Supplementary Material for "Causal Discovery and Forecasts in Nonstationary Environments with State-Space Models"

Biwei Huang [1] **Kun Zhang** [1] **Mingming Gong** [1 2] **Clark Glymour** [1]

S1. Proof of Theorem 1

Before the proof, we first give Lemma S1, which gives the identifiability of parameters in varying coefficients regression models. This result will be used in the proof of Theorem 1.

Lemma S1 ((Wall, 1987))**.** *The varying coefficients regression model takes on the following form:*

$$\begin{cases} y_t = \sum_i b_{i,t} x_{i,t} + e_t, \\ b_{i,t} = \alpha_{i,0} + \sum_p \alpha_{i,p} b_{i,t-p} + \epsilon_{i,t}, \end{cases} \tag{1}$$

where y_t is the scalar valued dependent variable and $x_{i,t}$ is the independent variable which we have observations. The additive error, e_t, represents a stationary zero mean white noise process, i.e., $E[e_t] = 0$ and $E[e_t e_{t'}] = \sigma_e^2 \delta_{tt'}$, $E[\epsilon_{i,t}] = 0$, where $\sigma_e^2 < \infty$ and $\delta_{tt'}$ is the delta function. Similarly, $E[\epsilon_{i,t}] = 0$ and $E[\epsilon_{i,t}\epsilon_{i,t'}] = \sigma_{\epsilon_i}^2 \delta_{tt'}$, for $\forall i \in \mathbf{N}^+$.

Then the parameters σ_e^2, $\alpha_{i,0}$, $\alpha_{i,p}$, $\sigma_{\epsilon_i}^2$, for $\forall i, p \in \mathbf{N}^+$ are globally identifiable.

Now we start to prove Theorem 1.

Proof of Theorem 1. The proof of Theorem 1 contains two phases. In the first phase, we identify the root variable and corresponding causal parameters. In the second phase, we identify the remaining causal graph and corresponding parameters in a recursive way.

Phase I Let $A_t = (I - B_t)^{-1}$. We define the following metric to characterize kurtosis of observed variables:

$$\begin{aligned} & E\big[x_{i,t} x_{j,t} x_{k,t+p} x_{l,t+p}\big] \\ = \; & E\big[\textstyle\sum_a A_{ia,t} E_{a,t} \sum_b A_{ib,t} E_{b,t} \sum_c A_{ic,t+p} E_{c,t+p} \sum_d A_{id,t+p} E_{d,t+p}\big] \\ = \; & E\big[\textstyle\sum_a \sum_b \sum_c \sum_d A_{ia,t} A_{ib,t} A_{ic,t+p} A_{id,t+p} E_{a,t} E_{b,t} E_{c,t+p} E_{d,t+p}\big] \\ = \; & \textstyle\sum_a \sum_c E\big[A_{ia,t} A_{ia,t} A_{ic,t+p} A_{ic,t+p} E_{a,t} E_{a,t} E_{c,t+p} E_{c,t+p}\big] \\ = \; & \sigma_a^2 \sigma_c^2 \textstyle\sum_a \sum_c E\big[A_{ia,t} A_{ia,t} A_{ic,t+p} A_{ic,t+p}\big], \end{aligned} \tag{2}$$

where the third equation holds because only when $a = b$ and $c = d$, the expectation is not zero.

By considering all combinations of i, j, k, and l, we can organize the above kurtosis in the matrix form,

$$\begin{aligned} & \mathbf{S}(t, t+p) \\ = \; & E\big[(A_t \Sigma_E A_t^T) \otimes (A_{t+p} \Sigma_E A_{t+p}^T)\big] \\ = \; & E\big[(I - B_t)^{-1} \Sigma_E (I - B_t)^{-T}) \otimes ((I - B_{t+p})^{-1} \Sigma_E (I - B_{t+p})^{-T})\big]. \end{aligned} \tag{3}$$

[1]Department of Philosophy, Carnegie Mellon University, Pittsburgh [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh. Correspondence to: Biwei Huang <biweih@andrew.cmu.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

where $\mathbf{S}(t, t+p)$ is a $n^2 \times n^2$ matrix, $\otimes$ denotes Kronecker product, and

$$\begin{aligned}&E\big[x_{i,t}x_{j,t}x_{k,t+p}x_{l,t+p}\big]\\&= \mathbf{S}(t,t+p)_{ijkl}\\&= E\big[\{(I-B_t)^{-1}\Sigma_E(I-B_t)^{-T})\}_{ij}\cdot\{(I-B_{t+p})^{-1}\Sigma_E(I-B_{t+p})^{-T})\}_{kl}\big].\end{aligned}$$

If the underlying graph is a DAG, then $(I-B_t)^{-1}$ can be reformulated as

$$(I-B_t)^{-1} = \sum_{r=0}^{m-1} B_t^r,$$

and thus

$$\mathbf{S}(t,t+p) = E\big[\big((\sum_{r=0}^{m-1} B_t^r)\Sigma_E(\sum_{r=0}^{m-1} B_t^{Tr})\big)\otimes\big((\sum_{r=0}^{m-1} B_{t+p}^r)\Sigma_E(\sum_{r=0}^{m-1} B_{t+p}^{Tr})\big)\big]. \tag{4}$$

Particularly, let us consider the case when $i=j=k=l$ and for notation simplicity, denote $S(t,t+p)_{ijkl}$ by $S(t,t+p)_i$. Then

$$S(t,t+p)_i = E[x_{i,t}^2 x_{i,t+p}^2]. \tag{5}$$

Let r_0 be the index of the root cause, and $\mathbf{V}_s = \mathbf{V}\backslash r_0$ denote the indices of the remaining processes, with $\mathbf{V} = \{1,\cdots,m\}$. Then we will have

$$\begin{aligned}&S(t,t+p)_{r_0} - S(t,t)_{r_0} = 0,\\&S(t,t+p)_{r_s} - S(t,t)_{r_s} < 0, \quad \forall r_s \in \mathbf{V}_s,\end{aligned} \tag{6}$$

for any $p \in \mathbf{N}^+$. The reason is that the root cause does not receive changing influences from other processes.

Let us now give the detailed proof procedure of (6). Suppose that we have known the causal order, denoted by π. Let us first see a few examples of the concrete representations of $S(t,t+p)_i$.

1. For the root cause $\pi(1)$, it is easy to get $S(t,t+p)_{\pi(1)} = (\sigma_{\pi(1)}^2)^2$, which is irrelevant to p. Thus,

$$S(t,t+p)_{\pi(1)} - S(t,t)_{\pi_{(1)}} = 0.$$

2. For $\pi(2)$, we have

$$\begin{aligned}&S(t,t+p)_{\pi(2)}\\=\;&(\sigma_{\pi(2)}^2)^2 + \sigma_{\pi(1)}^2\sigma_{\pi(2)}^2 E[b_{\pi(2)\pi(1),t+p}^2] + \sigma_{\pi(1)}^2\sigma_{\pi(2)}^2 E[b_{\pi(2)\pi(1),t}^2] + (\sigma_{\pi(1)}^2)^2 E[b_{\pi(2)\pi(1),t}^2 b_{\pi(2)\pi(1),t+p}^2].\end{aligned}$$

Thus,

$$\begin{aligned}&S(t,t+p)_{\pi(2)} - S(t,t)_{\pi(2)}\\=\;&(\sigma_{\pi(1)}^2)^2\Big(E[b_{\pi(2)\pi(1),t}^2 b_{\pi(2)\pi(1),t+p}^2] - E[b_{\pi(2)\pi(1),t}^2 b_{\pi(2)\pi(1),t}^2]\Big)\\=\;&(\sigma_{\pi(1)}^2)^2\Big(E[b_{\pi(2)\pi(1),t}^2\cdot(\alpha_{\pi(2)\pi(1)}^p b_{\pi(2)\pi(1),t} + \alpha_{\pi(2)\pi(1)}^{p-1}\epsilon_{t+1} + \cdots + \alpha_{\pi(2)\pi(1)}^0\epsilon_{t+p})^2] - E[b_{\pi(2)\pi(1),t}^2 b_{\pi(2)\pi(1),t}^2]\Big)\\=\;&(\sigma_{\pi(1)}^2)^2\cdot 2(\alpha_{\pi(2)\pi(1)}^{2p} - 1)\tfrac{(w_{\pi(2)\pi(1)})^2}{(1-\alpha_{\pi(2)\pi(1)}^2)^2}\\<\;&0,\end{aligned}$$

where $w_{\pi(2)\pi(1)}$ is the noise variance in the autoregressive model of $b_{\pi(2)\pi(1),t}$. $S(t,t+p)_{\pi(2)} - S(t,t)_{\pi(2)} < 0$ always holds because $\alpha_{\pi(2)\pi(1)}^{2p} - 1 < 0, \forall p \in \mathbf{N}^+$, since $\alpha_{\pi(2)\pi(1)} \in (-1,1)$.

