Technische Universität München
TUM School of Computation, Information and Technology

Nonparametric and Semiparametric Structure Learning in High-Dimensional Directed and Undirected Graphical Models

Konstantin Göbler

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr. Seyed Jalal Etesami

Prüfende der Dissertation:
1. Prof. Mathias Drton, Ph.D.
2. Assoc. Prof. Sebastian Weichwald, Ph.D.

Die Dissertation wurde am 05.09.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 10.12.2025 angenommen.

To my daughter.

Abstract

A primary objective across many scientific fields is to understand and characterize the dependencies among the phenomena being studied. In some cases, the goal is to uncover the dependency structure among a possibly high-dimensional set of variables; in others, it is to infer causal relations without access to randomized controlled trials or other interventional data. In this thesis, we use directed and undirected graphical models in high-dimensional settings to address challenges that frequently arise in real-world applications. One such challenge is the presence of mixed data in high-dimensional structure learning problems. We propose a latent Gaussian copula framework that incorporates polyserial and polychoric correlations, providing a procedure for estimating high-dimensional mixed graphs under sparsity. For directed models, we first address the scarcity of benchmark datasets by developing a semisynthetic data generation pipeline to facilitate the evaluation of causal discovery algorithms. Second, we extend additive noise models to the case of grouped variables. In this context, we establish structure identifiability and propose GroupRESIT, an order-based causal discovery algorithm that uses a multi-response group sparse additive model for pruning. These contributions aim to support the continued development of structure learning methods for complex data in practical applications.

Zusammenfassung

Ein zentrales Ziel in vielen wissenschaftlichen Disziplinen ist es, Abhängigkeiten zwischen den zu untersuchenden Phänomenen zu verstehen und zu charakterisieren. In manchen Fällen liegt das Interesse auf dem Erlernen der Abhängigkeitsstruktur einer möglicherweise hochdimensionalen Menge von Variablen, in anderen auf der Identifikation kausaler Zusammenhänge, ohne Zugang zu randomisierten Kontrollstudien oder anderen Interventionsdaten zu haben. In dieser Arbeit verwenden wir gerichtete und ungerichtete grafische Modelle in hochdimensionalen Anwendungsbereichen, um Herausforderungen zu adressieren, die häufig in realen Anwendungen auftreten. Eine solche Herausforderung ist das Vorliegen gemischter Daten in hochdimensionalen Strukturlernproblemen. Wir entwerfen einen latent Gaußschen Copula-Ansatz, der polyserielle und polychorische Korrelationen verwendet und ein Verfahren zum Lernen von hochdimensionalen gemischten Graphen bereitstellt. Für gerichtete Modelle gehen wir zunächst auf den Mangel an Benchmark-Datensätzen ein und entwickeln eine semisynthetische Datengenerierungspipeline, um das Benchmarking kausaler Lernverfahren zu erleichtern. Zweitens erweitern wir additive Fehlermodelle sodass sie in der Lage sind gruppierte Variablen zu modellieren. In diesem Zusammenhang stellen wir Ergebnisse zur Strukturidentifizierbarkeit vor und konstruieren GroupRESIT, einen ordnungsbasierten Algorithmus zum kausalen Lernen, der eine zugeschnittene Form von Sparse Additive Model für den Pruning-Schritt verwendet. Insgesamt, sollen diese Beiträge die Weiterentwicklung von Strukturlernverfahren in realen Anwendungen unterstützen.

Acknowledgments

First, I would like to sincerely thank my supervisor Mathias Drton for his guidance and support throughout my entire time at TUM and, in particular, for his patience and firm belief in me that helped me move past challenges in my research. I am very grateful for giving me the opportunity to join your group, where I was able to learn so much, both as a researcher and as a human wandering through this world.

I am also deeply grateful to Tobias Windisch. Working with you showed me how rewarding and productive collaboration can be. No question or suggestion ever felt too far-fetched, and I'm thankful for the opportunities we've had to work together.

I would also like to thank my wonderful colleagues and friends in the Mathematical Statistics group. I couldn't have hoped for a more supportive, joyful, and intellectually stimulating environment, perhaps precisely because of our wide range of backgrounds and research interests. A very special thanks goes to Andrea Grant for her help with administrative matters, for keeping us on our toes, and for supporting all our coffee endeavors, even though you don't even like coffee one bit.

I would also like to express my gratitude to my Bosch colleagues, both at VM and at CR. Through our collaboration, I developed a deeper appreciation for writing clean, effective code and gained valuable experience with real-world problems.

Lastly, I want to thank my friends and family for their tremendous support, you know who you are, and I could not have done it without you.

List of Contributed Articles

Core publications as (sole) first author:

1. Göbler, K., Drton, M., Mukherjee, S., and Miloschewski, A. (2024a). High-dimensional undirected graphical models for arbitrary mixed data. *Electron. J. Stat.*, 18(1):2339–2404

2. Göbler, K., Windisch, T., Drton, M., Pychynski, T., Roth, M., and Sonntag, S. (2024b). `causalAssembly`: Generating realistic production data for benchmarking causal discovery. In Locatello, F. and Didelez, V., editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR

3. Göbler, K., Windisch, T., and Drton, M. (2025). Nonlinear causal discovery for grouped data. In Chiappa, S. and Magliacane, S., editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1453–1475. PMLR

Contents

1 Introduction **1**

2 Preliminaries **4**

2.1 Undirected Graphical Models 4

2.2 Directed Acyclic Graphical Models 6

2.2.1 Structural Equation Models 7

2.2.2 The Causal Viewpoint 8

3 High-dimensional undirected graphical models for arbitrary mixed data **10**

3.1 Sparse Gaussian Graphical Models 10

3.2 The Nonparanormal Graphical Model 11

3.3 The Nonparanormal for Mixed Data 15

4 Generating Realistic Production Data for Benchmarking Causal Discovery **18**

4.1 Causal Discovery 18

4.1.1 Constraint-based methods 19

4.1.2 Score-based methods 19

4.1.3 Functional form restricting methods 20

5 Nonlinear causal discovery for grouped data **24**

5.1 Identifiability 24

5.2 Pruning 28

6 Conclusion **30**

7 Summary of Core Publications **31**

7.1 High-dimensional undirected graphical models for arbitrary mixed data 31

7.2 `causalAssembly`: Generating semi-synthetic real data for benchmarking causal discovery . 32

7.3 Nonlinear causal discovery for grouped data 33

Bibliography **34**

A Core Publications **43**

A.1 High-dimensional undirected graphical models for arbitrary mixed data 43

A.2 `causalAssembly`: Generating Semi-Synthetic Real Data for Benchmarking Causal Discovery 110

A.3 Nonlinear Causal Discovery for Grouped Data 145

1 Introduction

A primary objective across many scientific fields is to understand and characterize the nature of dependencies among the phenomena being studied. In statistics, considerable research focuses on analyzing events that occur jointly. For instance, consider investigating associations between the severity of a particular disease and various potential risk factors. Identifying these associations helps to create accurate predictive models for disease occurrence. Furthermore, in many medical contexts, available data on biomarkers, risk factors, patient characteristics, and medical histories often significantly outnumber actual patient observations. As a result, prediction models must either handle high-dimensional input effectively or rely on preliminary variable selection strategies.

Beyond accurately predicting disease occurrence, one might also be interested in identifying effective treatments or cures. In this pursuit, an observed dependence between a disease and a biomarker or risk factor might suggest that the biomarker or risk factor directly causes the disease. If true, manipulating this factor could offer a path to curing or preventing the disease. However, the observed dependence might alternatively result from the disease causing changes in the biomarker or risk factor. In such a scenario, manipulating these factors would have no therapeutic benefit. Moreover, it is also possible that both the disease and the biomarker or risk factor are influenced by a common, unmeasured confounder, again rendering direct manipulation ineffective. Clearly, inferring causal relationships from observational data is a challenging task. This raises a fundamental question: in the absence of controlled experiments, is it even possible to reliably identify causal relationships from observational data alone? Generally, the short answer is no. Nonetheless, under certain assumptions, some or all causal relationships within a system may indeed be identifiable solely from observational data.

In this thesis, we use undirected and directed probabilistic graphical models to represent dependencies among variables of interest. Probabilistic graphical models lie at the intersection of probability theory and graph theory, where structural features of the graphs encode conditional independences among the associated random quantities. Our focus is on the case where the graph structure is *not* known a priori and must be estimated or learned from samples of some underlying distribution — a problem known as *structure learning* (Drton and Maathuis, 2017).

