DAG Estimation for Multivariate Time-Series using Posterior Projections

Arkaprava Roy, Anindya Roy and Subhashis Ghosal
University of Florida, University of Maryland Baltimore County,
and North Carolina State University

Abstract

In multivariate time series analysis, understanding the underlying causal relationships among variables is of interest for various applications. Directed acyclic graphs (DAGs) provide a powerful framework for representing such causal dependencies. This paper proposes a novel approach for modeling multivariate time series where conditional independencies and causal structure are encoded by a DAG. The proposed model further allows structural properties such as stationarity to be easily accommodated in the model. Given the application, we further extend the model for matrix-variate time series. We take a Bayesian route of inference and a “projection-posterior" based efficient computational algorithm is developed. Posterior convergence properties of the proposed method are established. The utility of the proposed method is demonstrated through simulation studies and real data analysis.

To study multivariate data, graphical models provide an appealing framework to characterize association by assessing the conditional dependence between two variables given the remaining variables. A graph $\mathcal{G}$ is commonly denoted as the pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ denotes a set of nodes and $\mathbf{E}$ denotes the set of edges. There is vast literature on embedding a probabilistic model to study the graph structure [Lauritzen, 1996, Koller and Friedman, 2009]. Under a parent ordering of the nodes, one of the most popular probabilistic approaches is the Gaussian-directed acyclic graphs (DAGs) [Babula et al., 2004, Shojaie and Michailidis, 2010]. Importantly, DAGs allow causal discoveries with some additional causal assumptions [Pearl et al., 2000].

However, the parent ordering is often unknown. Thus, the DAG-ness specification leads to the class of $\mathbf{B}$ such that it could be permuted (by simultaneous equal row and column permutations) to strict lower triangularity. Most of the works on DAG estimation rely on the structural equation model (SEM), which is also the building block for our proposed method. Shimizu et al. [2006] proposed the LINGAM approach where an unrestricted estimate of $\mathbf{W}$ is post-processed into a DAG specification. There are other approaches that follow two-step procedures where the directional order is determined first, and then the coefficients are estimated [Bühlmann et al., 2014].

Recently Zheng et al. [2018] proposed a continuous constrained optimization-based approach, called NOTEAR using an algebraic characterization of DAG. Consequently, several other continuous characterizations of DAG constraint have been proposed [Yu et al., 2019, Bello et al., 2022]. These approaches provide a computationally efficient estimation avenue of the DAG employing augmented

Lagrangian, and further allow adding additional penalties in the objective function to induce sparsity and other structural properties.

Several approaches have been proposed for a fully Bayesian DAG estimation under both known [Altomare et al., 2013, Ni et al., 2015] and unknown parent ordering [Zhou et al., 2023]. In the case of unknown parent ordering, the MCMC-based computational algorithms are in general expensive. To alleviate that, two-stage procedures are also often adopted where the parent ordering is identified in the first stage, and under that ordering, the DAG is estimated next Shojaie and Michailidis [2010], Altomare et al. [2013]

Beyond the SEM framework, DAG can also be estimated using the PC-algorithm [Spirtes and Glymour, 1991] that estimates the completed partially directed acyclic graph (CPDAG). Kalisch and Bühlman [2007] studied the high dimensional uniform consistency of the PC algorithm. However, it relies on the faithfulness assumption and is sensitive to individual failures of conditional independence tests. Another approach is the score-based [Heckerman et al., 1995] where an M-estimator for some score function is defined, and it is based on measuring how well the graph fits the data. Due to the large search space, the problem is in general NP-hard and thus, it resorts to local search.

There are some recent developments in DAG estimation for functional data Lee and Li [2022], Yang and Suzuki [2022], Zhou et al. [2023]. However, it is not straightforward to ensure maintaining properties such as stationarity which is often desirable for time-series data. Another important property is causality in a time direction which is different from the above-mentioned causal associations. A time series $(\mathbf{y}_t : t = 0, 1, \dots)$ is called *causal* if for all t, $\mathbf{y}_t$ can be expressed as $\mathbf{H}_t(\mathbf{e}_t, \mathbf{e}_{t-1}, \dots)$ for some process $(\mathbf{e}_t : t = 0, \pm 1, \pm 2, \dots)$, where $\mathbf{H}_t$ is a vector-valued function possibly dependent on t. However, our primary focus here is to estimate causal associations among the component variables within the multivariate time series data. In this context, another popular measure of causality is the Granger causality [Granger, 1969] where a vector autoregressive model is fitted, and the elements in coefficient matrices are used to quantify Granger causal effects. However, our focus is to estimate a 'contemporaneous stationary causal' structure, where the causal structure is encoded in the marginal precision matrix of a stationary Gaussian time series. In this paper, we characterize the contemporaneous precision matrix of a stationary time series to estimate the causal graph and our approach also allows specification of the multivariate time series to be stationary and also causal.

In this work, our focus is on multivariate time series data as commonly found in applications such as economics [Imbens, 2020], finance [Ji et al., 2018], neuroscience [Ramsey et al., 2010], etc and our aim is to investigate the directional relationships among the time series. In different applications, such estimates give rise to varying forms of inferences. Specifically, a multivariate time series can be represented as a DAG at each time point [Dahlhaus and Eichler, 2003, Ben-David et al., 2011, Zuo and Kita, 2012, Deng et al., 2013]. Such a graphical dependence corresponds to a strictly lower-triangular adjacency matrix $\mathbf{B}$ under an ordering of the nodes $V_1, \dots, V_n$, such that V_i cannot be a child of V_j if $i < j$. If the process is stationary, then the DAG structure stays invariant over time. Such a causal form of dependence in a multivariate 'stationary' time series $(\mathbf{y}_t : t = 0, 1, \dots)$ may be encapsulated through a linear transform, modeling the 'causal residual process' $\mathbf{y}_t - \mathbf{W}\mathbf{y}_t$ by one not involving the graph. The operator $\mathbf{W}$ decorrelates the components in $\mathbf{y}_t$ at each time t over the nodes. We can learn the ordering of the nodes based on Kuipers and Moffa [2017] following

Friedman and Koller [2003].

In this work, we model p-dimensional residual $\mathbf{y}_t - \mathbf{W}\mathbf{y}_t$ as a vector of independent univariate stationary time series in the same spirit as in the orthogonally rotated univariate time series (OUT) model of Roy et al. [2024]. Like the OUT model, the proposed model also closely resembles Dynamic Factor Model (DFM) and possesses symmetric autocovariances. Our model also allows us a 'contemporaneous' DAG estimation, leading to the identification of marginal causal relationships among the variables. The proposed characterization also leads to a stationary multivariate time-series model. We impose the DAG-ness constraint on $\mathbf{W}$, we induce a "projection-posterior" by imposing the DAG-ness constraint on samples from the posterior distribution of the $\mathbf{W}$ applying the continuous constraint from Zheng et al. [2018] along with an adaptive LASSO penalty [Zou, 2006]. Recently, Pamfil et al. [2020] applied the NOTEAR approach on multivariate time series within a structural vector-autoregressive model (SVAR).

The rest of the article is organized as follows.

1 Projection posterior for DAG estimation in multivariate time-series

We consider the structural equation model from Shimizu et al. [2006], but allow the error to be an array of stationary processes. Specifically,

$$\mathbf{y}_t - \mathbf{W}\mathbf{y}_t = \mathbf{D}^{1/2}\mathbf{z}_t, \tag{1}$$

To estimate the directed acyclic graphical association in $\mathbf{y}_t$, we need $\mathbf{W}$ such that it can be permuted into a strictly lower triangular matrix by simultaneous equal row and column permutations. However, the imposition of such restrictions is hard computationally. Instead of taking a fully Bayesian approach, we thus take a projection-posterior-based route. In the model, we only restrict $\mathbf{W}$ to have zeroes in the diagonal.

To complete our model characterization, *i-th* component process in $\mathbf{z}_t$ is modeled in the frequency domain and parametrized in terms of a spectral density f_i. We then impose priors on these parameters, with details provided in Sections 1.4. Section 2 provides the steps to get the 'unrestricted' posterior samples of our model parameters $\{\mathbf{W}, \mathbf{D}, f_1(\cdot), \dots, f_p(\cdot)\}$. The samples of $\mathbf{W}$ obtained here do not satisfy the DAG-ness constraint, but are sparse due to the horse-shoe prior, and have zeroes in the diagonal as specified by the model construction. These samples are then passed through an immersion map that transforms these sparse unrestricted matrices with zero diagonals into sparse matrices satisfying the DAG constraint.

We set immersion map as a solution of the following adaptive LASSO-based [Zou, 2006] loss $\vartheta : \mathbf{W} \to \mathbf{W}^* := \arg\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{W}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y}\|_2^2 + \lambda\|\mathbf{C} \odot \boldsymbol{\beta}\|_1 + \alpha h(\boldsymbol{\beta}) + \frac{\rho}{2}h^2(\boldsymbol{\beta})$ We set $\mathbf{C} = 1/|\hat{\boldsymbol{\beta}}_A|^\zeta$ where $\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y}\|_2^2 + \alpha h(\boldsymbol{\beta}) + \frac{\rho}{2}h^2(\boldsymbol{\beta})$. Although ζ is often set based on cross-validation, we find that $\zeta = 1$ works well in all of our numerical experiments. Hence, this is our default setting. The above loss is based on Zheng et al. [2018] but employs an adaptive LASSO penalty [Xu et al., 2022]. The λ is set based on a crossvalidation applied to the LASSO loss without the augmented Lagrangian $\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{Y} - (\boldsymbol{\beta} \oslash \mathbf{C})\mathbf{Y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, where $\oslash$ stands for element-wise divide.

To apply gradient-based optimization using `lbfgs`, we rewrite the adaptive LASSO loss as $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2:\beta_{\ell,i,j}\geq 0} \frac{1}{Tp}\|\mathbf{W}^{(t)}\mathbf{Y} - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\mathbf{Y}\|_2^2 + \lambda\sum_{i,j} c_{i,j}\beta_{1,i,j} + \lambda\sum_{i,j} c_{i,j}\beta_{2,i,j} + \alpha h(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + \frac{\varrho}{2}h^2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and set $\mathbf{W}^{(t)} = \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2$.