3. For $\pi(3)$, we have

$$\begin{aligned}
& S(t,t+p)_{\pi(3)} \\
= \; & (\sigma^2_{\pi(1)})^2 E[b^2_{\pi(3)\pi(1),t} b^2_{\pi(3)\pi(1),t+p}] \\
& +4(\sigma^2_{\pi(1)})^2 E[b_{\pi(2)\pi(1),t} b_{\pi(2)\pi(1),t+p}] E[b_{\pi(3)\pi(1),t} b_{\pi(3)\pi(1),t+p}] E[b_{\pi(3)\pi(1),t} b_{\pi(3)\pi(2),t+p}] \\
& +(\sigma^2_{\pi(1)})^2 E[b^2_{\pi(2)\pi(1),t} b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t} b^2_{\pi(3)\pi(2),t+p}] + (\sigma^2_{\pi(2)})^2 E[b^2_{\pi(3)\pi(2),t} b^2_{\pi(3)\pi(2),t+p}] \\
& +(\sigma^2_{\pi(3)})^2 + (\sigma^2_{\pi(1)})^2 E[b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t+p}] \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(2)} E[b^2_{\pi(3)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t+p}] + \sigma^2_{\pi(1)}\sigma^2_{\pi(3)} E[b^2_{\pi(3)\pi(1),t}] \\
& +(\sigma^2_{\pi(1)})^2 E[b^2_{\pi(2)\pi(1),t}] E[b^2_{\pi(3)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t}] + \sigma^2_{\pi(1)}\sigma^2_{\pi(2)} E[b^2_{\pi(3)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t+p}] \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(2)} E[b^2_{\pi(3)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t+p}] \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(2)} E[b^2_{\pi(3)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t}] + \sigma^2_{\pi(1)}\sigma^2_{\pi(2)} E[b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t+p}] \\
& +\sigma^2_{\pi(2)}\sigma^2_{\pi(3)} E[b^2_{\pi(3)\pi(2),t}] + \sigma^2_{\pi(1)}\sigma^2_{\pi(3)} E[b^2_{\pi(3)\pi(1),t+p}] \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(3)} E[b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t+p}] + \sigma^2_{\pi(2)}\sigma^2_{\pi(3)} E[b^2_{\pi(3)\pi(2),t+p}].
\end{aligned}$$

Thus,

$$\begin{aligned}
& S(t,t+p)_{\pi(3)} - S(t,t)_{\pi(3)} \\
= \; & (\sigma^2_{\pi(1)})^2 \Big(E[b^2_{\pi(3)\pi(1),t} b^2_{\pi(3)\pi(1),t+p}] - E[b^2_{\pi(3)\pi(1),t} b^2_{\pi(3)\pi(1),t}] \Big) \\
& +4(\sigma^2_{\pi(1)})^2 \Big(E[b_{\pi(2)\pi(1),t} b_{\pi(2)\pi(1),t+p}] E[b_{\pi(3)\pi(1),t} b_{\pi(3)\pi(1),t+p}] E[b_{\pi(3)\pi(2),t} b_{\pi(3)\pi(2),t+p}] \\
& \qquad - E[b_{\pi(2)\pi(1),t} b_{\pi(2)\pi(1),t}] E[b_{\pi(3)\pi(1),t} b_{\pi(3)\pi(1),t}] E[b_{\pi(3)\pi(2),t} b_{\pi(3)\pi(2),t}] \Big) \\
& +(\sigma^2_{\pi(1)})^2 \Big(E[b^2_{\pi(2)\pi(1),t} b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t} b^2_{\pi(3)\pi(2),t+p}] - E[b^2_{\pi(2)\pi(1),t} b^2_{\pi(2)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t} b^2_{\pi(3)\pi(2),t}] \Big) \\
& +(\sigma^2_{\pi(2)})^2 \Big(E[b^2_{\pi(3)\pi(2),t} b^2_{\pi(3)\pi(2),t+p}] - E[b^2_{\pi(3)\pi(2),t} b^2_{32,t}] \Big) \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(2)} \Big(E[b^2_{\pi(3)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t+p}] - E[b^2_{\pi(3)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t}] \Big) \\
& +\sigma^2_{\pi(1)}\sigma^2_{\pi(1)} \Big(E[b^2_{\pi(2)\pi(1),t+p}] E[b^2_{\pi(3)\pi(2),t+p}] - E[b^2_{\pi(2)\pi(1),t}] E[b^2_{\pi(3)\pi(2),t}] \Big) \\
< \; & 0
\end{aligned}$$

Generally, we have the following form for variable with the nth $(n > 1)$ order:

$$\begin{aligned}
& S(t,t+p)_{\pi(n)} \\
= \; & \Big(\textstyle\sum_{i=1}^{\pi(n)-1} (\sum_{k>1,l} p^{(k,l)}_{\pi(n)i,t})^2 \sigma^2_i + \sigma^2_{\pi(n)} \Big) \cdot \Big(\sum_{i'=1}^{\pi(n)-1} (\sum_{k'>1,l'} p^{(k',l')}_{\pi(n)i',t+p})^2 \sigma^2_{i'} + \sigma^2_{\pi(n)} \Big),
\end{aligned} \tag{7}$$

where $p^{(k,l)}_{\pi(n)i,t}$ indicates the multiplication of causal coefficients of the lth directed path [1] from node i to node $\pi(n)$ with path length k. For example, for path with length 1, $p^{(1)}_{\pi(n)i,t} = b_{\pi(n)i,t}$; for path with length 2, $p^{(2)}_{\pi(n)i,t} = b_{\pi(n)j,t} b_{ji,t}$, with $i < j < \pi(n)$; for path with length 3, $p^{(3)}_{\pi(n)i,t} = b_{\pi(n)j_1,t} b_{j_1 j_2,t} b_{j_2 i,t}$, with $i < j_2 < j_1 < \pi(n)$. The longest path length from node i to node $\pi(n)$ is $\pi(n) - i$.

By expanding (7), we have

$$\begin{aligned}
& S(t,t+p)_{\pi(n)} \\
= \; & (\sigma^2_{\pi(n)})^2 + \textstyle\sum_{i<\pi(n)} E[b^2_{\pi(n)i,t} b^2_{\pi(n)i,t+p}] (\sigma^2_i)^2 \\
& + \textstyle\sum_{i<\mathbf{j}<\pi(n)} E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t+p}] E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t+p}] (\sigma^2_i)^2 \\
& + \textstyle\sum_{i<\mathbf{j},\mathbf{j}'<\pi(n),\mathbf{j}\neq\mathbf{j}'} 4E[b_{\pi(n)\mathbf{j},t} b_{\pi(n)\mathbf{j},t+p}] E[b_{\mathbf{j}i,t} b_{\mathbf{j}i,t+p}] E[b_{\pi(n)\mathbf{j}',t} b_{\pi(n)\mathbf{j}',t+p}] E[b_{\mathbf{j}'i,t} b_{\mathbf{j}'i,t+p}] (\sigma^2_i)^2 \\
& + \textstyle\sum_{i<\mathbf{j}<\pi(n),i'<\mathbf{j}'<\pi(n),\mathbf{j}\neq\mathbf{j}'\cup i\neq i'} E[b^2_{\pi(n)\mathbf{j},t} b^2_{\mathbf{j}i,t} b^2_{\pi(n)\mathbf{j}',t+p} b^2_{\mathbf{j}'i',t+p}] \sigma^2_i \sigma^2_{i'}
\end{aligned}$$

[1] A directed path is defined as a sequence of edges (or arcs) which connect a sequence of vertices, and the edges are all directed in the same direction.

where $\mathbf{j}$ and $\mathbf{j}'$ can be a series of indices, $\mathbf{j} = (j_1, j_2, \cdots, j_k)$ with $\pi(n) > j_1 > j_2 > \cdots i$, and $\mathbf{j}' = (j'_1, j'_2, \cdots, j'_k)$ with $\pi(n) > j'_1 > j'_2 > \cdots i$, and

$$
\begin{aligned}
&E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t+p}]E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t+p}] = E[b^2_{\pi(n)j_1,t} b^2_{\pi(n)j_1,t+p}]E[b^2_{j_1 j_2,t} b^2_{j_1 j_2,t+p}] \cdots E[b^2_{j_k i,t} b^2_{j_k i,t+p}],\\
&E[b_{\pi(n)\mathbf{j},t} b_{\pi(n)\mathbf{j},t+p}]E[b_{\mathbf{j}i,t} b_{\mathbf{j}i,t+p}] = E[b_{\pi(n)j_1,t} b_{\pi(n)j_1,t+p}]E[b_{j_1 j_2,t} b_{j_1 j_2,t+p}] \cdots E[b_{j_k i,t} b_{j_k i,t+p}],\\
&E[b_{\pi(n)\mathbf{j}',t} b_{\pi(n)\mathbf{j}',t+p}]E[b_{\mathbf{j}'i,t} b_{\mathbf{j}'i,t+p}] = E[b_{\pi(n)j'_1,t} b_{\pi(n)j'_1,t+p}]E[b_{j'_1 j'_2,t} b_{j'_1 j'_2,t+p}] \cdots E[b_{j'_k i,t} b_{j'_k i,t+p}],\\
&E[b^2_{\pi(n)\mathbf{j},t} b^2_{\mathbf{j}i,t} b^2_{\pi(n)\mathbf{j}',t+p} b^2_{\mathbf{j}'i',t+p}] = E[b^2_{\pi(n)j_1,t} b^2_{j_1 j_2,t} \cdots b^2_{j_k i,t} b^2_{n j'_1,t+p} b^2_{j'_1 j'_2,t+p} \cdots b^2_{j'_k i',t+p}].
\end{aligned}
$$