Structure learning is particularly challenging in high-dimensional settings. In the part of this thesis concerned with undirected graphical models, we use the term *high-dimensional* in its classical sense, namely when the number of variables exceeds (or is of the same order as) the number of samples. In the context of directed probabilistic graphical models, however, the difficulty arises for a different reason: the number of possible directed acyclic graphs (DAGs) grows super-exponentially with the number of nodes. For example, while there are 25 possible DAGs on three nodes, the number already increases to 3,781,503 on six nodes. With around 21 nodes, the number of DAGs becomes comparable to the estimated number of atoms in the observable universe ($\approx 10^{80}$) (Robinson, 1973; OEIS Foundation Inc., 2025). Therefore, we use *high-dimensional* here more loosely

to refer to situations where the number of variables is large — regardless of the sample size — since even moderate numbers of variables already make causal structure learning extremely challenging.

While high-dimensional undirected graphical models have been studied extensively in fully continuous or fully discrete settings, much less attention has been given to the mixed-data case, where both types of variables are present. This gap is important because many real-world applications involve mixed data. For instance, in our introductory example concerning the severity of a particular disease and various potential risk factors, it is common to observe both continuous measurements (e.g., blood pressure) and discrete variables (e.g., presence or absence of a disease) when determining treatment strategies. In this thesis, we address this gap by introducing a modeling framework that enables high-dimensional undirected graph learning in the presence of arbitrary mixed data.

Directed graphical models admit a natural causal interpretation through the use of structural equation models (SEMs) (Bollen, 1989), which posit noisy functional relationships among the variables under study. Each variable is assigned a subset of the remaining variables — its parents — governed by some functional mechanism and an associated stochastic noise term. The parent set of a variable corresponds exactly to the set of nodes with directed edges pointing into it in the associated DAG. These assignments reflect the inherent asymmetry of cause-effect relationships: changes in any of the parents lead to changes in the effect variable, but not vice versa. In particular, interventions in such a system of equations can be modeled by altering the functional mechanism for precisely those variables that are subject to the intervention (Pearl, 2009).

One fundamental challenge in the advancement and algorithmic development of causal structure learning methods—also called causal discovery—is the ability to accurately assess the quality of the learned causal relationships. However, for most real-world problems involving complex observational data sources, the ground truth causal relationships are unknown. As a result, novel procedures are often evaluated using simplistic simulation setups, which tend to have only limited generalizability to real-world settings (Reisach et al., 2021; Ng et al., 2024). In this thesis, we address this challenge by constructing a semisynthetic data generation pipeline that can produce data of varying (high) dimension and complexity for benchmarking causal discovery algorithms. The *semisynthetic* aspect comes from using real data from a highly automated production line, which — guided by domain knowledge — serves as the basis for learning conditional distributions that ensure the resulting observational distribution is Markovian.

Another essential challenge in causal discovery is the question of structure identifiability (Spirtes and Zhang, 2016). Even if we assume that the DAG associated with a given SEM is a perfect map of the joint distribution, multiple DAGs may induce the same observational distribution, allowing recovery of the underlying ground-truth DAG only up to its Markov equivalence class (Spirtes et al., 2000). A unique DAG beyond the Markov equivalence class can be identified only if the functional form of the causal mechanisms is appropriately restricted. One such restriction is to impose additivity of the stochastic noise term in the corresponding SEM, leading to an *additive noise model* (ANM). Intuitively, ANMs break symmetries between causal directions that persist under more general functional forms. In this thesis, we explore this restriction in the context of *grouped* variables, where the causal entities of interest are random vectors rather than scalar random variables. The resulting structure identifiability results motivate a causal discovery algorithm for grouped data whose computational complexity is polynomial in the number of groups.

Contributions of the thesis. This thesis comprises three core articles. The first article (Göbler et al., 2024a) addresses the problem of high-dimensional undirected graph learning in the presence of mixed data — that is, data comprising both continuous and discrete measurements. By modeling discrete variables as instances of latent continuous ones, we introduce a latent Gaussian copula framework. A key insight is that classical statistical measures, specifically polyserial and polychoric correlations, integrate naturally into this framework. We leverage this connection to derive theoretical properties for the resulting estimators and demonstrate their applicability in high-dimensional settings.

In the context of directed acyclic graphical models, the second article (Göbler et al., 2024b) tackles the challenge of scarcely available real data ground truth graphs when benchmarking causal discovery methods. Drawing on detailed domain knowledge from a fully automated industrial production line, we construct plausible ground truth causal relationships within each station. We then apply sparse additive models to estimate causal dependencies across production stations. To enable sampling from a joint distribution that is Markov with respect to the ground truth graph, we generate semisynthetic data by fitting conditional distributions using distributional random forests (Ćevid et al., 2022).

The third article (Göbler et al., 2025) focuses on nonlinear causal discovery for grouped data. While bivariate identifiability of nonlinear additive noise models has been established in the scalar case (Hoyer et al., 2008), we extend this result to the group setting. Building on these theoretical insights, we develop GroupRESIT, an extension of the RESIT algorithm (Peters et al., 2014) adapted to grouped variables. In particular, we propose a novel pruning strategy tailored to the group setting. We call this pruning strategy MURGS (Multi-response Group Sparse Additive Models) and derive a closed-form backfitting update.

Organization of the thesis. Chapter 2 reviews foundational concepts in undirected and directed graphical models, as well as structural equation models and their causal interpretation. Chapter 3 adopts a high-dimensional perspective under sparsity and discusses Gaussian graphical models, the Nonparanormal distribution, and the latent Gaussian copula model, which form the basis of the first core article (Göbler et al., 2024a). Chapter 4 focuses on graphical models based on directed acyclic graphs and addresses the scarcity of reliable ground-truth graphs for benchmarking causal discovery methods. Several classes of causal discovery algorithms are introduced here and featured in the second core article (Göbler et al., 2024b). Finally, Chapter 5 presents additive noise models and reviews identifiability results in the scalar case. It further discusses DAG pruning via sparse regression techniques for order-based causal discovery algorithms, motivating an extension of the framework to grouped data, which is the focus of the third core article (Göbler et al., 2025).

2 Preliminaries

Graphical models are used throughout many scientific disciplines to model the statistical relationship among variables of interest through graphs. More specifically, a graphical model is a set of multivariate joint distributions that are subject to certain conditional independences (Lauritzen, 1996). These independences are captured in a graph $G = (V, E)$, where the node set $V = [p]$, with $p \in \mathbb{N}$, indexes the variables $X_v, v \in V$ and the edge set $E \subset V \times V$ represents conditional dependencies. In particular, X_v and X_w are required to be conditionally independent given $\mathbf{X}_C = (X_u : u \in C)$, i.e. $X_v \perp\!\!\!\perp X_w \mid \mathbf{X}_C$, if all paths between the nodes v and w are suitably blocked by nodes in C (Drton and Maathuis, 2017).

2.1 Undirected Graphical Models

Undirected Graphical Models (UGMs), also known as Markov Random Fields (MRFs), are a class of graphical models in which the edge set E consists of unordered pairs (u, v) for distinct $u, v \in V$. We sometimes denote the presence of an edge between u and v by writing $u - v \in E$. Additionally, we denote the neighbors of node $v \in V$ as $ne_G(v) = \{u \in V : u - v \in E\}$.

The random vector $\mathbf{X} = (X_v : v \in V)$ is said to satisfy the *pairwise* Markov property (PMP) with respect to G if

$$X_v \perp\!\!\!\perp X_w \mid \mathbf{X}_{V\setminus\{v,w\}}, \tag{2.1}$$

whenever $(v, w) \notin E$. Further, $\mathbf{X}$ is said to satisfy the *local* Markov property (LMP) with respect to G if

$$X_v \perp\!\!\!\perp \mathbf{X}_{V\setminus(ne_G(v)\cup\{v\})} \mid \mathbf{X}_{ne_G(v)}, \tag{2.2}$$

for all $v \in V$. Lastly, $\mathbf{X}$ is said to satisfy the *global* Markov property (GMP) with respect to G if

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C, \tag{2.3}$$

for all pairwise disjoint subsets $A, B, C \subset V$ such that A and B are separated by C, meaning that every path between a node in A and a node in B passes through a node in C. It is well known that

$$\text{GMP} \implies \text{LMP} \implies \text{PMP}. \tag{2.4}$$

The reverse implications hold, in particular, when $\mathbf{X}$ has a strictly positive joint density with respect to the counting or Lebesgue measure (Lauritzen, 1996).

