Nonlinear Causal Discovery for Grouped Data

Konstantin Göbler[1,4] **Tobias Windisch**[3] **Mathias Drton**[1,2]

[1]Department of Mathematics, TUM School of Computation, Information and Technology, Technical University of Munich
[2]Munich Center for Machine Learning
[3]University of Applied Sciences Kempten, Faculty of Mechanical Engineering
[4]Robert Bosch GmbH

Abstract

Inferring cause-effect relationships from observational data has gained significant attention in recent years, but most methods are limited to scalar random variables. In many important domains, including neuroscience, psychology, social science, and industrial manufacturing, the causal units of interest are groups of variables rather than individual scalar measurements. Motivated by these applications, we extend nonlinear additive noise models to handle random vectors, establishing a two-step approach for causal graph learning: First, infer the causal order among random vectors. Second, perform model selection to identify the best graph consistent with this order. We introduce effective and novel solutions for both steps in the vector case, demonstrating strong performance in simulations. Finally, we apply our method to real-world assembly line data with partial knowledge of causal ordering among variable groups.

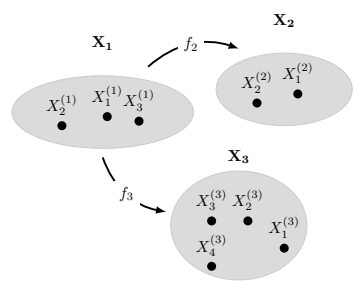


Figure 1: An example of a grouped additive noise model with three groups and varying group sizes.

1 INTRODUCTION

Many techniques have been developed to infer causal relations among random variables from observational data [Spirtes et al., 2000, Chickering, 2003, Shimizu et al., 2006, Peters et al., 2014, Zheng et al., 2018]. While most of these procedures focus on structure learning among scalar variables, there has been steadily increasing interest in settings where groups of measurements constitute the causal entities of interest [see Wahl et al., 2024, for an overview]. For instance in neuroscience, causal connections between brain regions rather than individual neurons are often of interest [Panzeri et al., 2017, Kohn et al., 2020, Semedo et al., 2020]. In earth and climate science, researchers are frequently interested in causal interactions among groups of measurements (e.g., wind speed, air pressure, etc.) across a number of grid locations spanning the planet [Runge et al., 2015, and references therein]. In psychology and social sciences studies often involve the measurement of certain psychological or societal traits based on several proxy variables that need to be treated jointly [Cronbach and Meehl, 1955, Campbell and Fiske, 1959, Antonoplis, 2022]. In industrial manufacturing, quality control systems often record several related measurements from automated process, e.g., industrial welding, injection molding, or staking and pressing, where causal relationships among several such process are the relevant causal items [Vuković and Thalmann, 2022, Kikuchi and Shimizu, 2023, Göbler et al., 2024].

There are a number of ways to incorporate known group structures in causal learning tasks. The most elementary approach is to apply dimension reduction, e.g., by taking the mean across all members in one group. Given the resulting set of scalar summary variables, standard causal discovery methods may be employed. This comes at the cost of severe information loss, and may even render conditional independence results useless [Wahl et al., 2024]. Another approach disregards the grouping structure during the learning tasks and applies causal discovery techniques to the group members. In a second step, the resulting graph estimate is appropriately coarsened to represent the variable groupings [see Rubenstein* et al., 2017, Chalupka et al., 2016, Parviainen and Kaski, 2017, Anand et al., 2023]. A last

approach seeks to treat the groups themselves as the causal quantities of interest and aims at performing causal learning on the groups directly [Janzing et al., 2010, Zscheischler et al., 2011, Entner and Hoyer, 2012, Wahl et al., 2023]. Our work belongs to the latter framework of group causal learning.

Motivated by the work of Peters et al. [2014] on nonlinear causal discovery in continuous additive noise models (ANMs) for multivariate data, we revisit the identifiability of causal directions in the group setting. In particular, we show that in general, the causal directions can be identified in the group setting. The statistical problems involved when operating with random vectors rather than scalar random variables become much more challenging. We construct an order independent version of the regression with subsequent independence test (RESIT) algorithm [Peters et al., 2014] in the group setting and propose efficient and flexible solutions to the estimation problems involved.

In particular, the model selection task after having obtained a causal order requires special attention. We propose a novel class of multi-response group sparse additive models (MURGS) for selecting relevant edges. MURGS is a standalone feature selection procedure specifically designed for the grouped setting and can be seamlessly integrated into the pruning phase of any order-based method [Teyssier and Koller, 2012]. For example, Entner and Hoyer [2012] estimate causal orderings using grouped linear non-Gaussian ANMs, and we anticipate that other order-based approaches — such as the score matching method introduced by Rolland et al. [2022] — can be extended to the group case.

In summary our main contributions are the following:

- We propose a flexible and fully nonparametric learning strategy to first obtain a causal order by means of neural networks and nonparametric independence tests. Then, for DAG pruning, we introduce MURGS, a class of multi-response sparse additive models to encourage sparsity on the group level and analytically derive a closed-form backfitting update for the corresponding block coordinate descent algorithm.
- We evaluate our proposed method on synthetic data and demonstrate superior performance compared with several other causal discovery algorithms. Further, we consider real world manufacturing data with partially known causal ordering that allows us to partly assess algorithmic performance.

The remainder of the paper is structured as follows. Section 2 introduces the group ANM and establishes identifiability results in the group setting. Next, in Section 3 we describe the two phases of the GroupRESIT algorithm including the development of MURGS for model selection. Section 4 presents the results of our synthetic experiments, and in Section 5 we apply our methods to real data from industrial manufacturing. We conclude in Section 6.

2 METHODOLOGY

2.1 PREREQUISITES

Throughout, we use the term *group* to refer to a set of scalar random variables that all belong to the same random vector. For integer p we define $[p] = \{1, \dots, p\}$. Further, define $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ as a tuple of p random vectors, where each $\mathbf{X}_g = (X_1^{(g)}, \dots, X_{d_g}^{(g)})$ is a vector in $\mathbb{R}^{d_g}$, with $d_g \in \mathbb{N}$, for $g \in [p]$. We refer to $\mathbf{X}_g$ as the g-th group. Denote by $P_{\mathbf{X}}$ the joint distribution on $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_p}$. If it exists, we denote the (joint) density of $\mathbf{X}$ by $p_{\mathbf{X}}$. Let $\mathcal{G} = (V, E)$ be a directed acyclic graph (DAG) with vertex set $V = [p]$ and edges $E \subseteq V \times V$. The vertex set indexes the set of random vectors in $\mathbf{X}$, and adopting the framework of graphical models, we assume that $P_{\mathbf{X}}$ factorizes according to $\mathcal{G}$ as follows

$$P_{\mathbf{X}}(\mathbf{X}) = \prod_{g=1}^{p} P_{\mathbf{X}_g}(\mathbf{X}_g \mid \mathbf{X}_{pa_{\mathcal{G}}(g)}),$$

where $pa_{\mathcal{G}}(g)$ refers to the parents of node g in G. We refer to Drton and Maathuis [2017] for more details on graphical models.

We call j a parent of g if $(j, g) \in E$. Further, we denote the set of non-descendants by $nd_{\mathcal{G}}(g) = V \setminus \big(\{g\} \cup \mathrm{desc}(g)\big)$ where the *descendants* of g are defined as $\mathrm{desc}_{\mathcal{G}}(g) = \{w \in V : \text{there exists a directed path from } g \text{ to } w\}$. Whenever the graph is clear from the context we omit the subscript. Note that the factorization agrees fully with the scalar case as long as $P_{\mathbf{X}}$ has a density with respect to some product measure. We may also express the above model in terms of structural equation models (SEM) in the grouped case:

Definition 1. *Let $\mathbf{X} = (\mathbf{X}_g)_{g \in [p]}$ and $\mathbf{N} = (\mathbf{N}_g)_{g \in [p]}$ be jointly distributed random vectors where $\mathbf{X}_g$ is d_g-dimensional and where $\mathbf{N}_g \perp\!\!\!\perp \mathbf{N}_h$ for all $g \neq h$. If there exists a graph $\mathcal{G}_0$ on $[p]$ and a sequence of vector-valued functions $\mathcal{F} = (f_1, \dots, f_p)$ such that*

$$\mathbf{X}_g = f_g(\mathbf{X}_{pa_{\mathcal{G}_0}(g)}, \mathbf{N}_g) \tag{1}$$

for all $g \in [p]$, then $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ is a grouped structural equation model *(GSEM).*

In case all groups are one-dimensional, the GSEM is just a standard SEM [Bollen, 1989]. Note that while Definition 1 requires the noise groups to be jointly independent, the model explicitly allows for dependence within each group.

2.2 CAUSAL MODELS FOR GROUPED DATA

One can generally not identify the true underlying DAG when given only observational data from the model in Defi-

nition 1 [see Peters et al., 2014, for an overview]. However, as we will see, restricting the functional form of the right-hand side in (1) to be additive in the noise vectors renders the model identifiable.

Definition 2. *Let $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ be a GSEM. If the functions in $\mathcal{F}$ are additive in the noise term, i.e., if*

$$\mathbf{X}_g = f_g(\mathbf{X}_{pa(g)}) + \mathbf{N}_g,$$

if $\mathbf{N}_g$ has a strictly positive density with respect to the Lebesgue measure for all $g \in [p]$, and if $\mathcal{G}_0$ is acyclic, then $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ is a group additive noise model *(GANM).*

In Definition 1, the noise vectors $\mathbf{N}_g$ and the predictors $\mathbf{X}_g$ are permitted to have different dimensions, whereas in Definition 2 their dimensions must coincide. Moreover, the graph underlying the SEM in Definition 1 is not required to be acyclic, while acyclicity is assumed in Definition 2. Figure 1 illustrates a GANM with three groups of varying sizes. In addition to the implicit assumption of causal sufficiency—i.e., that no latent variables are present due to the joint independence of the noise terms—the framework for GANMs also requires the notion of causal minimality.

Definition 3. *Let $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ be a GANM. We say that $P_{\mathbf{X}}$ satisfies* causal minimality *if all functions in $\mathcal{F}$ are non-constant in any of their arguments.*

Suppose that the function class $\mathcal{F} \subseteq C^3(\mathbb{R}^{d_{pa(g)}}, \mathbb{R}^{d_g})$, where $d_{pa(g)} = \sum_{j \in pa(g)} d_j$, consists of nonlinear functions—more precisely, for each output coordinate $k \in [d_g]$ and each input-dimension $i \in [d_{pa(g)}]$ there exists some $\mathbf{x} \in \mathbb{R}^{d_{pa(g)}}$ such that $\partial^2 f_k / \partial x_i^2(x) \neq 0$ or $\partial^3 f_k / \partial x_i^3(x) \neq 0$. Then, under causal minimality, identifiability for the bivariate scalar case has been established by Hoyer et al. [2008]. Consider a bivariate GANM where the two groups have size 1 respectively, i.e.,

$$\begin{aligned} X_1 &= N_1 \\ X_2 &= f_2(X_1) + N_2, \end{aligned} \tag{2}$$

where $X_1 \perp\!\!\!\perp N_2$. Identifiability follows from observing that a regression $\mathbb{E}[X_2 \mid X_1] = f_2(X_1)$ along the causal direction leads to independence among the residuals and the predictor X_1. For general nonlinear functions f, the regression in the anti-causal direction does not lead to independence among the residuals and X_2. Indeed, Hoyer et al. [2008] derive a specific differential equation that the triple (f_2, P_{X_1}, P_{N_2}) needs to satisfy for the backwards model to exist. A similar differential equation can be obtained in the bivariate group case.

Condition 1. *The triple $(f_g, P_{\mathbf{X}_j}, P_{\mathbf{X}_g})$ does not satisfy the following differential equation:*

$$\begin{aligned} & D_{\mathbf{x}_j} \mathbf{H}_\xi(\mathbf{x}_j) \Big[(D_{\mathbf{x}_j \mathbf{x}_j} \pi_1(\mathbf{x}_j, \mathbf{x}_g))^{-1} D_{\mathbf{x}_j \mathbf{x}_g} \pi_1(\mathbf{x}_j, \mathbf{x}_g) \Big] \\ =& D_{\mathbf{x}_j} D_{\mathbf{x}_j \mathbf{x}_g} \pi_1(\mathbf{x}_j, \mathbf{x}_g) - \Big[D_{\mathbf{x}_j} \big(\mathbf{H}_{f_g}(\mathbf{x}_j)[\nabla \nu(\mathbf{u})] \big) \\ & - D_{\mathbf{x}_j} \big(\mathbf{J}_{f_g}(\mathbf{x}_j)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_g}(\mathbf{x}_j) \big) \Big] \\ & \Big[(D_{\mathbf{x}_j \mathbf{x}_j} \pi_1(\mathbf{x}_j, \mathbf{x}_g))^{-1} D_{\mathbf{x}_j \mathbf{x}_g} \pi_1(\mathbf{x}_j, \mathbf{x}_g) \Big], \end{aligned} \tag{3}$$

where $\mathbf{u} := \mathbf{x}_g - f_g(\mathbf{x}_j)$ and further $\nu := \log p_{\mathbf{N}_g}$, $\xi := \log p_{\mathbf{X}_j}$, with arguments $\mathbf{u}, \mathbf{x}_j$, respectively. Additionally, $\pi_1(\mathbf{x}_j, \mathbf{x}_g) := \log p_{\mathbf{X}_j, \mathbf{X}_g}(\mathbf{x}_j, \mathbf{x}_g)$ and

$$\begin{aligned} D_{\mathbf{x}_j \mathbf{x}_j} \pi_1(\mathbf{x}_j, \mathbf{x}_g) = & \mathbf{H}_\xi(\mathbf{x}_j) - \mathbf{H}_{f_g}(\mathbf{x}_j)[\nabla \nu(\mathbf{u})] \\ & + \mathbf{J}_{f_g}(\mathbf{x}_j)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_g}(\mathbf{x}_j), \end{aligned}$$

and $D_{\mathbf{x}_j \mathbf{x}_g} \pi_1(\mathbf{x}_j, \mathbf{x}_g) = -\mathbf{J}_{f_g}(\mathbf{x}_j)^\top \mathbf{H}_\nu(\mathbf{u})$. $\mathbf{J}$ and $\mathbf{H}$ denote Jacobian and Hessian matrices or tensors, respectively.