Then

$$
\begin{aligned}
& S(t,t+p)_{\pi(n)} - S(t,t)_{\pi(n)}\\
= \ & \textstyle\sum_{i<\pi(n)} \Big(E[b^2_{\pi(n)i,t} b^2_{\pi(n)i,t+p}] - E[b^2_{\pi(n)i,t} b^2_{\pi(n)i,t}] \Big) (\sigma_i^2)^2\\
& + \textstyle\sum_{i<\mathbf{j}<\pi(n)} \Big(E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t+p}]E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t+p}] - E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t}]E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t}] \Big) (\sigma_i^2)^2\\
& + \textstyle\sum_{i<\mathbf{j},\mathbf{j}'<\pi(n),\mathbf{j}\neq\mathbf{j}'} 4 \Big(E[b_{\pi(n)\mathbf{j},t} b_{\pi(n)\mathbf{j},t+p}]E[b_{\mathbf{j}i,t} b_{\mathbf{j}i,t+p}]E[b_{\pi(n)\mathbf{j}',t} b_{\pi(n)\mathbf{j}',t+p}]E[b_{\mathbf{j}'i,t} b_{\mathbf{j}'i,t+p}]\\
& \qquad - E[b^2_{\pi(n)\mathbf{j},t}]E[b^2_{\mathbf{j}i,t}]E[b^2_{\pi(n)\mathbf{j}',t}]E[b^2_{\mathbf{j}'i,t}] \Big) (\sigma_i^2)^2\\
& + \textstyle\sum_{i<\mathbf{j}<\pi(n),i'<\mathbf{j}'<\pi(n),\mathbf{j}\neq\mathbf{j}'\cup i\neq i'} \Big(E[b^2_{\pi(n)\mathbf{j},t} b^2_{\mathbf{j}i,t} b^2_{n\mathbf{j}',t+p} b^2_{\mathbf{j}'i,t+p}] - E[b^2_{\pi(n)\mathbf{j},t} b^2_{\mathbf{j}i,t} b^2_{\pi(n)\mathbf{j}',t} b^2_{\mathbf{j}'i,t}] \Big) \sigma_i^2 \sigma_{i'}^2,
\end{aligned}
$$

where

$$
E[b^2_{\pi(n)i,t} b^2_{\pi(n)i,t+p}] - E[b^2_{\pi(n)i,t} b^2_{\pi(n)i,t}] = 2(\alpha^{2p}_{\pi(n)i} - 1)\tfrac{(w_{\pi(n)i})^2}{(1-\alpha^2_{\pi(n)i})^2} < 0,
$$

$$
\begin{aligned}
& E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t+p}]E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t+p}] - E[b^2_{\pi(n)\mathbf{j},t} b^2_{\pi(n)\mathbf{j},t}]E[b^2_{\mathbf{j}i,t} b^2_{\mathbf{j}i,t}]\\
& = \big((1+2\alpha^{2p}_{\pi(n)j})(1+2\alpha^{2p}_{ji}) - 9\big) \cdot \tfrac{(w_{\pi(n)j})^2}{(1-\alpha^2_{\pi(n)j})^2} \tfrac{(w_{ji})^2}{(1-\alpha^2_{ji})^2} < 0,
\end{aligned}
$$

$$
\begin{aligned}
& E[b_{\pi(n)\mathbf{j},t} b_{\pi(n)\mathbf{j},t+p}]E[b_{\mathbf{j}i,t} b_{\mathbf{j}i,t+p}]E[b_{\pi(n)\mathbf{j}',t} b_{\pi(n)\mathbf{j}',t+p}]E[b_{\mathbf{j}'i',t} b_{\mathbf{j}'i',t+p}] - E[b^2_{\pi(n)\mathbf{j},t}]E[b^2_{\mathbf{j}i,t}]E[b^2_{\pi(n)\mathbf{j}',t}]E[b^2_{\mathbf{j}'i',t}]\\
& = (\alpha^p_{\pi(n)j}\alpha^p_{\mathbf{j}i}\alpha^p_{\pi(n)\mathbf{j}'}\alpha^p_{\mathbf{j}'i'} - 1)E[b^2_{\pi(n)\mathbf{j},t}]E[b^2_{\mathbf{j}i,t}]E[b^2_{\pi(n)\mathbf{j}',t}]E[b^2_{\mathbf{j}'i',t}] < 0,
\end{aligned}
$$

$$
E[b^2_{n\mathbf{j},t} b^2_{\mathbf{j}i,t} b^2_{n\mathbf{j}',t+p} b^2_{\mathbf{j}'i,t+p}] - E[b^2_{n\mathbf{j},t} b^2_{\mathbf{j}i',t} b^2_{n\mathbf{j}',t} b^2_{\mathbf{j}'i,t}] \leq 0.
$$

Thus,

$$
S(t,t+p)_{\pi(n)} - S(t,t)_{\pi(n)} < 0 \text{ for } n > 1. \tag{8}
$$

Hence, $S(t,t+p)_{\pi(n)} - S(t,t)_{\pi(n)} < 0$ for $n > 1$, and $S(t,t+p)_{\pi(1)} - S(t,t)_{\pi(1)} = 0$. Therefore, we can identify the root cause. Since $S(t,t+p)_{\pi(1)} = \sigma^4_{\pi(1)}$, the corresponding parameter $\sigma^2_{\pi(1)}$ is also identifiable.

Phase II After identifying the root cause, in the second phase, we then identify the causal model over the remaining variables in a recursive way.

Suppose that we have known the causal order π. Then for each $x_{\pi(i)}$, its generating process can be reformulated as:

$$
\begin{aligned}
x_{\pi(i),t} &= b^T_{\pi(i),t} x_{\pi_{1,\cdots,i-1},t} + e_{\pi(i),t},\\
b_{\pi(i),t} &= A_{\pi(i)} b_{\pi(i),t-1} + \epsilon_{\pi(i),t},
\end{aligned} \tag{9}
$$

where $x_{\pi_{1,\cdots,i-1},t} = [x_{\pi(1),t}, \cdots, x_{\pi(i-1),t}]$, representing potential causes to $x_{\pi(i),t}$.

Lemma S1 has shown that parameters of the varying coefficients regression model are identifiable. Thus, parameters in (9) are identifiable; that is, the corresponding parameters $A_{\pi(i)}$, $\sigma^2_{e_{\pi(i)}}$(the variance of $e_{\pi(i),t}$), and $\sigma^2_{\epsilon_{\pi(i)}}$(the variance of $\epsilon_{\pi(i),t}$) are all identifiable, given the causal order.

We define a node's *level* in a acyclic graph the number of nodes in the directed path from the root to the node. For instance, the root has level 1, and any one of its adjacent nodes has level 2.

Suppose that we have identified the causal model of variables at the first n levels. Next, we will identify variables at the $(n+1)$th level and their corresponding parameters. Let $\mathbf{V}_n$ represent variable indices of the first n processes, and let $\bar{\mathbf{V}}_n = \mathbf{V} \backslash \mathbf{V}_n$. In the following, we will show that for node $r_s \in \bar{\mathbf{V}}_n$ which is at the $(n+1)$th level, $S(t, t+p)_{r_s}$ can be totally explained by a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$, but not for other nodes.