Whenever an edge is absent between two nodes in G, the pairwise Markov property implies conditional independence between those nodes given all other variables. The smallest graph (in the sense of containing the fewest edges) for which $\mathbf{X}$ satisfies the pairwise Markov property is called the *conditional independence graph* of $\mathbf{X}$ (Lauritzen, 1996).

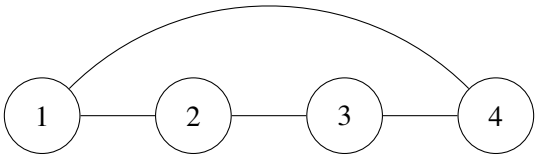


Figure 2.1 Undirected graph with four nodes.

Suppose that $\mathbf{X}$ possesses a density with respect to the product measure $\mu = \bigotimes_{v \in V} \mu_v$ on $\mathbb{R}^V$. Let $C(G)$ denote the set of all cliques in G; that is, a subset $C \subset V$ belongs to $C(G)$ if every pair of distinct nodes in C is connected by an edge. The distribution of $\mathbf{X}$ is said to factorize according to G if its density can be written as

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{C \in C(G)} \psi_C(\mathbf{x}_C), \tag{2.5}$$

where the functions ψ_C are referred to as *potential functions* and are not required to have any particular probabilistic interpretation. In practice, when working with such nonparametric UGMs, one typically imposes additional structure on either the graph G or on the functional form of the potential functions.

Consequently, by leaving the graph G unconstrained, UGMs can be defined by selecting appropriate families of potential functions. In particular, assuming $\mathbf{X}$ follows a multivariate Gaussian distribution $N(\mu, \mathbf{\Sigma})$ we obtain Gaussian graphical models (GGMs) where

$$\left(\mathbf{\Sigma}^{-1}\right)_{v,w} = 0 \quad \text{for all } (v, w) \notin E. \tag{2.6}$$

In other words, the conditional independences among the random variables as well as the existance of the edges in G are determined by the zero pattern of the inverse covariance matrix, i.e, the concentration matrix, $\mathbf{\Omega} = (\omega_{v,w}) \coloneqq \mathbf{\Sigma}^{-1}$. Equation (2.6) states that $\omega_{v,w} = 0$ for $v \neq w$ if and only if X_v and X_w are conditionally independent given all other variables.

Example 2.1. Let $\mathbf{X} = (X_1, \ldots, X_4)$ follow a four-dimensional Normal distribution with mean zero, covariance matrix $\mathbf{\Sigma}$ and concentration matrix $\mathbf{\Omega} = (\omega_{v,w})$. The density of $\mathbf{X}$ has the following form

$$p_{\mathbf{X}}(x) = \frac{1}{\sqrt{(2\pi)^4 \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2} \sum_{v,w=1}^{4} \omega_{v,w} x_v x_w\right), \tag{2.7}$$

and factorizes according to the graph in Figure 2.1 if and only if $\omega_{1,3} = \omega_{2,4} = 0$.

Estimating the conditional independence graph in Gaussian graphical models (GGMs) is equivalent to covariance selection (Dempster, 1972). Moreover, because GGMs belong to the class of regular exponential families, the selection of conditional independence graphs is supported by an extensive body of statistical theory (Buhl, 1993; Edwards, 2000; Drton and Perlman, 2007).

Let $\mathbf{S} = (s_{v,w})$ denote the sample covariance matrix for an i.i.d. Gaussian sample of size n. The maximum likelihood estimator (MLE) for the precision matrix $\mathbf{\Omega}$ is obtained by maximizing the log-likelihood function

$$\ell(\mathbf{\Omega}) = \log\det(\mathbf{\Omega}) - \operatorname{tr}(\mathbf{S}\mathbf{\Omega}), \tag{2.8}$$

subject to the constraints $\mathbf{\Omega} \succ 0$ and $\omega_{v,w} = 0$ for all $(v, w) \notin E$.

When the sample size exceeds the number of nodes (i.e., $n > p$), the sample covariance matrix $\mathbf{S}$ is almost surely positive definite, ensuring that $\ell(\mathbf{\Omega})$ has a unique maximizer with probability one. In high-dimensional settings where $n < p$, the estimation problem can remain tractable by imposing additional structural constraints, such as *sparsity*, which restricts each node to be incident to only a small number of edges (Yuan and Lin, 2007).

2.2 Directed Acyclic Graphical Models

Graphical models may also be formulated in terms of directed graphs $G = (V, E)$. However, in this case the edge set E contains ordered pairs (v, w) only. We sometimes write $v \to w$ for $(v, w) \in E$. We denote the parents of node $w \in V$ as $pa_G(w) = \{v \in V : (v, w) \in E\}$ and its descendants as $de_G(w) = v \in V : w \to \cdots \to v$ in G. Then, the set of non-descendants can be written as $nd_G(w) = V \setminus de_G(w)$. If the directed graph G does *not* contain any directed cycles (i.e., $v \to \cdots \to v$ for some $v \in V$), then G is called a *directed acyclic graph* (DAG).

In a graphical model with DAG G, we require that each variable $X_v, v \in V$ is conditionally independent of its non-descendants $X_{nd_G(v)\setminus pa_G(v)}$ given its parents $X_{pa_G(v)}$. This is referred to as the *local Markov property* with respect to DAG G (Lauritzen, 1996). We say that the distribution of $\mathbf{X}$, i.e., $P_{\mathbf{X}}$, satisfies the *global Markov property* with respect to G if $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C$ for all triples of pairwise disjoint subsets $A, B, C \subset V$ such that C *d-separates* A and B in G, i.e., $\mathbf{X}_A \perp_G \mathbf{X}_B \mid \mathbf{X}_C$. Note that A and B are said to be *d-separated* by C if every path between nodes in A and B is blocked by C. A path between nodes v_1 and v_n is blocked by a set C whenever there exists a node v_k such that one of the two conditions hold:

1. $v_k \in C$ and $v_{k-1} \to v_k \to v_{k+1}$ or $v_{k-1} \leftarrow v_{k+1} \leftarrow v_k$ or $v_{k-1} \leftarrow v_{k+1} \to v_k$.

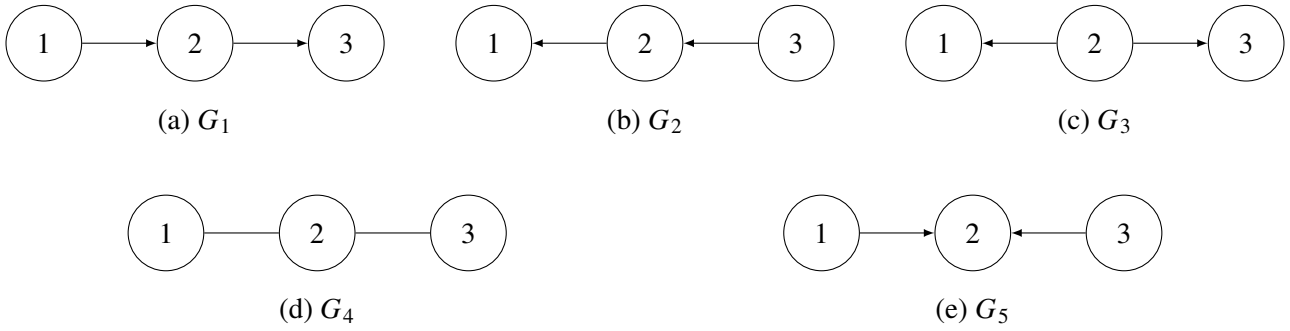
2. $v_{k-1} \to v_k \leftarrow v_{k+1}$ and neither v_k nor any of its descendants $de_G(v_k)$ are in C.