Remark. Note that when $\mathbf{J}_{f_g}(\mathbf{x}_j)$ is full rank and $\mathbf{H}_\nu(\mathbf{u})$ is positive definite, we can fully isolate the third-order derivative tensor $D_{\mathbf{x}_j} \mathbf{H}_\xi(\mathbf{x}_j)$:

$$\begin{aligned} D_{\mathbf{x}_j} \mathbf{H}_\xi(\mathbf{x}_j) = & D_{\mathbf{x}_j} (D_{\mathbf{x}_j \mathbf{x}_g} \pi_1)(D_{\mathbf{x}_j \mathbf{x}_g} \pi_1)^{-1} D_{\mathbf{x}_j \mathbf{x}_j} \pi_1 \\ & - \Big[D_{\mathbf{x}_j} (\mathbf{H}_{f_g}(\mathbf{x}_j)[\nabla \nu(\mathbf{u})]) \\ & - D_{\mathbf{x}_j} (\mathbf{J}_{f_g}(\mathbf{x}_j)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_g}(\mathbf{x}_j)) \Big], \end{aligned}$$

where we have suppressed the arguments $\mathbf{x}_j$ and $\mathbf{x}_g$ of π_1. Observe that $\mathbf{H}_\xi(\mathbf{x}_j)$ enters only via the term $D_{\mathbf{x}_j \mathbf{x}_j} \pi_1$ on the right hand side. In fact, this equation bears a direct resemblance to the scalar form derived by Hoyer et al. [2008], highlighting that our result is a natural generalization of the scalar case. In general, Eq. (3) describes a directional projection of $D_{\mathbf{x}_j} \mathbf{H}_\xi(\mathbf{x}_j)$ onto the directions defined by the columns of the matrix $(D_{\mathbf{x}_j \mathbf{x}_j} \pi_1)^{-1} D_{\mathbf{x}_j \mathbf{x}_g} \pi_1$. The dimensions d_{x_j} and d_{x_g} determine the range of the resulting tensor contraction. A detailed exploration of the role of the group sizes and implications for the form of the triple $(f_g, P_{\mathbf{X}_j}, P_{\mathbf{X}_g})$ is deferred to future work.

Definition 4. *Consider a* bivariate GANM *given by the equations*

$$\mathbf{X}_j = \mathbf{N}_j, \quad \mathbf{X}_g = f_g(\mathbf{X}_j) + \mathbf{N}_g,$$

for $\{j, g\} = \{1, 2\}$. If the corresponding triple $(f_g, P_{\mathbf{X}_j}, P_{\mathbf{X}_g})$ satisfies Condition 1, we call this model an identifiable bivariate *GANM.*

Theorem 2.1. *Let $P_{\mathbf{X}}$ be the joint distribution of $\mathbf{X}$ generated by an* identifiable bivariate *GANM $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ and suppose that causal minimality holds. Then, the graph $\mathcal{G}_0$ is identifiable from $P_{\mathbf{X}}$.*

All proofs for this section are provided in Appendix A. As demonstrated by Peters et al. [2014], extending the analysis

from two to multiple variables can be achieved by appropriately constraining the involved distributions and functions so that, locally, the problem reduces to the bivariate case.

Corollary 1. *Let $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ be a GANM with p variables. Suppose that for each node $g \in [p]$, each parent $j \in pa(g)$, and every set S satisfying*

$$pa(g) \setminus \{j\} \subseteq S \subseteq nd(g) \setminus \{g, j\},$$

there exists a realization $\mathbf{x}_S \in \mathbb{R}^{|S|}$ for which the conditional distribution of $\mathbf{X}_S$ has strictly positive density with respect to the Lebesgue measure, and the triple

$$\left(f_g(\mathbf{x}_{pa(g)\setminus\{j\}}, \mathbf{x}_j), P_{\mathbf{X}_j, \mathbf{X}_g | \mathbf{X}_S = \mathbf{x}_S}, P_{\mathbf{X}_j}\right)$$

satisfies Condition 1. Here, the arguments $\mathbf{x}_{pa(g)\setminus\{j\}}$ are held fixed, making the function $f_g(\mathbf{x}_{pa(g)\setminus\{j\}}, \mathbf{x}_j)$ depend solely on $\mathbf{x}_j$.

Then the graph $\mathcal{G}_0$ is identifiable from the joint distribution $P_{\mathbf{X}}$.

Clearly, fixing all arguments of f_g except $\mathbf{X}_j$ also leads to a bivariate GANM. Corollary 1 suggests that this is not enough. Instead, we need to put restrictions on the conditional distribution of $\mathbf{X}_j$. The following result guides the estimation strategy described in the ensuing section.

Lemma 2.2. *Let $(\mathbf{X}, \mathbf{N}, \mathcal{F}, \mathcal{G}_0)$ be a GSEM. For all $S \subseteq nd_{\mathcal{G}_0}(g)$, we have $\mathbf{N}_g \perp\!\!\!\perp \mathbf{X}_S$.*

In particular, if $g \in [p]$ is a sink node, i.e., a node without any descendants, its non-descendants is the set of all other nodes in $\mathcal{G}_0$. Thus, its corresponding noise vector $\mathbf{N}_g$ is independent of all other variables $\mathbf{X}_{[p]\setminus\{g\}}$.

3 GROUPED RESIT

In the context of ANMs, Peters et al. [2014] introduce the Regression with Subsequent Independence Test (RESIT) algorithm to learn the structure of the underlying DAG from observational data. In what follows, we denote the underlying true DAG by $\mathcal{G}_0$. RESIT has two phases. The **first phase** infers a causal order among the variables involved. A permutation $\pi : [p] \to [p]$ is a valid causal order for a DAG $\mathcal{G}$ if $\pi(g) < \pi(j)$ for all $g \in nd_{\mathcal{G}}(j)$. Motivated by Lemma 2.2, the following steps are performed $p - 1$ times:

(i) Regress each variable on all others. Then measure independence between the corresponding residuals and all other variables.

(ii) The variable that accounts for its *least dependent residual* is considered as a sink node and removed from the set of variables.

The **second phase** seeks to select the *best* DAG among those that agree with the causal order found in the first phase. Several model selection techniques may be used for this purpose. Peters et al. [2014] propose to *greedily test away* edges that are present in the super-DAG associated with the inferred causal order. The strategy we pursue involves sparse model selection techniques to prune extraneous edges.

In our setting, we can apply the same two-phase procedure to obtain a group DAG estimate of $\mathcal{G}_0$. Yet, all involved estimation and testing procedures need to solve much harder statistical problems. We propose solutions to the estimation tasks in PHASE I and II (Algorithms 1 and 2).

3.1 LEARNING A CAUSAL ORDER

Algorithm 1: GroupRESIT algorithm PHASE I

Input : IID samples of p-many jointly distributed random vectors $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$.
Output : Learned causal order π.
Initialize : $S := \{1, \ldots, p\}, \quad \pi := [\cdot]$.
repeat
  for $g \in S$ **do**
    (i) Regress $\mathbf{X}_g = (X_1^{(g)}, \ldots, X_{d_g}^{(g)})$ on $\{\mathbf{X}_j\}_{j \in S \setminus \{g\}}$
    (ii) Measure the independence between the residuals $\mathbf{R}_g = (R_1^{(g)}, \ldots, R_{d_g}^{(g)})$ and the remaining groups $\{\mathbf{X}_j\}_{j \in S \setminus \{g\}}$
  end
  Identify the group g^* that accounts for its least dependent residual
  $S := S \setminus \{g^*\}$
  $\pi := [g^*, \pi]$
until *until* $S = \emptyset$;

The first phase (Algorithm 1) consists of repeatedly solving regression problems on progressively smaller predictor sets. In each iteration, we train multi-output deep neural networks ϕ_θ whose output layers are sized to match the number of response variables in each group.

After training is completed, we compute the residuals $\mathbf{R}_g = \phi_\theta(\mathbf{X}_g) - \mathbf{X}_g$, and identify the group whose residuals exhibit the weakest dependence on the remaining variables. This group is then designated as a *sink node* and is subsequently removed from the set of variables.

We measure dependence via the HSIC [Gretton et al., 2005], i.e., $\text{HSIC}((\mathbf{X}_j)_{j \in S \setminus \{g\}}, \mathbf{R}_g)$. When equipped with characteristic kernels [Fukumizu et al., 2007], the HSIC equals zero if and only if the random quantities involved are (unconditionally) independent. In particular, the HSIC is applicable to random vectors of general dimension.

Remark. We emphasize that direct comparisons of HSIC

values are meaningful only when all models under consideration share the same dimension and all involved quantities are measured on an identical scale.

3.2 MURGS FOR MODEL SELECTION

Having obtained a causal order π, we construct a super-DAG $\mathcal{G}^\pi$ where nodes get assigned all of their predecessors in the causal order as parents. Formally, the set of parents for node $j \in [p]$ in $\mathcal{G}^\pi$ is $pa_\pi(j) = \{g \in [p] : \pi(g) < \pi(j)\}$. The goal of PHASE II is to find a not too large super-DAG $\mathcal{G}$ of $\mathcal{G}_0$ with $\mathcal{G} \subseteq \mathcal{G}^\pi$. We advocate to use sparse model selection techniques for this procedure. To that end, we tailor multi-task sparse additive models to the given setting. The resulting model class is of independent interest as it generalizes the sparse group lasso to the multi-task setting. In this paper, we use a common design matrix across all tasks rendering the models multi-response in nature.

In the original RESIT, Peters et al. [2014] propose to iteratively remove nodes from the potential parent set by greedily cycling through the following steps. First, remove a potential parent node from the regression set and second, test whether residuals are still independent and restore the node in question if this is not the case. Thus, the significance level of the test acts as a tuning parameter for the model selection procedure. As pointed out in Peters et al. [2014], such a procedure strongly depends on the order in which the independence tests are carried out. Accordingly, type-one errors lead to extraneous edges in the final DAG estimate. Instead, we use feature selection via sparse additive models to prune edges in $\mathcal{G}^\pi$.

Algorithm 2: GroupRESIT algorithm PHASE II

Input : π
Output : Learned DAG $\mathcal{G}$
Initialize : pa_π.
for $j \in \pi$ **do**
- Use MURGS and regress $\mathbf{X}_j$ on $\{\mathbf{X}_g\}_{g \in pa_\pi(j)}$
- Obtain the parent set $pa_{\mathcal{G}}(j) := \{g \in pa_\pi(j) : f^{(k)}_{j,g,h} \neq 0\}$

end

More explicitly, we are interested in the best sparse approximation of the regression function $\mathbb{E}[\mathbf{X}_j \mid \mathbf{X}_{pa_{\mathcal{G}^\pi}(j)} = \mathbf{x}]$ of the form

$$\mathbb{E}[X_k^{(j)} \mid \mathbf{X}_{pa_{\mathcal{G}^\pi}(j)} = \mathbf{x}] \approx \sum_{g \in pa_{\mathcal{G}^\pi}(j)} \sum_{h \in [d_g]} f^{(k)}_{j,g,h}(x_h^{(g)}),$$

for $k \in [d_j]$. The first index j identifies the node whose incoming edges are subject to pruning. The second index g corresponds to the parent groups, and the third index h to the individual entries within each parent group. Finally,

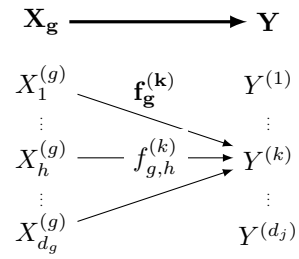


Figure 2: Overview of MURGS notation.

the superscript (k) selects the respective entry within the response group. To facilitate feature selection, we introduce a regularization functional that encourages a common sparsity pattern shared by both predictor and response group elements.

Provided we have found such a functional we remove spurious edges among possible parents $g \in pa_{\mathcal{G}^\pi}(j)$ to obtain a set of relevant edges

$$E^\pi = \{(g, j) : f^{(k)}_{j,g,h} \neq 0, \forall k \in [d_j] \text{ and } h \in [d_g]\}.$$

Algorithm 2 summarizes the procedure.

3.3 SIMULTANEOUS SPARSE BACKFITTING

Before stating the problem more formally, we introduce some relevant notation. If some random variable Z has a distribution P_Z, and f is a function of z, we denote its $L_2(P_Z)$ norm by $\|f\|^2 := \int_{\mathcal{Z}} f^2(z) dP_Z = \mathbb{E}[f^2]$. For an n-dimensional vector $\nu = (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n$ we define $\|\nu\|_2^2 = \frac{1}{n}\sum_{i=1}^n \nu_i^2$ and $\|\nu\|_\infty = \max_{i \in [n]} |\nu_i|$. Consider the p-dimensional random vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$. Denote by $\mathcal{H}_i, i \in [p]$ the Hilbert subspace $L_2(P_{Z_i})$ of P_{Z_i}-measurable functions $f_i(z_i)$ of the scalar variable Z_i with $\mathbb{E}[f_i(z_i)] = 0$. Hence, $\mathcal{H}_i$ is equipped with the inner product $\langle f_i, g_i \rangle = \mathbb{E}[f_i(Z_i) \cdot g_i(Z_i)]$. Whenever a quantity is estimated from a finite sample of size n, we denote this estimate with a hat.

Furthermore, to ease notation in what follows, we take a closer look at node $j \in [p]$ and its parents $pa_j := pa_{\mathcal{G}^\pi}(j)$. We define $Y^{(k)} := X_k^{(j)}$ such that we may drop the corresponding subscripts, i.e. $f^{(k)}_{g,h} := f^{(k)}_{j,g,h}$. Further, we define $f^{(k)}(x) = \sum_{g \in pa_j} \sum_{h \in [d_g]} f^{(k)}_{g,h}(x_h^{(g)})$ for $k \in [d_j]$. Denote by $\mathcal{L}_{f^{(k)}}(x, y) = (y^{(k)} - f^{(k)}(x))^2$ the quadratic loss. We often write $f^{(k)}_{g,h} := f^{(k)}_{g,h}(X_h^{(g)})$. For group $g \in pa_j$, we denote by $\mathbf{f}_g^{(k)} = (f^{(k)}_{g,1}, \ldots, f^{(k)}_{g,d_g})^T$ the vector of group component functions and let $\|\mathbf{f}_g^{(k)}\| = \sqrt{\sum_{h=1}^{d_g} \|f^{(k)}_{g,h}\|^2}$. Figure 2 visualizes the notation.