Here $h(\boldsymbol{\beta}) = \mathrm{trace}(e^{\boldsymbol{\beta}\odot\boldsymbol{\beta}}) - m$ is the constraint ensuring DAG-ness, which is proposed in Zheng et al. [2018]. Along the lines of constraint-based DAG characterization, there are now a few more choices in Yu et al. [2019], Bello et al. [2022]. However, after trying out all of these choices, $h(\boldsymbol{\beta}) = \mathrm{trace}(e^{\boldsymbol{\beta}\odot\boldsymbol{\beta}}) - m$ performs the best in our time-series setting.

Algorithm 1 Projection posterior samples for DAG

```
1) Set adaptive LASSO weights ($\mathbf{C}$): $\mathbf{C} = 1/|\hat{\boldsymbol{\beta}}_A|^{\zeta}$ [Zou, 2006] where $\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y}\|_2^2 + \alpha h(\boldsymbol{\beta}) + \frac{\varrho}{2}h^2(\boldsymbol{\beta})$ with $\zeta = 1$

2) Selection of $\lambda$: Select $\lambda$ via crossvalidation based on $\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{Y} - (\boldsymbol{\beta} \oslash \mathbf{C})\mathbf{Y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$.

3) Projected sample (for $t = 1, \ldots, B$): $\mathbf{W}^{(t)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{Tp}\|\mathbf{W}_U^{(t)}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y}\|_2^2 + \lambda\|\mathbf{C}\odot\boldsymbol{\beta}\|_1 + \alpha h(\boldsymbol{\beta}) + \frac{\varrho}{2}h^2(\boldsymbol{\beta})$, where $\mathbf{W}_U^{(t)}$ is the $t$-$th$ 'unrestricted' posterior sample of $\mathbf{W}$ i.e. $\vartheta(\mathbf{W}_U^{(t)}) = \mathbf{W}^{(t)}$
4) Set some small threshold $H$. Get the DAG-adjacency matrix $\hat{\mathbf{A}}$ setting $a_{i,j} = 1$ if $\left(\frac{1}{B}\sum_{t=1}^{B} \mathbf{1}\{|w_{i,j}^{(t)}| > H\} > 0.5\right)$.
```

In Step 2, we use `cv.glmnet` function from R package `glmnet`. The thresholding value in Step 4 is set at $H = 0.3$ as default in all of our numerical experiments.

1.1 Extension to matrix-variate time-series

We now extend our model in (1) to the case, where $p \times S$ dimensional matrix-variate data $\mathbf{Y}_t$ is observed at each timepoint t motivated by our application in Section 5. In this case, we are only interested in a causal association among the row variables of $\mathbf{Y}_t$. In case one is interested in DAGs for both of the two directions in $\mathbf{Y}_t$, a variation is discussed in Section 1.1.2 Specifically, $\mathbf{Y}_t = \{\mathbf{y}_t^{(s)}\}_{s=1}^{S}$ with p-dimensional vector-valued $\mathbf{y}_t^{(s)}$'s, and we are interested in a common DAG characterizing the causal associations within $\mathbf{Y}_t^{(s)}$. To achieve this, we consider a two-layer model, combining the model from the previous section and the OUT model [Roy et al., 2024] in layer 2 on the residuals.

$$
\begin{aligned}
\mathbf{y}_t^{(s)} - \mathbf{W}\mathbf{y}_t^{(s)} =& \mathbf{D}^{1/2}\mathbf{x}_t^{(s)}, \quad s = 1, \ldots, S \\
\mathbf{x}_{k,t} =& \mathbf{V}\mathbf{U}\mathbf{z}_{k,t}, \quad k = 1, \ldots, p,
\end{aligned}
$$

where $\mathbf{z}_{k,t} = \{Z_{s,j,t}\}_{s=1}^{S}$ are independent univariate time-series and $\mathbf{x}_{k,t} = {x_{k,t}^{(s)}}_{s=1}^{S}$. We again assume that all latent processes $(Z_{s,j,t} : t = 1, 2, \ldots)$, $j = 1, \ldots, p$, $s = 1, \ldots, S$ are stationary with spectral density function $f_{s,j}(\omega)$. As before we set $f_{s,j}(\omega) = \sum_{k=1}^{K} \theta_{s,j,k} B_k^*(|\omega|/\pi), \quad \theta_{s,j,k} \geq 0, \quad \sum_{k=1}^{K} \theta_{s,j,k} = 1/2.$ and $\theta_{s,j,k} = \Psi(\kappa_{s,j,k})/\{2\sum_{l=1}^{K} \Psi(\kappa_{s,jl})\}$ with a rank R CP-tensor decomposition structure $\kappa_{s,j,k} = \sum_{r=1}^{R} \chi_{sr} \xi_{j,r} \eta_{k,r}, \quad s = 1, \ldots, S,\ j = 1, \ldots, p,\ k = 1, \ldots, K$. Here, $\mathbf{V}$ is the spherical coordinate representation of Cholesky factorizations of the correlation matrix across the groups. Thus $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ is a correlation matrix. Finally, $\mathbf{U}$ is an orthogonal matrix.

The above model marginally at each time-point assumes a matrix-normal distribution on $\mathbf{R}_t = \{\mathbf{r}_t^{(1)}; \ldots; \mathbf{r}_t^{(S)}\}$, where $\mathbf{r}_t^{(s)} = \mathbf{y}_t^{(s)} - \mathbf{W}\mathbf{y}_t^{(s)}$. Specifically, marginally, $\mathbf{R}_t \sim \text{Matrix-Normal}(\mathbf{0}, \mathbf{D}, \mathbf{V}\mathbf{V}^{\mathrm{T}})$.

1.1.1 Modeling of V

We model $\mathbf{V}$ following Zhang et al. [2011]. Let $\mathbf{V}^{S\times S}$ be a lower triangular matrix such that $\mathbf{Q} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$. The form of $\mathbf{V}$ is

$$\mathbf{V} \quad = \quad \begin{pmatrix} v_{1,1} & 0 & \ldots & 0 \\ v_{2,1} & v_{2,2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ v_{S,1} & v_{S,2} & \ldots & v_{S,S} \end{pmatrix}.$$

The restriction that $\mathbf{Q}$ is a correlation matrix is satisfied by the following parameterization

$$\begin{aligned} & v_{1,1} = 1, \\ & v_{2,1} = \cos\theta_{2,1},\ v_{2,2} = \sin\theta_{2,1}, \\ & v_{3,1} = \cos\theta_{3,1},\ v_{3,2} = \sin\theta_{3,1}\cos\theta_{3,2},\ v_{3,3} = \sin\theta_{3,1}, \\ & v_{\ell,1} = \cos(\theta_{\ell,1}), \\ & v_{\ell,k} = \sin\theta_{\ell,1}\sin\theta_{\ell,2}\ldots\sin\theta_{\ell,k-1}\cos\theta_{\ell,k-1}, \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } k = 2, \ldots, (\ell - 1), \\ & v_{\ell,\ell} = \sin\theta_{\ell,1}\sin\theta_{\ell,2}\ldots\sin\theta_{\ell,k-1}\sin\theta_{\ell,\ell-1}, \end{aligned}$$

where $\ell = 4, \ldots, S$, and $0 \leq \theta_{j,i} \leq \pi$ for $1 \leq i < j \leq S - 1$ which make the diagonal entries to be positive. A sparsity-inducing prior is applied on $\theta_{i,j}$'s and illustrated in Section 1.4.

1.1.2 Extension to a two-way DAG model for $\mathbf{Y}_t$

Unlike the previous subsection, estimating DAGs characterizing causal association among the variables in both rows and columns of $\mathbf{Y}_t$ may be of interest. Then we can formulate the model as $\mathbf{Y}_t - \mathbf{W}_1\mathbf{Y}_t\mathbf{W}_2 = \mathbf{D}_1\mathbf{Z}_t\mathbf{D}_2$ with diagonal $\mathbf{D}_1$ and $\mathbf{D}_2$ which can be re-written in the form of (1) as,

$$\mathbf{y}_t - (\mathbf{W}_1 \otimes \mathbf{W}_2)\mathbf{y}_t = \mathbf{D}^{1/2}\mathbf{z}_t, \tag{2}$$

where $\mathbf{y}_t$ is column-wise vectorization of $\mathbf{Y}_t$ and $\otimes$ stands for the kronecker product. The DAG-ness constraint can be imposed on the two coefficient matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, allowing us to characterize causal associations among the variables in columns and rows, respectively. We can again generate unrestricted posterior samples of $\mathbf{W}_1$ and $\mathbf{W}_2$ and subsequently project those in the DAG-space.

1.2 Likelihood

The likelihood will be evaluated based on the distributions of the univariate time-series $\mathbf{z}_{k,t}$'s.

$$\begin{aligned} \mathbf{S}_1 &= \operatorname{Diag}(f_1(\omega_1), \ldots, f_p(\omega_1)), \\ \mathbf{S}_{2k+1} = \mathbf{S}_{2k} &= \operatorname{Diag}(f_1(\omega_k), \ldots, f_p(\omega_k)), \;\; k = 1, \ldots, \lfloor (T-1)/2 \rfloor, \\ \mathbf{S}_T &= \operatorname{Diag}(f_1(\omega_{T/2}), \ldots, f_p(\omega_{T/2})), \text{ for even } T. \end{aligned} \tag{3}$$

1.3 Modeling univariate error processes

Since our final data application relies on the model in Section 1.1, we discuss the model for the error processes for the multi-group case directly. As in Roy et al. [2024], the latent univariate processes $(z_{s,k,t} : t = 1, 2, \ldots)$, $s = 1, \ldots, S,\ k = 1, \ldots, p$, are assumed to be independent and stationary. We again model these univariate time series in the spectral domain and model the spectral density of $(Z_{s,k,t} : t = 1, 2, \ldots)$, defined by $f_{s,k}(\omega) = \gamma_j(0) + 2\sum_{h=1}^{\infty} \gamma_{s,k}(h)\cos(h\omega)$, $\omega \in [-\pi, \pi]$. Furthermore, it is symmetric about 0 and uniquely determines the distribution of the time series since the autocovariances $(\gamma_{s,k}(h))$ are given by the inverse Fourier coefficients $\gamma_{s,k}(h) = \int_{-\pi}^{\pi} f_\ell(\omega)\cos(h\omega)d\omega$, $h = 1, 2, \ldots$. To ensure unit marginal variances, we need $\int_{-\pi}^{\pi} f_{s,k}(\omega)\cos(h\omega)d\omega = \gamma_\ell(0) = 1$, as all $Z_{s,k,t}$. Hence, spectral densities $f_{s,k}$, $s = 1, \ldots, S,\ k = 1, \ldots, p$, are symmetric probability densities on $[-\pi, \pi]$. To model the spectral densities nonparametrically, B-spline bases are convenient because of their shape and order-preserving properties. Writing $B_j^* = B_j / \int_0^1 B_j(u)du$ for the normalized B-splines with a basis consisting of J many B-splines, we may consider a model indexed by the vector of spline coefficients $\boldsymbol{\theta}_{s,k} = (\theta_{s,k,1}, \ldots, \theta_{s,k,j})$ given by

$$\varphi(\omega; \boldsymbol{\theta}_{s,k}) = \sum_{j=1}^{J} \theta_{s,k,j} B_j^*(|\omega|/\pi), \quad \theta_{s,k,j} \geq 0, \quad \sum_{j=1}^{J} \theta_{s,k,j} = 1/2. \tag{4}$$

We consider the following convenient representation that avoids the restriction of nonnegativity and sum constraint:

$$\theta_{s,k,j} = \Psi(\kappa_{s,k,j}) / \{2 \sum_{j=1}^{J} \Psi(\kappa_{s,k,j})\}, \tag{5}$$

where $\Psi(u) = (1 + u/(1 + |u|))/2$ is a link function monotonically mapping the real line to the unit interval [Roy et al., 2024].

