = 1, ..., n$. This is a constrained nonlinear programming problem and we solve it using interior point methods.

After updating $\mu_{i,c}$ and $w_{i,c}$, σ can be updated by maximizing $\mathcal{L}_2$, which gives

$$\sigma_{i,c}^2 = \frac{\sum_t \sum_{j=1}^k \left\langle \tilde{e}_{t,i+n(j-1)}^2 - 2\mu_{i,c}\tilde{e}_{t,i+n(j-1)} \right\rangle_{p(\tilde{e}_{t,i+n(j-1)}, z_{t,i+n(j-1)}=c|\mathbf{x}_t, \mathbf{x}_{t-1})}}{\sum_t \sum_{j=1}^k p(z_{t,i+n(j-1)} = c|\mathbf{x}_t, \mathbf{x}_{t-1})} + \mu_{i,c}^2, \tag{S14}$$

Since there is no analytic solution to $\mathbf{A}$, we update $\mathbf{A}$ using conjugate gradient descent algorithm. The gradient of $\mathcal{L}_3$ with respect to $\mathbf{A}$ is given by

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{A})}{\partial \mathbf{A}_{ij}} = -\frac{1}{2}\sum_t \Bigg\{ & Tr\left[-2(\Lambda^{-1}(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})\tilde{\mathbf{x}}_{t-1}^\intercal)^\intercal \sum_{r=0}^{k-1} \mathbf{A}^r\mathbf{J}^{ij}\mathbf{A}^{k-1-r}\right] \\ & -2\Bigg\{ Tr\left[-(\Lambda^{-1}\mathbf{L}\left\langle \tilde{\mathbf{e}}_t \right\rangle \mathbf{x}_t^\intercal)^\intercal \sum_{r=0}^{k-1}\mathbf{A}^r\mathbf{J}^{ij}\mathbf{A}^{k-1-r}\right] \\ & + \sum_{l=1}^{k-1} Tr\left[(\Lambda^{-1}(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})\left\langle \tilde{\mathbf{e}}_{t,l}^\intercal \right\rangle)^\intercal \sum_{r=0}^{l-1}\mathbf{A}^r\mathbf{J}^{ij}\mathbf{A}^{l-1-r}\right]\Bigg\} \\ & + Tr\left(\langle \tilde{\mathbf{e}}_t\tilde{\mathbf{e}}_t^\intercal\rangle \frac{\partial U}{\partial \mathbf{A}_{ij}}\right)\Bigg\}, \end{aligned} \tag{S15}$$

where $U = \mathbf{L}^\intercal\Lambda^{-1}\mathbf{L}$ and $\mathbf{J}^{ij}$ is a matrix whose ij-th element is 1 and all the other elements are 0. U is composed of $k * k$ blocks of $n * n$ matrices. Each sub-matrix is $U_{mn} = (\mathbf{A}^m)^\intercal\Lambda^{-1}\mathbf{A}^n, m = 0, ..., k-1, n = 0, ..., k-1$. The gradient of each sub-matrix U_{mn} is

$$\begin{aligned}\frac{\partial (U_{mn})_{kl}}{\partial \mathbf{A}_{ij}} &= Tr\left[\left(mat_{i'j'}\frac{\partial((\mathbf{A}^m)^\intercal\Lambda^{-1}\mathbf{A}^n)_{kl}}{\partial \mathbf{A}^m_{i'j'}}\right)^\intercal \frac{\partial \mathbf{A}^m}{\partial \mathbf{A}_{ij}}\right] \\ &+ Tr\left[\left(mat_{i'j'}\frac{\partial((\mathbf{A}^m)^\intercal\Lambda^{-1}\mathbf{A}^n)_{kl}}{\partial \mathbf{A}^n_{i'j'}}\right)^\intercal \frac{\partial \mathbf{A}^n}{\partial \mathbf{A}_{ij}}\right] \\ &= Tr\left[\left(mat_{i'j'}(\delta_{kj'}(\Lambda^{-1}\mathbf{A}^n)_{i'l})\right)^\intercal \sum_{r=0}^{m-1}\mathbf{A}^r\mathbf{J}^{ij}\mathbf{A}^{m-1-r}\right] \\ &+ Tr\left[\left(mat_{i'j'}(\delta_{lj'}((\mathbf{A}^m)^\intercal\Lambda^{-1})_{ki'})\right)^\intercal \sum_{r=0}^{n-1}\mathbf{A}^r\mathbf{J}^{ij}\mathbf{A}^{n-1-r}\right], \end{aligned} \tag{S16}$$

where $mat_{i'j'}f(i', j')$ is a matrix whose $i'j'$-th element is $f(i', j')$.

References

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