The main objective is to perform feature selection among groups of variables. We formulate a regularization scheme

to encourage joint functional sparsity. However, the component functions are allowed to vary among response group members and predictor groups while sharing a common sparsity pattern. To achieve this, consider the following regularization functional

$$\Phi^j(f) = \sum_{g \in pa_j} \sqrt{d_g} \max_{k \in [d_j]} \|\mathbf{f}_g^{(k)}\|.$$

The functional $\Phi^j(f)$ combines the sum of sup-norms regularization with the functional version of the $\ell 1/\ell 2$ norms. Similar to the group lasso [Yuan and Lin, 2006], the $\ell 1/\ell 2$ norm induces sparsity at the group level. In turn, the sup-norm penalty encourages sparsity among the d_j response group components. A group of component functions across $k \in [d_j]$ is removed if and only if all involved smooth functions are estimated to be zero. On the other hand, if a component function group g is important with a positive sup-norm for some response k, no additional penalty is imposed on the $d_j - 1$ remaining ones. For feature selection purposes, this is desirable as all function groups remain in the model as long as the sup-norm is positive. This is in contrast to imposing the sum of ℓ_2 norms across the d_j responses which is often applied in multi-task settings [see e.g. Argyriou et al., 2006, Liu et al., 2009b, Wang et al., 2011, Li et al., 2020].

If the response is a scalar random variable, MURGS reduces to the group sparse additive model treated in Yin et al. [2012]. In turn, if all groups $g \in pa_j$ contain only singletons this model reduces to the sparse additive model proposed by [Liu et al., 2007, Ravikumar et al., 2009]. If only the response variable is vector-valued and the remaining groups are singletons this model reduces to the multi-response sparse additive model discussed by Liu et al. [2008].

MURGS can be cast as a penalized M-estimator [Negahban et al., 2012] through the following optimization problem

$$\hat{\mathbf{f}} = \min_{\mathbf{f}: f_{g,h}^{(k)} \in \mathcal{H}_{g,h}} \left\{ \frac{1}{2n} \sum_{\substack{k \in [d_j], \\ i \in [n]}} \mathcal{L}_{f^{(k)}}(\mathbf{x}_i, y_i^{(k)}) + \lambda \Phi^j(f) \right\}$$

with $\lambda > 0$ a regularization parameter.

Block-Coordinate Descent Algorithm In order to solve the optimization problem above, we employ a block-coordinate descent algorithm [see e.g. Hastie et al., 2015]. First, we derive the population version of the estimation problem. This leads to a range of sub-problems in each iteration that can be solved by means of a soft-thresholding update. Similar solutions in multi-task scalar settings have been found in the linear case as well as for sparse additive models [Liu et al., 2009a, 2008]. As is common in backfitting algorithms, we obtain a finite sample version of the algorithm by replacing the conditional expectations with nonparametric smoothers.

Algorithm 3: SoftThresholding update for group j

Input : Partial residual $\hat{R}_g^{(k)}$ for $k \in [d_j]$, smoother matrices $\{S_{g,h} : h \in [d_g]\}$, and tuning parameter λ.

Output : $\hat{\mathbf{f}}_g^{(k)} = (\hat{f}_{g,h}^{(k)})_{h \in [d_g]}$ for $k \in [d_j]$.

Estimate $P_h R_g^{(k)}$ by smoothing: $\hat{P}_h^{(k)} = S_{g,h} \hat{R}_g^{(k)}$.

Estimate $s_g^{(k)} = \|\mathbf{Q} R_g^{(k)}\|$ by:

$\hat{s}_g^{(k)} = \left(1/n \sum_{h \in [d_g]} \|\hat{P}_h^{(k)}\|^2\right)^{1/2}$.

if $\sum_{k \in K} s_g^{(k)} \leq \sqrt{d_g}\lambda$ **then**

  Set $\hat{\mathbf{f}}_g^{(k)} = 0$ for all $k \in [d_j]$.

else

  Order the indices according to $\hat{s}_g^{(k_1)} \geq \hat{s}_g^{(k_2)} \geq \cdots \geq \hat{s}_g^{(k_{d_j})}$.

  Set $m^* = \arg\max_m \frac{1}{m} \left(\sum_{l=1}^{m^*} \hat{s}_g^{(k_l)} - \sqrt{d_g}\lambda \right)$

$$\hat{f}_{g,h}^{(k_i)} = \begin{cases} \hat{P}_h^{(k_i)} & \text{for } i > m^* \\ \frac{1}{m^*} \left[\sum_{l=1}^{m^*} \hat{s}_g^{(k_l)} - \sqrt{d_g}\lambda \right] \frac{\hat{P}_h^{(k_i)}}{\hat{s}_g^{(k_i)}} & \text{o.w.} \end{cases}$$

end

Center $\hat{f}_{g,h}^{(k)}$ by subtracting its mean.

Consider the partial residual $R_g^{(k)} = Y^{(k)} - \sum_{g' \neq g} \sum_{h \in [d_{g'}]} f_{g',h}^{(k)}$ and assume that functions in group g can be fixed. Then the optimization problem on the population level cuts down to

$$\mathbf{f}_g = \underset{\mathbf{f}_g : f_{g,h}^{(k)} \in \mathcal{H}_{g,h}}{\arg\min} \left\{ \frac{1}{2} \mathbb{E}\Big[\sum_{k=1}^{d_j} \big(R_g^{(k)} - \sum_{h \in [d_g]} f_{g,h}^{(k)} \big)^2 \Big] + \lambda \sqrt{d_g} \max_{k \in [d_j]} \|\mathbf{f}_g^{(k)}\| \right\}. \quad (4)$$

Now, we are ready to state the population block update.

Theorem 3.1. *Denote $P_h = \mathbb{E}[\cdot \mid X_h^{(g)}]$ the conditional expectation operator, $\mathbf{Q} = (P_h)_{h \in [d_g]}$ and $s_g^{(k)} = \|\mathbf{Q} R_g^{(k)}\|$. Assume that $\mathbb{E}[f_{g,h'}^{(k)} \mid X_h^{(g)}] = 0$ for all $h' \neq h$, i.e., the covariance among the component functions within groups is zero. Order the indices according to $s_g^{(k_1)} \geq s_g^{(k_2)} \geq \cdots \geq s_g^{(k_{d_j})}$. Then the solution to Eq.* (4) *has coordinate functions given by*

$$f_{g,h}^{(k_i)} = P_h^{(k_i)} R_g^{(k_i)}$$

if $i > m^*$ and by

$$f_{g,h}^{(k_i)} = \frac{1}{m^*} \left[\sum_{l=1}^{m^*} s_g^{(k_l)} - \sqrt{d_g}\lambda \right]_+ \frac{P_h^{(k_i)} R_g^{(k_i)}}{s_g^{(k_i)}}$$

if $i \leq m^*$. Here, $h \in [d_g]$ and

$$m^* = \arg\max_{m \in [d_j]} \frac{1}{m} \left(\sum_{l=1}^{m} s_g^{(k_l)} - \sqrt{d_g}\lambda \right),$$

with $[\cdot]_+$ denoting the positive part function.

The proof involves calculus of variations in Hilbert spaces and is given in Appendix B. The zero covariance assumption is crucial to obtain a closed form update [see Foygel and Drton, 2010]. However, for feature selection we are primarily interested in the case where the sup-norm subdifferential evaluated at $(\|\mathbf{f}_g^{(1)}\|, \ldots, \|\mathbf{f}_g^{(d_j)}\|)^T$ is the zero vector. It turns out that the condition for this case holds in general without any assumptions on the conditional expectation. The following result makes this explicit:

Proposition 1. *$\|\mathbf{f}_g^{(k)}\| = 0$ for all $k \in [d_j]$ if and only if $\sum_{k=1}^{d_j} \|\mathbf{Q}R_g^{(k)}\| \leq \lambda\sqrt{d_g}$.*

Once the stationary condition is derived, the proof to the proposition is straight-forward and can also be found in Appendix B. Based on Theorem 3.1, Algorithms 3 and 4 detail the backfitting algorithm for MURGS in the finite sample setting. We choose the regularization parameter λ

Algorithm 4: Backfitting algorithm

Input : Data $\mathbf{X} = \{\mathbf{X}_g \in \mathbb{R}^{n \times d_g} : g \in pa_j\}$, $Y \in \mathbb{R}^{n \times d_j}$, regularization parameter λ.

Output : Fitted functions $\hat{\mathbf{f}} = (\hat{f}_{g,h}^{(k)})_{g \in pa_j, h \in [d_g], k \in [d_j]}$.

Initialize : $\hat{\mathbf{f}} = \mathbf{0}$ pre-compute the smoother matrices $\{S_{g,h} \in \mathbb{R}^{n \times n} : h \in d_g, g \in pa_j\}$.

for $g \in pa_j$ *until convergence* **do**

(i) Update partial residual $\hat{R}_g^{(k)} = Y^{(k)} - \sum_{g' \neq g} \sum_{h \in [d_{g'}]} \hat{f}_{g',h}^{(k)}$

(ii) $\hat{\mathbf{f}}_g^{(k)} \leftarrow \text{SoftThresholding}(\hat{R}_g^{(k)}, S_{g,:}, \lambda)$

end

by the generalized cross validation (GCV) criterion from Liu et al. [2007] and adapted to the multi-response setting by Liu et al. [2008]. In the multi-response case, the GCV criterion is given by

$$\text{GCV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{k=1}^{d_j} \mathcal{L}_{\hat{f}^{(k)}}(\mathbf{x}_i, y_i^{(k)})}{(n^2 d_j^2 - (nd_j)\text{df}(\lambda))^2},$$

where $\text{df}(\lambda) = d_j \sum_{g \in d_i} \nu_g I(\sum_{k=1}^{d_j} \|\mathbf{f}_g^{(k)}\| \neq 0)$ and $\nu_g = \sum_{h \in [d_g]} \text{tr}(S_{g,h})$ denotes the effective degrees of freedom for the local linear smoother $S_{g,h}$.

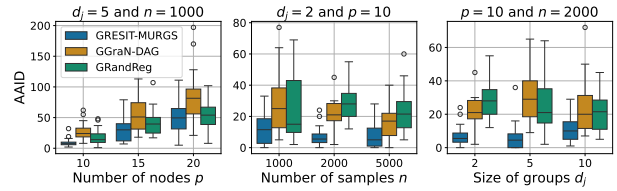


Figure 3: Performance of the algorithms under varying dimensions of the GANM.

4 EXPERIMENTS

Setup In this section we assess and compare the performance of GroupRESIT with MURGS (*GRESIT-MURGS*), GroupRESIT with greedy independence testing (*GRESIT-IND*), a grouped version of the PC algorithm [Spirtes et al., 1993] (GPC) and a grouped version of GraN-DAG [Lachapelle et al., 2020] (*GGraN-DAG*). Furthermore, we report a baseline algorithm that picks a causal order at random and subsequently applies MURGS (*GRandReg*). Implementation details—including our modifications to the PC algorithm and GraN-DAG to accommodate the group setting, along with a description of the metrics, synthetic data, and a table of hyperparameters—are provided in Section C.

Results Results from our experiments are shown in Figures 3 and 4. All metrics are averaged over 20 independent simulation runs. Synthetic data is generated from GANMs where nonlinear functions are generated from weighted sums of Gaussian processes.

Focusing on AAID, Figure 3 illustrates the algorithms' performance across a range of node sizes p, sample sizes n and group sizes d_j. Across all settings, *GRESIT-MURGS* consistently outperforms *GGraN-DAG*, although it too exhibits challenges when applied to very large graphs with limited sample sizes. The performance difference between *GRESIT-MURGS* and *GGraN-DAG* is particularly pronounced when the group size is varied. Indeed, *GRESIT-MURGS* retains its good performance across different group sizes.

In Figure 4, we present a comparison across all considered metrics for two fixed choices of p, d_g, and n. The first row displays classification metrics, where higher values indicate better performance, while the second row shows graph distances, where lower values indicate more accurate graph recovery. In every case, *GRESIT-MURGS* outperforms the other methods. Notably, *GPC* deteriorates as the group size increases, and *GGraN-DAG* suffers similarly, albeit to a lesser extent. Regarding graph distances, the AAID metric is particularly informative, as it reflects the quality of the estimated causal order; here, both GroupRESIT procedures exhibit a clear advantage. In contrast, the pruning phase in *GRESIT-IND* proves ineffective, as evidenced by high SHD.

Overall, the combination of flexible neural networks and

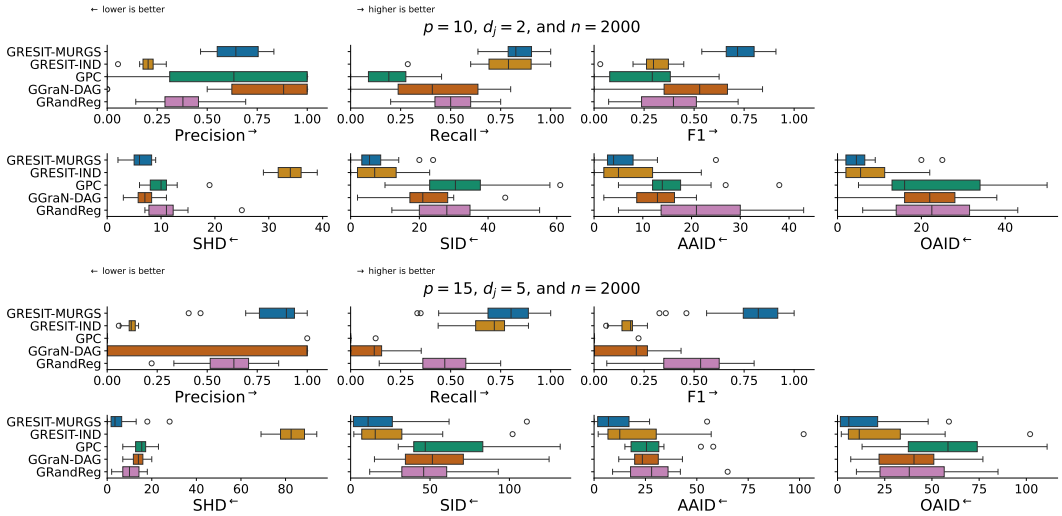


Figure 4: Simulation results based on 20 repetitions.

nonparametric independence tests proves highly effective in estimating a causal order, provided that the sample size is sufficiently large. Moreover, MURGS demonstrates a robust capability for feature selection even in the presence of large group sizes and numerous candidate parents.

