Structure Learning for Groups of Variables in Nonlinear Time-Series Data with Location-Scale Noise

Genta Kikuchi GENTA.KIKUCHI.J8G@JP.DENSO.COM
DENSO CORPORATION, Kariya, Aichi 448-8661, Japan
Graduate School of Data Science, Shiga University, Japan

Shohei Shimizu* SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Graduate School of Data Science, Shiga University, Japan
Center for Advanced Intelligence Project, RIKEN, Japan

Abstract

Learning causal structures from observational data has recently attracted considerable attention. Although many studies have focused on uncovering the connections between scalar random variables, estimation algorithms for groups of variables—particularly for multiple groups of variables—remain scarce. This paper proposes a novel differentiable algebraic constraint that can be used along with existing continuous optimization-based structure-learning algorithms to learn the causal relationships among groups of variables. Considering the complex functional relationships among variables in real-world scenarios, we propose a structure-learning algorithm for nonlinear time-series data with location-scale noise. Experimental results for synthetic and real-world data indicate that the proposed group acyclicity constraint significantly increases the estimation accuracy for the causal relationship among the groups of variables and verify the effectiveness of the proposed structure-learning algorithm.

Keywords: Causal discovery, structural causal models, time series, variable groups, continuous optimization.

1. Introduction

Numerous methods have been developed for estimating causal relationships using observational data (Spirtes et al., 2000; Pearl, 2009; Shimizu et al., 2006; Peters et al., 2014). Although the typical problem involves inferring the relationships between individual random scalar variables, there are many cases in which the relationships among groups of variables are of interest. For example, in neuroscience, researchers have focused on the relationships between brain regions (Smith et al., 2011). In the manufacturing domain, variables with relatively strong correlations were observed in multiple measurements obtained from the same machine. The analysis is typically performed by selecting one variable per group (Marazopoulou et al., 2016) or by calculating, for example, the sum of the variables in the same group (Scheines and Spirtes, 2008). These approaches can significantly reduce the computation time by reducing the number of variables, but they generally degrade the performance of causal discovery methods. This is caused by a change in the conditional dependencies between variables (Scheines and Spirtes, 2008) or the cancellation of dependence (Wahl et al., 2022).

* This work was partially supported by ONR N00014-20-1-2501.

© 2023 G. Kikuchi & S. Shimizu.

Several studies have been performed on structural learning among groups of variables. Parviainen and Kaski (2017) proposed an estimation method in which a directed acyclic graph is constructed over the individual variables and used to infer connections between the groups. This approach uses conditional-independence-based structure-learning methods; thus, causal relationships cannot be distinguished using the same set of conditional independencies. In the research regarding on functional causal models, Kawahara et al. (2010) proposed GroupLiNGAM—an estimation method for discovering the grouping of variables and the causal relationship among groups of variables. Subsequently, Entner and Hoyer (2012) proposed GroupDirectLiNGAM, which is an efficient estimation method when the assignment of groups for each variable is known. These methods have the advantage that unmeasured confounders can exist in each group; however, they assume that the functional relationship is linear. In the method of Khemakhem et al. (2021), normalizing flows are used to capture the nonlinearity of the data; however, the method is limited to infer a causal direction between two groups and is not straightforward to apply to three or more groups. To the best of our knowledge, no estimation method exists for the class of functional causal models capable of both multiple groups and beyond linear functional relationships.

Besides the problem of inferring the structure among multiple groups, from the complexity of data observed in real-world, we need to consider temporal dependencies as well as nonlinear functional relationships and the heteroscedasticity induced by the modulated variance of the noise (location-scale noise). Existing works (Hyvärinen et al., 2010; Peters et al., 2013; Immer et al., 2022) partially model these characteristics, while Gong et al. (2022) proposed Rhino, a functional causal model that addresses all the characteristics together. However, Rhino does not consider the location-scale noise on the instantaneous effects that are often observed in practice (Tagasovska et al., 2020).

In this paper, we propose a structure-learning method for groups of variables based on observational data. Assuming that there is no closed loop in the relationship between the groups, we derive novel differentiable algebraic constraints that characterize the causal structure among the groups of variables. This is a natural extension of the well-known algebraic constraint that characterizes acyclicity over the relationships of individual variables (Zheng et al., 2018); therefore, the proposed constraint can be applied to continuous optimization-based structure-learning methods that use continuous optimization. We further propose a functional causal model and corresponding estimation method to capture the nonlinear functional relationships as well as the temporal dependencies, and heteroscedasticity induced by location-scale noise. The remainder of this paper is organized as follows. In Section 2, we introduce the problem of structure learning among groups of variables and explain the existing algebraic constraints and structure-learning methods. In Section 3, we describe the proposed algebraic constraint and estimation method for nonlinear time-series data with location-scale noise in Section 3. The results for synthetic data as well as real-world data are presented in Section 4. Finally, Section 5 concludes the paper.

2. Background

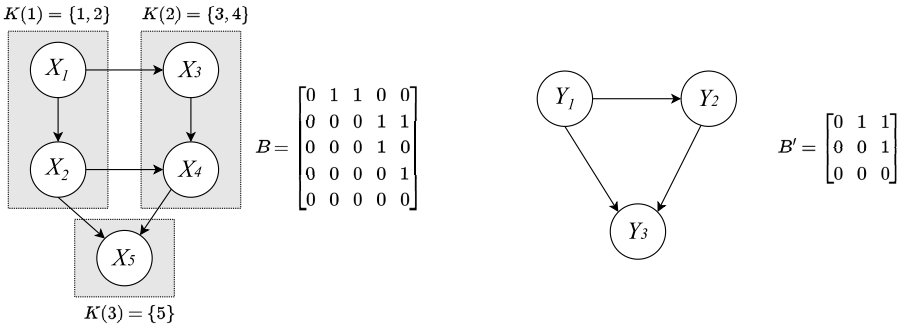
2.1. Problem Definition

Consider a set of P variables $X = \{X_1, ..., X_P\}$ and assume that the data-generating process of X can be represented by a directed graph $\mathcal{G}$, which induces a joint distribution $\mathcal{L}(X)$

over X. $\mathcal{G}$ is parameterized by the adjacency matrix $B \in \{0,1\}^{P\times P}$, where $[B]_{i,j} = 1$ if and only if a direct connection from X_i to X_j exists.