If $P_{\mathbf{X}}$ satisfies the global Markov property with respect to the DAG G then we say that G is an *independence map* (I-Map) of $P_{\mathbf{X}}$. If the conditional independence statements in $\mathbf{X}$ imply the d-separation statements in the graph we say that $\mathbf{X}$ is *faithful* with respect to G. We call G a *perfect map* of $P_{\mathbf{X}}$ if both the global Markov property and faithfulness hold. Similar to UGMs, the distribution of $\mathbf{X}$ is said to factorize according to G if its density can be written as

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{v \in V} p_{\mathbf{X}}(x_v \mid \mathbf{x}_{pa_G(v)}), \tag{2.9}$$

where $p_{\mathbf{X}}(x_v \mid \mathbf{x}_{pa_G(v)})$ is the conditional density of X_v given its parents. The global and local Markov properties coincide. In fact, if the density of $\mathbf{X}$ exists, global and local Markov properties are equivalent to the factorization property (Verma and Pearl, 1990a). This implies that we can construct graphical models based on DAGs by specifying conditional densities $p_{\mathbf{X}}(x_v \mid \mathbf{x}_{pa_G(v)})$ for $v \in V$.

Note that for a given distribution $P_{\mathbf{X}}$, the global Markov property may hold with respect to multiple DAGs. For example, the complete graph — which encodes no independences — is trivially an I-map for every distribution over $\mathbf{X}$. However, from the perspectives of interpretability, statistical efficiency, and computational efficiency, it is generally desirable to find a DAG that captures as many independences in $P_{\mathbf{X}}$ as possible. Such a DAG is known as a *minimal independence map* (minimal I-map); removing *any* edge from this DAG would yield a graph G' that no longer satisfies the global Markov property with respect to $P_{\mathbf{X}}$. However, minimal I-maps need



(a) G_1 (b) G_2 (c) G_3

(d) G_4 (e) G_5

Figure 2.2 Markov-equivalent DAGs in a-c. G_4 is the CPDAG representing the MEC of G_1, G_2 and G_3. The DAG G_5 in (e) is not Markov-equivalent to any of the other DAGs and is also the only DAG in the MEC of G_5.

not be unique for a given distribution. Whether one can distinguish among minimal I-maps is an identifiability issue that, as discussed in Chapters 4 and 5, usually demands extra assumptions.

Two DAGs G and G' are called *Markov equivalent* to one another if their d-separation statements coincide. In this case, we write $G \approx_{\mathcal{M}} G'$. Furthermore, the set of all Markov equivalent DAGs is called the *Markov equivalence class* (MEC) of G (Andersson et al., 1997). All DAGs in the MEC of G share the same skeleton (the set of edges where arrowheads are ignored) and v-structures (triples of the form $u \to w \leftarrow v$, where u and v are not joined by an edge) (Verma and Pearl, 1990b). Hence, Markov-equivalent DAGs can not be distinguished from one another using only samples from $P_{\mathbf{X}}$.

Example 2.2. Consider the DAGs shown in Figure 2.2. The graphs G_1, G_2, and G_3 are Markov-equivalent, as they share the same skeleton and have no v-structures. Their only d-separation statement is that node 1 is d-separated from node 3 by node 2. In contrast, G_4 is not Markov-equivalent to any of the other DAGs. Although it shares the same skeleton, it differs by having a v-structure and by implying that nodes 1 and 3 are unconditionally d-separated and become d-connected when conditioning on node 2.

A MEC can be uniquely represented by a *completed partially directed acyclic graph* (CPDAG), comprising both directed and undirected edges. Within a MEC, all DAGs share the skeleton and the set of v-structures; consequently, the directed edges in a CPDAG correspond precisely to these v-structures and to edges whose orientation is consistent across all DAGs in the MEC. Undirected edges in a CPDAG reflect cases where there is at least one conflicting orientation among DAGs in the MEC. For instance, in Figure 2.2, the CPDAG G_4 visually represents the MEC containing DAGs G_1, G_2, and G_3. Conversely, the DAG and CPDAG G_5 in subfigure (e) coincide, indicating that the MEC contains exactly one DAG.

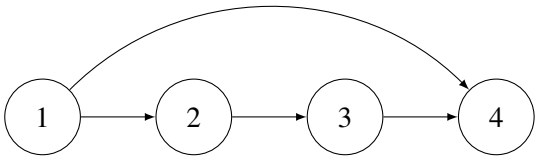
2.2.1 Structural Equation Models

Graphical models associated with some DAG G may also be viewed through the lense of structural equation models (SEM) (Bollen, 1989). Let $\mathbf{N} = (N_v : v \in V)$ be a vector of jointly independent noise variables and f_v measureable functions in some function space $\mathcal{F}$. Then the distribution of $\mathbf{X}$ where

$$X_v = f_v(\mathbf{X}_{pa_G(v)}, N_v), \quad v \in V, \tag{2.10}$$

is Markov with respect to the DAG G (Pearl, 2009, Theorem 1.4.1). The converse is also true, i.e., if $\mathbf{X}$ is Markov with respect to a DAG G then there exist jointly independent noise terms $\mathbf{N}$ and functions f_v such that

DAG



Factorization

$$p_{\mathbf{X}}(x_1, \ldots, x_4) =$$
$$p_{\mathbf{X}}(x_1) p_{\mathbf{X}}(x_2 \mid x_1)$$
$$p_{\mathbf{X}}(x_3 \mid x_2) p_{\mathbf{X}}(x_4 \mid x_1, x_3)$$

Structural Eq.

$$X_1 = N_1$$
$$X_2 = f_2(X_1, N_2)$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_1, X_3, N_4)$$

Figure 2.3 Demonstrating the factorization property (middle) of the distribution of $(X_1, \ldots, X_4)$ with respect the DAG (left) and the corresponding SEM (right).

Eq. (2.10) holds (Druzdzel and Van Leijen, 2001, Theorem 2). The graph of a SEM is obtained by drawing directed edges between the parents and their direct effects, i.e., from each node appearing on the right-hand side of Eq. (2.10) to the left-hand-side of the equation. Figure 2.3 demonstrates the factorization property of the density of $\mathbf{X}$ and the SEM corresponding to a DAG on four nodes.

2.2.2 The Causal Viewpoint

From the construction of SEMs, it is clear that they naturally admit a causal interpretation: the parents $\mathbf{X}_{pa_G(v)}$ of each variable X_v are treated as its direct causes. These causal relationships are then succinctly summarized by the associated DAG G. Answering causal questions is fundamental to virtually all parts of science. Such causal questions are inertly about the *mechanism* behind the data. In the context of a SEM we can determine how $\mathbf{X}_{pa_G(v)}$ affect X_v in the sense that changes in any of the parents may cause X_v to change but not the other way around. This asymmetry is an important feature of SEMs as we can not invert these relationships as we could for ordinary algebraic equations. Consequently, we can model experimental interventions by, e.g., altering the structural equations for those variables that are affected by a given intervention. Finally, this allows predictions in changed environments from which causal effects can be estimated (Pearl, 2009).

Intervening on a variable X_v involves modifying its generating mechanism — specifically, altering the function $f_v(\cdot)$ defined in Eq. (2.10) — while preserving all other equations in the system. A common example is assigning X_v a fixed value x'_v within its support, known as a *do-intervention*, denoted by $do(X_v = x'_v)$. Such an intervention produces a new structural equation model (SEM) that induces an *interventional distribution*, denoted $P^I_{\mathbf{X}}$, on the variables $\mathbf{X}$:

$$P^I_{\mathbf{X}} = \prod_{v \notin I} P_{\mathbf{X}}(X_v \mid \mathbf{X}_{pa_G(v)}) \prod_{v \in I} \mathbb{I}(X_v = x'_v). \tag{2.11}$$

More broadly, *soft interventions* (Peters et al., 2017), also known as *mechanism changes* (Tian and Pearl, 2001), modify the causal mechanisms by replacing the original conditional distributions with new ones, without necessarily eliminating the direct causal effects of parent variables. In this general scenario, the interventional distribution is expressed as:

$$P^I_{\mathbf{X}} = \prod_{v \notin I} P_{\mathbf{X}}(X_v \mid \mathbf{X}_{pa_G(v)}) \prod_{v \in I} P^I_{\mathbf{X}}(X_v \mid \mathbf{X}_{pa_G(v)}). \tag{2.12}$$

Two important generalizations of SEMs frequently arise in causal analysis. First, the assumption of independence among noise terms $\mathbf{N}$ can be relaxed, leading to *semi-Markovian SEMs* (Pearl, 2009). Such models

effectively represent systems with hidden confounding, violating the assumption of *causal sufficiency*, which requires observation of all variables generated by the underlying SEM. Second, one may relax the acyclicity assumption. While acyclicity is natural when variables represent measurements taken sequentially at specific time points, feedback loops frequently occur when systems are observed in aggregated or equilibrium forms.