5 REAL MANUFACTURING DATA APPLICATION

We use real-world data from a highly automated production line. Each unit travels along a conveyor belt and passes through several process cells $C_1, \ldots, C_6$ (see Figure 5). In each cell, various automated manufacturing processes—such as staking, welding, etc.—are performed. For example, cell C_1 executes ten distinct tasks, whereas cell C_4 performs only two. For every process, a set of physical measurements is extracted and recorded, denoted by $\mathbf{X}_i$ for cell C_i. Their dimensionality range from one to 13. In a staking process, for instance, measurements like the press-in force and the maximum attained force are obtained to ensure production quality. These individual process measurements form the causal groups of interest. We observe 19 groups, which together comprise 121 features. The full dataset contains approximately $35,000$ parts, of which we randomly sample $5,000$ for computational tractability. Access to complete causal ground truth in real-world applications is typically prohibitive or unattainable [Göbler et al., 2024]. However, in production lines, partial causal ordering is available through domain knowledge due to the sequential nature of unit flow: processes in cell C_i can affect only those in cell C_j with $i \leq j$. Thus, we can partially assess learned graphs by their edge orientation.

Figure 5 shows edges learned by *GRESIT-MURGS*. Notably, only one edge conflicts with the cell-implied partial order.

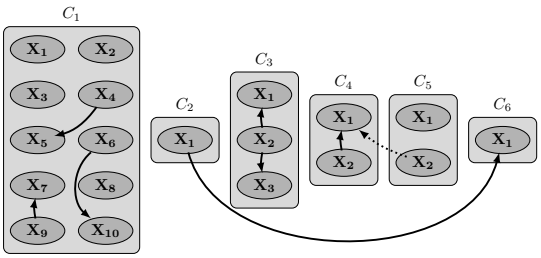


Figure 5: Learned causal edges from the real-world dataset using *GRESIT-MURGS*

.

Conversely, *GPC* yields many undirected edges and two directed edges, one violating the cell arrangement, while *GGraN-DAG* identifies just two edges in cell C_3. Lastly, applying *GroupDirectLiNGAM*Entner and Hoyer [2012] with MURGS pruning (SectionC.3) results exclusively in edges violating the partial ordering (see Section D). Thus, *GRESIT-MURGS* produces the most meaningful causal structure in this setting.

6 CONCLUSIONS

In this work, we develop methodology and procedures for nonlinear causal discovery for grouped data. Building upon the RESIT algorithm introduced by Peters et al. [2014], we propose *MURGS*, a pruning procedure to perform model selection after a causal order has been learned. We derive a closed-form backfitting update within a block coordinate descent framework. Extensive experiments on synthetic data demonstrate that *GRESIT* combined with *MURGS* achieves superior performance, clearly outperforming other state-of-the-art causal discovery algorithms. More-

over, evaluation on real-world manufacturing data with partially known causal structure further substantiates the practical applicability of our method. We have implemented all methods presented in this paper in a Python library, which is publicly available at https://github.com/boschresearch/gresit.

The current limitations of our work include the assumptions on nonlinearity and differentiability imposed on the function class $\mathcal{F}$. Next to rather standard assumptions (acyclicity and i.i.d. data) we assume the absence of unobserved confounding. Addressing these limitations presents promising directions for future work.

Overall, our contributions offer a flexible framework for causal discovery in grouped data under the GANM, laying the groundwork for both theoretical developments and practical applications in complex real-world systems.

Acknowledgements

The authors gratefully acknowledge Andrew McCormack, Martin Roth, and Hongjian Shi for their insightful discussions and valuable feedback. We also thank the anonymous referees whose thoughtful suggestions greatly improved the clarity and quality of this work. Mathias Drton acknowledges funding from the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts.

References

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):12172–12179, 2023. doi: 10.1609/aaai.v37i10.26435. URL https://ojs.aaai.org/index.php/AAAI/article/view/26435.

Stephen Antonoplis. Studying socioeconomic status: Conceptual problems and an alternative path forward. *Perspectives on Psychological Science*, 18(2):275–292, 2022. doi: 10.1177/17456916221093615. URL http://dx.doi.org/10.1177/17456916221093615.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/0afa92fc0f8a9cf051bf2961b06ac56b-Paper.pdf.

Kenneth A. Bollen. *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1989. doi: 10.1002/9781118619179. URL https://doi.org/10.1002/9781118619179. A Wiley-Interscience Publication.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi: 10.1037/h0046016. URL http://dx.doi.org/10.1037/h0046016.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 361–369, Cadiz, Spain, 2016. PMLR. URL https://proceedings.mlr.press/v51/chalupka16.html.

David Maxwell Chickering. Optimal structure identification with greedy search. volume 3, pages 507–554. 2003. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717. Computational learning theory.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741–3782, 2014.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957. URL http://dx.doi.org/10.1037/h0040957.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393, 2017. doi: 10.1146/annurev-statistics-060116-053803. URL https://doi.org/10.1146/annurev-statistics-060116-053803.

Doris Entner and Patrik O. Hoyer. Estimating a causal order among groups of variables in linear models. In Alessandro E. P. Villa, Włodzisław Duch, Péter Érdi, Francesco Masulli, and Günther Palm, editors, *Artificial Neural Networks and Machine Learning – ICANN 2012*, pages 84–91, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

P. Erdös and A. Rényi. *On the evolution of random graphs*, pages 38–82. Princeton University Press, 2011. doi: 10.1515/9781400841356.38. URL http://dx.doi.org/10.1515/9781400841356.38.

Massimo Fornasier and Holger Rauhut. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613, 2008. doi: 10.1137/0606668909. URL https://doi.org/10.1137/0606668909.

Rina Foygel and Mathias Drton. Exact block-wise optimization in group lasso and sparse group lasso for linear regression, 2010. URL https://arxiv.org/abs/1010.3320.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/3a0772443a0739141292a5429b952fe6-Paper.pdf.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Martin Roth, and Steffen Sonntag. causalAssembly: Generating realistic production data for benchmarking causal discovery. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR, 2024. URL https://proceedings.mlr.press/v236/gobler24a.html.

Daniel Greenfeld and Uri Shalit. Robust learning with the Hilbert-schmidt independence criterion. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3759–3768. PMLR, 2020. URL https://proceedings.mlr.press/v119/greenfeld20a.html.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, volume 3734 of *Lecture Notes in Comput. Sci.*, pages 63–77. Springer, Berlin, 2005. doi: 10.1007/11564089_7. URL https://doi.org/10.1007/11564089_7.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*, volume 143 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015. The lasso and generalizations.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: A gadjid for causal structure learning. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=jO5UNNrjJr.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning*, pages 479–486, Madison, WI, USA, 2010. Max-Planck-Gesellschaft, International Machine Learning Society.

Genta Kikuchi and Shohei Shimizu. Structure learning for groups of variables in nonlinear time-series data with location-scale noise. In Erich Kummerfeld, Sisi Ma, Eric Rawls, and Bryan Andrews, editors, *Proceedings of the 2023 Causal Analysis Workshop Series*, volume 223 of *Proceedings of Machine Learning Research*, pages 20–39. PMLR, 2023. URL https://proceedings.mlr.press/v223/kikuchi23a.html.

Adam Kohn, Anna I. Jasper, João D. Semedo, Evren Gokcen, Christian K. Machens, and Byron M. Yu. Principles of corticocortical communication: Proposed schemes and design considerations. *Trends in Neurosciences*, 43(9):725–737, 2020. doi: https://doi.org/10.1016/j.tins.2020.07.001. URL https://www.sciencedirect.com/science/article/pii/S016622362030165X.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Lei Li, Deborah Chang, Lei Han, Xiaojian Zhang, Joseph Zaia, and Xiu-Feng Wan. Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza a(h1n1) virus using mutations and variations in glycosylation of hemagglutinin. *BMC Bioinformatics*, 21(1), 2020. doi: 10.1186/s12859-020-3527-5. URL http://dx.doi.org/10.1186/s12859-020-3527-5.

Han Liu, Larry Wasserman, John Lafferty, and Pradeep Ravikumar. Spam: Sparse additive models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/42e7aaa88b48137a16a1acd04ed91125-Paper.pdf.

Han Liu, Larry Wasserman, and John Lafferty. Nonparametric regression and classification with joint sparsity constraints. In D. Koller, D. Schuurmans, Y. Bengio,

and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/6faa8040da20ef399b63a72d0e4ab575-Paper.pdf.

Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 649–656, 2009a. doi: 10.1145/1553374.1553458. URL http://dx.doi.org/10.1145/1553374.1553458.

Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 339–348, 2009b.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 745–752, 2009. doi: 10.1145/1553374.1553470. URL https://doi.org/10.1145/1553374.1553470.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012. doi: 10.1214/12-STS400. URL https://doi.org/10.1214/12-STS400.

Stefano Panzeri, Christopher D. Harvey, Eugenio Piasini, Peter E. Latham, and Tommaso Fellin. Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507, 2017. doi: https://doi.org/10.1016/j.neuron.2016.12.036. URL https://www.sciencedirect.com/science/article/pii/S0896627316310091.

Pekka Parviainen and Samuel Kaski. Learning structures of Bayesian networks for variable groups. *Internat. J. Approx. Reason.*, 88:110–127, 2017. doi: 10.1016/j.ijar.2017.05.006. URL https://doi.org/10.1016/j.ijar.2017.05.006.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Comput.*, 27(3):771–799, 2015. doi: 10.1162/neco_a_00708. URL https://doi.org/10.1162/neco_a_00708.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 589–598, 2011.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053, 2014.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):1009–1030, 2009. doi: 10.1111/j.1467-9868.2009.00718.x. URL https://doi.org/10.1111/j.1467-9868.2009.00718.x.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 1–13. NeurIPS Proceedings, 2021.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998. doi: 10.1007/978-3-642-02431-3. URL https://doi.org/10.1007/978-3-642-02431-3.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753. PMLR, 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

P. K. Rubenstein*, S. Weichwald*, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 11, 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf. *equal contribution.

Jakob Runge, Vladimir Petoukhov, Jonathan F. Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6(1), 2015. doi: 10.1038/ncomms9502. URL http://dx.doi.org/10.1038/ncomms9502.

João D Semedo, Evren Gokcen, Christian K Machens, Adam Kohn, and Byron M Yu. Statistical methods for dissecting interactions between brain areas. *Current Opinion in Neurobiology*, 65:59–69, 2020. doi: https://doi.org/10.1016/j.conb.2020.09.009.

URL https://www.sciencedirect.com/science/article/pii/S0959438820301367. Whole-brain interactions between neural circuits.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL http://jmlr.org/papers/v12/shimizu11a.html.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. doi: 10.1007/978-1-4612-2748-9. URL https://doi.org/10.1007/978-1-4612-2748-9.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks, 2012. URL https://arxiv.org/abs/1207.1429.

Kento Uemura, Takuya Takagi, Kambayashi Takayuki, Hiroyuki Yoshida, and Shohei Shimizu. A multivariate causal discovery based on post-nonlinear model. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 826–839. PMLR, 2022. URL https://proceedings.mlr.press/v177/uemura22a.html.

Matej Vuković and Stefan Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1), 2022. doi: 10.3390/jmmp6010010. URL https://www.mdpi.com/2504-4494/6/1/10.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Vector causal inference between two groups of variables. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):12305–12312, 2023. doi: 10.1609/aaai.v37i10.26450. URL https://ojs.aaai.org/index.php/AAAI/article/view/26450.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1), 2024. doi: doi:10.1515/jci-2023-0041. URL https://doi.org/10.1515/jci-2023-0041.

Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L. Risacher, Andrew J. Saykin, Li Shen, and For the Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2011. doi: 10.1093/bioinformatics/btr649. URL https://doi.org/10.1093/bioinformatics/btr649.

Junming Yin, Xi Chen, and Eric P Xing. Group sparse additive models. *Proc. Int. Conf. Mach. Learn.*, 2012: 871–878, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006. doi: 10.1111/j.1467-9868.2005.00532.x. URL https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655, 2009.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Jakob Zscheischler, Dominik Janzing, and Kun Zhang. Testing whether linear equations are causal: a free probability theory approach. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 839–848, 2011.

Nonlinear Causal Discovery for Grouped Data (Supplementary Material)

Konstantin Göbler[1,4] **Tobias Windisch**[3] **Mathias Drton**[1,2]

[1]Department of Mathematics, TUM School of Computation, Information and Technology, Technical University of Munich
[2]Munich Center for Machine Learning
[3]University of Applied Sciences Kempten, Faculty of Mechanical Engineering
[4]Robert Bosch GmbH

A PROOFS OF IDENTIFIABILITY

A.1 PROOF OF THEOREM 2.1

Proof. If $\mathcal{G}_0$ is the empty graph we have $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$. If the graph is not empty, and we have $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ causal minimality is violated. Hence, we assume that the graph is not empty and $\mathbf{X}_1 \not\perp\!\!\!\perp \mathbf{X}_2$. The joint density has the following form

$$p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2) = p_{\mathbf{X}_1}(\mathbf{x}_1)p_{\mathbf{N}_2}(\mathbf{x}_2 - f_2(\mathbf{x}_1)).$$

Let us assume $\mathcal{G}_0$ is not identifiable from $P_{\mathbf{X}}$ alone, i.e., there must exist a backward model of the same form

$$p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2) = p_{\mathbf{X}_2}(\mathbf{x}_2)p_{\mathbf{N}_1}(\mathbf{x}_1 - f_1(\mathbf{x}_2)).$$

Define

$$\pi_1(\mathbf{x}_1,\mathbf{x}_2) := \nu(\mathbf{x}_2 - f_2(\mathbf{x}_1)) + \xi(\mathbf{x}_1) \tag{5}$$

and

$$\pi_2(\mathbf{x}_1,\mathbf{x}_2) := \tilde{\nu}(\mathbf{x}_1 - f_1(\mathbf{x}_2)) + \eta(\mathbf{x}_2), \tag{6}$$

where $\tilde{\nu} := \log p_{\mathbf{N}_1}$ and $\eta := \log p_{\mathbf{X}_2}$. Clearly, we have that $\pi_1(\mathbf{x}_1,\mathbf{x}_2) = \pi_2(\mathbf{x}_1,\mathbf{x}_2) = \log p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)$.