We denote $r_s \in \bar{\mathbf{V}}_n$ at the $(n+1)$th level as $\pi(n+1)$. Then we have

$$\begin{aligned}
& S(t,t+p)_{\pi(n+1)} \\
= \ & E\big[x^2_{\pi(n+1),t} x^2_{\pi(n+1),t+p}\big] \\
= \ & E\big[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t} + e_{\pi(n+1),t})^2 (b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p} + e_{\pi(n+1),t+p})^2\big] \\
= \ & E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2 (b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2] + E[e_{\pi(n+1)^2,t}] E[e_{\pi(n+1)^2,t+p}] \\
& + E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2] E[e^2_{\pi(n+1),t+p}] + E[(b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2] E[e^2_{\pi(n+1),t}] \\
= \ & E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2 (b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2] + (\sigma^2_{e_{\pi(n+1)}})^2 + E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2] \sigma^2_{e_{\pi(n+1)}} \\
& + E[(b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2] \sigma^2_{e_{\pi(n+1)}} \\
= \ & \textstyle\sum_{i,j} E[b^2_{i,t} b^2_{j,t+p}] E[x^2_{i,t} x^2_{j,t+p}] + 2 \sum_{i,j\neq k} E[b^2_{i,t} b_{j,t+p} b_{k,t+p}] E[x^2_{i,t} x_{j,t+p} x_{k,t+p}] + (\sigma^2_{e_{\pi(n+1)}})^2 \\
& + 2 \textstyle\sum_{i,j\neq k} E[b^2_{i,t+p} b_{j,t} b_{k,t}] E[x^2_{i,t+p} x_{j,t} x_{k,t}] + 4 \sum_{i\neq j, k\neq l} E[b_{i,t} b_{j,t} b_{k,t+p} b_{l,t+p}] E[x_{i,t} x_{j,t} x_{k,t+p} x_{l,t+p}] \\
& + \textstyle\sum_i E[b^2_{i,t}] E[x^2_{i,t}] \sigma^2_{e_{\pi(n+1)}} + \sum_{i\neq j} 2E[b_{i,t} b_{j,t}] E[x_{i,t} x_{j,t}] \sigma^2_{e_{\pi(n+1)}} + \sum_i E[b^2_{i,t+p}] E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n+1)}} \\
& + \textstyle\sum_{i\neq j} 2E[b_{i,t+p} b_{j,t+p}] E[x_{i,t+p} x_{j,t+p}] \sigma^2_{e_{\pi(n+1)}} \\
= \ & \textstyle\sum_{i\neq j} E[b^2_{i,t}] E[b^2_{j,t+p}] E[x^2_{i,t} x^2_{j,t+p}] + \sum_i E[b^2_{i,t} b^2_{i,t+p}] E[x^2_{i,t} x^2_{i,t+p}] + (\sigma^2_{e_{\pi(n+1)}})^2 \\
& + \textstyle\sum_{i\neq j} 4E[b_{i,t} b_{i,t+p}] E[b_{j,t} b_{j,t+p}] E[x_{i,t} x_{i,t+p} x_{j,t} x_{j,t+p}] + \sum_i E[b^2_{i,t}] E[x^2_{i,t}] \sigma^2_{e_{\pi(n+1)}} \\
& + \textstyle\sum_i E[b^2_{i,t+p}] E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n+1)}} \\
= \ & \textstyle\sum_{i\neq j} E[b^2_{i,t}] E[b^2_{j,t}] E[x^2_{i,t} x^2_{j,t+p}] + \sum_i \frac{1+2\alpha_i^{2p}}{3} E[b^4_{i,t}] E[x^2_{i,t} x^2_{i,t+p}] + (\sigma^2_{e_{\pi(n+1)}})^2 \\
& + \textstyle\sum_{i\neq j} 4\alpha_i^p \alpha_j^p E[b^2_{i,t}] E[b^2_{j,t}] E[x_{i,t} x_{i,t+p} x_{j,t} x_{j,t+p}] + \sum_i E[b^2_{i,t}] E[x^2_{i,t}] \sigma^2_{e_{\pi(n+1)}} + \sum_i E[b^2_{i,t}] E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n+1)}} \\
= \ & \textstyle\sum_{i\neq j} \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x^2_{i,t} x^2_{j,t+p}] + \sum_i \frac{1+2\alpha_i^{2p}}{3} 3(\frac{\sigma_i^2}{1-\alpha_i^2})^2 E[x^2_{i,t} x^2_{i,t+p}] + (\sigma^2_{e_{\pi(n+1)}})^2 \\
& + \textstyle\sum_{i\neq j} 4\alpha_i^p \alpha_j^p \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x_{i,t} x_{i,t+p} x_{j,t} x_{j,t+p}] + \sum_i \frac{\sigma_i^2}{1-\alpha_i^2} E[x^2_{i,t}] \sigma^2_{e_{\pi(n+1)}} + \sum_i \frac{\sigma_i^2}{1-\alpha_i^2} E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n+1)}}
\end{aligned} \tag{10}$$

with $1 \leq i, j \leq \pi(n)$, $E[b^2_{i,t}] = \frac{\sigma_i^2}{1-\alpha_i^2}$, and $E[b^4_{i,t}] = 3(\frac{\sigma_i^2}{1-\alpha_i^2})^2$. We can see that $S(t, t+p)_{\pi(n+1)}$ is determined by corresponding parameters in the causal models of $x_{\mathbf{V}_n,t}$ and a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$. We denote the set of parameters by $\theta_{\pi(n+1)}$. We can find a $\theta_{\pi(n+1)}$, so that $\forall p$, $S(t, t+p)_{\pi(n+1)}$ can be totally explained by a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$.

Next we show that for other nodes $r_s \in \bar{\mathbf{V}}_n$ which are at the n'th level with $n' > n+1$, $S(t,t+p)_{r_s}$ can not be totally explained by a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$. For $x_{\pi(n')}$, its potential causes are $x_{\mathbf{V}_n,t} \cup z$; here we use z to denote the set of variables which are from the $(n+2)$th to $(n'-1)$th level. Then

$$
\begin{array}{rl}
& S(t,t+p)_{\pi(n')} \\
= & E\big[x^2_{\pi(n'),t} x^2_{\pi(n'),t+p}\big] \\
= & E\big[(b^T_{\pi(n'),t} x_{\mathbf{V}_n,t} + b^T_{z,t} z_t + e_{\pi(n'),t})^2 (b^T_{\pi(n'),t+p} x_{\mathbf{V}_n,t+p} + b^T_{z,t} z_t + e_{\pi(n'),t+p})^2\big] \\
= & E[(b^T_{\pi(n'),t} x_{\mathbf{V}_n,t})^2 (b^T_{\pi(n'),t+p} x_{\mathbf{V}_n,t+p})^2] + (\sigma^2_{e_{\pi(n')}})^2 + E[(b^T_{\pi(n'),t} x_{\mathbf{V}_n,t})^2] \sigma^2_{e_{\pi(n')}} \\
& + E[(b^T_{\pi(n'),t+p} x_{\mathbf{V}_n,t+p})^2] \sigma^2_{e_{\pi(n')}} + E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2 (b^T_{z,t+p} z_{t+p})^2] + E[(b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2 (b^T_{z,t} z_t)^2] \\
& + E[(b^T_{z,t} z_t)^2] \sigma^2_{e_{\pi(n')}} + E[(b^T_{z,t+p} z_{t+p})^2] \sigma^2_{e_{\pi(n')}} + E[(b^T_{z,t} z_t)^2 (b^T_{z,t+p} z_{t+p})^2] \\
& + 4E[(b^T_{\pi(n+1),t} x_{\mathbf{V}_n,t})^2 (b^T_{\pi(n+1),t+p} x_{\mathbf{V}_n,t+p})^2 (b^T_{z,t} z_t)^2 (b^T_{z,t+p} z_{t+p})^2] \\
= & \sum_{i \neq j} E[b^2_{i,t}] E[b^2_{j,t}] E[x^2_{i,t} x^2_{j,t+p}] + \sum_i \frac{1+2a_i^{2p}}{3} E[b^4_{i,t}] E[x^2_{i,t} x^2_{i,t+p}] + \sum_{i \neq j} 4a_i^p a_j^p E[b^2_{i,t}] E[b^2_{j,t}] E[x_{i,t} x_{i,t+p} x_{j,t} x_{j,t+p}] \\
& + \sum_i E[b^2_{i,t}] E[x^2_{i,t}] \sigma^2_{e_{\pi(n')}} + \sum_i E[b^2_{i,t}] E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n')}} + (\sigma^2_{e_{\pi(n')}})^2 + \sum_{i',j} E[b^2_{i',t}] E[b^2_{j,t}] E[x^2_{i',t} x^2_{j,t+p}] \\
& + \sum_{i'} \frac{1+2a_{i'}^{2p}}{3} E[b^4_{i',t}] E[x^2_{i',t} x^2_{i',t+p}] + \sum_{i',j} 4a_{i'}^p a_j^p E[b^2_{i',t}] E[b^2_{j,t}] E[x_{i',t} x_{i',t+p} x_{j,t} x_{j,t+p}] \\
& + \sum_{i'} E[b^2_{i',t}] E[x^2_{i',t}] \sigma^2_{e_{\pi(n')}} + \sum_{i'} E[b^2_{i',t}] E[x^2_{i',t+p}] \sigma^2_{e_{\pi(n')}} + \sum_{i,j'} E[b^2_{i,t}] E[b^2_{j',t}] E[x^2_{i,t} x^2_{j',t+p}] \\
& + \sum_{i \neq j'} 4a_i^p a_{j'}^p E[b^2_{i,t}] E[b^2_{j',t}] E[x_{i,t} x_{i,t+p} x_{j',t} x_{j',t+p}] + \sum_{i',j'} E[b^2_{i',t}] E[b^2_{j',t}] E[x^2_{i',t} x^2_{j',t+p}] \\
& + \sum_{i' \neq j'} 4a_{i'}^p a_{j'}^p E[b^2_{i',t}] E[b^2_{j',t}] E[x_{i',t} x_{i',t+p} x_{j',t} x_{j',t+p}] \\
= & \sum_{i \neq j} \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x^2_{i,t} x^2_{j,t+p}] + \sum_i \frac{1+2a_i^{2p}}{3} 3(\frac{\sigma_i^2}{1-\alpha_i^2})^2 E[x^2_{i,t} x^2_{i,t+p}] + \sum_{i \neq j} 4a_i^p a_j^p \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x_{i,t} x_{i,t+p} x_{j,t} x_{j,t+p}] \\
& + \sum_i \frac{\sigma_i^2}{1-\alpha_i^2} E[x^2_{i,t}] \sigma^2_{e_{\pi(n')}} + \sum_i \frac{\sigma_i^2}{1-\alpha_i^2} E[x^2_{i,t+p}] \sigma^2_{e_{\pi(n')}} + (\sigma^2_{e_{\pi(n')}})^2 \\
& + \sum_{i',j} \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x^2_{i',t} x^2_{j,t+p}] + \sum_{i'} \frac{1+2a_{i'}^{2p}}{3} 3(\frac{\sigma_{i'}^2}{1-\alpha_{i'}^2})^2 E[x^2_{i',t} x^2_{i',t+p}] \\
& + \sum_{i',j} 4a_{i'}^p a_j^p \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} \frac{\sigma_j^2}{1-\alpha_j^2} E[x_{i',t} x_{i',t+p} x_{j,t} x_{j,t+p}] + \sum_{i'} \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} E[x^2_{i',t}] \sigma^2_{e_{\pi(n')}} + \sum_{i'} \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} E[x^2_{i',t+p}] \sigma^2_{e_{\pi(n')}} \\
& + \sum_{i,j'} \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_{j'}^2}{1-\alpha_{j'}^2} E[x^2_{i,t} x^2_{j',t+p}] + \sum_{i,j'} 4a_i^p a_{j'}^p \frac{\sigma_i^2}{1-\alpha_i^2} \frac{\sigma_{j'}^2}{1-\alpha_{j'}^2} E[x_{i,t} x_{i,t+p} x_{j',t} x_{j',t+p}] \\
& + \sum_{i',j'} \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} \frac{\sigma_{j'}^2}{1-\alpha_{j'}^2} E[x^2_{i',t} x^2_{j',t+p}] + \sum_{i' \neq j'} 4a_{i'}^p a_{j'}^p \frac{\sigma_{i'}^2}{1-\alpha_{i'}^2} \frac{\sigma_{j'}^2}{1-\alpha_{j'}^2} E[x_{i',t} x_{i',t+p} x_{j',t} x_{j',t+p}]
\end{array}
\tag{11}
$$