3 High-dimensional undirected graphical models for arbitrary mixed data

The first core article of this thesis (Göbler et al., 2024a) addresses high-dimensional undirected graph learning for general mixed data types — that is, data comprising both continuous and discrete variables with arbitrary levels. In what follows, we build the theoretical foundations step by step. Already in Section 2.1, we introduced Gaussian graphical models (GGMs) in the standard setting where the sample size n exceeds the number of variables (nodes) p. Now, in Section 3.1, we shift our focus to the high-dimensional scenario, characterized by having more variables than observations ($n < p$). Next, in Section 3.2 we introduce the *nonparanormal* graphical model, discuss different estimation strategies and their associated theoretical properties. Lastly, in Section 3.3 we introduce the problem of general mixed data in the context of high-dimensional nonparanormal graphical models and discuss our ansatz in Göbler et al. (2024a).

We focus on high-dimensional problems since they commonly arise across various scientific disciplines and practical applications, including remote sensing, computational biology (especially gene expression arrays), fMRI data analysis, spectroscopic imaging, and medical diagnostics. In such settings, traditional statistical approaches often become inadequate. For instance, the maximum likelihood estimator (MLE) of the precision matrix $\mathbf{\Omega}$ (see Eq. 2.8) no longer exists when the sample size n is smaller than the dimension p (Buhl, 1993). Consequently, additional structural assumptions or regularization methods become essential to derive well-defined estimators.

A prevalent modeling assumption in high-dimensional covariance estimation involves shrinkage estimators. Shrinkage estimators aim to improve estimation by combining the sample covariance matrix $\mathbf{S}$ with a structured target matrix. A natural choice for the target is the identity matrix, which corresponds to assuming equal variances and zero covariances among variables (Ledoit and Wolf, 2004). Alternatively, one may assume the covariance or concentration matrix to be banded, i.e., entries decay in numerical size depending on their distance to the diagonal (Bickel and Levina, 2008). This assumption is often appropriate when studying time-series data (Furrer and Bengtsson, 2007).

3.1 Sparse Gaussian Graphical Models

In this part of the thesis, we focus on constraints in the form of *sparsity* in the precision matrix. Since the non-zero entries in the precision matrix correspond to the edges of a GGM, sparsity in the precision matrix effectively requires that the corresponding undirected graph has relatively few edges. More specifically we study high-dimensional scaling where the number of nodes in the graph p as well as the maximum node degree d —

the maximum number of edges incident to a single node — are allowed to grow as a function of the sample size n.

There exist several well-established penalization techniques for estimating sparse GGMs. In seminal work, Meinshausen and Bühlmann (2006) proposed and analyzed a neighborhood selection procedure utilizing the lasso estimator (Tibshirani, 1996), demonstrating consistency under high-dimensional scaling. Neighborhood selection involves identifying, for each variable $v \in [p]$, the smallest set of neighbors $ne(v)$ that satisfies the local Markov property, i.e.,

$$X_v \perp\!\!\!\perp \mathbf{X}_{V\setminus(ne_G(v)\cup\{v\})} \mid \mathbf{X}_{ne_G(v)}. \tag{3.1}$$

Edges are subsequently inferred by aggregating through union or intersection rules.

An alternative approach focuses directly on penalizing the Gaussian log-likelihood. Specifically, much research has concentrated on estimators that maximize the Gaussian log-likelihood regularized by the ℓ_1 norm of the off-diagonal entries of the precision matrix, defined by the optimization problem:

$$\widehat{\mathbf{\Omega}} \in \underset{\mathbf{\Omega}\succeq 0}{\arg\max} \left\{\log\det(\mathbf{\Omega}) - \operatorname{tr}(\mathbf{S}\mathbf{\Omega}) - \lambda\rho_1(\mathbf{\Omega})\right\}, \tag{3.2}$$

where $\rho_1(\mathbf{\Omega}) = \sum_{i\neq j}|\omega_{ij}|$ (d'Aspremont et al., 2008; Friedman et al., 2007; Yuan and Lin, 2007). As a log-determinant optimization problem, it can be addressed using interior-point methods (Boyd and Vandenberghe, 2004), or, more efficiently, via first-order block coordinate descent methods developed by d'Aspremont et al. (2008) and Friedman et al. (2007). The latter authors coined this efficient algorithm the *graphical lasso*.

Rothman et al. (2008) studied convergence rates under bounded eigenvalue conditions and established convergence in both Frobenius and spectral norms at the rate $O\left(\sqrt{\frac{(s+p)\log p}{n}}\right)$, where s denotes the number of edges in G. Elementwise ℓ_∞-norm convergence was given by Ravikumar et al. (2011), who also generalized earlier model selection consistency results from Lam and Fan (2009). Specifically, while Lam and Fan (2009) required the number of edges s to scale at most as $O(\sqrt{p})$ and the sample size n to scale at least as $\Omega((s+p)\log p)$ to recover the zero pattern in $\mathbf{\Omega}$ with high probability, Ravikumar et al. (2011) provided an improved lower bound on the required sample size of the form $n = \Omega(d^2 \log p)$ if $d \leq s + p$.

Alternatives to ℓ_1 regularization have also been considered. Addressing the bias issue of ℓ_1 penalization, Lam and Fan (2009) investigated nonconvex regularizers including the hard-thresholding penalty $\rho_{\text{thresh}}(\theta) = \lambda^2 - (|\theta| - \lambda)^2\mathbb{1}_{|\theta|<\lambda}$ and the SCAD penalty $\rho_{\text{SCAD}}(\theta) = \lambda\mathbb{1}_{\theta\leq\lambda} + (\alpha\lambda - \theta)_+\mathbb{1}_{\theta>\lambda}/(\alpha - 1)$, for some $\alpha > 2$. Using linear programming Yuan (2010) propose a graphical version of the Dantzig selector (Candes and Tao, 2007) and Cai et al. (2011) propose constrained ℓ_1-minimization for inverse matrix estimation (CLIME). Each of these methods has been rigorously analyzed, providing convergence rates under high-dimensional scaling and establishing conditions for consistent model selection.

3.2 The Nonparanormal Graphical Model

Although the normality assumption can be relaxed for the sole purpose of precision matrix estimation, it remains crucial when relating sparse precision matrix estimators to sparse undirected graphical models. Yet, the joint normality assumption tends to be too restrictive in practice where data is often skewed or heavy-tailed.

As an illustration of this issue, we consider the study by Pappe et al. (2025), in which the effect of a dietary intervention—specifically, a whole-food plant-based diet—on key periodontal parameters was examined in

Table 3.1 Number of times out of 20 target variables the Null hypothesis of normality gets rejected at significance level 0.05 with and without Bonferroni correction.

	Critical value	**Lilliefors**	**Shapiro-Wilks**	**D'Agostino and Pearson**
Raw data	0.05	10	14	11
	0.05/20	5	10	7
Log transformed	0.05	8	8	7
	0.05/20	5	8	6

patients with elevated cardiometabolic risk (see also Dressler et al., 2022). A total of 36 participants were enrolled in a single-blinded randomized clinical trial. Table 3.1 reports the number of rejections of the Null hypothesis of normality, at the 0.05 significance level, for several appropriate tests applied to 20 continuous metabolic and periodontal parameters[1] measured at baseline.

Although log-transformations reduce the number of rejected Null hypotheses, the issue of non-normality persists for approximately one-third of the parameters, after applying multiple testing corrections. Transforming variables to achieve (approximate) normality has a long-standing tradition in statistical modeling (Box and Cox, 1964). However, relying on parametric transformations may introduce model misspecification. Consequently, we explore a nonparametric transformation approach designed specifically to mitigate the non-normality challenge in the estimation of undirected graphical models.