A dimension reduced formulation can be achieved by assuming a CANDECOMP/PARAFAC (CP) decomposition-based low-rank decomposition,

$$\kappa_{s,k,j} = \sum_{r=1}^{R} \xi_{s,r} \chi_{k,r} \eta_{j,r}, \quad s = 1, \ldots, S,\ k = 1, \ldots, p,\ j = 1, \ldots, J, \tag{6}$$

in which case the dimension reduces to $R(p + S + J)$.

1.4 Prior specification

We now illustrate the prior distributions for each parameter involved in our model.

- Horseshoe prior on $\mathbf{W}$: For $i \neq j$, we let $w_{i,j} \sim \text{Normal}(0, d_{i,i}\lambda_{i,j}^2\tau^2)$ with $\lambda_{i,j}, \tau \sim C^+(0,1)$, where C^+ stands for the half-Cauchy distribution. The half-Cauchy priors from Carvalho et al. [2010] admit scale mixture representation, proposed in Makalic and Schmidt [2015]. Specifically, it leads to the following mixture representation, $\lambda_{i,j}^2 \sim \text{Inverse-Gamma}(1/2, 1/\nu_{i,j})$, $\tau^2 \sim \text{Inverse-Gamma}(1/2, 1/\xi)$ and $\nu_{i,j}, \xi \sim \text{Inverse-Gamma}(1/2, 1)$ which makes the Gibbs sampling from the posterior distribution straightforward. Marginally $\lambda_{i,j}, \tau \sim C^+(0,1)$.

- $\theta_{i,j}$: We put a soft-thresholding prior to promote shrinkage on $\theta_{i,j}$ at $\pi/2$. $\theta_{i,j} = sign(\theta_{i,j}^* - \pi/2)(|\theta_{i,j}^* - \pi/2| - \lambda)_+ + \pi/2$ and $\frac{\theta_{i,j}^*}{\pi} \sim \text{Beta}(a,a)$ or $\theta_{i,j}^* = \pi(1 + \theta_{i,j}^u/(1 + |\theta_{i,j}^u|))/2$ and $\theta_{i,j}^u \sim \text{Normal}(0, \sigma_T^2)$. Finally, we set a uniform prior for the soft-thresholing parameter $\lambda \sim \text{Unif}[\lambda_L, \lambda_U]$.

- $\mathbf{D} \sim G$: The stick-breaking prior for the unknown distribution $G = \sum_{i=k}^M \pi_k \delta(m_k)$ with m_k following an Inverse-Gaussian distribution [Chhikara, 1988] with density function $\pi_d(t) \propto t^{-3/2}e^{-(t-\mu_d)^2/(2t)}$, $t > 0$, for some $\mu_d > 0$. We put weakly informative mean-zero normal prior with large variance on μ_d. and $p_k = v_k \prod_{j=1}^{k-1} v_j$ with $v_j \sim$Beta$(1, v)$ and $v \sim$Gamma(a_v, b_v).

- The CP-decomposition parameters: Since an appropriate value of the rank R is unknown, we consider an indirect automatic selection of R through cumulative shrinkage priors [Bhattacharya and Dunson, 2011]. 1) For all $r = 1, \ldots, R$, we put independent priors $\xi_{s,r}|v_{s,r}, \tau_r \sim \text{N}(0, v_{s,r}^{-1}\tau_r^{-1})$, $v_{s,r} \sim \text{Gamma}(\nu_1, \nu_1)$, $\tau_r = \prod_{i=1}^r \Delta_i$, where $\Delta_1 \sim \text{Gamma}(\kappa_1, 1)$ and $\Delta_i \sim \text{Gamma}(\kappa_2, 1)$, $i \geq 2$. The parameters $v_{j,r}$, $j = 1, \ldots, p$, $r = 1, 2, \ldots$, control local shrinkage of the elements in $\xi_{j,r}$, whereas τ_r controls column shrinkage of the rth column. 2) Next, let $\boldsymbol{\eta}_r = (\eta_{1,r}, \ldots, \eta_{J,r}) \sim \text{N}_K(0, \sigma_\kappa \mathbf{P}^{-1})$, $\sigma_\xi, \sigma_\kappa \sim \text{Inv-Ga}(c_1, c_1)$. The $\mathbf{P}$ is the second-order difference matrix to impose smoothness. Specifically, $\mathbf{P} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$, where $\mathbf{Q}$ is a $K \times (K+2)$ matrix such that $\mathbf{Q}\boldsymbol{\eta}_r$ computes the second differences in $\boldsymbol{\eta}_r$. The above prior thus induces smoothness in the coefficients because it penalizes $\sum_{j=1}^J (\Delta^2\boldsymbol{\eta}_r)^2 = \boldsymbol{\eta}_r^{\mathrm{T}}\mathbf{P}\boldsymbol{\eta}_r$, the sum of squares of the second-order differences in $\boldsymbol{\eta}_r$. On J, we put a prior with Poisson-like tail $e^{-J \log J}$. 3) Finally, another cumulative shrinkage prior is set for $\chi_{k,r}$.

2 Posterior sampling

Here, we discuss unrestricted posterior sampling steps for all the parameters.

- Update $\mathbf{W}$: Each row in $\mathbf{W}$ enjoys full conditional Gaussian posterior.

- Update $\mathbf{D}$: We introduce latent indicator variables z_ℓ's for each diagonal entry $d_{\ell,\ell}$ in $\mathbf{D}$. The posterior of the atoms m_k's are Generalized Inverse-Gaussian and are updated

- Update $\mathbf{V}$: We consider adaptive Metropolis-Hastings [Haario et al., 2001] to update the $\theta^*_{i,j}$'s and the thresholding parameter λ is updated based on a random-walk Metropolis-Hastings in log-scale with a Jacobian adjustment.
- Updating the spectral densities: We consider gradient-based Langevin Monte Carlo (LMC) to update the parameters involving $\kappa_{s,k,j}$'s as in Roy et al. [2024].

3 Consistency

For simplicity, we focus on the simpler model without any multi-group component for theoretical analysis and make the following assumptions on the true parameters.

(A1) The true $\|\mathbf{D}_0^{-1/2}\mathbf{W}_0\|_\infty$ has an upper bound E_T and has s non-zero entries and $\mathbf{D}_0$ bounded up and below. There exists a permutation matrix $\mathbf{P}$ such that $\mathbf{P}\mathbf{W}_0\mathbf{P}^{\mathrm{T}}$ is strictly lower triangular and thus $\|\mathbf{P}(\mathbf{I}-\mathbf{W}_0)\mathbf{P}^{\mathrm{T}}\|_{op} = 1$.

(A2) The true spectral densities $f_{0,1}, \dots, f_{0,p}$ of the latent processes are positive and Hölder continuous of smoothness index α for some $\alpha > 0$ (cf., Definition C.4 of Ghosal and Van der Vaart [2017]) such that for some uniformly bounded sequence θ^*_{jk} of the form (5) and (6), it holds that

$$\max_{1\le j\le p}\{\sup\{|f_{0,j}(\omega) - \sum_{k=1}^{K}\theta^*_{jk}B^*_k(\omega)| : \omega \in [0,1]\} \lesssim K^{-\alpha}, \tag{7}$$

that is, the approximation ability of the splines for the true spectral densities is not affected if the coefficients are restricted by the dimension-reduction condition.

(A3) Growth of dimension: $\log p \lesssim \log T$.

Due to the complexity in nonparametric time-series modeling, we need to impose the following two restrictions in the supports of $\mathbf{D}$, $\mathbf{W}$, and the spectral density parameters as in Theorem 3 of Roy et al. [2024].

Conditions on prior:

1. Diagonal entries $d_1, \dots, d_p$ and the range of the functions $f_1, \dots, f_p$ lie in a fixed, compact subinterval of $(0, \infty)$.
2. Eigenvalues of $(\mathbf{I}-\mathbf{W})^{\mathrm{T}}(\mathbf{I}-\mathbf{W})$ are bounded between two fixed positive real values in $(0, \infty)$.

The last condition is not too restrictive, as eigenvalues of $(\mathbf{I}-\mathbf{W}_0)$ are all 1. We first define $\mathbf{\Omega} = (\mathbf{I}-\mathbf{W})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{I}-\mathbf{W})$ and similarly $\mathbf{\Sigma} = (\mathbf{I}-\mathbf{W})^{-1}\mathbf{D}\{(\mathbf{I}-\mathbf{W})^{\mathrm{T}}\}^{-1}$. Let,

$$L_T(\boldsymbol{\beta}) = \frac{1}{T}\|\mathbf{W}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y}\|_2^2 + \zeta_T\sum_{i,j}c_{i,j}|\beta_{ij}| + \frac{\rho}{2}|h(\boldsymbol{\beta})|^2 + \alpha h(\boldsymbol{\beta}), \tag{8}$$

where posterior of $\mathbf{W}$ is unrestricted. We also assume $\zeta_T/T \to \zeta_0$.