We assume that each variable belongs to one of $M(M \leq P)$ groups and that the assigned group of each variable is known. Let $K = \{K(1), ..., K(M)\}$ be a set of index sets for each group. Let $Y = \{Y_1, ..., Y_M\}$ be a supervertex obtained by contracting the variables belonging to the same group on $\mathcal{G}$. In this paper, we refer to a graph on X as a *variable graph* and a graph on Y as a *group graph*, and we call the corresponding adjacency matrices $B \in \{0,1\}^{P\times P}$ and $B^{'} \in \{0,1\}^{M\times M}$ *variable adjacency matrices* and *group adjacency matrices* , respectively. $B^{'}$ encapsulates the connections between the groups, where $[B^{'}]_{k,l} = 1$ if and only if $\exists [B]_{i\in K(k),\ j\in K(l)} = 1$. We further assume that the group graph of $\mathcal{G}$ is a directed acyclic graph (DAG), where we call $\mathcal{G}$ *group-acyclic* given the grouping K.

The goal is to estimate B' from $\mathcal{L}(X)$, which we call the corresponding graph, *group DAG*. Many existing structure-learning methods perform estimation under $M = P$, i.e. the number of groups is equal to that of variables; we call this the corresponding graph, *variable DAG*. An example of a variable DAG with grouping K and the corresponding group DAG is shown in Figure 1. If $M < P$ and f_j are linear and N_j represents additive non-Gaussian noises that are independent of each other over the groups, this problem is equivalent to that of Entner and Hoyer (2012).



(*a*) Variable DAG (*b*) Group DAG

Figure 1: Example of a variable DAG and a group DAG

2.2. Algebraic Characterization of DAGs and NOTEARS Method

A recent breakthrough in the structure learning of variable DAGs is the algebraic characterization of DAGs using the trace of a matrix exponential (Zheng et al., 2018). Zheng et al. show that a (weighted) adjacency matrix $W \in \mathbb{R}^{P\times P}$ represents a DAG if and only if the following equation holds:

$$h(W) = \mathrm{tr}\left(e^{W\circ W}\right) - P = 0, \tag{1}$$

where $\circ$ denotes the elementwise product. Using Equation (1) as a constraint, Zheng et al. formulated a structure-learning problem as a continuous optimization problem and proposed an estimation algorithm for linear data (NOTEARS). Let $\mathbf{X} \in \mathbb{R}^{N\times P}$ be a dataset

comprising N independent and identically distributed observations. NOTEARS solves the following constrained optimization problem:

$$\min_{W} \frac{1}{2N}\|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda\|W\|_1 \quad \text{subject to} \quad h(W) = 0, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_1 = \|\text{vec}(\cdot)\|_1$ is the vector L^1-norm , and λ denotes the penalty factor. The constrained optimization problem is converted into an unconstrained optimization problem using the augmented Lagrangian method, followed by optimization using L-BFGS (Byrd et al., 1995).

2.3. Variants of NOTEARS Method

The optimization problem (2) for linear DAGs can be generalized to nonlinear relationships (Zheng et al., 2020) as follows:

$$\min_{\theta} F(\mathbf{X}, \theta) + \|\theta\|_1 \quad \text{subject to} \quad h(W(\theta)) = 0, \tag{3}$$

where $F(\mathbf{X}, \theta)$ is the loss function, θ is the model parameter, and $W(\theta)$ is the weighted adjacency matrix calculated from the model parameter. The choice of $F(\mathbf{X}, \theta)$ and the calculation of $W(\theta)$ depend on the model. Zheng et al. (2020) proposed NOTEARS-MLP, which leverages multilayer perceptrons (MLPs) to capture nonlinear relationships. $W(\theta)$ for NOTEARS-MLP was calculated using the weights of the first layers of the MLPs. Sun et al. (2021) proposed NTS-NOTEARS, which extended NOTEARS-MLP to exploit temporal and instantaneous dependencies by using convolutional neural networks (CNN). Similarly, $W(\theta)$ for NTS-NOTEARS is calculated using the kernel weights of the CNNs. Kikuchi (2023) proposed a continuous optimization-based estimation algorithm for the location-scale noise model (LSNM) (Immer et al., 2022), modeling the conditional expectation as well as the conditional variance of each variable with an MLP. They reported that the scale sensitivity of NOTEARS (Reisach et al., 2021) is mitigated by estimating the conditional variance and utilizing the log-likelihood for the score function.

The key parts of the above methods are the design of the data modeling and the usage of the algebraic constraint. In Section 3.1, we derive an algebraic constraint for estimating group DAGs.

3. Proposed Method

We introduce an algebraic constraint for characterizing group DAGs in Section 3.1. In Section 3.2, we propose a functional causal model using time-series data with location-scale noise.

3.1. Group DAG Constraint

We extend the variable adjacency matrix B and group adjacency B' given in Section 2.1 to the weighted adjacency matrix. Suppose that we have a weighted adjacency matrix $W \in \mathbb{R}^{P\times P}$ that represents the connection strength between individual variables. The

corresponding weighted group adjacency matrix $W^{'} \in \mathbb{R}^{M \times M} \geq 0$ is calculated as follows:

$$[W^{'}]_{k,l} = \begin{cases} 0 & \text{if } k = l, \\ \sum_{i \in K(k)} \sum_{j \in K(l)} [W \circ W]_{i,j} & \text{else} \end{cases}. \tag{4}$$

Here, $[W^{'}]_{k,l}$ denotes the total amount of squared connection strengths from the variables in groups k to the variables in group l. We calculate the squared values to avoid cancellation of the dependencies, because the values of W can be either positive or negative. We do not claim that this calculation is optimal, we can also use the absolute values of W to calculate $W^{'}$. The diagonal elements of $W^{'}$ are set to zero not to restrict the connections within the same group. An example of a calculation using Equation (4) is given in Appendix A.

By substituting $W^{'}$ into the algebraic constraint h (1), we obtain a constraint for the group DAGs, which we call the *group DAG constraint*:

$$h(W^{'}) = \mathrm{tr}(e^{W^{'} \circ W^{'}}) - M = 0. \tag{5}$$

Clearly, the group DAG constraint is satisfied if and only if the corresponding variable graph is group-acyclic.

Corollary 1 (Acyclicity of group DAGs) *A variable graph $\mathcal{G}$ with grouping K represented by a weighted adjacency matrix W is group-acyclic if and only if the constraint in Equation (5) is satisfied.*

Proof From Equation (4), we can see that $[W^{'}]_{k,l} > 0$ if and only if $\exists [W]_{i \in K(k), j \in K(l)} \neq 0$; thus, $W^{'}$ represents the weighted adjacency matrix of the group graph of $\mathcal{G}$. From the results in (Zheng et al., 2018), $\mathrm{tr}(e^{W^{'} \circ W^{'}}) - M = 0$ is satisfied if and only if the group graph of $\mathcal{G}$ is acyclic, which implies that $\mathcal{G}$ is group-acyclic. ■

By replacing the group DAG constraint with the algebraic constraint (1), we can estimate the structure among the groups of variables by using an existing method that uses the existing algebraic DAG constraint. The group DAG constraint (5) requires the graph to be group-acyclic but does not constrain the acyclicity within each group. As described in Appendix B.1, we conducted a numerical experiment for a case with a closed loop in each group. The estimation accuracy was similar to that for a case without closed loops.