with $1 \leq i,j \leq n$, $n+1 \leq i',j' < n'$. Denote by $st(x_{\mathbf{V}_n,t}, \theta_{\pi(n+1)}, p)$ the sum of first six terms in the last equality in (11) and $st(x_{\mathbf{V}_{n'-1},t}, \theta_{\pi(n')}, p)$ the remaining parts. Provided that $st(x_{\mathbf{V}_{n'-1},t}, \theta_{\pi(n')}, p) = a \cdot st(x_{\mathbf{V}_n,t}, \theta_{\pi(n+1)}, p)$ does not hold for any $a \in \mathbf{R}$, $S(t,t+p)_{\pi(n')}$ cannot be determined by parameters in $\theta_{\pi(n+1)}$ and statistics of $x_{\mathbf{V}_n,t}$. Thus, we can not find a set of parameters, so that $S(t,t+p)_{\pi(n')}$ can be totally explained by a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$.

Thus, we can determine the process which is at the $(n+1)$th level, and its parameters are identifiable according to (9) and Lemma S1. Therefore, we can identify the causal model up to the $(n+1)$th level.

Repeating this procedure until we go through all processes, we have the identifiability of the whole causal model.

□

S2. Proof of Corollary 1

Proof. Since in Theorem 1, we have shown that the instantaneous causal order is identifiable, and for lagged causal relations, their causal order is fixed: from past to future, it reduces to a parameter identification problem.

For variable with ordering $\pi(i)$, it has the following varying coefficients model:

$$
\begin{cases}
x_{\pi(i),t} = \sum_{j=1}^{\pi(i-1)} b_{\pi(i)j,t} x_{j,t} + \sum_{s=1}^{s_l} \sum_{k=1}^{m} c^{(s)}_{\pi(i)k,t} x_{k,t-s} + e_{\pi(i),t}, \\
b_{\pi(i)j,t} = \alpha_{\pi(i)j,0} + \alpha_{\pi(i)j,1} b_{\pi(i)j,t-1} + \epsilon_{\pi(i)j,t}, \\
c^{(s)}_{\pi(i)j,t} = \gamma^{(s)}_{\pi(i)j,0} + \sum_{r=1} \gamma^{(s)}_{\pi(i)j,r} c^{(s)}_{\pi(i)j,t-r} + \nu^{(s)}_{\pi(i)j,t},
\end{cases}
\tag{12}
$$

where the instantaneous causal order π is known.

According to Lemma S1, the parameters in the varying coefficients model are identifiable. Thus, the parameters in (12) are identifiable. Combined with the result from Theorem 1 that the causal order is identifiable, therefore, the whole causal model is identifiable. □

S3. M Step in SAEM algorithm

In this section, we give detailed derivations of the M step in the SAEM algorithm. We consider three scenarios separately: with the change of both causal strengths and noise variances (Section S3.1), with the change of only causal strengths (Section S3.2), and with both instantaneous and time-lagged causal relationships (Section S3.3).

S3.1. With the Change of Both Causal Strengths and Noise Variances

The proposed time-varying causal network is defined as:

$$\begin{cases} X_t &= (I - B_t)^{-1} E_t, \\ b_{ij,t} &= \alpha_{ij,0} + \sum\limits_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t}, \\ h_{i,t} &= \beta_{i,0} + \sum\limits_{q=1}^{q_l} \beta_{iq} h_{i,t-q} + \eta_{i,t}, \end{cases} \tag{13}$$

for $t = max(p_l, q_l), \cdots, T$, where $X_t = (x_{1,t}, \cdots, x_{m,t})^{\mathrm{T}}$, B_t is an $m \times m$ causal adjacency matrix with entries $b_{ij,t}$, and $E_t = (e_{1,t}, \cdots, e_{m,t})^{\mathrm{T}}$ with $e_{i,t} \sim \mathcal{N}(0, \sigma_{i,t}^2)$, $h_{i,t} = \log(\sigma_{i,t}^2)$, $\epsilon_{ij,t} \sim \mathcal{N}(0, w_{ij})$, and $\eta_{i,t} \sim \mathcal{N}(0, v_i)$.

The model defined in equation (13) can be regarded as a nonlinear state space model, with causal coefficients and the logarithm of noise variances being latent variables $Z = \big\{\{b_{ij}\}_{ij}, \{h_i\}_i\big\}$, and model parameters $\theta = \big\{\{\alpha_{ij,0}\}_{ij}, \{\alpha_{ij,p}\}_{ij,p}, \{\beta_{i,0}\}_i, \{\beta_{i,q}\}_{i,q}, \{w_{ij}\}_{ij}, \{v_i\}_i\big\}$. Therefore, it can be transformed to a standard nonlinear state space model estimation. Particularly, we exploit an efficient stochastic approximation expectation maximization (SAEM) algorithm (Delyon et al., 1999), combining with conditional particle filters with ancestor sampling (CPF-AS) in the E step (Lindsten et al., 2012; Lindsten, 2013), for model estimation.

SAEM computes the E step by Monte Carlo integration and uses a stochastic approximation update of the quantity $\mathcal{Q}$:

$$\hat{\mathcal{Q}}_k(\theta) = (1 - \lambda_k)\hat{\mathcal{Q}}_{k-1}(\theta) + \lambda_k \sum_{j=1}^{M} \frac{\omega_T^{(k,j)}}{\sum_l \omega_T^{(k,l)}} \log p_\theta(X_{1:T}, B_{1:T}^{(k,j)}, h_{1:T}^{(k,j)}), \tag{14}$$

with

$$\begin{aligned} &\log p_\theta(X_{1:T}, B_{1:T}^{(k,j)}, h_{1:T}^{(k,j)}) \\ = \; &\textstyle\sum_{t=1}^{T} \log p_\theta(X_t | B_t^{(k,j)}, h_t^{(k,j)}) + \sum_{t=p_l+1}^{T} \log p_\theta(B_t^{(k,j)} | B_{t-1}^{(k,j)}, \cdots, B_{t-p_l}^{(k,j)}) + \sum_{t=1}^{p_l} \log p_\theta(B_t^{(k,j)}) \\ &\textstyle+ \sum_{t=q_l+1}^{T} \log p_\theta(h_t^{(k,j)} | h_{t-1}^{(k,j)}, \cdots, h_{t-q_l}^{(k,j)}) + \sum_{t=1}^{q_l} \log p_\theta(h_t^{(k,j)}), \end{aligned} \tag{15}$$

where $B_t^{(k,j)}$ is the sampled jth particle of B_t at the kth iteration, and $h_t^{(k,j)}$ is the sampled jth particle of h_t at the kth iteration.