Consistency-proof steps will be as: 1) **Part 1** showing $\Pi(\|\mathbf{W}-\mathbf{W}_0\|_2 > \gamma \mid \mathbf{Y}) \to 0$ implies $\Pi(\|\mathbf{W}^*-\mathbf{W}_0\|_2 > \epsilon \mid \mathbf{Y}) \to 0$ for some $\epsilon \asymp \gamma$ and 2) **Part 2** show $\Pi(\|\mathbf{W}-\mathbf{W}_0\|_2 > \gamma \mid \mathbf{Y}) \to 0$.

3.1 Part 1

Let $\boldsymbol{\delta}_T = \boldsymbol{\beta} - \mathbf{W}_0 = \boldsymbol{\delta}_T = \eta_T \boldsymbol{\delta}$, where $\boldsymbol{\delta} = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,p} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{p,1} & \delta_{p,2} & \cdots & \delta_{p,p} \end{pmatrix}$, $\|\boldsymbol{\delta}\|_F \leq d$, where d is a constant.

$$\begin{aligned} D_T(\boldsymbol{\delta}_T) =& L_T(\mathbf{W}_0 + \boldsymbol{\delta}_T) - L_T(\mathbf{W}_0) \\ =& l_T(\mathbf{W}_0 + \boldsymbol{\delta}_T) - l_T(\mathbf{W}_0) + \zeta_T \sum_i \sum_j \left\{ |w_{0,i,j} + \eta_T \delta_{i,j}| - |w_{0,i,j}| \right\} + \\ & \frac{\rho}{2} \left\{ |h(\mathbf{W}_0 + \boldsymbol{\delta}_T)|^2 - |h(\mathbf{W}_0)|^2 \right\} + \alpha \left\{ h(\mathbf{W}_0 + \boldsymbol{\delta}_T) - h(\mathbf{W}_0) \right\}, \end{aligned}$$

where $l_T(\boldsymbol{\beta}) = \frac{1}{T}\|\mathbf{WY} - \boldsymbol{\beta}\mathbf{Y}\|_2^2$.

If $\mathbf{W}^* = \arg\min_{\boldsymbol{\beta}} L_T(\boldsymbol{\beta})$, then $\mathbf{W}^* - \mathbf{W}_0 = \arg\min_{\boldsymbol{\delta}_T} D_T(\boldsymbol{\delta}_T)$.

Let $\mathbf{W}_0$ denote the underlying true adjacency matrix. And then, we define

$$\begin{aligned} \mathcal{A} &= \{(i,j) : w_{0,i,j} \neq 0\}, \\ \mathcal{A}_c &= \{(i,j) : w_{0,i,j} = 0\}, \end{aligned}$$

which means that $\mathcal{A}$ collects the indices for terms whose true parameters are nonzero, and $\mathcal{A}_c$ contains the indices for terms that do not exist in the underlying true model.

Let $a_{T,i,j} = \zeta_T c_{i,j}$, $(i,j) \in \mathcal{A}$, we can derive

$$\begin{aligned} D_T(\boldsymbol{\delta}) \geq & l_T(\mathbf{W}_0 + \boldsymbol{\delta}_T) - l_T(\mathbf{W}_0) + \sum_{(i,j)\in\mathcal{A}} a_{T,i,j} \left\{ |w_{0,i,j} + \eta_T \delta_{i,j}| - |W_{0,i,j}| \right\} + \\ & \frac{\rho}{2} \left\{ |h(w_{0,i,j} + \eta_T \delta_{i,j})|^2 - |h(w_{0,i,j})|^2 \right\} + \alpha \left\{ h(w_{0,i,j} + \eta_T \delta_{i,j}) - h(w_{0,i,j}) \right\} \\ \geq & l_T(\mathbf{W}_0 + \boldsymbol{\delta}_T) - l_T(\mathbf{W}_0) + \sum_{(i,j)\in\mathcal{A}} a_{T,i,j} |\delta_{T,i,j}| + \\ & \frac{\rho}{2} \left\{ |h(\mathbf{W}_0 + \boldsymbol{\delta}_T)|^2 - |h(\mathbf{W}_0)|^2 \right\} + \alpha \left\{ h(\mathbf{W}_0 + \boldsymbol{\delta}_T) - h(\mathbf{W}_0) \right\} \\ = & \nabla l_T(\mathbf{W}_0)^{\mathrm{T}}(\boldsymbol{\delta}_T) + \frac{1}{2}(\boldsymbol{\delta}_T)^{\mathrm{T}}[\nabla^2 l_T(\mathbf{W}_0)](\boldsymbol{\delta}_T) \\ & - \sum_{(i,j)\in\mathcal{A}} a_{T,i,j} |\delta_{T,i,j}| + \frac{\rho}{2} \left\{ |h(\mathbf{W}_0 + \boldsymbol{\delta}_T)|^2 - |h(\mathbf{W}_0)|^2 \right\} + \alpha \left\{ h(\mathbf{W}_0 + \boldsymbol{\delta}_T) - h(\mathbf{W}_0) \right\} \end{aligned}$$

Let $\mathbf{X} = \mathbf{WY}$. Then, for the first term,

$$\begin{aligned} \mathrm{trace}\{\nabla l_T(\mathbf{W}_0)^{\mathrm{T}}(\boldsymbol{\delta}_T)\} =& \mathrm{trace}\{\left(\nabla l_T(\mathbf{W}_0)^{\mathrm{T}}\right) \boldsymbol{\delta}_T\} \\ =& -2\frac{1}{T}\mathrm{trace}\left\{ \left((\mathbf{X} - \mathbf{W}_0\mathbf{Y})\mathbf{Y}^{\mathrm{T}} \right) \boldsymbol{\delta}_T \right\}. \end{aligned}$$

For second term,

$$\mathrm{trace}\left\{ \boldsymbol{\delta}_T^{\mathrm{T}} [\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}] \boldsymbol{\delta}_T \right\} > 0.$$

For third term,

$$-\sum_{(i,j)\in\mathcal{A}} a_{T,i,j}|\delta_{T,i,j}| \geq -\eta_T A\frac{\zeta_T}{T}d,$$

where $a_{T,i,j} \leq A\frac{\zeta_T}{T}$ for some ζ_T and constant A. The last two terms can be combined and since $h(\mathbf{W}) \geq 0$, $\forall \mathbf{W} \in \mathbb{R}^{p\times p}$ with $h(\mathbf{W}_0) = 0$, we have

$$[h(\mathbf{W}_0 + \boldsymbol{\delta}_T) - h(\mathbf{W}_0)]\left\{\frac{\rho}{2}\left[h(\mathbf{W}_0 + \boldsymbol{\delta}_T) + h(\mathbf{W}_0)\right] + \alpha\right\} \geq 0$$

Thus,

$$\begin{aligned}
D_T(\boldsymbol{\delta}_T) \geq & -2\text{trace}\left\{\left(\frac{1}{T}(\mathbf{X} - \mathbf{W}_0\mathbf{Y})\mathbf{Y}^{\mathrm{T}}\right)\boldsymbol{\delta}_T\right\} + \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}[\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}]\boldsymbol{\delta}_T\right\} - \sum_{(i,j)\in\mathcal{A}} a_{T,i,j}|\delta_{T,i,j}| \\
\geq & -2\text{trace}\left\{\left|\frac{1}{T}(\mathbf{X} - \mathbf{W}_0\mathbf{Y})\mathbf{Y}^{\mathrm{T}}\right||\boldsymbol{\delta}_T|\right\} + \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}[\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}]\boldsymbol{\delta}_T\right\} - \sum_{(i,j)\in\mathcal{A}} a_{T,i,j}|\delta_{T,i,j}| \\
\geq & -2\text{trace}\left\{\left|\frac{1}{T}(\mathbf{W} - \mathbf{W}_0)\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right||\boldsymbol{\delta}_T|\right\} + \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}\boldsymbol{\Sigma}_{0,\mathbf{Y}}\boldsymbol{\delta}_T\right\} \\
& - \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}|\boldsymbol{\Sigma}_{0,\mathbf{Y}} - \frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}|\boldsymbol{\delta}_T\right\} - \sum_{(i,j)\in\mathcal{A}} a_{T,i,j}|\delta_{T,i,j}| \\
\geq & -2\|\mathbf{W} - \mathbf{W}_0\|_2\left\|\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right\|_2\|\boldsymbol{\delta}_T\|_2 + \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}\boldsymbol{\Sigma}_{0,\mathbf{Y}}\boldsymbol{\delta}_T\right\} \\
& - \text{trace}\left\{\boldsymbol{\delta}_T^{\mathrm{T}}|\boldsymbol{\Sigma}_{0,\mathbf{Y}} - \frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}|\boldsymbol{\delta}_T\right\} - \sum_{(i,j)\in\mathcal{A}} a_{T,i,j}|\delta_{T,i,j}|
\end{aligned}$$

$$\|\mathbf{W}-\mathbf{W}_0\|_2\|\tfrac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\|_2\|\boldsymbol{\delta}_T\|_2 \leq \|\mathbf{W}-\mathbf{W}_0\|_2\|\boldsymbol{\Sigma}_{0,\mathbf{Y}}\|_2\|\boldsymbol{\delta}_T\|_2 + p\|\mathbf{W}-\mathbf{W}_0\|_2\|\tfrac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \boldsymbol{\Sigma}_{0,\mathbf{Y}}\|_\infty\|\boldsymbol{\delta}_T\|_2.$$

Let $\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ converges to $\boldsymbol{\Sigma}_{0,\mathbf{Y}}$ such that each entry is of order $\theta_T = \|\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \boldsymbol{\Sigma}_{0,\mathbf{Y}}\|_\infty$. For a large enough T, we have $\theta_T < \lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}})/2 - c$ with probability more than 1/2 for some small constant c.

Under the null, each row i of $(\mathbf{I} - \mathbf{W}_0)\mathbf{Y}$ follows $\text{MVN}(\mathbf{0}, d_i\boldsymbol{\Sigma}_{0,i})$, independently.