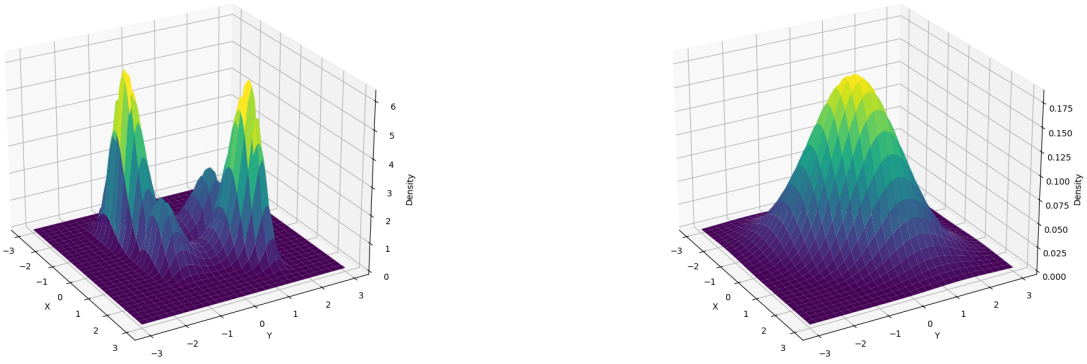
More specifically, Liu et al. (2009) propose to replace the original random variables $\mathbf{X} = (X_1, \ldots, X_p)$ with transformed ones, namely, $f(\mathbf{X}) = (f_1(X_1), \ldots, f_p(X_p))$ and assume that $f(\mathbf{X})$ is multivariate Gaussian. The result is a nonparametric extension of the Normal the so called *nonparanormal*: $\mathbf{X}$ follows a p-dimensional nonparanormal distribution if there exists univariate monotone increasing transformation functions $f = \{f_j\}_{j=1}^p$ such that the transformed variables $f(\mathbf{X})$ follow a multivariate Normal with mean 0 and covariance matrix $\mathbf{\Sigma}$. Whithout loss of generality, we set the diagonals of $\mathbf{\Sigma}$ to be equal to 1. More concisely, we say that $\mathbf{X} \sim NPN_p(\mathbf{\Sigma}, f)$ if $f(\mathbf{X}) \sim N_p(0, \mathbf{\Sigma})$.

If the transformation functions are monotone and differentiable then the nonparanormal density takes the following form:

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\det(\mathbf{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2} f(\mathbf{x})^\top \mathbf{\Sigma}^{-1} f(\mathbf{x})\right\} \prod_{j=1}^{p} \det(f_j'(\mathbf{x}_j)), \tag{3.3}$$

where $f_j'(\mathbf{x}_j)$ is the corresponding Jacobian. The form of the nonparanormal density suggests that the undirected graph of the nonparanormal model is encoded in $\mathbf{\Omega} \equiv \Sigma^{-1}$, as for the parametric Normal. This is because the density clearly factorizes with respect to the undirected graph associated with $\mathbf{\Omega}$ such that the global Markov property holds. In other words estimating the undirected graph of the nonparanormal model is equivalent to finding the sparsity pattern of the precision matrix $\mathbf{\Omega}$. Figure 3.1 illustrates the difference of a bivariate nonparanormal density and a parametric Normal density with the same covariance matrix. The nonparanormal

[1] The periodontal parameters include bleeding on probing, plaque index, clinical attachment level, and periodontal inflamed surface area. The metabolic parameters include systolic and diastolic blood pressure, fasting glucose, glycated haemoglobin percentage, high- and low-density lipoprotein levels, body mass index, and blood cholesterol.



(a) **(b)**

Figure 3.1 Densities of a two-dimensional nonparanormal with $f_j = \text{sign}(x)|x|^\alpha$, where $j \in \{1, 2\}$ and $\alpha_1 = 1.5$ and $\alpha_2 = 2.5$ in **(a)** and a two-dimensional parametric normal distribution with the same mean $\mu = (0, 0)$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 1 & .5 \\ .5 & 1 \end{smallmatrix}\right)$ and consequently the same underlying undirected graph in **(b)**.

allows for much more complex data structures while the independence relations are still encoded in the precision matrix.

Now, let $F_j(x)$ denote the marginal distribution function of X_j and observe that since $f_j(X_j)$ is standard Gaussian, we have

$$F_j(x) = P(X_j \leq x) = P(f_j(X_j) \leq f_j(x)) = \Phi(f_j(x)), \tag{3.4}$$

where $\Phi(\,\cdot\,)$ denotes the cumulative distribution function of the standard normal. Then, Eq. 3.4 implies that

$$f_j(x) = \Phi^{-1}(F_j(x)). \tag{3.5}$$

Liu et al. (2009) show that the nonparanormal family is equivalent to the Gaussian copula family (Klaassen and Wellner, 1997) when the transformation functions are differentiable. The semiparametric Gaussian copula graphical model (GCGM) has seen increased attention in recent years due to its well-balanced blend between model flexibility and interpretability (Dobra and Lenkoski, 2011; Abegaz and Wit, 2015; Mohammadi et al., 2017; Behrouzi and Wit, 2019; Hermes et al., 2024).

Estimation Now, suppose we observe a n-sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$. Liu et al. (2009) study the most natural estimation scheme for the GCGM where first the transformation functions are estimated and second well-established GGM estimation methods — such as the ones outlined in Section 3.1 — are applied to the transformed data. This idea is in spirit similar to sparse additive regression models (Ravikumar et al., 2009) where univariate transformation functions are used within a regularized risk minimization scheme. In case of the nonparanormal graphical model, a two-step procedure is more appropriate.

First, from our earlier derivation in Eq. 3.5, a natural way forward is to replace $F_j(t)$ by its empirical counterpart, the empirical distribution function

$$\widehat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_{ij} \leq t\}. \tag{3.6}$$

In order to avoid infinite values for the largest and smallest values of x_{ij} and to strike a balance between bias and variance, Liu et al. (2009) use a truncated or Winsorized[2] version of the empirical distribution function, i.e.

$$\widehat{f}_j(t) = \Phi^{-1}(W_{\delta_n}[\widehat{F}_j(t)]), \tag{3.7}$$

where

$$W_{\delta_n}(u) \equiv \delta_n \mathbb{I}(u < \delta_n) + u\mathbb{I}(\delta_n \leq u \leq (1-\delta_n)) + (1-\delta_n)\mathbb{I}(u > (1-\delta_n)).$$

The truncation constant δ_n can be chosen in several ways. For instance, in the low-dimensional regime, Klaassen and Wellner (1997) derive efficiency results by setting $\delta_n = 1/(n+1)$. For the high-dimensional regime, Liu et al. (2009) propose to use $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$ in order achieve the desired rate of convergence pertaining to the estimate of Ω and to graph recovery.

Second, the nonparanormal estimate of the precision matrix and its zero pattern can be obtained by replacing the usual sample correlation matrix with $\widehat{\mathbf{S}}^f$ in any of the methods outlined in Section 3.1, where $\widehat{\mathbf{S}}^f = (\widehat{S}^f_{jk})$ denotes the correlation matrix of the transformed data, i.e.,

$$\widehat{S}^f_{jk} = \begin{cases} \frac{\frac{1}{n}\sum_{i=1}^n \widehat{f}_j(x_{ij})\widehat{f}_k(x_{ik})}{\sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{f}_j^2(x_{ij})}\sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{f}_k^2(x_{ik})}} & \text{if } j \neq k. \\ 1 & \text{if } j = k. \end{cases} \tag{3.8}$$

When p is restricted to a polynomial order of n Liu et al. (2009) establish the following rate of convergence:

$$\|\widehat{\mathbf{S}}^f - \mathbf{\Sigma}\|_\infty = O_P\left(\sqrt{\frac{\log p \log^2 n}{n^{1/2}}}\right).$$

However, as discussed in Section 3.1, for GGMs, the dimension p is permitted to grow nearly exponentially with the sample size n. To address this, Liu et al. (2012) and Xue and Zou (2012) propose employing rank-based estimators, which circumvent direct estimation of the transformations. This approach enables improvement in convergence rates, aligning them with those obtained in the parametric GGM setting.

Let r_{ij} be the rank of x_{ij} among the observations $(x_{1j}, \ldots, x_{nj})$ for $j \in [p]$. Further, observe that $\bar{r}_j = 1/n \sum_{i=1}^n r_{ij} = n + 1/2$. Liu et al. (2012) consider Spearman'rho (Spearman, 1904) and Kendall's tau (Kendall, 1938) defined by

$$\widehat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2 \sum_{i=1}^n (r_{ik} - \bar{r}_k)^2}},$$
$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \operatorname{sign}(x_{ij} - x_{i'j})(x_{ik} - x_{i'k}),$$

respectively. Both these nonparametric correlation measures are invariant under monotone transformations. Thus, evaluating them on $\mathbf{X}$ is equivalent to evaluating them on $f(\mathbf{X})$. Since the GCGM entails $(f_j(X_j), f_k(X_k))$ to be bivariate normal with correlation parameter Σ_{jk} we can draw from classical results mapping ρ_{jk} and τ_{jk} to Σ_{jk}. Here, ρ_{jk} and τ_{jk} refer the population versions of Spearman's rho and Kendall's tau, i.e., $\rho_{jk} :=$

[2] John Tukey popularized the term "Winsorizing", i.e., the robustizing technique of replacing outliers by the nearest non-outlier value (Tukey, 1960). The technique is named after the "engeneer-turned-physiologist-turned-statistician" Charles Winsor.