3.2. TS-LSNM: Structure Learning for Nonlinear Time-Series Data with Location-Scale Noise

3.2.1. Model Formulation

We propose a time-series location-scale noise model (TS-LSNM)—a functional causal model that extends the LSNM (Immer et al., 2022) to capture temporal dependencies. Suppose that we have a set of P time series $X^t = \{X_1^t, X_2^t, ..., X_P^t\}$ with joint distribution $\mathcal{L}(X^t)$. Let $\mathrm{PA}_j^t \subseteq X^t \backslash X_j^t$ denote a set of instantaneous parents of X_j^t at time t and $\mathrm{PA}_j^{t-\tau} \subseteq X^{t-\tau}$ denote a set of lagged parents, which is a set of variables with a direct connection from the previous timestep $X_i^{t-\tau}$ to X_j^t. TS-LSNM is defined as follows:

$$X_j^t = f_j\left(\mathrm{PA}_j^t, ..., \mathrm{PA}_j^{t-L}\right) + s_j\left(\mathrm{PA}_j^t, ..., \mathrm{PA}_j^{t-L}\right) N_j^t, \quad j = 1, ..., P \tag{6}$$

where $L \geq 0$ represents the maximum time lag of the model and N_j^t is a noise term that is mutually independent over j and t, which implies that there are no latent confounders. N_j^t is identically distributed in t. f_j and the scaling function $s_j > 0$ are twice differentiable functions, where s_j induces heteroscedasticity by scaling the variance of the noise terms. We further assume causal stationarity (Runge, 2018), causal minimality (Hoyer et al., 2008), and joint distribution of X^t satisfy the causal Markov property with respect to the underlying graph. If $L = 0$, TS-LSNM is reduced to the LSNM (Immer et al., 2022).

The structural identifiability of TS-LSNM comes from the identifiability results of the LSNM and the time-series model with independent noise (TiMINo) (Strobl and Lasko, 2022; Immer et al., 2022; Peters et al., 2012). LSNM belongs to an identifiable functional model class (IFMOC) (Peters et al., 2012), therefore is structurally identifiable under the assumption of causal minimality, no cycles, and no hidden confounders. TS-LSNM is a special case of TiMINo. Because we can recover the underlying graph of TiMINo if the data-generating functions come from the IFMOC (Peters et al. (2013), Theorem 1 (i)), TS-LSNM is also structurally identifiable from the joint distribution $\mathcal{L}(X^t)$.

3.2.2. Estimation of TS-LSNM

We adopt a negative log-likelihood for the loss function to estimate TS-LSNM. We consider TS-LSNM (6) and set $\tilde{\mathrm{PA}}_j = \cup_{\tau=0}^{L} \mathrm{PA}_j^{t-\tau}$ and $\tilde{N}_j^t := s_j(\tilde{\mathrm{PA}}_j) N_j^t$. Given N observations, i.e., $\mathbf{X}^t = \{\mathbf{x}^{t,(n)}\}_{n=1}^N$, the log-likelihood of TS-LSNM is defined as follows:

$$\log \mathcal{L}(\mathbf{X}^t) = -\frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \sigma_j^{t,(n)} + \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \tilde{p}_j \left(\frac{\mathrm{x}_j^{t,(n)} - f_j(\tilde{\mathrm{PA}}_j^{(n)})}{\sigma_j^{t,(n)}} \right), \quad (7)$$

where $\tilde{p}_j$ denotes the probability density functions of noise $\tilde{N}_j^t$ standardized to unit variance, and $(\sigma_j^{t,(n)})^2$ represents the conditional variance of the *n-th* sample before standardization.

To model TS-LSNM, we leverage CNNs to capture temporal dependencies (Sun et al., 2021) and model each variable's conditional expectation and variance separately, as presented in (Kikuchi, 2023). Therefore, we create two CNNs to estimate $f_j(\tilde{\mathrm{PA}}_j^{(n)})$ and $\sigma_j^{t,(n)}$ for each target variable, resulting in $2P$ CNNs. The first layer of each CNN is a convolutional layer with kernel size S, a stride of 1, and no padding, where the parameters are expressed as ϕ, which is a set of weight matrices of shape $P \times (L+1)$. The corresponding weights with respect to the target variable of the instantaneous step $\tau = 0$ are set to zero to avoid estimating X_j^t using its own value. The remaining layers are fully connected layers with parameter ψ, which is a set of weight matrices. The estimations $\hat{f}_j(\tilde{\mathrm{PA}}_j^{(n)})$ and $\hat{\sigma}_j^{t,(n)}$ are given by CNNs with parameters $\theta_j^A = (\phi_j^A, \psi_j^A)$ and $\theta_j^C = (\phi_j^C, \psi_j^C)$, respectively:

$$\hat{f}_j(\tilde{\mathrm{PA}}_j^{(n)}) = \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^A), \quad (8)$$

$$\hat{\sigma}_j^{t,(n)} = \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C). \quad (9)$$

Following (Sun et al., 2021), the weighted adjacency matrix for TS-LSNM is calculated using the weights of the convolutional layers of CNNs. Let $\phi_j^A(\tau)$ and $\phi_j^C(\tau)$ denote the collection of the τ-*th* column of the S convolutional weights. For example, the *i-th* element

of each $\phi_j^A(L-\tau)$ represents the connection strength with respect to the conditional expectation from $X_i^{t-\tau}$ to X_j^t. We calculate two weighted adjacency matrices $W^\tau(\theta^A)$ and $W^\tau(\theta^C)$ for each time lag τ, representing the overall connection strengths with respect to the conditional expectations and the variances, respectively. $W^\tau(\theta^A)$ and $W^\tau(\theta^C)$ are calculated as follows:

$$[W^\tau(\theta^A)]_{i,j} = \|\text{i-th element across all } \phi_j^A(\tau)\|_2, \tag{10}$$

$$[W^\tau(\theta^C)]_{i,j} = \|\text{i-th element across all } \phi_j^C(\tau)\|_2. \tag{11}$$