Considering first the backwards model (Eq. (6)) we derive the gradient $\nabla_{\mathbf{x}_1}\pi_2(\mathbf{x}_1,\mathbf{x}_2)$ with respect to $\mathbf{x}$, i.e.,

$$\nabla_{\mathbf{x}_1}\pi_2(\mathbf{x}_1,\mathbf{x}_2) = \nabla\tilde{\nu}(\mathbf{u}), \tag{7}$$

where $\tilde{\mathbf{u}} := \tilde{\mathbf{u}}(\mathbf{x}_1,\mathbf{x}_2) := \mathbf{x}_1 - f_1(\mathbf{x}_2)$. Then, the second derivatives take the following form

$$D_{\mathbf{x}_1\mathbf{x}_2}\pi_2(\mathbf{x}_1,\mathbf{x}_2) = -\mathbf{J}_{f_1}(\mathbf{X}_2)^\top \mathbf{H}_{\tilde{\nu}}(\tilde{\mathbf{u}}), \tag{8}$$

and

$$D_{\mathbf{x}_1\mathbf{x}_1}\pi_2(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{H}_{\tilde{\nu}}(\tilde{\mathbf{u}}), \tag{9}$$

with Jacobian $\mathbf{J}_{f_1} \in \mathbb{R}^{d_{x_1}\times d_{x_2}}$ and Hessian $\mathbf{H}_{\tilde{\nu}} \in \mathbb{R}^{d_{x_1}\times d_{x_1}}$. Since we have assumed that f_1 and f_2 are three times continuously differentiable the Hessian is symmetric (Schwarz's theorem) and invertible such that

$$\begin{aligned} Q_2(\mathbf{x}_1,\mathbf{x}_2) &:= \left(D_{\mathbf{x}\mathbf{x}}\pi_2(\mathbf{x}_1,\mathbf{x}_2)\right)^{-1}\left(D_{\mathbf{x}_1\mathbf{x}_2}\pi_2(\mathbf{x}_1,\mathbf{x}_2)\right)^\top \\ &= \mathbf{H}_{\tilde{\nu}}(\tilde{\mathbf{u}})^{-1}(-\mathbf{J}_{f_1}(\mathbf{x}_2)^\top\mathbf{H}_{\tilde{\nu}}(\tilde{\mathbf{u}}))^\top \\ &= -\mathbf{J}_{f_1}(\mathbf{x}_2), \end{aligned}$$

which does not depend on $\mathbf{x}_1$.

Now, we repeat the above steps for Eq. (5), i.e.,

$$\nabla_{\mathbf{x}_1}\pi_1(\mathbf{x}_1,\mathbf{x}_2) = -\mathbf{J}_{f_2}(\mathbf{x}_1)^\top\nabla\nu\big(\mathbf{u}\big) + \nabla\xi(\mathbf{x}_1), \tag{10}$$

and

$$D_{\mathbf{x}_1\mathbf{x}_2}\pi_1(\mathbf{x}_1,\mathbf{x}_2) = -\mathbf{J}_{f_2}(\mathbf{x}_1)^\top\mathbf{H}_\nu\big(\mathbf{u}\big), \tag{11}$$

where $\mathbf{J}_{f_2}(\mathbf{x}_1) \in \mathbb{R}^{d_{x_2}\times d_{x_1}}$, $\nabla\xi(\mathbf{x}_1) \in \mathbb{R}^{d_{x_1}}$, and $\mathbf{H}_\nu\big(\mathbf{u}\big) \in \mathbb{R}^{d_{x_2}\times d_{x_2}}$. Finally,

$$\begin{aligned} D_{\mathbf{x}_1\mathbf{x}_1}\pi_1(\mathbf{x}_1,\mathbf{x}_2) &= \mathbf{H}_\xi(\mathbf{x}_1) - \mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu\big(\mathbf{u}\big)] && (12) \\ &\quad + \mathbf{J}_{f_2}(\mathbf{x}_1)^\top\mathbf{H}_\nu\big(\mathbf{u}\big)\mathbf{J}_{f_2}(\mathbf{x}_1), && (13) \end{aligned}$$

where $\mathbf{H}_\xi(\mathbf{x}_1) \in \mathbb{R}^{d_{x_1}\times d_{x_1}}$ and the Hessian $\mathbf{H}_{f_2} \in \mathbb{R}^{d_{x_2}\times d_{x_1}\times d_{x_1}}$ is a third-order tensor such that the contraction with the vector $\nabla\nu\big(\mathbf{u}\big)$ may be written as $\mathbf{H}_{f_2}[\nabla\nu\big(\mathbf{u}\big)]$, i.e.

$$\big(\mathbf{H}_{f_2}[\nabla\nu\big(\mathbf{u}\big)]\big)_{ik} = \sum_{j=1}^{d_{x_2}} \frac{\partial^2 f_j(\mathbf{x}_1)}{\partial x_{1i}\partial x_{1k}}\ \big[\nabla\nu\big(\mathbf{u}\big)\big]_j.$$

We find

$$\begin{aligned} Q_1(\mathbf{x}_1,\mathbf{x}_2) &:= (D_{\mathbf{x}_1\mathbf{x}_1}\pi_1(\mathbf{x}_1,\mathbf{x}_2))^{-1}(D_{\mathbf{x}_1\mathbf{x}_2}\pi_1(\mathbf{x}_1,\mathbf{x}_2)) \\ &= \Big[\mathbf{H}_\xi(\mathbf{x}_1) - \mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu(\mathbf{u})] \\ &\quad + \mathbf{J}_{f_2}(\mathbf{x}_1)^\top\mathbf{H}_\nu\big(\mathbf{u}\big)\mathbf{J}_{f_2}(\mathbf{x}_1)\Big]^{-1} \\ &\quad [-\mathbf{J}_{f_2}(\mathbf{x}_1)^\top\mathbf{H}_\nu(\mathbf{u})]. \end{aligned}$$

Recall that $Q_2(\mathbf{x}_1,\mathbf{x}_2) = Q_2(\mathbf{x}_2)$ is essentially a function only of $\mathbf{x}_2$ such that $D_{\mathbf{x}_1}Q_2(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{0} \in \mathbb{R}^{d_{x_1}\times d_{x_1}\times d_{x_2}}$.

Since $Q_1(\mathbf{x}_1, \mathbf{x}_2) = Q_2(\mathbf{x}_1, \mathbf{x}_2)$, taking the derivative with respect to $\mathbf{x}_1$ yields

$$\begin{aligned} D_{\mathbf{x}_1} Q_1(\mathbf{x}_1, \mathbf{x}_2) &= D_{\mathbf{x}_1}\big[(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1(\mathbf{x}_1, \mathbf{x}_2))^{-1} \\ &\quad (D_{\mathbf{x}_1\mathbf{x}_2}\pi_1(\mathbf{x}_1, \mathbf{x}_2))\big] \\ &= \mathbf{0} \in \mathbb{R}^{d_{x_1} \times d_{x_1} \times d_{x_2}}. \end{aligned}$$

Note that each slice $\partial Q_1(\mathbf{x}_1, \mathbf{x}_2)/\partial x_{1k}$ is matrix-valued, and d_{x_1}-many such slices exist. Thus, we apply the product rule component-wise for each slice of the tensor and subsequently stack them together, i.e.

$$\begin{aligned} D_{\mathbf{x}_1} Q_1(\mathbf{x}_1, \mathbf{x}_2) &= D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1(\mathbf{x}_1, \mathbf{x}_2)^{-1}) \\ &\quad D_{\mathbf{x}_1\mathbf{x}_2}\pi_1(\mathbf{x}_1, \mathbf{x}_2) \\ &\quad + (D_{\mathbf{x}_1\mathbf{x}_1}\pi_1(\mathbf{x}_1, \mathbf{x}_2))^{-1} \\ &\quad D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_2}\pi_1(\mathbf{x}_1, \mathbf{x}_2)). \end{aligned}$$

Using the identity

$$D_{\mathbf{x}_1}(\mathbf{A})^{-1} = -(\mathbf{A})^{-1} D_{\mathbf{x}_1}\mathbf{A}(\mathbf{A})^{-1},$$

where $\mathbf{A} := \mathbf{A}(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{d_{x_1} \times d_{x_1}}$. Therefore, we obtain

$$\begin{aligned} &-(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 \\ &\qquad = -(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_2}\pi_1), \end{aligned}$$

where we drop the arguments of $\pi_1(\mathbf{x}_1, \mathbf{x}_2)$, i.e., $\pi_1 := \pi_1(\mathbf{x}_1, \mathbf{x}_2)$ to improve readability. Making sure that all of the matrix-tensor products act on the appropriate indices of the tensors, the expression simplifies to

$$D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 = D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_2}\pi_1).$$

Note that only $D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)$ contains third order derivatives of the log marginal ξ. Let us state this more explicitly.

$$\begin{aligned} D_{\mathbf{x}_1}(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1) &= D_{\mathbf{x}_1}\Big[\mathbf{H}_\xi(\mathbf{x}_1) - \mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu(\mathbf{u})] \\ &\quad + \mathbf{J}_{f_2}(\mathbf{x}_1)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_2}(\mathbf{x}_1)\Big] \\ &= D_{\mathbf{x}_1}\mathbf{H}_\xi(\mathbf{x}_1) - D_{\mathbf{x}_1}(\mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu(\mathbf{u})]) \\ &\quad + D_{\mathbf{x}_1}(\mathbf{J}_{f_2}(\mathbf{x}_1)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_2}(\mathbf{x}_1)). \end{aligned}$$

And finally we have

$$\begin{aligned} &D_{\mathbf{x}_1}\mathbf{H}_\xi(\mathbf{x}_1)(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 \\ &\quad = D_{\mathbf{x}_1} D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 \Big[D_{\mathbf{x}_1}\big(\mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu(\mathbf{u})]\big) \\ &\quad - D_{\mathbf{x}_1}\big(\mathbf{J}_{f_2}(\mathbf{x}_1)^\top \mathbf{H}_\nu(\mathbf{u}) \mathbf{J}_{f_2}(\mathbf{x}_1)\big)\Big] \\ &\quad (D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1} D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 \end{aligned} \tag{14}$$

where the remaining second order derivatives of the log marginal ξ are contained in the expression for $D_{\mathbf{x}_1\mathbf{x}_1}\pi_1$. This contradicts the assumption that $P_\mathbf{X}$ is generated from a *identifiable bivariate* GANM. □

A.2 PROOF OF COROLLARY 1

Proof. Suppose there are two identifiable GANMs that both induce the distribution $P_\mathbf{X}$ with DAGs $\mathcal{G}_0$ and $\mathcal{G}_0'$, respectively. For any two groups $\mathbf{X}_Q, \mathbf{X}_R$ that satisfy Proposition 29 in Peters et al. [2014] we consider the set of parents without R in $\mathcal{G}_0$, i.e, $pa_Q := pa_{\mathcal{G}_0}(Q) \setminus R$ and the set of parents without Q in $\mathcal{G}_0'$, i.e, $pa_R := pa_{\mathcal{G}_0'}(R) \setminus Q$. Denote their union by $S := pa_Q \cup pa_R$. For any $s = (q, r)$ we write $\mathbf{X}_Q^* = \mathbf{X}_Q \mid_{S=s}$ and $\mathbf{X}_R^* = \mathbf{X}_R \mid_{R=r}$. From Lemma 2.2, we have that $\mathbf{N}_Q \perp\!\!\!\perp \mathbf{X}_R, \mathbf{X}_S$ and $\mathbf{N}_R \perp\!\!\!\perp \mathbf{X}_Q, \mathbf{X}_S$. Thus, by Lemma 2 in Peters et al. [2011], we have the following bivariate GANM in $\mathcal{G}_0$

$$\mathbf{X}_Q^* = f_Q(\mathbf{X}_q, \mathbf{X}_R^*) + \mathbf{N}_Q, \quad \mathbf{N}_Q \perp\!\!\!\perp \mathbf{X}_R^*,$$

and in $\mathcal{G}_0'$

$$\mathbf{X}_R^* = f_R(\mathbf{X}_r, \mathbf{X}_Q^*) + \mathbf{N}_R, \quad \mathbf{N}_R \perp\!\!\!\perp \mathbf{X}_Q^*,$$

However this is a contradiction since in Corollary 1 we can choose any $s = (q, r)$ that ensures that the bivariate GANMs are identifiable. □

A.3 PROOF OF LEMMA 2.2

Proof. Write $\mathbf{S} = (S_1, \ldots, S_k)$ such that $\mathbf{S} = (f_{S_1}(\mathbf{X}_{pa(S_1)}, \mathbf{N}_{S_1}), \ldots, f_{S_k}(\mathbf{X}_{pa(S_k)}, \mathbf{N}_{S_k}))$. It takes finitely many steps to recursively substitute the parents of $S_i, i \in [k]$ by the corresponding structural equation such that $\mathbf{S} = f(\mathbf{N}_{A_1}, \ldots, \mathbf{N}_{A_l})$ with $\{A_1, \ldots, A_l\}$ the set of all ancestors of nodes in $\mathbf{S}$ that do not contain the node g. The statement follows from the joint independence of the noise variables across groups. □

B PROOF OF BACKFITTING UPDATE

Lemma B.1. *The stationary condition of Problem* (4) *with respect to* $\mathbf{f}_g$ *is given by*

$$f_{g,h}^{(k)} + \sum_{h' \in [d_g]: h' \neq h} P_h f_{g,h'}^{(k)} - P_h R_g^{(k)} + \lambda\sqrt{d_g} u^{(k)} v_h^{(k)} = 0, \tag{15}$$

for all $h \in [d_g]$ *where* $u^{(k)}$ *are scalars and* $\mathbf{v}_g^{(k)} = (v_h^{(k)})_{h \in [d_g]}$ *is a vector of measurable functions of* $X_h^{(g)}$, *with*

$$(u^{(1)}, \ldots, u^{(d_j)})^T \in \partial\|\cdot\|_\infty\big|_{(\|\mathbf{f}_g^{(1)}\|, \ldots, \|\mathbf{f}_g^{(d_j)}\|)^T}, \tag{16}$$

and

$$\mathbf{v}_g^{(k)} \in \partial\|\mathbf{f}_g^{(k)}\|, \tag{17}$$

for $k = 1, \ldots d_j$. *The subdifferential of the sup-norm evaluated at* $(\|\mathbf{f}_g^{(1)}\|, \ldots, \|\mathbf{f}_g^{(d_j)}\|)^T$ *lies in the* d_j*-dimensional Euclidean space.*

Proof. Since both the loss function and the regularization term are convex, the solution to the objective function in (4) can be characterized by the Karush-Kuhn-Tucker conditions. We investigate the subdifferential for the loss and the regularization term separately. For readability, we omit the argument of the component function when it is clear from the context. Starting with the loss function we define

$$L(\mathbf{f}_g^{(k)}) := \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{d_j}\left(R_g^{(k)} - \sum_{h\in[d_g]} f_{g,h}^{(k)}\right)^2\right].$$