$\text{corr}(F_j(X_j), F_k(X_k))$ and $\tau_{jk} \coloneqq \text{corr}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k))$, where $\widetilde{X}_j$ and $\widetilde{X}_k$ are independent copies of X_j and X_k, respectively. Then, for $\mathbf{X} \sim NPN_p(\boldsymbol{\Sigma}, f)$ we have according to Kruskal (1958)

$$\Sigma_{jk} = 2 \sin\left(\frac{\pi}{6}\rho_{jk}\right) = \sin\left(\frac{\pi}{2}\tau_{jk}\right). \tag{3.9}$$

From this, the estimators $\widehat{\mathbf{S}}^{\rho} = (\widehat{S}^{\rho}_{jk})$ and $\widehat{\mathbf{S}}^{\tau} = (\widehat{S}^{\tau}_{jk})$ for the unknown correlation matrix $\boldsymbol{\Sigma}$ are obtained by plugging in the corresponding sample versions into Eq. 3.9, i.e.,

$$\begin{aligned} \widehat{S}^{\rho}_{jk} &= \begin{cases} 2 \sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if} j = k, \end{cases} \\ \widehat{S}^{\tau}_{jk} &= \begin{cases} \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if} j = k. \end{cases} \end{aligned} \tag{3.10}$$

Importantly, Liu et al. (2012) show that despite the semi-parametric nature of $\widehat{\mathbf{S}}^{\rho}$ and $\widehat{\mathbf{S}}^{\tau}$, they both achieve the optimal parametric exponential concentration rate, i.e.,

$$\|\widehat{\mathbf{S}}^{\rho} - \boldsymbol{\Sigma}\|_{\infty} = \|\widehat{\mathbf{S}}^{\tau} - \boldsymbol{\Sigma}\|_{\infty} = O_P\left(\sqrt{\frac{\log p}{n}}\right). \tag{3.11}$$

Subsequently, as before, the rank-based nonparanormal estimate of the precision matrix and its associated sparsity pattern can be obtained by replacing the sample correlation matrix with $\widehat{\mathbf{S}}^{\rho}$ or $\widehat{\mathbf{S}}^{\tau}$ in any of the methods discussed in Section 3.1. Since these rank-based procedures attain the optimal parametric convergence rate, they can be reliably used in place of standard GGMs.

3.3 The Nonparanormal for Mixed Data

Until now, we have focused on continuous data only. In many settings however, both continuous and discrete variables are present. For instance in medical statistics, one often encounters (ordinal) discrete measurements pertaining to patients' medical histories and continuous measurements of patients' health conditions such as blood pressure or heart rate. Learning undirected graphs from such mixed data has seen much less attention than the purely continuous or purely discrete case.

There are three main lines of research that address high-dimensional graphical modeling for mixed data. The earliest approach is the conditional Gaussian model for joint modeling of categorical and continuous variables, first proposed by Lauritzen and Wermuth (1989). While conceptually elegant, this model suffers from a key limitation: the number of parameters grows exponentially with the number of discrete variables, rendering it impractical in high-dimensional settings. To mitigate this issue, Lee and Hastie (2015) — and later Cheng et al. (2017), who included interaction terms — proposed a pairwise mixed graphical model as a tractable special case.

A second approach extends neighborhood selection, as introduced in Section 3.1, to the mixed data setting. This strategy models each variable conditionally on all others, often using a separate generalized linear model for each node (Chen et al., 2015; Yang et al., 2014, 2018).

The third and most relevant line of work for our purposes builds upon GGMs and GCGMs, as introduced in Sections 3.1 and 3.2, extending these frameworks to mixed data types. To handle such mixed data, Fan et al.

(2017) introduce the latent Gaussian copula graphical model (latent GCGM). Specifically, the latent GCGM accommodates scenarios involving both binary and continuous variables by assuming that each observed binary variable arises as a thresholded version of a latent continuous variable. These latent variables, together with the directly observed continuous variables, then jointly follow a GCGM.

Formally, suppose that $\mathbf{X}$ partitions into discrete variables $\mathbf{X}_1$ with dimension p_1 — possibly ordinal — and continuous variables $\mathbf{X}_2$ with dimension p_2. Then $\mathbf{X}$ is said to follow a latent GCGM if there exists a p_1-dimensional random vector $\mathbf{Z}_1 = (Z_1, \ldots, Z_{p_1})^\top$ such that $\mathbf{Z} \coloneqq (\mathbf{Z}_1, \mathbf{X}_1) \sim NPN_p(\mathbf{\Sigma}, f)$ and

$$X_i = j \quad \text{if } \gamma_i^{j-1} \leq Z_i < \gamma_i^j \quad \forall i \in [p_1] \text{ and } j \in [l_i + 1], \tag{3.12}$$

where $\gamma_i^0 < \gamma_i^1 < \cdots < \gamma_i^{l_i+1}$ is a sequence of unknown threshold constants with $\gamma_i^0 = -\infty$ and $\gamma_i^{l_i+1} = \infty$ and $l_i + 1$ the number of levels of the discrete variable X_i. In short, we write $\mathbf{X} \sim LNPN(\mathbf{\Sigma}, f, \gamma)$, where $\gamma = (\gamma_1, \ldots, \gamma_{p_1})$ and $\gamma_i = (\gamma_i^1, \ldots, \gamma_i^{l_i})$. Note that if $\mathbf{Z} \sim N(0, \mathbf{\Sigma})$ the model reduces to a latent Gaussian model, i.e., $LN(0, \mathbf{\Sigma}, \gamma)$.

Remark 3.1. The latent GCGM is particularly well-suited for handling ordinal discrete data, as it innately assumes the existence of underlying latent continuous variables generating the observed discrete values. A straightforward example arises in medical studies, where a patient's age might be recorded within discrete intervals or brackets. In contrast, when dealing with nominal discrete variables, it is advisable to first convert these into binary indicators corresponding to each nominal category before applying the latent GCGM framework.

Fan et al. (2017) proposed a special case of the latent GCGM that specifically addresses a mix of binary and continuous variables. In this simplified setting, the authors derive bridge functions that link the latent correlation Σ_{jk} to the population Kendall's tau τ_{jk}. The derivation considers three distinct cases:

(I) Both X_j and X_k are continuous;

(II) One variable (either X_j or X_k) is binary, and the other is continuous;

(III) Both X_j and X_k are binary.

For case (I), the bridge function coincides with the one previously employed by Liu et al. (2012) and Xue and Zou (2012), as provided in Eq. (3.9). For cases (II) and (III), closed-form expressions for the bridge functions $F(\Sigma_{jk}, \gamma_j)$ and $F(\Sigma_{jk}, \gamma_j, \gamma_k)$ generally do not exist. Nevertheless, Fan et al. (2017) demonstrate that, for fixed thresholds, these bridge functions are strictly increasing in Σ_{jk}. Consequently, equations of the form $F(\Sigma_{jk}, \gamma_j) = \tau_{jk}$ and $F(\Sigma_{jk}, \gamma_j, \gamma_k) = \tau_{jk}$ have unique solutions, which can be efficiently computed numerically. Moreover, Fan et al. (2017) establish concentration inequalities yielding the same rate of convergence as stated in Eq. (3.11). Thus, the theoretical properties of the latent GCGM are essentially equivalent to those obtained if the latent variables $Z_1, \ldots, Z_{p_1}$ were directly observed.

Yet, when considering a discrete mix beyond the binary case, Fan et al. (2017) advocate for the binarization of all discrete variables. Accordingly, this idea was further developed by Feng and Ning (2019), who proposed initially binarizing all discrete variables to generate preliminary estimators, which are then combined through a weighted aggregate. Quan et al. (2018) extended the binary latent Gaussian copula model by considering scenarios involving continuous, binary, and ternary variables. However, this extension highlights a significant

drawback: complexity grows substantially as discrete variables with distinct state spaces are introduced. Specifically, for a mix involving continuous variables and discrete variables spanning k distinct state spaces, $\binom{k+2}{2}$ different bridge functions are required.