To obtain a weighted adjacency matrix $W^\tau(\theta^A, \theta^C)$ that represents the connection strengths of both the conditional expectation and the variance, we calculate $W^\tau(\theta^A, \theta^C) = W^\tau(\theta^A) + W^\tau(\theta^C)$. Element i, j of $W^\tau(\theta^A, \theta^C)$ indicates the overall connection strength from $X_i^{t-\tau}$ to X_j^t. $W^0(\theta^A, \theta^C)$ represents the dependency structure of the current timestep t, and $W^\tau(\theta^A, \theta^C)$ $(0 < \tau <= L)$ represent the time-lagged dependencies from time $t-\tau$ to t. Because we only constrain the connections of the instantaneous step to be acyclic, we use $W^0(\theta^A, \theta^C)$ as the input of the algebraic constraint. Finally, using the log-likelihood (7) with constraint (1) and regularization terms with respect to the model weights $\theta^A = (\theta_1^A, ..., \theta_P^A)$ and $\theta^C = (\theta_1^C, ..., \theta_P^C)$, we obtain the following constrained optimization problem for TS-LSNM:

$$\min_{\theta^A, \theta^C} F(\mathbf{X}^t, \theta^A, \theta^C) \quad \text{subject to} \quad h(W^0(\theta^A, \theta^C)) = 0, \tag{12}$$

where

$$\begin{aligned} F(\mathbf{X}^t, \theta^A, \theta^C) = & \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C) \\ & - \frac{1}{N-L} \sum_{n=L+1}^{N} \sum_{j=1}^{P} \log \tilde{p}_j \left(\frac{\mathrm{x}_j^{t:t-L,(n)} - \mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^A)}{\mathrm{CNN}(\mathbf{x}^{t:t-L,(n)}; \theta_j^C)} \right) \\ & + \sum_{j=1}^{P} \left(\lambda_1 \|\phi_j^A\|_1 + \lambda_1 \|\phi_j^C\|_1 + \frac{1}{2}\lambda_2 \|\theta_j^A\|_2 + \frac{1}{2}\lambda_2 \|\theta_j^C\|_2 \right). \end{aligned}$$

Here, λ_1 and λ_2 are regularization parameters. Following Kikuchi (2023), we leverage the approximation of log probability $\log \tilde{p}_j$. We choose the approximated function from the two candidates according to whether the variable is super-Gaussian or sub-Gaussian (Hyvärinen et al., 2001; Hyvärinen and Oja, 1998) during the optimization. Although the approximation assumes that the noise terms follow a non-Gaussian distribution and that the probability density functions are symmetric, the results of the numerical experiments presented in Section 4.1.3 indicate that a relatively high estimation accuracy can be obtained even when the noise terms follow a Gaussian or Gumbel distribution.

We convert the constrained optimization problem (12) into an unconstrained optimization problem using the augmented Lagrangian method and employ L-BFGS (Byrd et al., 1995) for optimization. After optimization, we estimate the weighted adjacency matrix as $\tilde{W^\tau}(\theta_A, \theta_C) = 1/2(W^\tau(\theta_A) + W^\tau(\theta_C))$ and round the small values to zero with a small threshold $w > 0$ to remove redundant edges and the remaining cycles in the graph (Zhou,

2009). Note that we can estimate group DAGs by TS-LSNM using the DAG constraint (5), where we calculate $W'^0(\theta_A, \theta_C)$ using Equation (4) and replace the constraint on Equation (12) to $h(W'^0(\theta_A, \theta_C)) = 0$.

4. Numerical Experiments

We performed numerical experiments to evaluate our method on synthetic data as well as real-world data. The results of the numerical experiments using synthetic data are presented in Section 4.1. Next, we compare the results of different models for real-world data collected from a manufacturing process in Section 4.2.

4.1. Synthetic Data

4.1.1. Setup

We generated synthetic data with the grouped variables as follows. Given the number of variables P and the number of groups M, we randomly assigned group $l \in \{1, ..., M\}$ to each variable such that the groups had an equal number of variables. We then randomly selected a parent variable for each group and generated an intragroup DAG with a tree structure having a depth of 1. This operation simulated the observation that variables in the same group had similar values. Subsequently, starting from the first group $k = 1$, we assigned a connection from the variables in subsequent groups l ($l > k$) to the variables in group k with a probability of 0.1, where at least one connection from group l to group k was established. Thus, we obtained an adjacency matrix $B \in \{0, 1\}^{P \times P}$ representing a group-acyclic graph. An example of a simulated variable DAG and the corresponding group DAG for $P = 12$ and $M = 3$ are shown in Figure 2. For the time-lagged effects, following Sun et al. (2021), we created a connection from $X_i^{t-\tau}$ to X_j^t with a probability of $1/P$, which indicates that on average, there was one connection from each $X_i^{t-\tau}$ to X_j^t.

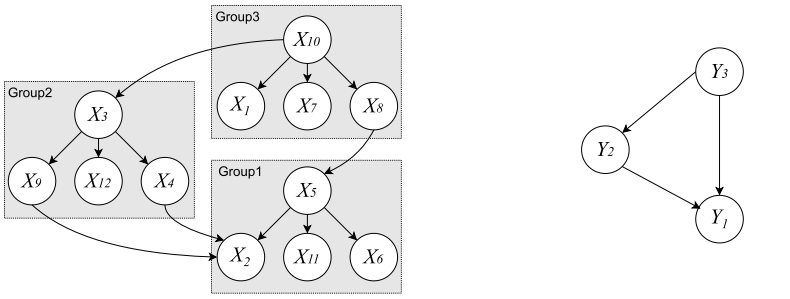
After generating the connections between the variables, each variable was generated using the following function based on the index models:

$$X_j^t = \tanh\left(f_j^{(1)}(\tilde{\mathrm{PA}}_j)\right) + \cos\left(f_j^{(2)}(\tilde{\mathrm{PA}}_j)\right) + \sin\left(f_j^{(3)}(\tilde{\mathrm{PA}}_j)\right) + s_j\left(\tilde{\mathrm{PA}}_j\right) N_j^t, \tag{13}$$

where $f_j^{(1)} = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [W_\tau^{(1)}]_{i,j} X_i^{t-\tau}$, $f_j^{(2)} = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [W_\tau^{(2)}]_{i,j} X_i^{t-\tau}$ and $f_j^{(3)} = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [W_\tau^{(3)}]_{i,j} X_i^{t-\tau}$. $pa^\tau(j)$ denotes the index set of $\mathrm{PA}_j^{t-\tau}$. For the scaling function s_j, for each j, we randomly selected a strictly positive nonlinear function from a set $\{1/(1 + \exp g(\tilde{\mathrm{PA}}_j)) + 0.5, \exp(g(\tilde{\mathrm{PA}}_j)), \tanh(g(\tilde{\mathrm{PA}}_j)) + 1.5\}$, where $g(\tilde{\mathrm{PA}}_j)) = \sum_{\tau=0}^{L} \sum_{i \in \mathrm{pa}^\tau(j)} [C_\tau]_{i,j} X_i^{t-\tau}$. The connection weights $W_\tau^{(1)}, W_\tau^{(2)}, W_\tau^{(3)}$ were sampled from $\pm U(0.5, 2.0)$, and C_τ was sampled from $\pm U(0.4, 0.8)$. Unless otherwise stated, the noise terms N_j^t were generated from the standard Gaussian distribution. We generated 2000 data points in total, scaled all the variables to zero-mean unit variance, and shuffled the column order.