Consider a perturbation of $L(\mathbf{f}_g^{(k)})$ along the direction

$$\psi_g^{(k)} = \left(\psi_{g,h}^{(k)} \in \mathcal{H}_{g,h}^{(k)}\right)_{h\in[d_g]}$$

such that

$$\begin{aligned}
&\lim_{\tau\to 0}\frac{L(\mathbf{f}_g^{(k)} + \tau\psi_g^{(k)}) - L(\mathbf{f}_g^{(k)})}{\tau} \\
&= \sum_{h\in[d_g]}\mathbb{E}\Big[(\sum_{h'\in[d_g]} f_{g,h'}^{(k)} - R_g^{(k)})\psi_{g,h}^{(k)}\Big] \\
&= \sum_{h\in[d_g]}\mathbb{E}\left[\mathbb{E}\Big[\sum_{h'\in[d_g]} f_{g,h'}^{(k)} - R_g^{(k)} \mid X_h^{(g)}\Big]\psi_{g,h}^{(k)}\right] \\
&= \sum_{h\in[d_g]}\left\langle\mathbb{E}\Big[\sum_{h'\in[d_g]} f_{g,h'}^{(k)} - R_g^{(k)} \mid X_h^{(g)}\Big], \psi_{g,h}^{(k)}\right\rangle.
\end{aligned}$$

The second equation follows from the law of iterated expectation and the third from expressing the resulting expectation as an inner product in the Hilbert space $\mathcal{H}_{g,h}^{(k)}$. The gradient of $L(\mathbf{f}_g^{(k)})$ is

$$\nabla L(\mathbf{f}_g^{(k)}) = \left[\mathbb{E}\Big[\sum_{h'\in[d_g]} f_{g,h'}^{(k)} - R_g^{(k)} \mid X_h^{(g)}\Big]\right]_{h\in[d_g]}.$$

The subdifferential of $\Phi_{\text{group}}^{d_j}(f)$ is given by

$$\partial\Phi_{\text{group}}^{d_j}(f) = \lambda\sqrt{d_g}u_{hk}v_{hk}, \quad \forall h\in[d_g],$$

where $(u_{h1},\dots,u_{hd_j})^T \in \partial\|\cdot\|_\infty|_{(\|\mathbf{f}_g^{(1)}\|,\dots,\|\mathbf{f}_g^{(d_j)}\|)^T}$ and $v_{hk} \in \partial\|\mathbf{f}_g^{(k)}\|$.

Isolating $f_{g,h}^{(k)}(X_h^{(g)}) = \mathbb{E}[f_{g,h}^{(k)}(X_h^{(g)}) \mid X_h^{(g)}]$ and using the conditional expectation operator $\mathbb{E}[\,\cdot \mid X_h^{(g)}]$ for the remaining terms yields the expression for the stationary condition. □

The following two Lemmas characterize the subdifferential of sup-norms (Lemma B.2, proof is provided in [Rockafellar and Wets, 1998, Chapter 8]) and that of the Euclidean norm (Lemma B.3).

Lemma B.2. *The subdifferential of $\|\cdot\|_\infty$ in $\mathbb{R}^{d_j}$ is*

$$\partial\|\cdot\|_\infty|_x = \begin{cases}\{\eta : \|\eta\|_1 \le 1\} & \text{if } \boldsymbol{x} = \mathbf{0} \\ \text{conv}\{\text{sign}(x_k)e_k : |x_k| = \|x\|_\infty\} & \text{o.w.},\end{cases} \tag{18}$$

where conv(A) denotes the convex hull of set A and e_k is the kth canonical unit vector in $\mathbb{R}^{d_j}$.

Lemma B.3. *The subdifferential of $\|\mathbf{f_g}\|$ is*

$$\partial\|f_g\| = \begin{cases}\{f_j/\|\mathbf{f_g}\|\}_{\mathbf{j}\in\mathbf{g}} & \text{if } \|\mathbf{f_g}\| \ne \mathbf{0} \\ \{\mathbf{v_g} : \|\mathbf{v_g}\| \le \mathbf{1}\} & \text{if } \|\mathbf{f_g}\| = \mathbf{0}\end{cases} \tag{19}$$

The proof proceeds by considering three cases for the sup-norm subdifferential evaluated at $(\|\mathbf{f}_g^{(1)}\|,\dots,\|\mathbf{f}_g^{(d_j)}\|)^T$: (1) $\|\mathbf{f}_g^{(k)}\| = 0$ for all $k = 1,\dots,d_j$; (2) there exists a unique k, such that $\|\mathbf{f}_g^{(k)}\| = \max_{k'=1,\dots,d_j}\|\mathbf{f}_g^{(k')}\|$; (3) There exist at least two $k \ne k'$, such that $\|\mathbf{f}_g^{(k)}\| = \|\mathbf{f}_g^{(k')}\| = \max_{m=1,\dots,d_j}\|\mathbf{f}_g^{(m)}\|$

We begin with the proof of Proposition 1, i.e., the case where $\sum_{k=1}^{d_j}\|\mathbf{Q}R_g^{(k)}\| \le \lambda\sqrt{d_g}$ and show that $\|\mathbf{f}_g^{(k)}\| = 0$ must be a solution.

Proof. From Lemma B.1 we know that if $\|\mathbf{f}_g^{(k)}\| = 0$ then $\|\mathbf{u}\|_1 \le 1$ and $\|\mathbf{v}_g^{(k)}\| \le 1$. It follows that

$$P_h R_g^{(k)} = \lambda\sqrt{d_g}u^{(k)}v_h^{(k)}$$

$$\sum_{k=1}^{d_j}\sqrt{\sum_{h\in[d_g]}\mathbb{E}[(P_h R_g^{(k)})^2]} \le \lambda\sqrt{d_g}.$$

On the other hand, we also know from Lemma B.1 that $\|\mathbf{f}_g^{(k)}\| = 0$ if and only if $\exists u^{(1)},\dots,u^{(d_j)}$ such that $\sum_{k=1}^{d_j}|u|^{(k)} \le 1$ and $\exists v_1^{(k)},\dots,v_{d_g}^{(k)}$ such that $\sqrt{\sum_{h=1}^{d_g}(v_h^2)^{(k)}} \le 1$ and

$$\lambda\sqrt{d_g}u^{(k)}v_h^{(k)} = P_h R_g^{(k)}.$$

Then, if $\sum_{k=1}^{d_j}\|\mathbf{Q}R_g^{(k)}\| \le \lambda\sqrt{d_g}$, choosing $u^{(k)}$ and $v_h^{(k)}$ as above guarantees that $\sum_{k=1}^{d_j}|u|^{(k)} \le 1$ and $\sum_{k=1}^{d_j}|u|^{(k)} \le 1$, therefore $\|\mathbf{f}_g^{(k)}\| = 0$. □

If we continue to allow the within-group covariances to be nonzero, we obtain the following known result.

Lemma B.4. *Consider the case where there exists a unique k at which the sup-norm is attained. Then the group SpAM by [Yin et al., 2012] is recovered.*

Proof. Note that we must have that $\sum_{k=1}^{d_j}\|\mathbf{Q}R_g^{(k)}\| > \lambda\sqrt{d_g}$ otherwise all $f_{g,h}^{(k)} = 0, \forall h\in[d_g], k\in[d_j]$.

Denote $k_1 \in [d_j]$ the unique k that attains the sup-norm, then $\|\mathbf{f}_g^{(k_1)}\| > \|\mathbf{f}_g^{(k)}\|$ for all $k \neq k_1$. Consequently, the subdifferential of the sup-norm becomes $\partial\|\cdot\|_\infty|_{(\|\mathbf{f}_g^{(1)}\|,\ldots,\|\mathbf{f}_g^{(d_j)}\|)^T} = e_{k_1}$, the k_1-th canonical vector in $\mathbb{R}^{d_j}$. Hence, from Lemma B.1 we have that

$$P_h R_g^{(k)} - \Big(f_{g,h}^{(k)} + \sum_{h' \in [d_g]: h' \neq h} P_h f_{g,h'}^{(k)}\Big) = \lambda\sqrt{d_g}\frac{f_{g,h}^{(k)}}{\|\mathbf{f}_g^{(k)}\|}\mathbb{1}_{\{k=k_1\}}. \quad (20)$$

If $k \neq k_1$ then Equation (20) implies

$$\mathbf{f}_g^{(k)} = \mathcal{I}^{-1}\mathbf{Q}R_g^{(k)},$$

where

$$\mathcal{I} = \begin{bmatrix} 1 & P_1 & \cdots & P_1 \\ P_2 & 1 & \cdots & P_2 \\ \vdots & \vdots & \ddots & \vdots \\ P_{d_g} & P_{d_g} & \cdots & 1 \end{bmatrix}$$

On the other hand, if $k = k_1$, we have

$$\mathbf{Q}R_g^{(k_1)} - \mathcal{I}\mathbf{f}_g^{(k_1)} = \lambda\sqrt{d_g}\frac{\mathbf{f}_g^{(k_1)}}{\|\mathbf{f}_g^{(k_1)}\|}$$

$$\left[\mathcal{I} + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}I_{d_g}\right]\mathbf{f}_g^{(k_1)} = \mathbf{Q}R_g^{(k_1)},$$

from which we obtain the group SpAM result discussed by Yin et al. [2012], i.e.,

$$\mathbf{f}_g^{(k_1)} = \left[\mathcal{I} + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}I_{d_g}\right]^{-1}\mathbf{Q}R_g^{(k_1)}.$$

□

For the remainder of the backfitting update derivation, assume that

$$P_h f_{g,h'}^{(k)} = \mathbb{E}[f_{g,h'}^{(k)} \mid X_h^{(g)}] = 0, \quad \forall h' \neq h.$$

This implies that the covariance of the within-group component functions is zero, i.e.,

$$\begin{aligned} Cov(f_{g,h}^{(k)}, f_{g,h'}^{(k)}) &= \mathbb{E}[f_{g,h}^{(k)} f_{g,h'}^{(k)}] \\ &= \mathbb{E}[f_{g,h}^{(k)}\mathbb{E}[f_{g,h'}^{(k)} \mid X_h^{(g)}]] \\ &= 0. \end{aligned}$$

Under this assumption, the stationary condition in Lemma B.1 simplifies to

$$f_{g,h}^{(k)} - P_h R_g^{(k)} + \lambda\sqrt{d_g}u^{(k)}v_h^{(k)} = 0.$$

Suppose $k_1 \in [d_j]$ is the unique k that attains the sup-norm (case 2). Then, we have the simplified expression

$$f_{g,h}^{(k)} = P_h^{(k)} R_g^{(k)}, \quad \forall k \neq k_1 \quad (21)$$

On the other hand, if $k = k_1$, the expression simplifies to

$$f_{g,h}^{(k_1)} - P_h^{(k_1)} R_g^{(k_1)} = \lambda\sqrt{d_g}\frac{f_{g,h}^{(k_1)}}{\|\mathbf{f}_g^{(k_1)}\|}$$

$$f_{g,h}^{(k_1)}\left[1 + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}\right] = P_h^{(k_1)} R_g^{(k_1)}$$

$$f_{g,h}^{(k_1)} = \left[1 + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}\right]^{-1} P_h^{(k_1)} R_g^{(k_1)}.$$

Taking the L_2-norm on both sides yields

$$\begin{aligned} \sqrt{\sum_{h \in [d_g]} \mathbb{E}[(f_{g,h}^{(k_1)})^2]} &= \|\mathbf{f}_g^{(k_1)}\| \\ &= \left[1 + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}\right]^{-1} \|\mathbf{Q}R_g^{(k_1)}\|. \end{aligned}$$

Solving for $\|\mathbf{f}_g^{(k_1)}\|$ gives us the following identity

$$\|\mathbf{f}_g^{(k_1)}\| = \|\mathbf{Q}R_g^{(k_1)}\| - \lambda\sqrt{d_g}.$$

Plugging into the simplified update for above finally yields

$$\begin{aligned} f_{g,h}^{(k_1)} &= \left[1 + \frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_1)}\|}\right]^{-1} P_h^{(k_1)} R_g^{(k_1)} \\ &= \left[1 + \frac{\lambda\sqrt{d_g}}{\|\mathbf{Q}R_g^{(k_1)}\| - \lambda\sqrt{d_g}}\right]^{-1} P_h^{(k_1)} R_g^{(k_1)} \\ &= \left[1 - \frac{\lambda\sqrt{d_g}}{\|\mathbf{Q}R_g^{(k_1)}\|}\right] P_h^{(k_1)} R_g^{(k_1)} \\ &= \left[\|\mathbf{Q}R_g^{(k_1)}\| - \lambda\sqrt{d_g}\right]\frac{P_h^{(k_1)} R_g^{(k_1)}}{\|\mathbf{Q}R_g^{(k_1)}\|} \end{aligned}$$

for all $h \in [d_g]$.