Then $\mathbb{E}\left(\frac{1}{T}(\mathbf{I} - \mathbf{W}_0)\mathbf{Y}\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{W}_0)^{\mathrm{T}}\right) = \mathbf{D}_0\text{diag}(\mathbf{s}_0)$, where $s_{0,i} = \text{trace}(\boldsymbol{\Sigma}_{0,i}) = T$ as $\boldsymbol{\Sigma}_{0,i}$'s are correlation matrices and $\mathbb{E}\left(\frac{1}{T}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) = \boldsymbol{\Sigma}_{0,\mathbf{Y}} = (\mathbf{I} - \mathbf{W}_0)^{-1}\mathbf{D}_0\{(\mathbf{I} - \mathbf{W}_0)^{\mathrm{T}}\}^{-1}$.

$$D_T(\boldsymbol{\delta}_T) \geq \|\boldsymbol{\delta}_T\|_2^2\lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) - 2\|\boldsymbol{\delta}_T\|_2\|\mathbf{W} - \mathbf{W}_0\|_2\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} - A\frac{\zeta_T}{T}p\|\boldsymbol{\delta}_T\|_2 - \theta_T\|\boldsymbol{\delta}_T\|_2^2$$

If $\{\|\mathbf{W} - \mathbf{W}_0\|_2 \leq \gamma\}$ with $\|\boldsymbol{\delta}_T\|_2\lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) > 4\gamma\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} + 2A\frac{\zeta_T}{T}p + 2\theta_T\|\boldsymbol{\delta}_T\|_2$, then $D_T(\boldsymbol{\delta}_T) \geq \|\boldsymbol{\delta}_T\|_2\left(\|\boldsymbol{\delta}_T\|_2\lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) - 2\gamma - A\frac{\zeta_T}{T}p - \theta_T\|\boldsymbol{\delta}_T\|_2\right) > 0$

Since, $D_T(\mathbf{0}) = 0$, thus $D_T(\boldsymbol{\delta}_T)$ does not achieve minima at least in the set $\{\boldsymbol{\delta}_T : \|\boldsymbol{\delta}_T\|_2 \lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) - 4\gamma\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} - 2A\frac{\zeta_T}{T}p - 2\theta_T\|\boldsymbol{\delta}_T\|_2 > 0\}$

Here $\mathbf{W}^* - \mathbf{W}_0$ is δ_T.

$\Pi(\|\mathbf{W}^* - \mathbf{W}_0\|_2 > \epsilon \mid \mathbf{Y}) \leq \Pi(\|\mathbf{W} - \mathbf{W}_0\|_2 > \gamma \mid \mathbf{Y})$, where $\epsilon = \frac{4\gamma\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} + 2Ap\frac{\zeta_T}{T}}{\lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) - 2\theta_T} \leq \frac{\gamma(4\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} + 2A)}{\lambda_{\min}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) - 2\theta_T}$ setting $\gamma\{\sqrt{p}\lambda_{\max}(\boldsymbol{\Sigma}_{0,\mathbf{Y}}) + p\theta_T\} \geq 2p\frac{\zeta_T}{T}$.

This completes the proof of the first part.

3.2 Part 2

We can rewrite our model as $\mathbf{y}_t = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{D}\mathbf{z}_t$ which is similar to the OUT model in Roy et al. [2024]. If we have a similar bound for the horseshoe prior like the OUT model, we show that $\Pi(\|(\mathbf{I} - \mathbf{W})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{I} - \mathbf{W}) - (\mathbf{I} - \mathbf{W}_0)^{\mathrm{T}}\mathbf{D}_0^{-1}(\mathbf{I} - \mathbf{W}_0)\|_2 > \varepsilon \mid \mathbf{Y}) \to 0$.

We apply Lemma A.4 from Song and Liang [2023] or Lemma 6 of Bernardo et al. [1998]. Let $\boldsymbol{\kappa} = \{\mathbf{W}, \mathbf{D}, f_1, \ldots, f_p\}$. It requires two conditions. For the testing condition, we use the test considered in Roy et al. [2024]. To show the ELBO bound condition $\{\int \frac{q_{T,\boldsymbol{\kappa}}}{q_{T,\boldsymbol{\kappa}^*}}\pi(\boldsymbol{\kappa})d\boldsymbol{\kappa} \geq \exp(-cT\epsilon_T^2)\}$, we apply the KL support result from Roy et al. [2024] and invoke Lemma 8.10 of Ghosal and Van der Vaart [2017].

We also define the event $B_T = \{$At least s_T many entries in $|\mathbf{W}\mathbf{D}^{-1/2}|$ is greater than $a_T\}$.

$$\begin{aligned}
&1 - P(-a \leq w_{i,j}/\sqrt{d_{i,i}} \leq a) \\
&= 2\int_0^\infty \left(1 - \Phi\left(\frac{a}{\alpha\lambda_{i,j}}\right)\right) f(\lambda_{i,j})d\lambda_{i,j} \\
&= 2\int_0^b \left(1 - \Phi\left(\frac{a}{\alpha\lambda_{i,j}}\right)\right) f(\lambda_{i,j})d\lambda_{i,j} + 2\int_b^\infty \left(1 - \Phi\left(\frac{a}{\alpha\lambda_{i,j}}\right)\right) f(\lambda_{i,j})d\lambda_{i,j} \\
&\leq 2\left(1 - \Phi\left(\frac{a}{\alpha b}\right)\right) + \int_b^\infty \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda \\
&= 2\left(1 - \Phi\left(\frac{a}{\alpha b}\right)\right) + \frac{2}{\pi}\left(\frac{\pi}{2} - \tan^{-1}(b)\right)
\end{aligned}$$

We will use the tail-bound of Gaussian $\Phi(-x) \leq (1/\sqrt{2\pi})(e^{-x^2/2}/x)$ for $x > 0$. Then, $\left(1 - \Phi\left(\frac{a}{\alpha b}\right)\right) \leq (1/\sqrt{2\pi})\left\{\frac{\exp\left(-\frac{1}{2}\frac{a^2}{\alpha^2 b^2}\right)}{\frac{a}{\alpha b}}\right\}$

- For $a = \epsilon_T/p^2, b = p^{-(1+\mu'')}$: Then $1 - P(-a \leq w_{i,j}/\sqrt{d_{i,i}} \leq a) \leq p^{-(1+\mu)}$ for some $0 < \mu < \mu''$

Lemma 1. *Let $\mathbf{W}_1, \mathbf{W}_2$ be $(p \times p)$-matrices with zero in the diagonal. Let $\mathbf{D}_1^{-1/2}, \mathbf{D}_2^{-1/2}$ be diagonal matrices with entries in $(0, B)$ including those of $\mathbf{D}_1^{-1/2}\mathbf{W}_1$, $\mathbf{D}_2^{-1/2}\mathbf{W}_2$ for some $B > 0$. Then for $\boldsymbol{\Omega}_1 = (\mathbf{I}_p - \mathbf{W}_1)^{\mathrm{T}}\mathbf{D}_1^{-1}(\mathbf{I}_p - \mathbf{W}_1)$, $\boldsymbol{\Omega}_2 = (\mathbf{I}_p - \mathbf{W}_2)^{\mathrm{T}}\mathbf{D}_2^{-1}(\mathbf{I}_p - \mathbf{W}_2)$, we have that $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_F \leq 2pB\{\|\mathbf{D}_1^{-1/2}\mathbf{W}_1 - \mathbf{D}^{-1/2}\mathbf{W}_2\|_{\mathrm{F}} + \|\mathbf{D}_1^{-1/2} - \mathbf{D}_2^{-1/2}\|_{\mathrm{F}}\}$.*

Proof.

$$\begin{aligned}
&\|(\mathbf{I}_p - \mathbf{W}_1)^{\mathrm{T}}\mathbf{D}_1^{-1}(\mathbf{I}_p - \mathbf{W}_1) - (\mathbf{I}_p - \mathbf{W}_2)^{\mathrm{T}}\mathbf{D}_2^{-1}(\mathbf{I}_p - \mathbf{W}_2)\|_{\mathrm{F}} \\
&\leq \{\|\mathbf{D}_1^{-1/2}\mathbf{W}_1 - \mathbf{D}^{-1/2}\mathbf{W}_2\|_{\mathrm{F}} + \|\mathbf{D}_1^{-1/2} - \mathbf{D}_2^{-1/2}\|_{\mathrm{F}}\}(\|\mathbf{D}_1^{-1/2} - \mathbf{D}_1^{-1/2}\mathbf{W}_1\|_{\mathrm{op}} + \|\mathbf{D}_2^{-1/2} - \mathbf{D}_2^{-1/2}\mathbf{W}_2\|_{\mathrm{op}}) \\
&\leq 2pB\{\|\mathbf{D}_1^{-1/2}\mathbf{W}_1 - \mathbf{D}^{-1/2}\mathbf{W}_2\|_{\mathrm{F}} + \|\mathbf{D}_1^{-1/2} - \mathbf{D}_2^{-1/2}\|_{\mathrm{F}}\}
\end{aligned}$$

□

Let $\mathbf{M}_t = (\mathbf{I} - \mathbf{W})^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{S}_t^{-1}\mathbf{D}^{-1/2}(\mathbf{I} - \mathbf{W})$ for $1 \leq t \leq T$. For any $\epsilon_T > 0$, $\Pi(\max_{1 \leq t \leq T} \|\mathbf{M}_t - \mathbf{M}_{t0}\|_{\mathrm{F}}^2 \leq \epsilon_T^2) > 0$.

Following the proof of Lemma 6 of Roy et al. [2024], we obtain the estimate

$$\begin{aligned}
\|\mathbf{M}_t - \mathbf{M}_{t0}\|_{\mathrm{F}} &\lesssim \max\{\|\mathbf{\Omega}^{1/2} - \mathbf{\Omega}_0^{1/2}\|_{\mathrm{F}}, \|\mathbf{S}_t^{-1} - \mathbf{S}_{t0}^{-1}\|_{\mathrm{F}}\} \\
&\lesssim \max\{\|\mathbf{\Omega} - \mathbf{\Omega}_0\|_{\mathrm{F}}, \|\mathbf{S}_t^{-1} - \mathbf{S}_{t0}^{-1}\|_{\mathrm{F}}\} \qquad (9)
\end{aligned}$$

uniformly for $t = 1, \ldots, T$. Then, using the prior independence of the parameters, it will suffice to show that for all sufficiently small $\epsilon > 0$,

(i) $-\log\Pi(\|\mathbf{\Omega} - \mathbf{\Omega}_0\|_{\mathrm{F}} \leq \epsilon) \lesssim (p + s)\log(p/\epsilon)$;

(ii) $-\log\Pi(\max\{\|\mathbf{S}_t^{-1} - \mathbf{S}_{t0}^{-1}\|_{\mathrm{F}} : 1 \leq t \leq p\} \leq \epsilon) \lesssim (p + \epsilon^{-1/\alpha})\log(1/\epsilon)$ for $K \asymp \epsilon^{-1/\alpha}$.