We used the structural Hamming distance (SHD) between the true and estimated group DAGs as an evaluation metric. Structural intervention distance (Peters and Bühlmann, 2015) showed similar results as SHD (Appendix B.2). As methods without a group DAG constraint do not necessarily return a group-acyclic DAG, we recursively remove edges with

the smallest absolute value from the estimated group adjacency matrix until we obtain a group-acyclic graph, which is analogous to the postprocessing in (Ng et al., 2020). Each experiment was performed 20 times.



(a) Variable DAG (b) Group DAG

Figure 2: Example of the simulated variable DAG and group DAG for $P = 12$ and $M = 3$

4.1.2. Nonlinear Data

We first performed an experiment to examine the effect of the group DAG constraint on nonlinear data with additive noise and no temporal dependencies, where we set $L = 0$ and $s_j = 1$ in Eq. (13). We compared the following four methods: NOTEARS-MLP (Zheng et al., 2020), NOTEARS-MLP using randomly selected variables for each group (NOTEARS-MLP-SEL), NOTEARS-MLP using the average value of the variables for each group (NOTEARS-MLP-AVE), and NOTEARS-MLP with a group DAG constraint (NOTEARS-MLP-ACY). All four methods had the same parameter settings of NOTEARS-MLP given in the original paper, i.e., $\lambda_1 = \lambda_2 = 0.01$, and $w = 0.3$, and the MLPs consisted of a single hidden layer with 10 nodes.

The results for different numbers of variables $P = \{10, 20, 30, 40\}$ and number of groups $M = \{5, 10\}$ are presented in Figure 3. As shown, NOTEARS-MLP-ACY outperformed the other methods, indicating the effectiveness of the group DAG constraint. Interestingly, NOTEARS-MLP-AVE and NOTEARS-MLP-SEL exhibited worse performance than NOTEARS-MLP, which indicates that aggregating the information of the groups leads to inferior results. The cases of $P = 10$ and $M = 10$ corresponded to the estimation of the group DAGs; thus, the four methods exhibited identical results.

4.1.3. Nonlinear Time-Series Data with Location-Scale Noise

Next, we performed an experiment on time-series data with location-scale noise to assess the performance of TS-LSNM and TS-LSNM with the group DAG constraint. We compared the following three methods: NTS-NOTEARS (Sun et al., 2021), the proposed TS-LSNM, and TS-LSNM with the group DAG constraint (TS-LSNM-ACY). The parameters for each method were determined by performing a grid search in the condition of $P = 20$, $M = 10$, and $N_j^t \sim U(-1/\sqrt{3}, 1/\sqrt{3})$ with parameter space $\lambda_1 = \lambda_2 \in \{0.05, 0.01, 0.005\}, w \in \{0.3, 0.2, 0.1\}$, resulting in $\lambda_1 = \lambda_2 = 0.01$ for all methods, $w = 0.2$ for NTS-NOTEARS

and TS-LSNM-ACY, and $w = 0.3$ for TS-LSNM. The number of hidden layers was set to 1, and the kernel size was set to 10. The maximum length L of the temporal dependencies of the data and models was set to 1.

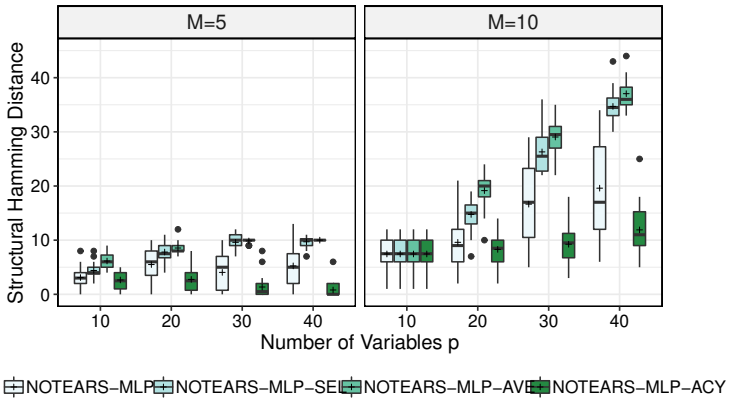


Figure 3: Results for nonlinear data obtained using different numbers of variables P and numbers of groups M

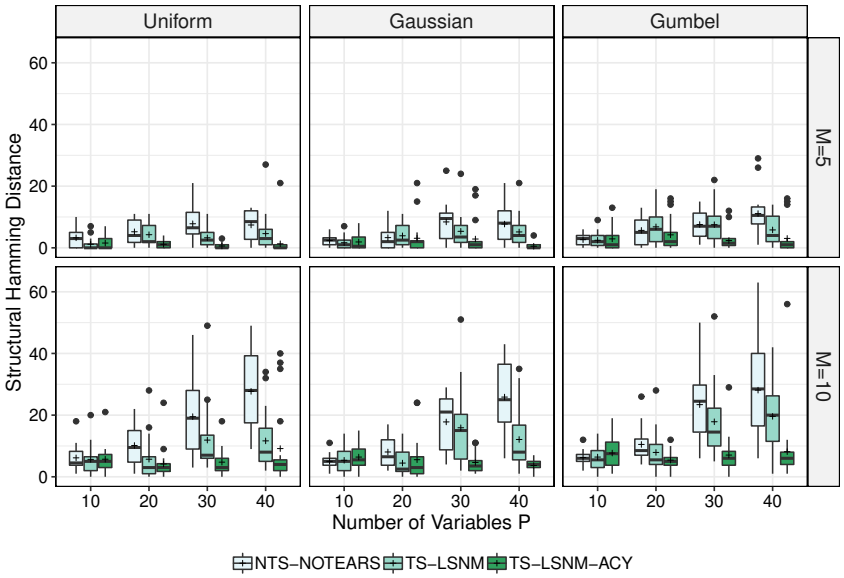


Figure 4: Results for nonlinear time-series data with location-scale noise obtained using different numbers of variables P, numbers of groups M, and noise distributions

The results for different numbers of variables $P = \{10, 20, 30, 40\}$, numbers of groups $M = \{5, 10\}$, and noise distributions $N_j^t \sim \{U(-1/\sqrt{3}, 1/\sqrt{3}), \mathcal{N}(0, 1), Gumbel(0, \sqrt{6}/\pi)\}$ are shown in Figure 4. Among the methods, TS-LSNM-ACY generally exhibited the best performance, followed by TS-LSNM. TS-LSNM outperformed NTS-NOTEARS because it captured the induced noise variances. Because of the DAG constraint, TS-LSNM-ACY achieved a relatively low SHD even if the number of variables increases. TS-LSNM-ACY is defeated only when $P = M$, indicating that the task involves estimation of the variable DAG. This suggests that the parameters should be adjusted according to the task to be solved.