Let $\alpha_0 = \{\alpha_{ij,0}\}_{ij}$, $\alpha_p = \{\alpha_{ij,p}\}_{ij}$ for $p = 1, \cdots, p_l$, $w = \{w_{ij}\}_{ij}$, and let $\beta_0 = \{\beta_{i,0}\}_i$, $\beta_q = \{\beta_{i,q}\}_{i,q}$ for $q = 1, \cdots, q_l$, $v = \{v_i\}_i$.

For presentation convenience, we reorganize the form of some parameters and latent variables. Let $\tilde{B}$ be an $m(m-1) \times 1$ vector, which is derived by stacking each column of B in sequence after removing diagonal entries. Let $\tilde{\alpha}_p$ be an $m(m-1) \times m(m-1)$ matrix, which is derived first by stacking each column of α in sequence after removing the diagonal entries and then diagonalize the vector into a matrix. The same operation is applied on α_0 to get $\tilde{\alpha}_0$, and on w to get $\tilde{w}$. $\tilde{\beta}_q$ is an $m \times m$ matrix, derived by diagonalize the vector β_q into a matrix. The same operations are applied on β_0 to get $\tilde{\beta}_0$. The reorganized parameters are $\tilde{\theta} = \{\tilde{\alpha}_0, \{\tilde{\alpha}_p\}_{p=1}^{p_l}, \tilde{w}, \tilde{\beta}_0, \{\tilde{\beta}_q\}_{q=1}^{q_l}, v\}$.

By inductive reasoning, equation (14) can be rewritten as

$$\hat{\mathcal{Q}}_k(\theta) = \sum_{i=1}^{k}\sum_{j=1}^{M}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot L^{(i,j)}, \tag{16}$$

where $L^{(i,j)} = \log p_\theta(X_{1:T}, B_{1:T}^{(i,j)}, h_{1:T}^{(i,j)})$. Each parameter is estimated by setting the corresponding partial derivative of the expected log-likelihood $\hat{\mathcal{Q}}_k$ to zero.

By taking the derivative of $\hat{\mathcal{Q}}_k(\theta)$ w.r.t $\tilde{\alpha}_p$, we have

$$\begin{aligned} & \tfrac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\alpha}_p} \\ = & \tfrac{\partial}{\partial \tilde{\alpha}_p} \textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot L^{(i,j)} \\ = & \textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})(1-\lambda_i) \cdot \omega_T^{(i,j)} \cdot \tfrac{\partial L^{(i,j)}}{\partial \tilde{\alpha}_p} \end{aligned} \tag{17}$$

with

$$\frac{\partial L^{(i,j)}}{\partial \tilde{\alpha}_p} = \sum_{t=p_l+1}^{T} \tilde{w}^{-1}(\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})\tilde{B}_{t-p}^{T(i,j)}. \tag{18}$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\alpha}_p} = 0$, and thus,

$$\begin{aligned} & \hat{\tilde{\alpha}}_p^{(k)} \\ = & \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}\sum_{t=p_l+1}^{T}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l\neq p}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})\tilde{B}_{t-p}^{T(i,j)} \Big) \\ & \cdot \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}\sum_{t=p_l+1}^{T}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot \tilde{B}_{t-p}^{(i,j)}\tilde{B}_{t-p}^{T(i,j)} \Big)^{-1} \\ = & \underbrace{\Big((1-\lambda_k)\hat{\tilde{\alpha}}_{p_a}^{(k-1)} + \lambda_k \sum_{j=1}^{M}\sum_{t=p_l+1}^{T} \omega_T^{(k,j)} \cdot (\tilde{B}_t^{(k,j)} - \tilde{\alpha}_0 - \sum_{l\neq p}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(k,j)})\tilde{B}_{t-p}^{T(k,j)} \Big)}_{\hat{\tilde{\alpha}}_{p_a}^{(k)}} \\ & \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{\alpha}}_{p_b}^{(k-1)} + \lambda_k \sum_{j=1}^{M}\sum_{t=p_l+1}^{T} \omega_T^{(k,j)} \cdot \tilde{B}_{t-p}^{(k,j)}\tilde{B}_{t-p}^{T(k,j)} \Big)^{-1}}_{\hat{\tilde{\alpha}}_{p_b}^{(k)}}. \end{aligned} \tag{19}$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\alpha}_0} = 0$, and thus

$$\begin{aligned} & \hat{\tilde{\alpha}}_0^{(k)} \\ = & \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}(T-p_l)\cdot(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\ & \cdot \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}\sum_{t=p_l+1}^{T}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (B_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot B_{t-l}^{(i,j)}) \Big) \\ = & \underbrace{\Big((1-\lambda_k)\hat{\alpha}_{0_a}^{(k-1)} + \lambda_k \sum_{j=1}^{M}(T-p_l)\cdot \omega_T^{(k,j)} \Big)^{-1}}_{\hat{\alpha}_{0_a}^{(k)}} \\ & \cdot \underbrace{\Big((1-\lambda_k)\hat{\alpha}_{0_b}^{(k-1)} + \lambda_k \sum_{j=1}^{M}\sum_{t=p_l+1}^{T} \omega_T^{(k,j)} \cdot (B_t^{(k,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{q_l}\tilde{\alpha}_l \cdot B_{t-l}^{(k,j)}) \Big)}_{\hat{\alpha}_{0_b}^{(k)}}. \end{aligned} \tag{20}$$

Take the derivative of $\hat{\mathcal{Q}}_k(\theta)$ w.r.t $\tilde{w}$:

$$\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{w}} = \sum_{i=1}^{k}\sum_{j=1}^{M}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot \frac{\partial L^{(i,j)}}{\partial \tilde{w}}, \tag{21}$$

where

$$\begin{aligned} &\frac{\partial L^{(i,j)}}{\partial \tilde{w}} \\ = \;& \textstyle\sum_{t=p_l+1}^{T}\big[-\frac{1}{2}\tilde{w}^{-T} + \frac{1}{2}\tilde{w}^{-T}(\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})(\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})^T\tilde{w}^{-T}\big]. \end{aligned} \tag{22}$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{w}} = 0$, and thus

$$\begin{aligned} &\hat{\tilde{w}}^{(k)} \\ = \;& \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}(T-p)\cdot(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)}\Big)^{-1} \\ &\cdot\Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}\sum_{t=p+1}^{T}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \\ &\quad\cdot(\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})(\tilde{B}_t^{(i,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(i,j)})^T\Big) \\ = \;& \underbrace{\Big((1-\lambda_k)\hat{\tilde{w}}_a^{(k-1)} + \lambda_k\sum_{j=1}^{M}(T-p)\cdot\omega_T^{(k,j)}\Big)^{-1}}_{\hat{\tilde{w}}_a^{(k)}} \\ &\underbrace{\cdot\Big((1-\lambda_k)\hat{\tilde{w}}_b^{(k-1)} + \lambda_k\sum_{j=1}^{M}\sum_{t=p_l+1}^{T}\omega_T^{(k,j)}\cdot(\tilde{B}_t^{(k,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(k,j)})(\tilde{B}_t^{(k,j)} - \tilde{\alpha}_0 - \sum_{l=1}^{p_l}\tilde{\alpha}_l \cdot \tilde{B}_{t-l}^{(k,j)})^T\Big)}_{\hat{\tilde{w}}_b^{(k)}}. \end{aligned} \tag{23}$$

Take the derivative of $\hat{\mathcal{Q}}_k(\theta)$ w.r.t β_0 and set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \beta_0} = 0$, and thus

$$\begin{aligned} &\hat{\beta}_0^{(k)} \\ = \;& \Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}(T-q)\cdot(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)}\Big)^{-1} \\ &\cdot\Big(\textstyle\sum_{i=1}^{k}\sum_{j=1}^{M}\sum_{t=q+1}^{T}(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)}\cdot(h_t^{(i,j)} - \tilde{\beta}_0 - \sum_{l=1}^{q_l}\tilde{\beta}_l \cdot h_{t-l}^{(i,j)})\Big) \\ = \;& \underbrace{\Big((1-\lambda_k)\hat{\beta}_{0_a}^{(k-1)} + \lambda_k\sum_{j=1}^{M}(T-q)\cdot\omega_T^{(k,j)}\Big)^{-1}}_{\hat{\beta}_{0_a}^{(k)}} \\ &\underbrace{\cdot\Big((1-\lambda_k)\hat{\beta}_{0_b}^{(k-1)} + \lambda_k\sum_{j=1}^{M}\sum_{t=q_l+1}^{T}\omega_T^{(k,j)}\cdot(h_t^{(k,j)} - \tilde{\beta}_0 - \sum_{l=1}^{q_l}\tilde{\beta}_l \cdot h_{t-l}^{(k,j)})\Big)}_{\hat{\beta}_{0_b}^{(k)}}. \end{aligned} \tag{24}$$