In the case where $m > 1$ entries $\|\mathbf{f}_g^{(k_1)}\|, \ldots, \|\mathbf{f}_g^{(k_m)}\|$ achieve the sup-norm also simplifies, i.e., for all $i \in [m]$ we have

$$P_h^{(k_i)} R_g^{(k_i)} = \lambda\sqrt{d_g}a_i\frac{f_{g,h}^{(k_i)}}{\|\mathbf{f}_g^{(k_i)}\|} + f_{g,h}^{(k_i)}.$$

Recall that $\|\mathbf{f}_g^{(k_1)}\| = \cdots = \|\mathbf{f}_g^{(k_m)}\|$ and taking the L_2

norm on both sides as well as summing over all m yields

$$\begin{aligned}\sum_{i=1}^{m}\|\mathbf{Q}R_g^{(k_i)}\| &= \sum_{i=1}^{m}\left\|\left[\frac{\lambda\sqrt{d_g}a_i}{\|\mathbf{f}_g^{(k_m)}\|}+1\right]f_{g,h}^{(k_i)}\right\| \\ &= \sum_{i=1}^{m}\left[\frac{\lambda\sqrt{d_g}a_i}{\|\mathbf{f}_g^{(k_m)}\|}+1\right]\|\mathbf{f}_g^{(k_m)}\| \\ &= \left[\frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_m)}\|}\sum_{i=1}^{m}a_i+m\right]\|\mathbf{f}_g^{(k_m)}\| \\ &= \lambda\sqrt{d_g}+m\|\mathbf{f}_g^{(k_m)}\|.\end{aligned}$$

Isolating $\|\mathbf{f}_g^{(k_m)}\|$ gives us the following identity

$$\|\mathbf{f}_g^{(k_m)}\| = \frac{1}{m}\left[\sum_{i=1}^{m}\|\mathbf{Q}R_g^{(k_i)}\|-\lambda\sqrt{d_g}\right] \quad \forall m\in[m].$$

Plugging this into the simplified update for case (3) yields

$$f_{g,h}^{(k_m)} = \left[1+\frac{\lambda\sqrt{d_g}}{\|\mathbf{f}_g^{(k_m)}\|}\right]^{-1}P_hR_g^{(k_m)} = \left[1+\frac{\lambda\sqrt{d_g}}{\frac{1}{m}\left[\sum_{i=1}^{m}\|\mathbf{Q}R_g^{(k_i)}\|-\lambda\sqrt{d_g}\right]}\right]^{-1}P_hR_g^{(k_m)}.$$

After a few algebraic manipulations, we obtain

$$f_{g,h}^{(k_m)} = \frac{1}{m}\left[\sum_{i=1}^{m}s_g^{(k_i)}-\lambda\sqrt{d_g}\right]\frac{P_hR_g^{(k_m)}}{s_g^{(k_m)}},$$

where $s_g^{(k_m)} = \|\mathbf{Q}R_g^{(k_m)}\|$ for all $m\in[m]$.

Now that we have the update for case (3), we still need to find the exact condition that the sup-norm is attained. This condition is given in Lemma B.5. The arguments follow a similar logic to Fornasier and Rauhut [2008].

Lemma B.5. *For some $m > 1$ precisely m entries $\|\mathbf{f}_g^{(k_1)}\|,\dots,\|\mathbf{f}_g^{(k_m)}\|$ attain the sup-norm $\max_{k=1,\dots,d_j}\|\mathbf{f}_g^{(k)}\|$ if and only if*

$$s_g^{(k_m)} \geq \frac{1}{m-1}\left[\sum_{i=1}^{m-1}s_g^{(k_i)}-\lambda\sqrt{d_g}\right]$$

and

$$s_g^{(k_{m+1})} < \frac{1}{m}\left[\sum_{i=1}^{m}s_g^{(k_i)}-\lambda\sqrt{d_g}\right].$$

Proof. Assume exactly m entries $\|\mathbf{f}_g^{(k_1)}\|,\dots,\|\mathbf{f}_g^{(k_m)}\|$ attain the sup-norm. Then, for all $m\in[m]$ we have

$$f_{g,h}^{(k_m)} = \frac{1}{m}\left[\sum_{i=1}^{m}s_g^{(k_i)}-\lambda\sqrt{d_g}\right]\frac{P_hR_g^{(k_m)}}{s_g^{(k_m)}}.$$

By Lemmas B.1 and B.2 we know that

$$P_h^{(k_i)}R_g^{(k_i)} = \lambda\sqrt{d_g}a_i\frac{f_{g,h}^{(k_i)}}{\|\mathbf{f}_g^{(k_i)}\|}+f_{g,h}^{(k_i)} \quad \forall i\in[m].$$

Isolating a_i after taking the L_2 norm on both sides we obtain

$$a_i = \frac{\|\mathbf{Q}R_g^{(k_i)}\|-\|\mathbf{f}_g^{(k_i)}\|}{\lambda\sqrt{d_g}} \quad \forall i\in[m].$$

Plugging the identity for $\|\mathbf{f}_g^{(k_m)}\|$ into a_m and since $a_m\geq 0$ we obtain

$$\|\mathbf{Q}R_g^{(k_m)}\| \geq \frac{1}{m-1}\left[\sum_{i=1}^{m-1}\|\mathbf{Q}R_g^{(k_i)}\|-\lambda\sqrt{d_g}\right].$$

Since $\|\mathbf{f}_g^{(k_m)}\| > \|\mathbf{f}_g^{(k_{m+1})}\|$ we must have that from Eq. (21) $\|\mathbf{Q}R_g^{(k_{m+1})}\| = \|\mathbf{f}_g^{(k_{m+1})}\| < \|\mathbf{f}_g^{(k_m)}\| = \frac{1}{m}\left[\sum_{i=1}^{m}\|\mathbf{Q}R_g^{(k_i)}\|-\lambda\sqrt{d_g}\right]$.

On the other hand, assuming that exactly $l\neq m$ entries $\|\mathbf{f}_g^{(k_1)}\|,\dots,\|\mathbf{f}_g^{(k_l)}\|$ attain the sup-norm, then following the same steps as above leads to contradiction. □

From Lemma B.5 we conclude that there exist exactly m^* entries that attain the sup-norm if and only if

$$m^* = \arg\max_m \frac{1}{m}\left[\sum_{i=1}^{m}\|\mathbf{Q}R_g^{(k_i)}\|-\lambda\sqrt{d_g}\right].$$

This concludes the proof of Theorem 3.1.

C SIMULATION DETAILS

We have developed a Python library that implements all the methodologies presented in this work, including our simulation study. This library is available under the following link `https://github.com/boschresearch/gresit`.

C.1 SYNTHETIC DATA GENERATION

Synthetic data used throughout the experiments is generated from GANMs (Definition 2) with varying dimension and group size. First, we construct the ground truth causal graph using the Erdös-Rényi model [Erdös and Rényi, 2011] with sparsity level set proportional to the number of nodes. In order to set up the nonlinear link functions f_g, we follow a strategy similar to previous work on scalar ANMs [Lachapelle et al., 2020, Uemura et al., 2022, Rolland et al., 2022]. We sample f_g from randomly weighted sums of Gaussian processes.

Additive noise is generated from multivariate Log-Normal distributions. First, multivariate Gaussians are generated

where mean vector entries and off-diagonal entries in the covariance matrix are sampled uniformly from the $[-0.8, 0.8]$ interval. Then the exponential is taken component-wise. While propagating through the graph, we adjust the scales of the variables in each equation to a signal-to-noise ratio of two. To reduce the risk of involuntary design patterns that might be picked up by the algorithms we employ [Reisach et al., 2021], data is always standardized.

C.2 METRICS

Evaluating causal graphs is a difficult task as the mere number of dissimilar edges between learned and the ground truth graph does not necessarily reflect how the graphs differ in their capability to answer causal queries. This fact is exacerbated by the problem that not all causal discovery algorithms return the same graphical object. Different causal assumptions posed by the corresponding routines lead to different types of causal graphs. Consequently, we report a number of evaluation metrics in order to capture strengths and weaknesses of the routines employed across a wide spectrum of tasks. Next to metrics that quantify edge recovery, we report recently proposed [Henckel et al., 2024] graph distances that count the number of wrongly inferred causal effects, determined by different identification strategies, when using the learned rather than the true DAG.

Precision, Recall and F_1 score

We start with binary classification metrics that are prominent in the machine learning literature. Put in a graphical context, *Precision* refers to the fraction of correctly determined edges among all identified edges. Recall, sets correctly identified edges in relation to all edges present in the ground truth. The F_1 score is the harmonic mean of Precision and Recall, i.e. $F_1 = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$.

Structural Hamming Distance (SHD)

The SHD counts the number of differing edges between to graphs.

Structural Interventional Distance (SID) [Peters and Bühlmann, 2015]:

The SID counts the number of incorrectly inferred interventional distributions from the learned graph when compared to the ground truth. For DAGs, this equates to counting parent sets in the learned DAG that are not valid adjustment sets in the ground truth DAG.

Ancestor Adjustment Identification Distance (AAID) [Henckel et al., 2024]

While parent adjustment is a valid adjustment strategy, there exist statistically more efficient adjustment sets. Based on this observation, Henckel et al. [2024] generalize the idea of the SID and present identification distances that count the number of incorrect identification formulas that arise when using some identification strategy. The SID then arises when parent adjustment is chosen as the identification strategy. However, SID may produce surprisingly large quantities when comparing two DAGs that have the same causal order. Choosing ancestor rather than parent adjustment leads to an adjustment strategy that returns a zero whenever estimated and ground truth DAG agree in terms of their causal orders.

Order Adjustment Identification Distance (OAID) [Henckel et al., 2024]

In order to emphasize the role played by the pruning step, we also provide a DAG to order distance. *OAID* arises when comparing the super-DAG $\mathcal{G}^\pi$ and the ground truth $\mathcal{G}_0$ in terms of their *AAID*. Consequently, *OAID* counts the number of incorrect identification formulas derived from the estimated causal order π.

C.3 ALGORITHMS

We employ the following causal discovery algorithms in our experiments.

GroupRESIT

Owing to the modular design of GroupRESIT, one may, in principle, combine various pairs of multi-response regression methods and vector independence tests in the first phase. To ensure broad applicability, we employ neural networks—specifically, multilayer perceptrons (MLPs) with hyperbolic tangent activation functions and early stopping—for the multi-response regression. We use the mean squared error (MSE) loss, which yielded better performance than the HSIC loss [Mooij et al., 2009, Greenfeld and Shalit, 2020]. For the subsequent independence test, we apply the empirical HSIC [Gretton et al., 2005] with Gaussian RBF kernels, using the median heuristic for bandwidth selection.

In the second phase, we compare the performance of MURGS with that of the greedy independence criterion employed in the original RESIT framework [Peters et al., 2014]. Any linear smoother can be used to estimate the conditional expectations during the backfitting procedure; in our experiments, we employ Gaussian kernel regression with a plug-in bandwidth $h = 0.6 \cdot \text{sd}(X) n^{-1/5}$. For greedy independence testing, we again apply the empirical HSIC, with p-values computed via the gamma approximation on a separate test dataset.

GroupGraN-DAG

GraN-DAG [Lachapelle et al., 2020] is a score based algorithm developed to handle nonlinear relations among variables. GraN-DAG utilizes the continuous acyclicity constraint first suggested by Zheng et al. [2018]. With the appropriate loss function, GraN-DAG can be tailored towards Gaussian nonlinear additive noise models. However, in order to ensure that GraN-DAG operates on a group level,

Table 1: Hyperparameter and tuning choices for all methods employed in this work.

Method	Parameters
GroupRESIT	`regression`: MLP with tanh activation; `n_epochs`=500; `lr`=0.01; `loss`=MSE; `batch_size`=500; `indep_test`=HSIC; `alpha`=0.01
MURGS	`smoother`=Gaussian kernel regression; `plugin bandwidth`=$0.6\,\mathrm{sd}(X)\,n^{-1/5}$
GroupPC	`indep_test`=Fisher's Z; `alpha`=0.05
GroupGraN-DAG	`hidden_num`=2; `hidden_dim`=10; `batch_size`=64; `lr`=0.001; `iterations`=100 000; `model_name`=NonLinGaussANM; `nonlinear`=leaky-ReLU; `optimizer`=RMSProp; `h_threshold`=1e-8; `lambda_init`=0.0; `mu_init`=0.001; `omega_lambda`=1e-4; `omega_mu`=0.9; `stop_crit_win`=100; `edge_clamp_range`=1e-4
GroupLiNGAM	`regression`=OLS; `indep_test`=HSIC

we adapt the micro-level acyclicity constraint to encourage acyclicity on the corresponding group DAG.

Recall that for variable groups $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ each group $\mathbf{X}_g \in \mathbb{R}^{d_g}$. Suppose we consider all group entries as scalar random variables such that we have $m = \sum_{g=1}^{p} d_g$ many micro variables. GraN-DAG enforces acyclicity via the trace exponential of a weighted adjacency matrix $A_\phi \in \mathbb{R}_{\geq 0}^{m \times m}$ that arises from the weights in the neural network. More specifically, the micro-acyclicity constraint amounts to $h(A_\phi) = tr(e^{A_\phi \circ A_\phi}) - m = 0$, where $\circ$ denotes the Hadamard product. Similar to Kikuchi and Shimizu [2023], we enforce acyclicity in the corresponding group DAG by the following weighted group adjacency matrix $A_\phi^{\text{group}} \in \mathbb{R}_{\geq 0}^{p \times p}$ where

$$(A_\phi^{\text{group}})_{gh} = \begin{cases} 0 & \text{if } g = h \\ \frac{1}{d_g d_h} \sum_{i \in [d_g]} \sum_{j \in [d_h]} (A_\phi)_{i,j} & \text{o.w.} \end{cases}. \tag{22}$$

By setting the diagonal to zero in A_ϕ^{group}, we ignore the graph structure induced by A_ϕ within the groups and only focus on the inter-group relations. As we want to enforce an acyclic group graph, we use the following modified constraint in the augmented Lagrangian method used in Lachapelle et al. [2020]

$$h(A_\phi^{\text{group}}) = tr(e^{A_\phi^{\text{group}} \circ A_\phi^{\text{group}}}) - p = 0. \tag{23}$$

In general, the final weighted adjacency matrix in *GroupGraN-DAG* is not sparse. Therefore, appropriate thresholds need to be set to enforce strict zeros. Unfortunately, this might lead to a clipped adjacency matrix that need not necessarily encode a DAG. In such cases, we continue to select thresholds until the resulting weighted adjacency matrix becomes acyclic.

GroupPC

The PC algorithm [Spirtes et al., 1993] performs conditional independence tests in a resource efficient way in order to remove edges from a fully connected undirected graph. Given the first phase, the algorithm orients as many edges as possible. We implement the stable version of the algorithm proposed by Colombo and Maathuis [2014]. In order to adapt the PC algorithm to the group setting, the involved tests need to be able to handle testing conditional independence among two random vectors given a set of random vectors. While Zhang and Hyvärinen [2009] extended the HSIC to conditional independence testing its prohibitively long runtime prevents us from using it in our experiments. Instead, we use the simple Fisher-Z scoring test and treat group entries individually. Then, we aggregate the coordinate-wise hypotheses based on scalar variables by using the union-intersection method. More precisely, in the skeleton-finding phase, we remove an edge if the union of the p-values of the individual tests is larger than the significance level α. Otherwise, the edge is retained. The significance level α of the involved test acts as a hyperparameter. The smaller α the larger the hurdle to keep an edge in the first phase such that sparser graphs will be returned. In general, the algorithm returns a completed partially directed acyclic graph (CPDAG). While in principle the new graph metrics developed by Henckel et al. [2024] return meaningful results for CPDAG to DAG comparisons, the same cannot be said for the remaining ones. In particular, comparability becomes difficult between those algorithms that return DAGs and the PC algorithm. Thus, we compute the SHD for each DAG consistent with the CPDAG and select the one with the smallest SHD.