Note that $\|\mathbf{I} - \mathbf{W}_0\|_{op} = 1$ The bound in (i) was established in the proof of Theorem 4.2 of Shi et al. [2021] for the sparse Cholesky decomposition prior for $\mathbf{\Omega}$ based on independent continuous shrinkage distributions on the entries of $\mathbf{W}$;

To establish the bound in (ii), since the true spectral densities are bounded away from 0, it suffices to prove that $-\log\Pi(\max\{\|f_t - f_{t0}\|_\infty : 1 \leq t \leq p\} \leq \epsilon) \lesssim (p + \epsilon^{-1/\alpha})\log(1/\epsilon)$. In view of (7), it suffices to use $K \asymp \epsilon^{-1/\alpha}$ terms in the B-spline basis expansion to control the bias within a multiple of ϵ. Hence, using the fact that the normalized B-splines are uniformly bounded by a multiple of K, it will be enough to estimate the prior concentration of the ϵ/K neighborhoods of a $R(p+K)$-dimensional vector of uniformly bounded entries in the Euclidean distance. The resulting estimate is $(\epsilon/K)^{R(p+K)}$. Since R is assumed to be not growing with T, this leads to (ii). Hence, the result follows.

Lemma 2 (Identifiability)**.** *If* $(\mathbf{W}_0, \mathbf{D}_0, f_{0,j})$ *and* $(\mathbf{W}_1, \mathbf{D}_1, f_{1,j})$ *impose distributionally equivalent densities on* $\mathbf{y}_t$*, then we must have* $\mathbf{W}_0 = \mathbf{W}_1$ *and* $\mathbf{D}_0 = \mathbf{D}_1$ *under the assumption that either the spectral densities or the entries in* $\mathbf{D}_0$ *or* $\mathbf{D}_1$ *are not constant.*

Proof. Under $(\mathbf{W}_0, \mathbf{D}_0, f_{0,j})$, we have $(\mathbf{I} - \mathbf{W}_0)\tilde{y}_t$ are component-wise independent and heteroscedastic. Here, $\tilde{y}_t$ is the Fourier transformation of the original time-series at *t-th* frequency.

$(\mathbf{I} - \mathbf{W}_1)\tilde{y}_t = \{(\mathbf{I} - \mathbf{W}_1)(\mathbf{I} - \mathbf{W}_0)^{-1}\}(\mathbf{I} - \mathbf{W}_0)\tilde{y}_t$ is also component-wise independent and heteroscedastic by assumption. Then using Gaussianity, $(\mathbf{I} - \mathbf{W}_1)(\mathbf{I} - \mathbf{W}_0)^{-1}$ must be diagonal.

Now if $(\mathbf{I} - \mathbf{W}_1)(\mathbf{I} - \mathbf{W}_0)^{-1} = \mathbf{D}_2$, which is diagonal, then $(\mathbf{I} - \mathbf{W}_1) = \mathbf{D}_2(\mathbf{I} - \mathbf{W}_0)$. Then comparing diagonal entries, we must have $\mathbf{D}_2 = \mathbf{I}$ as $\mathbf{W}_1$ and $\mathbf{W}_0$ have zero entries in the diagonal. Thus $\mathbf{W}_1 = \mathbf{W}_0$. Since, $\frac{1}{T}\sum_j f_{k,j}(\omega_j) \to 1$ for $k = 0, 1$, we have $\mathbf{D}_0 = \mathbf{D}_1$. □

Lemma 3. *If a continuous function h is injective on an open set U, then h^{-1} is continuous on $h(U)$.*

Let $h : (\mathbf{W}, \mathbf{D}) \to (\mathbf{I} - \mathbf{W})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{I} - \mathbf{W})$ and $V = \{\boldsymbol{\Omega} : \|\boldsymbol{\Omega} - h(\mathbf{W}_0, \mathbf{D}_0)\|_2 < \varepsilon\}$. Here h is continuous and injective. For ε small, $h^{-1}(V)$ is continuous due to identifiability and thus $\|\mathbf{W} - \mathbf{W}_0\|_2 < \delta$ for $\mathbf{W} \in h^{-1}(V)$. Hence, $\Pi(\|(\mathbf{I}-\mathbf{W})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{I}-\mathbf{W}) - (\mathbf{I}-\mathbf{W}_0)^{\mathrm{T}}\mathbf{D}_0^{-1}(\mathbf{I}-\mathbf{W}_0)\|_2 > \varepsilon \mid \mathbf{Y}) \to 0$ implies $\Pi(\|\mathbf{W} - \mathbf{W}_0\|_2 > \delta \mid \mathbf{Y}) \to 0$. This completes the proof of the second part.

4 Simulation

To evaluate the performance of our proposed model, we run three simulation experiments. We generate the simulated datasets following the proposed model with $p = 40, S = 15$, and $T = 32$ or 48 with three different choices for the distributions of the univariate time series. The specific differences are described at the beginning of each subsection. For all the cases, the sparse DAG matrix, $\mathbf{W}$, is generated using `randDAG` of `pcalg` [Alain Hauser and Peter Bühlmann, 2012] with two possible values for the 'Expected neighbors' as 2 and 4. This leads to different levels of sparsity in $\mathbf{W}$. The weights for the edges are generated from $\mathrm{Unif}((-2, -0.5) \cup (0.5, 2))$.

The entries in $\mathbf{D}$ are generated as an absolute value of $\mathrm{Normal}(7, 2)$. The across-state correlation matrix $\mathbf{V}$ is generated in two steps. First, a precision matrix is generated using g-Wishart where the underlying graph is simulated by combining three small-worlds, each with 5 nodes similar to Roy et al. [2024]. Then, we take its inverse and subsequently scale to get the $\mathbf{V}$. We compare our methods with NO-TEARS [Zheng et al., 2018] in terms of both estimation of $\mathbf{W}$ and identification of edges. We also compare with PC, and LINGAM but only in terms of identification of DAG edges. There is a polynomial version of Zheng et al. [2018] and is implemented in R package `gnlearn`. However, it worked very poorly and thus omitted. Since the data has time dependence, we also tried to marginally decorrelate the data first and then apply algorithms like NO-TEARS, PC, or LINGAM. However, the performance got either worse or remained relatively the same in comparison to no adjustment. It may be due to the fact that the overall covariance of the data is not separable as different latent time series possess different covariance kernels. Hence, the presented results are based on the direct application of the alternative methods to the simulated data.

We also fit the NOTEARS method with adaptive LASSO penalty and employ Algorithm 1 replacing $\mathbf{W}^{(t)}\mathbf{Y}$ with $\mathbf{Y}$ in Step 3. Thresholding similar to step 4 is also applied to estimate the DAG. We call this A-NOTEARS, and it works better than the original NOTEARS with LASSO penalty. The rank-PC is fitted following [Harris and Drton, 2013].

4.1 Simulation setting 1

The latent univariate stationary series are generated from Gaussian Processes with exponential kernel. These Gaussian processes only differ in the range parameter, which is uniformly generated from (0, 10).

Table 1: Estimation MSE in estimating $\mathbf{W}$ when $\mathbf{z}_\ell$'s are generated following exponential covariance.

Time points	Expected neighbors = 2			Expected neighbors = 4		
	DAG-OUT	A-NOTEARS	LINGAM	DAG-OUT	A-NOTEARS	LINGAM
32	0.004	0.02	0.02	0.04	0.08	0.09
48	0.003	0.02	0.01	0.04	0.07	0.07

Table 2: MCC when $\mathbf{z}_\ell$'s are generated following Gaussian process with exponential kernel.

Time points	Expected neighbors = 2				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.91	0.59	0.62	0.36	0.29
48	0.92	0.61	0.59	0.41	0.39
	Expected neighbors = 4				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.61	0.46	0.41	0.28	0.22
48	0.62	0.48	0.45	0.34	0.30

4.2 Simulation setting 2

In this section, the exponential covariance is replaced by the covariance kernel $K(t_1, t_2) = \sum_{h=1}^{M} a_h \cos(h\pi|t_1 - t_2|)$, where $t_1, t_2 \in [0, 1]$ and set $a_h = 1/h^2$. For different univariate time series, M is sampled from $\{1, \ldots, T\}$ uniformly at random. Different values of M induce different degrees of smoothness.

Table 3: Estimation MSE in estimating $\mathbf{W}$ when $\mathbf{z}_\ell$'s are generated following cosine covariance kernel.

Time points	Expected neighbors = 2			Expected neighbors = 4		
	DAG-OUT	A-NOTEARS	LINGAM	DAG-OUT	A-NOTEARS	LINGAM
32	0.02	0.03	0.02	0.05	0.10	0.10
48	0.01	0.02	0.02	0.04	0.08	0.08

4.3 Simulation setting 3

The latent univariate stationary series are generated from ARMA(1,1) model as $z_{i,t} = \phi z_{i,t-1} + \theta\epsilon_{i,t-1} + \epsilon_{i,t}$ with $\epsilon_{i,t-1} \sim \text{Normal}(0, \sigma_e^2)$ with $\sigma_e^2 = \frac{1-\phi^2}{1+2\theta\phi+\theta^2}$. We generate $\theta, \phi \sim \text{Unif}((0.9, 1) \cup (-1, -0.9))$ in order to generate time series with strong dependence.

Table 4: MCC when $\mathbf{z}_\ell$'s are generated following Gaussian process with cosine covariance kernel.

Time points	Expected neighbors = 2				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.81	0.46	0.53	0.27	0.29
48	0.83	0.50	0.54	0.42	0.37
	Expected neighbors = 4				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.55	0.41	0.38	0.24	0.18
48	0.63	0.45	0.41	0.25	0.32

Table 5: Estimation MSE in estimating the precision matrix of dimension 40×40 when $\mathbf{z}_\ell$'s are generated following causal ARMA(1,1) models.

Time points	Expected neighbors = 2			Expected neighbors = 4		
	DAG-OUT	A-NOTEARS	LINGAM	DAG-OUT	A-NOTEARS	LINGAM
32	0.01	0.02	0.02	0.05	0.10	0.13
48	0.01	0.02	0.02	0.03	0.06	0.12

Table 6: MCC when $\mathbf{z}_\ell$'s are generated following causal ARMA(1,1) models.

