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

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

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

$$\mathcal{L}_3 = \sum_t \int p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{e}}_t) d\tilde{\mathbf{e}}_t,$$
$$= -\frac{1}{2}\sum_t \Big\{ \big[(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})^\intercal \Lambda^{-1} (\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})\big] - 2(\tilde{\mathbf{x}}_t - \mathbf{A}^k\tilde{\mathbf{x}}_{t-1})^\intercal \Lambda^{-1}\mathbf{L} \left\langle \tilde{\mathbf{e}}_t \right\rangle_{p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta'))}$$
$$+Tr\left(\mathbf{L}^\intercal \Lambda^{-1}\mathbf{L} \left\langle \tilde{\mathbf{e}}_t\tilde{\mathbf{e}}_t^\intercal \right\rangle_{p(\tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta')}\right) + \ln|\Lambda| + n\ln 2\pi \Big\}, \tag{S12}$$
$$\mathcal{L}_4 = -\sum_t\sum_{\mathbf{z}_t} \int p(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') \ln p(\mathbf{z}_t, \tilde{\mathbf{e}}_t|\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}, \Theta') d\tilde{\mathbf{e}}_t, \tag{S13}$$

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

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

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

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

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.