Take the derivative of $\hat{\mathcal{Q}}_k(\theta)$ w.r.t $\tilde{\beta}_q$ and set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\beta}_q} = 0$, and thus

$$
\begin{aligned}
&\hat{\tilde{\beta}}_q^{(k)} \\
=\ & \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=q_l}^{T} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (h_t^{(i,j)} - \beta_0 - \sum_{l\neq q} \beta_l h_{t-l}^{(i,j)}) h_{t-q}^{T(i,j)} \Big) \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=q_l}^{T} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot h_{t-q}^{(i,j)} h_{t-q}^{T(i,j)} \Big)^{-1} \\
=\ & \underbrace{\Big((1-\lambda_k)\hat{\tilde{\beta}}_{q_a}^{(k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=q_l+1}^{T} \omega_T^{(k,j)} \cdot (h_t^{(k,j)} - \beta_0 - \sum_{l\neq q} \beta_l h_{t-l}^{(k,j)}) h_{t-q}^{T(k,j)} \Big)}_{\hat{\tilde{\beta}}_{q_a}^{(k)}} \\
& \underbrace{\cdot \Big((1-\lambda_k)\hat{\tilde{\beta}}_{q_b}^{(k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=q_l+1}^{T} \omega_T^{(k,j)} \cdot h_{t-q}^{(k,j)} h_{t-q}^{T(k,j)} \Big)^{-1}}_{\hat{\tilde{\beta}}_{q_b}^{(k)}}.
\end{aligned}
\tag{25}
$$

Take the derivative of $\hat{\mathcal{Q}}_k(\theta)$ w.r.t v and set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial v} = 0$, and thus

$$
\begin{aligned}
&\hat{v}^{(k)} \\
=\ & \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} (T-q_l) \cdot (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=q_l+1}^{T} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \\
& \qquad \cdot (h_t^{(i,j)} - \beta_0 - \textstyle\sum_{l=1}^{q_l} \tilde{\beta}_l h_{t-l}^{(i,j)})(h_t^{(i,j)} - \beta_0 - \sum_{l=1}^{q_l} \tilde{\beta}_l h_{t-l}^{(i,j)})^T \Big) \\
=\ & \underbrace{\Big((1-\lambda_k)\hat{v}_a^{(k-1)} + \lambda_k \sum_{j=1}^{M} (T-q_l) \cdot \omega_T^{(k,j)} \Big)^{-1}}_{\hat{v}_a^{(k)}} \\
& \underbrace{\cdot \Big((1-\lambda_k)\hat{v}_b^{(k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=q_l+1}^{T} \omega_T^{(k,j)} \cdot (h_t^{(k,j)} - \beta_0 - \sum_{l=1}^{q_l} \tilde{\beta}_l h_{t-l}^{(k,j)})(h_t^{(k,j)} - \beta_0 - \sum_{l=1}^{q_l} \tilde{\beta}_l h_{t-l}^{(k,j)})^T \Big)}_{\hat{v}_b^{(k)}}.
\end{aligned}
\tag{26}
$$

S3.2. With the Change of Only Causal Strengths

If we assume that the variance of E_t does not change across time, then we have the following time-varying causal network:

$$
\begin{cases}
X_t = (I - B_t)^{-1} E_t, \\
b_{ij,t} = \alpha_{ij,0} + \sum_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t},
\end{cases}
\tag{27}
$$

for $t = p_l, \cdots, T$, where $X_t = (x_{1,t}, \cdots, x_{m,t})^{\mathrm{T}}$, B_t is an $m \times m$ causal adjacency matrix with entries $b_{ij,t}$, $E_t = (e_{1,t}, \cdots, e_{m,t})^{\mathrm{T}}$ with $e_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$, and $\epsilon_{ij,t} \sim \mathcal{N}(0, w_{ij})$. Let R be an $m \times m$ diagonal matrix with diagonal entries σ_i^2. The latent variables are $Z = \big\{\{b_{ij}\}_{ij}\big\}$, and model parameters are $\theta = \big\{\{\alpha_{ij,0}\}_{ij}, \{\alpha_{ij,p}\}_{ij,p}, \{w_{ij}\}_{ij}, \{\sigma_i\}_i\big\}$.

We update R by setting $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{R}} = 0$, and thus

$$\begin{aligned}
&\hat{R}^{(k)} \\
= \;& \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} T \cdot (1-\lambda_k)(1-\lambda_{k-1}) \cdots (1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=1}^{T} (1-\lambda_k)(1-\lambda_{k-1}) \cdots (1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (I - \tilde{B}_t^{(i,j)}) X_t X_t^T (I - \tilde{B}_t^{(i,j)})^T \Big) \\
= \;& \underbrace{\Big((1-\lambda_k)\hat{\tilde{R}}_a^{(k-1)} + \lambda_k \sum_{j=1}^{M} T \cdot \omega_T^{(k,j)} \Big)^{-1}}_{\hat{\tilde{R}}_a^{(k)}} \\
& \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{R}}_b^{(k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=1}^{T} \omega_T^{(k,j)} \cdot (I - \tilde{B}_t^{(k,j)}) X_t X_t^T (I - \tilde{B}_t^{(k,j)})^T \Big)}_{\hat{\tilde{R}}_b^{(k)}}.
\end{aligned} \tag{28}$$

The other parameters are updated in the same way as those is Section S3.1.

S3.3. With Both Contemporaneous and Time-Lagged Causal Relationships

It is easy to extend to the case with both contemporaneous and time-lagged causal relationships:

$$\begin{cases}
X_t &= (I - B_t)^{-1} (\sum_s^{s_l} C_t^{(s)} X_{t-s} + E_t), \\
b_{ij,t} &= \alpha_{ij,0} + \sum_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t}, \\
h_{i,t} &= \beta_{i,0} + \sum\limits_{q=1}^{q_l} \beta_{iq} h_{i,t-q} + \eta_{i,t}, \\
c_{ij,t}^{(s)} &= \gamma_{ij,0}^{(s)} + \sum\limits_{r=1}^{r_l} \gamma_{ij,r}^{(s)} c_{ij,t-r}^{(s)} + \nu_{ij,t}^{(s)},
\end{cases} \tag{29}$$

for $t = max(p_l, q_l), \cdots, T$, where $X_t = (x_{1,t}, \cdots, x_{m,t})^{\mathrm{T}}$, B_t is an $m \times m$ causal adjacency matrix with entries $b_{ij,t}$, $C_t^{(s)}$ is an $m \times m$ matrix with entries $c_{ij,t}^{(s)}$ for $s = 1, \cdots, s_l$, and $E_t = (e_{1,t}, \cdots, e_{m,t})^{\mathrm{T}}$ with $e_{i,t} \sim \mathcal{N}(0, \sigma_{i,t}^2)$, $h_{i,t} = \log(\sigma_{i,t}^2)$, $\epsilon_{ij,t} \sim \mathcal{N}(0, w_{ij})$, $\eta_{i,t} \sim \mathcal{N}(0, v_i)$, and $\nu_{ij,t}^{(s)} \sim \mathcal{N}(0, u_{ij}^{(s)})$. The latent variables are $Z = \big\{ \{b_{ij}\}_{ij}, \{c_{ij}^{(s)}\}_{ij,s}, \{h_i\}_i \big\}$, and model parameters are $\theta = \big\{ \{\alpha_{ij,0}\}_{ij}, \{\alpha_{ij,p}\}_{ij,p}, \{\beta_{i,0}\}_i, \{\beta_{i,q}\}_{i,q}, \{w_{ij}\}_{ij}, \{v_i\}_i, \{\gamma_{ij,0}^{(s)}\}_{ij,s}, \{\gamma_{ij,r}^{(s)}\}_{ij,r,s}, \{u_{ij}\}_{ij}^{(s)} \big\}$.

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\gamma}_r^{(s)}} = 0$, and thus

$$\begin{aligned}
&\hat{\tilde{\gamma}}_r^{(s,k)} \\
&= \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} (1-\lambda_k)(1-\lambda_{k-1}) \cdots (1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (\tilde{C}_t^{(s,i,j)} - \gamma_0^{(s)} - \sum_{l \neq r} \gamma_l^{(s)} \tilde{C}_{t-l}^{(s,i,j)}) \tilde{C}_{t-r}^{T(s,i,j)} \Big) \\
&\quad \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} (1-\lambda_k)(1-\lambda_{k-1}) \cdots (1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot \tilde{C}_{t-r}^{(s,i,j)} \tilde{C}_{t-r}^{T(s,i,j)} \Big)^{-1}. \\
&= \underbrace{\Big((1-\lambda_k)\hat{\tilde{\gamma}}_{r_a}^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} \omega_T^{(k,j)} \cdot (\tilde{C}_t^{(s,k,j)} - \gamma_0^{(s)} - \sum_{l \neq r} \gamma_l^{(s)} \tilde{C}_{t-l}^{(s,k,j)}) \tilde{C}_{t-r}^{T(s,k,j)} \Big)}_{\hat{\tilde{\gamma}}_{r_a}^{(s,k)}} \\
&\quad \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{\gamma}}_{r_b}^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} \omega_T^{(k,j)} \cdot \tilde{C}_{t-r}^{(s,k,j)} \tilde{C}_{t-r}^{T(s,k,j)} \Big)^{-1}}_{\hat{\tilde{\gamma}}_{r_b}^{(s,k)}}
\end{aligned} \tag{30}$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{u}^{(s)}} = 0$, and thus