Time points	Expected neighbors = 2				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.77	0.60	0.53	0.28	0.34
48	0.82	0.63	0.40	0.39	0.37
	Expected neighbors = 4				
	DAG-OUT	A-NOTEARS	LINGAM	PC	rank-PC
32	0.61	0.44	0.32	0.24	0.25
48	0.65	0.48	0.40	0.28	0.38

5 QWI data analysis

We study causal associations among different age and earning groups based on three earning-related variables, listed in Table 12. The data set is the Quarterly Workforce Indicators data published by US Census Bureau. We use the EarnS variable from Year 1993 to Year 1997 to form 6 earning/salary groups (as shown in Table 8), following the convention that group i on an average earns more than group $i-1$. Then we fit our proposed model on the data from 1998 to 2018. As a pre-processing step, we remove a linear trend and then mean-centered and normalized. Here $S = 3$ due to three responses. We take Age × Earning groupings, forming $8 \times 6 = 48$ combinations in total, which is our p. The estimated directed acyclic graph (DAG) provides causal associations among these 48 combinations of ages and salaries. We further study the resulting collapsed DAGs among the salary groups and age groups, separately.

Table 7: Variables considered as multi-task

	Variable	Explanation
1	EarnS	Average monthly earnings of employees with stable jobs
2	EarnHirAS	Average monthly earnings for workers who started a job that turned into a job lasting a full quarter
3	EarnHirNS	Average monthly earnings of newly stable employees

Table 8: Industries with their corresponding salary groups. Higher salary groups earn more

Salary group	Industry names
1	Agriculture, Forestry, Fishing and Hunting
6	Mining, Quarrying, and Oil and Gas Extraction
6	Utilities
4	Construction
4	Manufacturing
4	Wholesale Trade
1	Retail Trade
3	Transportation and Warehousing
5	Information
6	Finance and Insurance
3	Real Estate and Rental and Leasing
5	Professional, Scientific, and Technical Services
5	Management of Companies and Enterprises
2	Administrative and Support and Waste Management and Remediation Services
2	Educational Services
3	Health Care and Social Assistance
2	Arts, Entertainment, and Recreation
1	Accommodation and Food Services
1	Other Services (except Public Administration)

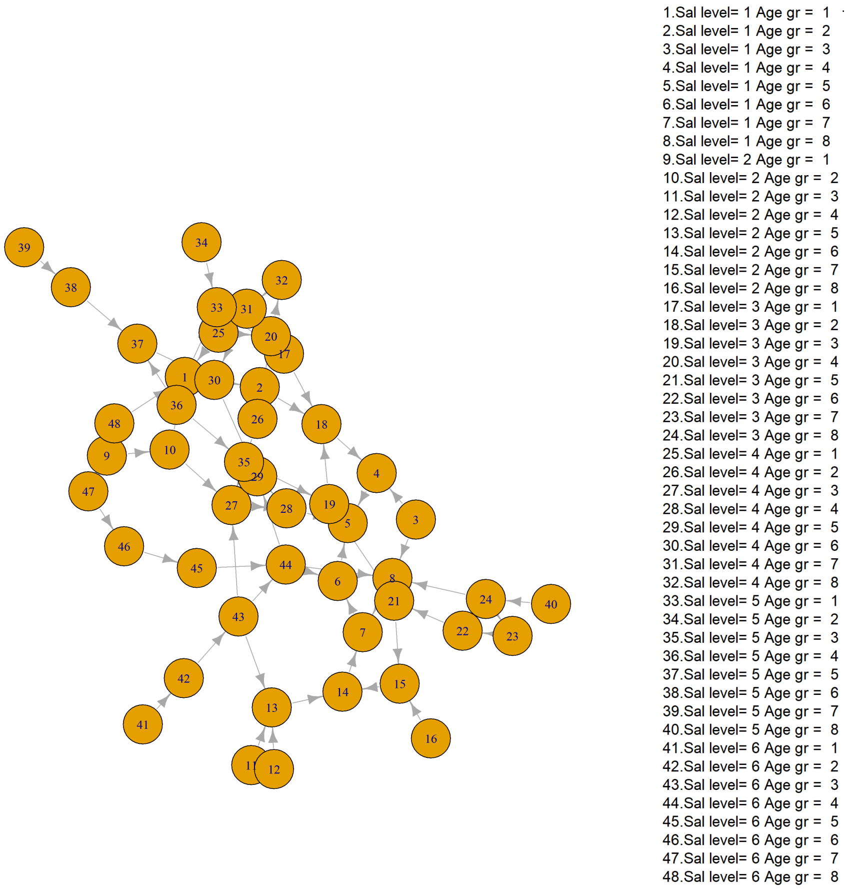
39
38
34
32
33
31
25
20
17
37
1
30
2
36
26
18
48
10
35
4
9
29
47
27
28
19
3
5
46
45
44
6
8
21
24
40
43
22
23
7
42
15
14
13
41
16
11
12
1.Sal level= 1 Age gr = 1
2.Sal level= 1 Age gr = 2
3.Sal level= 1 Age gr = 3
4.Sal level= 1 Age gr = 4
5.Sal level= 1 Age gr = 5
6.Sal level= 1 Age gr = 6
7.Sal level= 1 Age gr = 7
8.Sal level= 1 Age gr = 8
9.Sal level= 2 Age gr = 1
10.Sal level= 2 Age gr = 2
11.Sal level= 2 Age gr = 3
12.Sal level= 2 Age gr = 4
13.Sal level= 2 Age gr = 5
14.Sal level= 2 Age gr = 6
15.Sal level= 2 Age gr = 7
16.Sal level= 2 Age gr = 8
17.Sal level= 3 Age gr = 1
18.Sal level= 3 Age gr = 2
19.Sal level= 3 Age gr = 3
20.Sal level= 3 Age gr = 4
21.Sal level= 3 Age gr = 5
22.Sal level= 3 Age gr = 6
23.Sal level= 3 Age gr = 7
24.Sal level= 3 Age gr = 8
25.Sal level= 4 Age gr = 1
26.Sal level= 4 Age gr = 2
27.Sal level= 4 Age gr = 3
28.Sal level= 4 Age gr = 4
29.Sal level= 4 Age gr = 5
30.Sal level= 4 Age gr = 6
31.Sal level= 4 Age gr = 7
32.Sal level= 4 Age gr = 8
33.Sal level= 5 Age gr = 1
34.Sal level= 5 Age gr = 2
35.Sal level= 5 Age gr = 3
36.Sal level= 5 Age gr = 4
37.Sal level= 5 Age gr = 5
38.Sal level= 5 Age gr = 6
39.Sal level= 5 Age gr = 7
40.Sal level= 5 Age gr = 8
41.Sal level= 6 Age gr = 1
42.Sal level= 6 Age gr = 2
43.Sal level= 6 Age gr = 3
44.Sal level= 6 Age gr = 4
45.Sal level= 6 Age gr = 5
46.Sal level= 6 Age gr = 6
47.Sal level= 6 Age gr = 7
48.Sal level= 6 Age gr = 8

Figure 1: Estimated directed acyclic graphical connections among different variables with different salary levels and age groups with $H = 0.2$.

Table 9: Age groups

Age-groups	1	2	3	4	5	6	7	8
Range	14-18	19-21	22-24	25-34	35-44	45-54	55-64	65-99

Table 10: Structured Hamming distance between different pairs of salary groups setting $H = 0.2$

	Sal level 1	Sal level 2	Sal level 3	Sal level 4	Sal level 5	Sal level 6
Sal level 1	0	4	5	4	4	3
Sal level 2	4	0	5	6	4	5
Sal level 3	5	5	0	5	6	4
Sal level 4	4	6	5	0	5	1
Sal level 5	4	4	6	5	0	5
Sal level 6	3	5	4	1	5	0

Table 11: Structured Hamming distance between different pairs of age groups with $H = 0.2$

	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5	Age group 6	Age group 7	Age group 8
Age group 1	0	4	4	3	3	3	3	3
Age group 2	4	0	4	1	1	1	1	2
Age group 3	4	4	0	3	3	3	3	3
Age group 4	3	1	3	0	0	0	0	2
Age group 5	3	1	3	0	0	0	0	2
Age group 6	3	1	3	0	0	0	0	2
Age group 7	3	1	3	0	0	0	0	2
Age group 8	3	2	3	2	2	2	2	0

The maximum number of possible edges in the graph is 1128 resulting in a dense graph. However, in practice, the proposed model is expected to estimate a much sparser graph with a few meaningful causal connections. Figure 2 illustrates the estimated DAG. The estimated DAG has approximately 55 connections (or 62 in case $H = 0.2$) that are deemed significant. Here, we see that most of the edges or causal associations are confined within a fixed salary group. Within a given salary group, the edges would suggest casual associations across different age groups. Such associations are predominant. This suggests that causal associations based on the outcomes quantifying earning numbers are primarily present within a given salary groups or between two adjacent salary groups. The number of within-age edges is very limited.

For a detailed inference, Table 10 shows the structured hamming distances between different salary groups based on the summarized number of edges. It is defined as the minimum number of single operations, such as deletions, insertions, and re-orientations, needed to transform one DAG into another DAG. We compute these distances using the `shd` function from the package `pcalg` [Alain Hauser and Peter Bühlmann, 2012]. The Structured Hamming distance (SHD) provides a measure of changes in the causal associations among the age groups as we move along the different salary groups. Table 11 shows the same for the age groups, measuring how the causal connections between the salary groups change from one age group to another.

The SHDs in Table 10 are larger in comparison to those in Table 11, suggesting causal associations among different salary groups stay relatively stable across different ages. The SHDs in Table 11 are often 0, specifically for higher age groups, meaning no differences in the estimated causal associations among the salary groups as they move from one age group to the other. Specifically, age groups 4 to 7 exhibit identical causal associations among the salary groups.

5.1 Employment-Quantity (EarnS-EmpTotal)

Labor economists are interested in studying the joint behavior or earnings and employment totals. We used the bivariate quarterly series of total employed in stable jobs and average monthly earnings from stable jobs. We extended the study of causal relations concerning earnings to the study of the causal structures of different age and salary groups with respect to earnings and total employment. Figure 2 shows the estimated graph of the 48 different age and salary group combinations. Interestingly, several path-like structures in the graph indicate a smooth monotone change of the causal structure within some age/salary groups. For example, in salary levels 5 and 6, the causal structure within the middle age groups (3-6) is a directed path.