GroupLiNGAM

We implement the *GroupDirectLiNGAM* algorithm of Entner and Hoyer [2012], which extends the direct estimation method for LiNGAM introduced by Shimizu et al. [2006, 2011] to handle vector-valued variables. Since Entner and Hoyer [2012] focus solely on the causal ordering step, we use MURGS to recover the full graph structure thereafter.

Table 1 reports all hyperparameters and tuning parameter choices for benchmark and real data results.

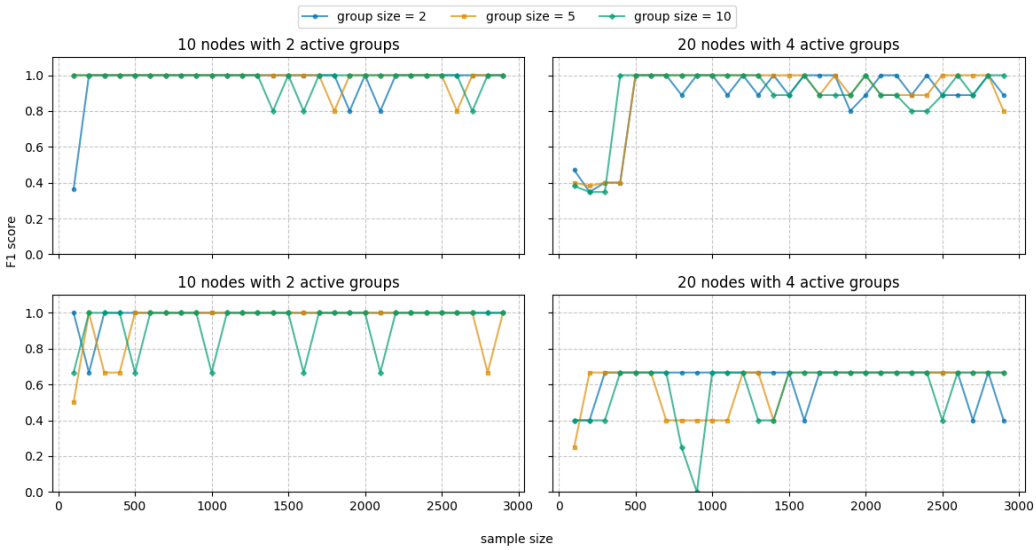


Figure 6: Feature selection capability of MURGS. Node size is $p = 10, 20$, with $2, 4$ active groups, respectively.

D ADDITIONAL RESULTS

MURGS experiments In this section we report results of a separate simulation study to assess the feature selection capability of MURGS. Figure 6 demonstrates in terms of F1 score (see Section C.2) how MURGS is able to select active groups across with increasing sample size. Encouragingly, group size does not seem to play a significant role in the performance of MURGS. The first row of Figure 6 is based on data generated according to the synthetic data generation strategy described in Sections 4 and C.1. We select sink nodes in a graph making sure that the number of parents is the same across group size. Then we generate data with increasing sample size and track the F1 score. Besides some numerical instabilities MURGS shows convergence behavior already on small sample sizes.

To test out the boundaries, we reiterate the experiments with a different data generation strategy (second row). Nonlinear functions are now generated via randomly initialized deep neural networks with ReLU activation function such that additive models will have a hard time approximating the regression functions. Indeed, MURGS struggles much more to recover the active nodes when the number of nodes in the graph is larger. Despite the challenging nonlinearity F1 score depicts high values throughout again without requiring extraordinarily high sample size.

Additional real data and synthetic data results Figures 7, 8, and 9 present supplementary results for the remaining algorithms in the real-data experiment (Section 5) that were omitted from the main text.

Additionally, the ensuing boxplots provide additional results regarding the synthetic experiment described in Sections 4 and C.1. The results plots are ordered by node size and group size.

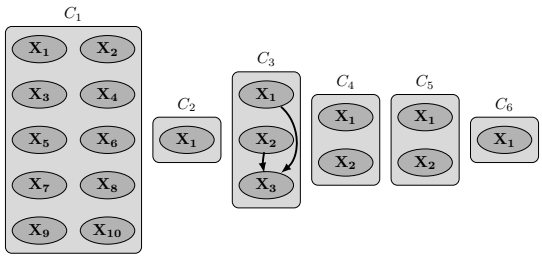


Figure 7: Learned causal edges from the real-world dataset using *GroupGraN-DAG*.

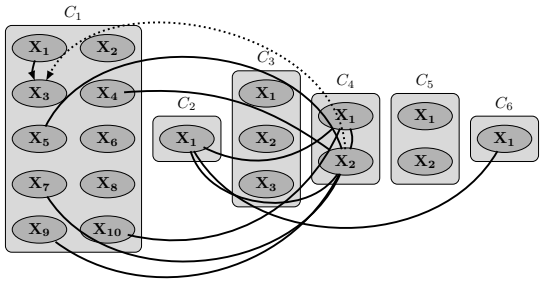


Figure 8: Learned causal edges from the real-world dataset using *GroupPC*.

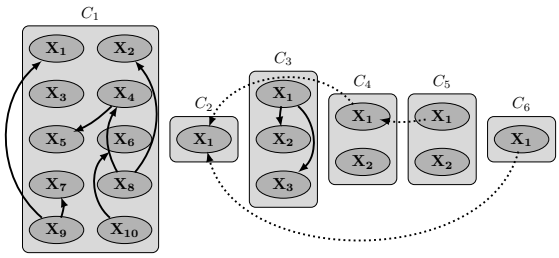
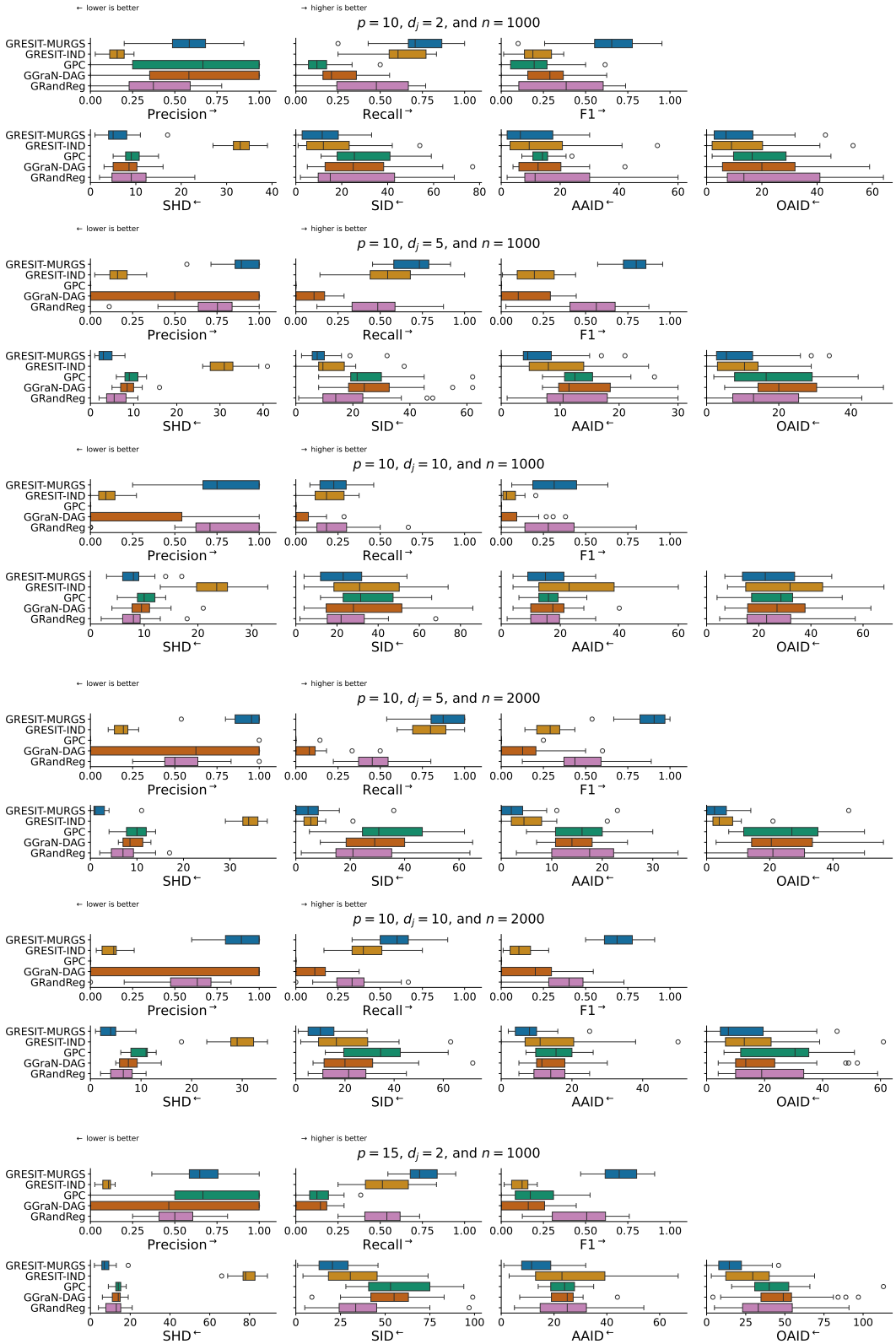
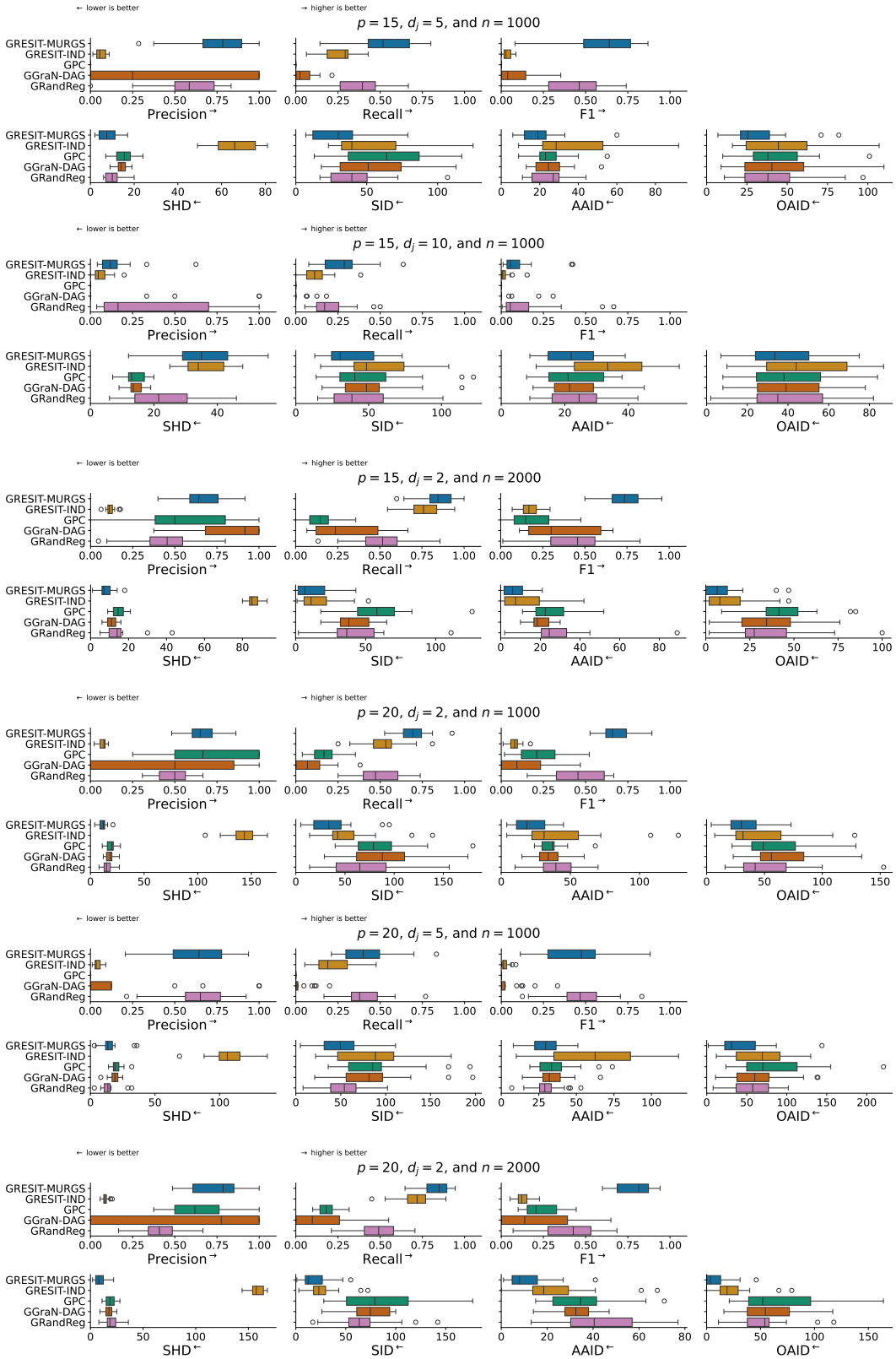


Figure 9: Learned causal edges from the real-world dataset using *GroupDirectLiNGAM*.


← lower is better
→ higher is better
$p = 10$, $d_j = 2$, and $n = 1000$
$p = 10$, $d_j = 5$, and $n = 1000$
$p = 10$, $d_j = 10$, and $n = 1000$
$p = 10$, $d_j = 5$, and $n = 2000$
$p = 10$, $d_j = 10$, and $n = 2000$
$p = 15$, $d_j = 2$, and $n = 1000$
GRESIT-MURGS
GRESIT-IND
GPC
GGraN-DAG
GRandReg
Precision →
Recall →
F1 →
SHD ←
SID ←
AAID ←
OAID ←


← lower is better
→ higher is better
$p = 15$, $d_j = 5$, and $n = 1000$
$p = 15$, $d_j = 10$, and $n = 1000$
$p = 15$, $d_j = 2$, and $n = 2000$
$p = 20$, $d_j = 2$, and $n = 1000$
$p = 20$, $d_j = 5$, and $n = 1000$
$p = 20$, $d_j = 2$, and $n = 2000$
GRESIT-MURGS
GRESIT-IND
GPC
GGraN-DAG
GRandReg
Precision →
Recall →
F1 →
SHD ←
SID ←
AAID ←
OAID ←