$$
\begin{aligned}
&\hat{\tilde{u}}^{(s,k)} \\
= \;& \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} (T-r_l)\cdot(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \\
& \quad \cdot (\tilde{C}_t^{(s,i,j)} - \tilde{\gamma}_0^{(s)} - \textstyle\sum_{l=1}^{r_l} \tilde{\gamma}_l \cdot \tilde{C}_{t-l}^{(s,i,j)})(\tilde{C}_t^{(s,i,j)} - \tilde{\gamma}_0^{(s)} - \sum_{l=1}^{p_l} \tilde{\gamma}_l \cdot \tilde{C}_{t-l}^{(s,i,j)})^T \Big) \\
= \;& \underbrace{\Big((1-\lambda_k)\hat{\tilde{u}}_a^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} (T-r_l)\cdot \omega_T^{(k,j)} \Big)^{-1}}_{\hat{\tilde{u}}_a^{(s,k)}} \\
& \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{u}}_b^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} \omega_T^{(k,j)} \cdot (\tilde{C}_t^{(s,k,j)} - \tilde{\gamma}_0^{(s)} - \sum_{l=1}^{r_l} \tilde{\gamma}_l^{(s)} \cdot \tilde{C}_{t-l}^{(s,k,j)})(\tilde{C}_t^{(s,k,j)} - \tilde{\gamma}_0^{(s)} - \sum_{l=1}^{p_l} \tilde{\gamma}_l^{(s)} \cdot \tilde{C}_{t-l}^{(s,k,j)})^T \Big)}_{\hat{\tilde{u}}_b^{(s,k)}}.
\end{aligned}
\tag{31}
$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{u}_c} = 0$, and thus

$$
\begin{aligned}
&\hat{\tilde{u}}_c^{(k)} \\
= \;& \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot \frac{1}{r_l} \sum_{t=1}^{r_l} (\tilde{C}_t^{(i,j)} - \mu_c)(\tilde{C}_t^{(i,j)} - \mu_c)^T \Big) \\
= \;& \underbrace{\Big((1-\lambda_k)\hat{\tilde{u}}_{c_a}^{(k-1)} + \lambda_k \sum_{j=1}^{M} \omega_T^{(k,j)} \Big)^{-1}}_{\hat{\tilde{u}}_{c_a}^{(k)}} \\
& \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{u}}_{c_b}^{(k-1)} + \lambda_k \sum_{j=1}^{M} \omega_T^{(k,j)} \cdot \frac{1}{r_l} \sum_{t=1}^{r_l} (\tilde{C}_t^{(k,j)} - \mu_c)(\tilde{C}_t^{(k,j)} - \mu_c)^T \Big)}_{\hat{\tilde{u}}_{c_b}^{(k)}}.
\end{aligned}
\tag{32}
$$

Set $\frac{\partial \hat{\mathcal{Q}}_k}{\partial \tilde{\gamma}_0^{(s)}} = 0$, and thus

$$
\begin{aligned}
&\hat{\tilde{\gamma}}_0^{(s,k)} \\
= \;& \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} (T-r_l)\cdot(1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \Big)^{-1} \\
& \cdot \Big(\textstyle\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{t=r_l+1}^{T} (1-\lambda_k)(1-\lambda_{k-1})\cdots(1-\lambda_{i+1})\lambda_i \cdot \omega_T^{(i,j)} \cdot (C_t^{(s,i,j)} - \tilde{\gamma}_0^{(s)} - \sum_{l=1}^{r_l} \tilde{\gamma}_l^{(s)} \cdot C_{t-l}^{(s,i,j)}) \Big) \\
= \;& \underbrace{\Big((1-\lambda_k)\hat{\tilde{\gamma}}_{0_a}^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} (T-r_l)\cdot \omega_T^{(k,j)} \Big)^{-1}}_{\hat{\tilde{\lambda}}_{0_a}^{(s,k)}} \\
& \cdot \underbrace{\Big((1-\lambda_k)\hat{\tilde{\gamma}}_{0_b}^{(s,k-1)} + \lambda_k \sum_{j=1}^{M} \sum_{t=q_l+1}^{T} \omega_T^{(k,j)} \cdot (C_t^{(s,k,j)} - \tilde{\gamma}_0^{(s)} - \sum_{l=1}^{q_l} \tilde{\gamma}_l^{(s)} \cdot C_{t-l}^{(s,k,j)}) \Big)}_{\hat{\tilde{\gamma}}_{0_b}^{(s,k)}}.
\end{aligned}
\tag{33}
$$

The other parameters are updated with the same way as those in Section S3.1.

S4. Conditional Particle Filter with Ancestor Sampling

The detailed procedure of conditional particle filter with ancestor sampling is summarized in Algorithm S1.

Algorithm S1 CPF with Ancestor Sampling

```
Let prespecified particles be $\mathring{Z}'_{1:T} = \{\mathring{Z}'_1, \cdots, \mathring{Z}'_T\}$.
Draw $\mathring{Z}_1^{(j)}$ from $\mathring{Z}_1^{(j)} \sim f_\theta(\mathring{Z}_1),\ j = 1, \cdots, M-1$.
Set $\mathring{Z}_1^{(M)} = \mathring{Z}'_1$.
Set $\omega_1^{(j)} = W_{\theta,1}(\mathring{Z}_1^{(j)})$ for $j = 1, \cdots, M$.
for $t = 2$ to $T$ do
    Draw $s_t^j$ with $P(s_t^j = i) \propto \omega_{t-1}^{(i)}$ for $j = 1, \cdots, M-1$.
    Draw $s_t^M$ with $P(s_t^M = j) \propto \omega_{t-1}^{(j)} f_\theta(\mathring{Z}'_t | \mathring{Z}_{t-1}^j)$.
    Draw $\mathring{Z}_t^{(j)} \sim f_\theta(\mathring{Z}_t | \mathring{Z}_{t-1}^{s_t^j})$ for $j = 1, \cdots, M-1$.
    Set $\mathring{Z}_t^{(M)} = \mathring{Z}'_t$.
    Set $\mathring{Z}_{1:t}^{(j)} = \{\mathring{Z}_{1:t-1}^{s_t^j}, \mathring{Z}_t^{(j)}\}$ for $j = 1, \cdots, M$.
    Set $\omega_t^{(j)} = W_{\theta,t}(\mathring{Z}_t^{(j)}, \mathring{Z}_{t-1}^{s_t^j})$ for $j = 1, \cdots, M$.
end for
```

S5. Sparsity Constraints

In practical problems, the causal connections may be sparse. In this section, we consider the sparsity constraints on causal adjacency matrix B_t and that on $b_{ij,t} - b_{ij,t-1}, \forall i, j$, which ensures smooth changes of $b_{ij,t}$ across time.

It is well known that lasso regularization for sparsity is biased. Thus, we utilize the smoothly clipped absolute deviation (SCAD) penalty, which has shown to be unbiased for the resulting estimator of significant parameters (Fan & Li, 2001). The SCAD penalty is given by

$$p_\lambda^{\text{SCAD}}(b_{ij,t}) = \begin{cases} \lambda |b_{ij,t}| & \text{if } |b_{ij,t}| \leq \lambda \\ -\frac{|b_{ij,t}|^2 - 2a\lambda|b_{ij,t}| + \lambda^2}{2(a-1)} & \text{if } \lambda < |b_{ij,t}| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |b_{ij,t}| > a\lambda \end{cases}$$

where a and λ are hyperparameters. The penalized log-likelihood for equation is modified as

$$\begin{aligned} & p_\lambda(X|B) \\ = \; & \textstyle\sum_{t=1}^T \log p(X_t | B_t, \sigma_t^2) - \sum_{t=1}^T \sum_{i,j} p_\lambda^{\text{SCAD}}(b_{ij,t}) \\ & \textstyle - \sum_{t=2}^T \sum_{i,j} p_\lambda^{\text{SCAD}}(b_{ij,t} - b_{ij,t-1}). \end{aligned} \tag{34}$$

Since adding sparsity on B_t and $b_{ij,t} - b_{ij,t-1}$ does not affect the derivatives of the parameters in the M step, we only need to modify the likelihood used in the E step, by replacing $p(X|B)$ with $p_\lambda(X|B)$.

References

Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. In *The Annals of Statistics*, volume 27(1), pp. 94–128, 1999.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. In *Journal of the American statistical Association*, volume 96(456), pp. 1348–1360, 2001.

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

Lindsten, F., Jordan, M. I., and Schön, T. B. Ancestor sampling for particle gibbs. In *Proceedings of the Conference on Neural Information Processing Systems*, 2012.

Wall, K. D. Identification theory for varying coefficient regression models. In *Journal of Time Series Analysis*, volume 8(3), pp. 359–371, 1987.