The SHD analysis of the nested groups of salary and age for the earnings and employment total together reveals a story similar to that for earnings-only analysis. Notably, the causal structure among the salary groups remains stable across the middle age groups, ages ranging from 25-55. However, in the extreme age groups, below 20 years old or above 65 years old, the causal structure among the salary groups can shift significantly from that for the middle age groups.

Table 12: Variables considered as multi-task

	Variable	Explanation
1	EmpS	Estimate of stable jobs - the number of jobs that are held on both the first and last day of the quarter with the same employer
2	EarnS	Average monthly earnings of employees with stable jobs

Table 13: Structured Hamming distance between different pairs of salary groups setting $H = 0.35$ with Employment-Quantity

	Sal level 1	Sal level 2	Sal level 3	Sal level 4	Sal level 5	Sal level 6
Sal level 1	0	7	4	5	7	6
Sal level 2	7	0	6	7	4	8
Sal level 3	4	6	0	6	7	5
Sal level 4	5	7	6	0	7	8
Sal level 5	7	4	7	7	0	6
Sal level 6	6	8	5	8	6	0

Table 14: Structured Hamming distance between different pairs of age groups with $H = 0.35$ with Employment-Quantity

	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5	Age group 6	Age group 7	Age group 8
Age group 1	0	6	7	6	7	6	8	8
Age group 2	6	0	5	5	6	6	5	6
Age group 3	7	5	0	1	2	2	3	2
Age group 4	6	5	1	0	1	3	2	3
Age group 5	7	6	2	1	0	4	3	4
Age group 6	6	6	2	3	4	0	5	4
Age group 7	8	5	3	2	3	5	0	5
Age group 8	8	6	2	3	4	4	5	0

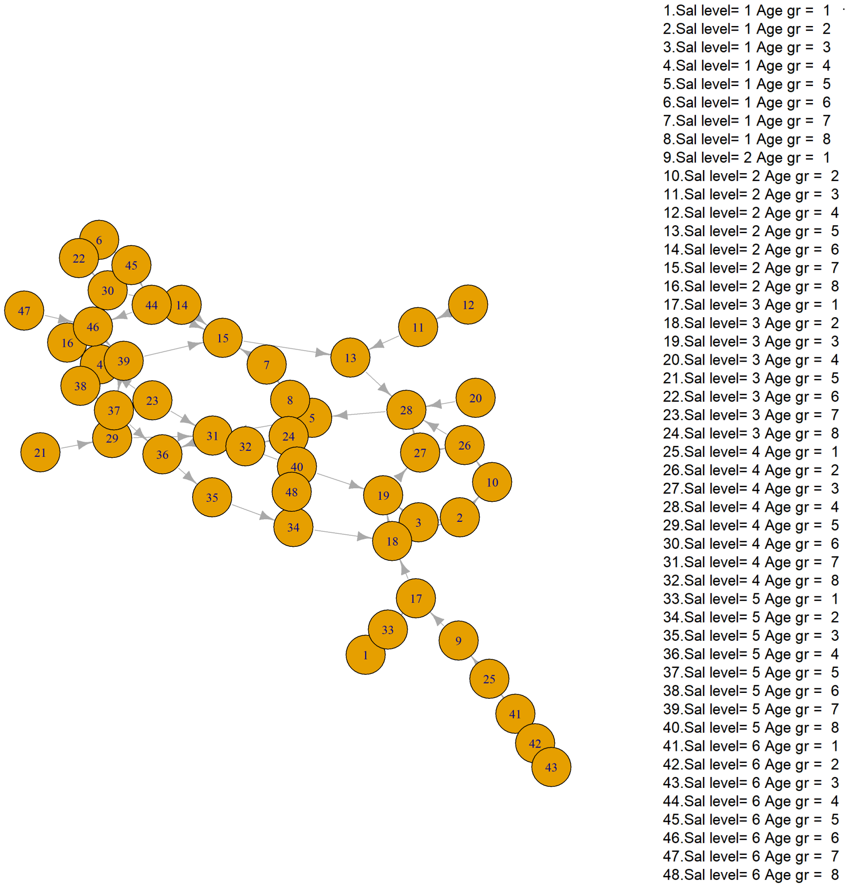
6
22
45
30
44
14
47
46
16
4
39
15
7
13
11
12
38
23
37
8
5
28
20
29
21
31
32
24
36
26
27
40
10
35
48
19
3
2
34
18
17
33
9
1
25
41
42
43
1.Sal level= 1 Age gr = 1
2.Sal level= 1 Age gr = 2
3.Sal level= 1 Age gr = 3
4.Sal level= 1 Age gr = 4
5.Sal level= 1 Age gr = 5
6.Sal level= 1 Age gr = 6
7.Sal level= 1 Age gr = 7
8.Sal level= 1 Age gr = 8
9.Sal level= 2 Age gr = 1
10.Sal level= 2 Age gr = 2
11.Sal level= 2 Age gr = 3
12.Sal level= 2 Age gr = 4
13.Sal level= 2 Age gr = 5
14.Sal level= 2 Age gr = 6
15.Sal level= 2 Age gr = 7
16.Sal level= 2 Age gr = 8
17.Sal level= 3 Age gr = 1
18.Sal level= 3 Age gr = 2
19.Sal level= 3 Age gr = 3
20.Sal level= 3 Age gr = 4
21.Sal level= 3 Age gr = 5
22.Sal level= 3 Age gr = 6
23.Sal level= 3 Age gr = 7
24.Sal level= 3 Age gr = 8
25.Sal level= 4 Age gr = 1
26.Sal level= 4 Age gr = 2
27.Sal level= 4 Age gr = 3
28.Sal level= 4 Age gr = 4
29.Sal level= 4 Age gr = 5
30.Sal level= 4 Age gr = 6
31.Sal level= 4 Age gr = 7
32.Sal level= 4 Age gr = 8
33.Sal level= 5 Age gr = 1
34.Sal level= 5 Age gr = 2
35.Sal level= 5 Age gr = 3
36.Sal level= 5 Age gr = 4
37.Sal level= 5 Age gr = 5
38.Sal level= 5 Age gr = 6
39.Sal level= 5 Age gr = 7
40.Sal level= 5 Age gr = 8
41.Sal level= 6 Age gr = 1
42.Sal level= 6 Age gr = 2
43.Sal level= 6 Age gr = 3
44.Sal level= 6 Age gr = 4
45.Sal level= 6 Age gr = 5
46.Sal level= 6 Age gr = 6
47.Sal level= 6 Age gr = 7
48.Sal level= 6 Age gr = 8

Figure 2: Estimated directed acyclic graphical connections among different variables with different salary levels and age groups with $H = 0.35$ with Employment-Quantity.

Funding

The authors would like to thank the National Science Foundation collaboative research grants DMS-2210280 (Subhashis Ghosal) / 2210281 (Anindya Roy) / 2210282 (Arkaprava Roy).

References

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Davide Altomare, Guido Consonni, and Luca La Rocca. Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics*, 69(2): 478–487, 2013.

Ronald A Babula, David A Bessler, and Warren S Payne. Dynamic relationships among us wheat-related markets: Applying directed acyclic graphs to a time series model. *Journal of Agricultural and Applied Economics*, 36(1):1–22, 2004.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Emanuel Ben-David, Tianxi Li, Helene Massam, and Bala Rajaratnam. High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv preprint arXiv:1109.4371*, 2011.

J Bernardo, J Burger, and ADEM Smith. Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. *Bayesian statistics*, 6, 1998.

Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, pages 2526–2556, 2014.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Raj Chhikara. *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, volume 95. CRC Press, 1988.

Rainer Dahlhaus and Michael Eichler. Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, pages 115–137, 2003.

Wanlu Deng, Zhi Geng, and Hongzhe Li. Learning local directed acyclic graphs based on multivariate time series data. *The Annals of Applied Statistics*, 7(3):1249, 2013.

Nir Friedman and Daphne Koller. Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50:95–125, 2003.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

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

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179, 2020.

Qiang Ji, Elie Bouri, Rangan Gupta, and David Roubaud. Network causality structures among bitcoin and other financial assets: A directed acyclic graph approach. *The Quarterly Review of Economics and Finance*, 70:203–213, 2018.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques.* MIT Press, 2009.

Jack Kuipers and Giusi Moffa. Partition mcmc for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

S.L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

Kuang-Yao Lee and Lexin Li. Functional structural equation model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):600–629, 2022.

Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

Yang Ni, Francesco C Stingo, and Veerabhadran Baladandayuthapani. Bayesian nonlinear model selection for gene regulatory networks. *Biometrics*, 71(3):585–595, 2015.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. Pmlr, 2020.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19 (2):3, 2000.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2): 1545–1558, 2010.

Arkaprava Roy, Anindya Roy, and Subhashis Ghosal. Bayesian inference for high-dimensional time series by latent process modeling. *arXiv preprint arXiv:2403.04915*, 2024.

Wenli Shi, Subhashis Ghosal, and Ryan Martin. Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error. *Electronic Journal of Statistics*, 15(2):4545–4579, 2021.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Qifan Song and Faming Liang. Nearly optimal Bayesian shrinkage for high-dimensional regression. *Science China Mathematics*, 66(2):409–442, 2023.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Danru Xu, Erdun Gao, Wei Huang, Menghan Wang, Andy Song, and Mingming Gong. On the sparse dag structure learning based on adaptive lasso. *arXiv preprint arXiv:2209.02946*, 2022.

Tianle Yang and Joe Suzuki. The functional lingam. In *International Conference on Probabilistic Graphical Models*, pages 25–36. PMLR, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

Saijuan Zhang, Douglas Midthune, Patricia M Guenther, Susan M Krebs-Smith, Victor Kipnis, Kevin W Dodd, Dennis W Buckman, Janet A Tooze, Laurence Freedman, and Raymond J Carroll. A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics*, 5:1456–1487, 2011.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Fangting Zhou, Kejun He, Kunbo Wang, Yanxun Xu, and Yang Ni. Functional Bayesian networks for discovering causality from multivariate functional data. *Biometrics*, 79(4):3279–3293, 2023.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Yi Zuo and Eisuke Kita. Stock price forecast using Bayesian network. *Expert Systems with Applications*, 39(8):6729–6737, 2012.