4.2. Ceramic Substrate Manufacturing Process Data

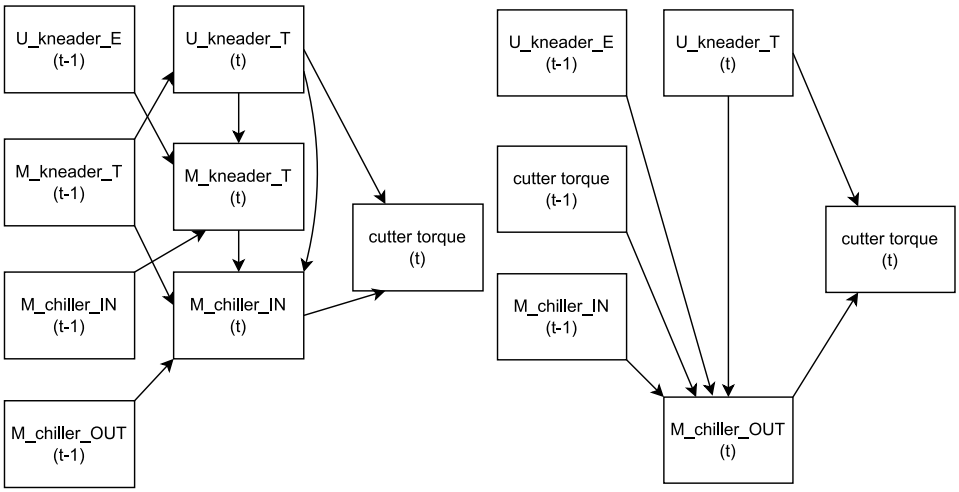
We compared the group DAGs obtained using NTS-NOTEARS and TS-LSNM-ACY with real-world data collected from the kneading process of a ceramic substrate manufacturing line. This process consisted of two kneaders (upper and middle) to mix the ingredients of the ceramic, each of which was cooled using a separate water-cooled chiller. The kneaded ingredients were cut to the same length and subjected to the baking process. The cutting torque is an important characteristic of the viscosity of the ceramic and is closely related to crack failure. The temperature, electricity (voltage and frequency), and pressure were measured at several positions of the kneaders and chillers, with a total of 19 variables and 2000 data points. We assigned groups to each variable according to domain knowledge and used them as groupings for TS-LSNM-ACY. Details regarding these groups are presented in Table 1. We incorporated prior knowledge that the cutter torque is the sink variable for both methods by restricting the corresponding kernel weights to zero (Sun et al., 2021). We used the parameter set obtained in Section 4.1.3. For the maximum time lag L, we used $L = 1$. We first fitted the models to $L = 5$ and then selected the value using the Frobenius norms of the estimated adjacency matrix for each time lag. Details are presented in Appendix C.

Table 1: Assigned groups for the ceramic manufacturing process data

Group ID	Name	Description	# of variables
1	U_chiller_T	Upper chiller temperature	1
2	U_kneader_T	Upper kneader temperature	3
3	U_kneader_E	Upper kneader electricity	3
4	M_chiller_IN	Water entering middle chiller	3
5	M_chiller_OUT	Water exiting middle chiller	2
6	M_kneader_T	Middle kneader temperature	3
7	M_kneader_E	Middle kneader electricity	3
8	Cutter torque	Cutting torque	1

The obtained group DAGs are shown in Figure 5, where groups irrelevant to the cutter torque are omitted. Both methods succeeded in recovering the connection between the temperature of the kneader (U_kneader_T) and the cutter torque, and their results

matched the domain knowledge. The result of TS-LSNM-ACY, which revealed an arrow from U_kneader_T to the cooling water flowing into the chiller (M_chiller_IN), was more consistent with the domain knowledge than the result of NTS-NOTEARS, which revealed a connection from U_kneader_T to the cooling water flowing out of the chiller (M_chiller_OUT), because M_chiller_OUT is expected to be controlled by the chiller. Moreover, the result of TS-LSNM-ACY revealed the correct physical phenomenon in which the chillers cool the kneaders, whereas the result of NTS-NOTEARS revealed no connection between the chillers and kneaders, which disagrees with expectations. Therefore, we conclude that TS-LSNM with the DAG constraint obtained better estimation results than NTS-NOTEARS.



(a) TS-LSNM-ACY (b) NTS-NOTEARS

Figure 5: Estimated group DAGs for ceramic substrate manufacturing process data

5. Conclusions and Remarks

We propose group DAG constraint—a novel differentiable algebraic constraint—to perform structure learning on groups of variables, assuming that the relationship among the groups is acyclic and the assignment of the groups is known in advance. The group DAG constraint offers not only the use of prior knowledge about the variable groups, but also can give more simplified results that are more comprehensive by estimating group DAGs. Furthermore, we propose TS-LSNM—a functional causal model that can handle nonlinear time-series data with location-scale noise. A corresponding estimation algorithm was developed and tested. We evaluated the performance of the group DAG constraint and TS-LSNM by performing numerical experiments on synthetic and real-world data acquired from the kneading process of a ceramic substrate manufacturing line, and the results indicated the effectiveness of the proposed methods. The effect when the assignment of the groups is incorrect, and the impact of the imbalance on group size is left to be our future work.

References

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Doris Entner and Patrik O Hoyer. Estimating a causal order among groups of variables in linear models. In *Artificial Neural Networks and Machine Learning–ICANN 2012: 22nd International Conference on Artificial Neural Networks, Lausanne, Switzerland, September 11-14, 2012, Proceedings, Part II 22*, pages 84–91. Springer, 2012.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

Patrik Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21:689–696, 2008.

Aapo Hyvärinen and Erkki Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *signal processing*, 64(3):301–313, 1998.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. *arXiv preprint arXiv:2210.09054*, 2022.

Yoshinobu Kawahara, Kenneth Bollen, Shohei Shimizu, and Takashi Washio. Grouplingam: Linear non-gaussian acyclic models for sets of variables. *arXiv preprint arXiv:1006.5041*, 2010.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

Genta Kikuchi. Differentiable causal discovery under heteroscedastic noise. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part I*, pages 284–295. Springer, 2023.

Katerina Marazopoulou, Rumi Ghosh, Prasanth Lade, and David Jensen. Causal discovery for manufacturing domains. *arXiv preprint arXiv:1605.04056*, 2016.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33, 2020.

Pekka Parviainen and Samuel Kaski. Learning structures of bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88:110–127, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Sch"olkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Alexander G Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Varsortability in additive noise models. *arXiv preprint arXiv:2102.13647*, 2021.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7): 075310, 2018.

Richard Scheines and Peter Spirtes. Causal structure search: Philosophical foundations and future problems. In *NIPS 2008 Workshop, Causality: Objectives and Assessment, Whistler, Canada*, 2008.

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.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for FMRI. *Neuroimage*, 54(2):875–891, 2011.

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *arXiv preprint arXiv:2205.13085*, 2022.

Xiangyu Sun, Guiliang Liu, Pascal Poupart, and Oliver Schulte. Nts-notears: Learning nonparametric temporal dags with time-series data and prior knowledge. *arXiv e-prints*, pages arXiv–2109, 2021.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pages 9311–9323. PMLR, 2020.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Vector causal inference between two groups of variables. *arXiv preprint arXiv:2209.14283*, 2022.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9492–9503, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

Shuheng Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22, 2009.

Appendix A. An Example of the Weighted Group Adjacency Matrix (4)

We give an example of the weighted group adjacency matrix calculated by Equation (4). Suppose we have a graph $\mathcal{G}$ on 5 variables with grouping $K(1) = \{1, 2\}, K(2) = \{3, 4\}, K(3) = \{5\}$, and a corresponding weighted adjacency matrix D:

$$D = \begin{bmatrix} 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & -5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Figure 6 shows the variable DAG and a group DAG with corresponding adjacency matrix B, weighted adjacency matrix D, group adjacency matrix $B^{'}$, and weighted group adjacency matrix $D^{'}$. The connection strengths are shown as the edge labels.

We will show that we get $D^{'}$ by substituting D to Equation (4):

$$[D^{'}]_{k,l} = \begin{cases} 0 & \text{if } k = l, \\ \sum_{i \in K(k)} \sum_{j \in K(l)} [\bar{D}]_{i,j} & \text{else} \end{cases},$$

where

$$\bar{D} = D \circ D = \begin{bmatrix} 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 4 & 16 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 25 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

For the off-diagonal elements ($k \neq l$), we get:

$$[D^{'}]_{1,2} = \sum_{i \in K(1)=\{1,2\}} \sum_{j \in K(2)=\{3,4\}} [\bar{D}]_{i,j} = [\bar{D}]_{1,3} + [\bar{D}]_{1,4} + [\bar{D}]_{2,3} + [\bar{D}]_{2,4} = 4 + 0 + 0 + 4 = 8,$$

$$[D^{'}]_{2,1} = \sum_{i \in K(2)=\{3,4\}} \sum_{j \in K(1)=\{1,2\}} [\bar{D}]_{i,j} = [\bar{D}]_{3,1} + [\bar{D}]_{3,2} + [\bar{D}]_{4,1} + [\bar{D}]_{4,2} = 0 + 0 + 0 + 0 = 0,$$

$$[D^{'}]_{1,3} = \sum_{i \in K(1)=\{1,2\}} \sum_{j \in K(3)=\{5\}} [\bar{D}]_{i,j} = [\bar{D}]_{1,5} + [\bar{D}]_{2,5} = 0 + 16 = 16,$$

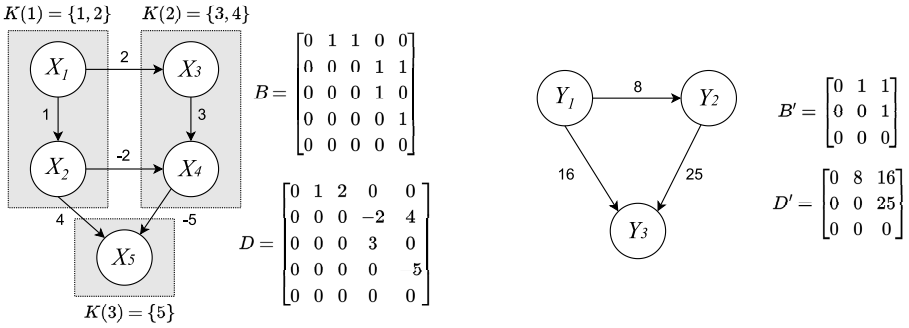
$$[D^{'}]_{3,1} = \sum_{i \in K(3)=\{5\}} \sum_{j \in K(1)=\{1,2\}} [\bar{D}]_{i,j} = [\bar{D}]_{5,1} + [\bar{D}]_{5,2} = 0 + 0 = 0,$$

$$[D^{'}]_{2,3} = \sum_{i \in K(2)=\{3,4\}} \sum_{j \in K(3)=\{5\}} [\bar{D}]_{i,j} = [\bar{D}]_{3,5} + [\bar{D}]_{4,5} = 0 + 25 = 25,$$

$$[D^{'}]_{3,2} = \sum_{i \in K(3)=\{5\}} \sum_{j \in K(2)=\{3,4\}} [\bar{D}]_{i,j} = [\bar{D}]_{5,3} + [\bar{D}]_{5,4} = 0 + 0 = 0.$$

Since the diagonal elements ($k = l$) of $D^{'}$ are zero, we obtain:

$$D^{'} = \begin{bmatrix} 0 & 8 & 16 \\ 0 & 0 & 25 \\ 0 & 0 & 0 \end{bmatrix}.$$



(a) Variable DAG (b) Group DAG

Figure 6: A variable DAG and a group DAG

Appendix B. Additional Experiments

B.1. Results for Data with Cycles in Each Group

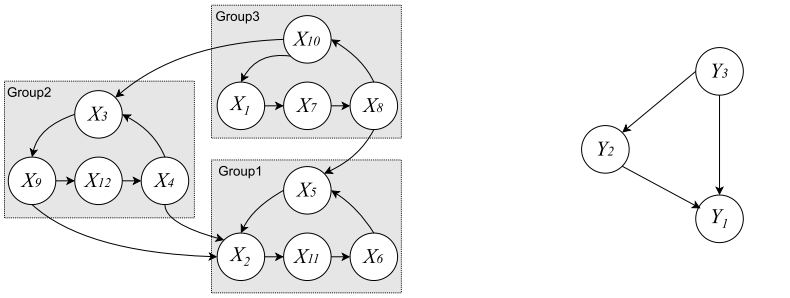
We report the results for data with cycles in each group. We generate a binary adjacency matrix B by creating connections between the variables using the procedure described in Section 4.1.1, except for the cyclic data, for which we created a single loop consisting of all the variables in the same group. A simple example of a variable DAG and the corresponding group DAG for $P = 12$ and $M = 4$ is shown in Figure 7.

By replacing the nonzero elements of B with independent realizations of $\pm U(0.5, 2.0)$, we generate data X using the following linear equation:

$$\begin{aligned} X &= XB + e \\ \Leftrightarrow X &= (I - B)^{-1} e, \end{aligned} \tag{14}$$

where I is an $P \times P$ identity matrix, and $e = (e_1, ..., e_P)$ denotes the noise terms generated independently from $U(-1/\sqrt{3}, 1/\sqrt{3})$. We generated 2000 data points and standardized all the columns to zero-mean unit variance. We compared TS-LSNM and TS-LSNM-ACY with $L = 0$, using the parameter settings presented in Section 4.1.2. NOTEARS-MLP was not used as the base model, because it exhibits unsatisfactory performance when we standardized the variables generated from the linear model owing to the assumption of Gaussian noise with equal variance.

The results for different numbers of variables $P = \{10, 20, 30, 40\}$ and graph types $\{\text{acyclic}, \text{cyclic}\}$ with $M = 10$ groups are presented in Figure 8. There was no significant difference in SHD between the acyclic and cyclic cases, indicating that the connections between the variables in the same group did not necessarily affect the estimation accuracy for the connections among the groups. Using the DAG constraint makes the estimation more robust to the number of variables in each group for both acyclic and cyclic cases.



(*a*) Variable DAG (*b*) Group DAG

Figure 7: Example of simulated variable DAG with cycles and corresponding group DAG for $P = 12$ and $M = 3$

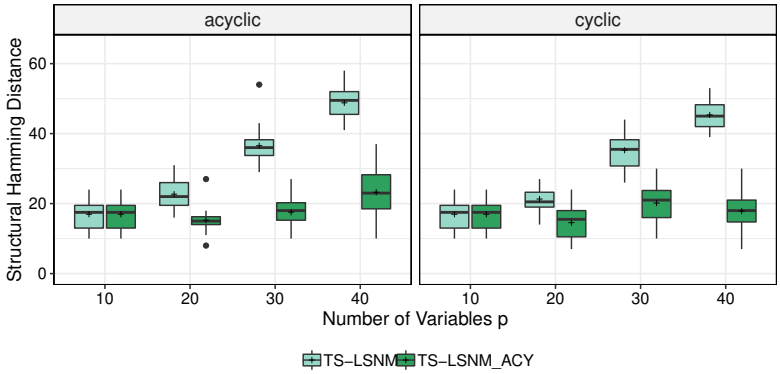


Figure 8: Results for acyclic and cyclic data generated from a linear model, with different numbers of variables: $P = \{10, 20, 30, 40\}$. The number of groups M was set to 10.

B.2. Structural Intervention Distance

We conducted a numerical experiment on nonlinear time-series data with location-scale noise with the same setting as Section 4.1.3, but only for the number of groups $M = 10$, number of variables $p = 40$, and uniform noise distribution. Here we used the structural intervention distance (SID) (Peters and Bühlmann, 2015) for the evaluation, where SID evaluates the number of wrongly estimated interventional distributions. Compared to SHD, SID prioritizes the causal order of the variables. The result is shown in Figure 9. We can see that the proposed TS-LSNM-ACY show the smallest SID followed by TS-LSNM, which is the same as the result of SHD shown in Figure 4.

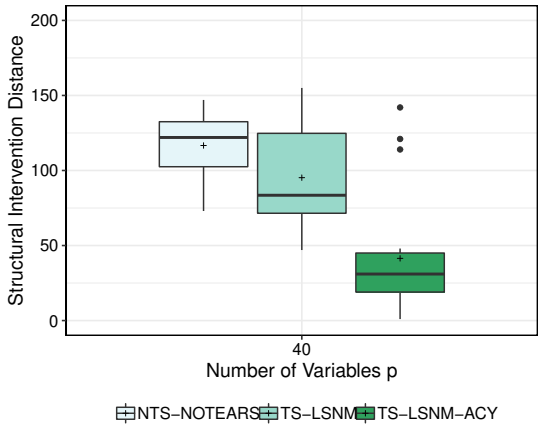


Figure 9: Structural intervention distance for time-series data with location-scale noise, on number of groups $M = 5$, number of variables $P = 40$, and uniform noise distribution

Appendix C. Selection of Maximum Time Lags L

In the numerical experiments on synthetic data, we assumed that the true time lags L were known in advance. However, in a real-world scenario, we must select an appropriate L value from the data. For the ceramic manufacturing process data described in Section 4.2, we fitted each model with a large time-lag value of $L = 5$ and estimated the weighted adjacency matrix $\tilde{W}^{\tau}(\theta^A, \theta^C)$ $(\tau = 0, ..., L)$. We then calculated the Frobenius norm of the estimated weighted adjacency matrix for each time lag τ.

The results are presented in Figure 10, where plateaus are observed for $L > 1$ for both methods. Therefore, we selected $L = 1$ for both methods and fitted the model again with $L = 1$. An alternative approach for determining L is to determine the value of the objective function F, although we must fit the model multiple times.

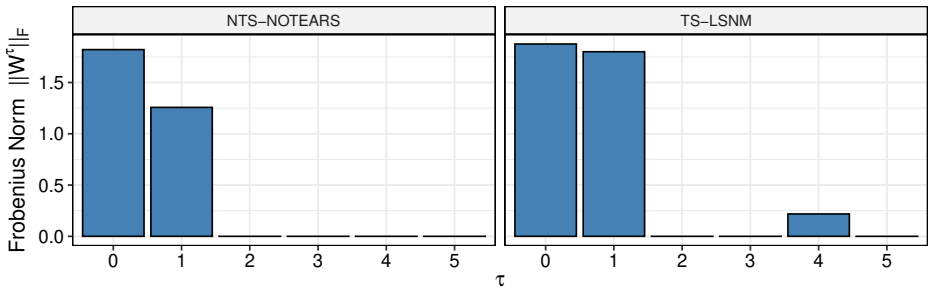


Figure 10: Frobenius norm of the estimated weighted adjacency matrix on each time lag τ ($\|\tilde{W}^\tau(\theta^A, \theta^C)\|_F$)