In the first core article (Göbler et al., 2024a), we address the estimation of the general latent GCGM by integrating classical statistical concepts — specifically, polychoric and polyserial correlations — with the latent GCGM framework. The foundations of these ideas date back to the early 1900s, when Pearson (1900, 1913) introduced the tetrachoric and biserial correlation coefficients. These concepts subsequently led to maximum likelihood estimators (MLEs) for the general case, in which discrete variables may assume an arbitrary number of states (Olsson, 1979; Olsson et al., 1982). This generality implies that it suffices to consider only the three previously outlined cases (I), (II), and (III), regardless of the number of the discrete variables with distinct state spaces.

Our theoretical analysis in (Göbler et al., 2024a) further reveals that the proposed estimators achieve optimal convergence rates in cases (I) and (III). For case (II), however, the additional modeling flexibility necessitates an initial step of estimating the transformation functions of the observed continuous variables, in a manner similar to that proposed by Liu et al. (2009). This requirement leads naturally to a two-step estimation procedure: first, we estimate the transformation functions for all continuous variables $X_{p_1+1}, \ldots, X_{p_2}$; second, we employ these estimated transformations in place of the observed continuous variables to define an *ad-hoc* polyserial correlation coefficient within the nonparanormal setting. Due to the first stage estimation of the transformation functions the rate in supremum norm for case (II) amounts to

$$\|\widehat{\mathbf{S}}^{\text{case (II)}} - \boldsymbol{\Sigma}\|_\infty = O_p\left(\sqrt{\frac{\log p \log n}{\sqrt{n}}}\right). \tag{3.13}$$

Overall, in our first core article (Göbler et al., 2024a), we provide comprehensive theoretical analysis, simulation studies, and an empirical evaluation using UK Biobank data on severe COVID-19 patients. Through this investigation, we rigorously characterize the properties of our proposed estimators and demonstrate their practical usability and broad applicability.

4 Generating Realistic Production Data for Benchmarking Causal Discovery

Methodological progress in machine learning has historically been closely tied to the availability and quality of meaningful benchmark datasets and tools (Langley, 2011; Longjohn et al., 2024). When benchmark datasets are both widely accessible and representative of real-world tasks, they play a crucial role in evaluating the performance of newly developed methods (Raji et al., 2021). As the machine learning community has grown, benchmarking repositories have become increasingly sophisticated, covering a broad range of domains and data types.

In contrast, benchmarking has advanced far more slowly in the field of causal discovery. One key reason is the inherent difficulty in obtaining ground truth causal relationships in real-world settings. Even when such relationships are available, there is often considerable uncertainty surrounding their accuracy and completeness (Ramsey and Andrews, 2018). The reason is that the stated ground truth causal model contains predictions about manipulated (i.e., intervened upon) populations that might not exist or be observable in the real world (Spirtes and Zhang, 2016). Additionally, access to high-quality datasets is frequently restricted due to privacy concerns or legal constraints, such as intellectual property protections.

In the second core article (Göbler et al., 2024b), we introduce a semisynthetic data generator for benchmarking causal discovery methods. It is based on a complex real-world dataset collected along a manufacturing assembly line. By analyzing the underlying physical processes, we identify a partial ground truth of causal relationships that is rich enough to estimate a full causal graph using nonparametric regression. To address confounding and privacy concerns, we employ distributional random forests (Ćevid et al., 2022) to estimate the conditional distributions implied by the ground truth. These are then combined into a joint distribution that strictly adheres to a causal model over the observed variables.

4.1 Causal Discovery

To clarify the essential properties of the benchmarking tool introduced in our second core article (Göbler et al., 2024b), we begin by outlining the most prominent classes of causal structure learning methods. Since the real-world data used in that work consist exclusively of continuous variables, our discussion of causal discovery methods is limited to algorithms designed for continuous data.

Recall from Section 2.2.2 that causal graphical models are a special class of graphical models in which edges represent direct causal effects. This causal interpretation enables the prediction of manipulated, or interventional, distributions and thereby the estimation of causal effects (Spirtes and Zhang, 2016). In observational settings —

i.e., when no experimental interventions are deliberately performed — the underlying causal graph is generally unknown and must be inferred from data. This task is known as causal discovery or causal structure learning.

A wide array of methodologies has been proposed to infer causal structures from observational joint distributions. As discussed in Section 2.2, when no assumptions are imposed on the functional form of the causal mechanisms, the underlying graph G can only be identified up to its MEC using observational data alone. Consequently, in the absence of additional assumptions it is natural to perform searches within the space of MECs rather than among individual DAGs.

In what follows, we discuss two main categories of approaches for causal discovery: *constraint-based* and *score-based* methods. Constraint-based methods formulate structure learning as a constraint satisfaction problem (Squires and Uhler, 2022). These methods leverage observed conditional independences and other structural constraints (see Peters et al., 2017, Chapter 9.5) to iteratively narrow down possible candidate graphs. Conversely, score-based methods approach structure learning as a combinatorial optimization problem. They assign scores to candidate graphs and aim to identify the graph that maximizes the score by efficiently traversing the relevant graph space.

4.1.1 Constraint-based methods

The most prominent constraint-based method is perhaps the PC Algorithm, named after the inventors Peter Spirtes and Clark Glymour (Spirtes et al., 1993, 2000). The algorithm has two phases and outputs a CPDAG. The first phase starts with a complete undirected graph and iteratively removes edges by testing conditional independences with conditioning sets of increasing cardinality. The second phase attempts to orient as many of the edges as possible found in the first phase. This is done by first determining v-structures and then applying the *Meek orientation rules* (Meek, 1995) to the remaining edges.

Being able to set up appropriate conditional independence (CI) tests plays a vital role in constraint-based methods with CI constraints such as the PC algorithm. Which CI method to employ depends on data characteristics, modeling assumptions and considerations regarding computational complexity. For instance, for multivariate Gaussian data zero partial correlation implies conditional independence such that hypothesis tests are readily constructed and are computationally efficient (Colombo and Maathuis, 2014, see e.g.). Utilizing the mapping between the rank-based measures Kendall's tau and Spearman's rho (Eq. (3.9) in Section 3.2) and the underlying Gaussian correlation parameter Harris and Drton (2013) obtain a computationally efficient version of the PC algorithm for the broader class of nonparanormal distributions. Conversely, nonparametric conditional independence testing (Zhang et al., 2011; Zheng et al., 2020) while often more appropriate in practice is notoriously hard in the sense that no *uniformly valid* CI test can have any statistical power (Shah and Peters, 2020). This means that CI tests in the nonparametric setting require further assumptions on the set of possible distributions (Squires and Uhler, 2022).

4.1.2 Score-based methods

A prominent example of score-based methods is greedy equivalence search (GES) (Chickering, 2003). In the first (forward) phase of the algorithm, single edges are added (greedily) to the initial graph estimate according to the largest score improvement. In the second (backwards) phase, individual edges are deleted from the final

graph estimate from the first phase until the score can no longer be improved. The search space for GES is the space of all CPDAGs since DAGs that belong to the same MEC are assigned the same score. This property of some score $S(G)$ is called *score equivalence*. Another desirable property of the score is *decomposability*, where any single edge operation only changes the score locally, such that recalculating the score after an update can be done fast and efficiently (see, e.g. Maathuis and Nandy, 2016). In cases where the score $S(G)$ satisfies score equivalence, decomposability, and *consistency* — i.e., the true CPDAG is assigned the highest score with probability tending to one — GES can be shown to be consistent.

Scores based on the marginal likelihood such as the Bayesian Information Criterion (BIC) satisfy the above properties for Gaussian or multinomial distributions. Taking this as a starting point, nonparametric extensions through nonparanormal distributions (Nandy et al., 2018) or Gaussian process priors (Friedman and Nachman, 2000) or more intricate scores altogether (Brenner and Sontag, 2013) have been developed.

Since finding the DAG with the highest score has been shown to be NP-complete (Chickering, 1996), score-based methods need to trade off computational efficiency against guarantees regarding algorithmic consistency (Squires and Uhler, 2022). On one end of this trade-off *exact* score based methods are based on the goal to exactly find some graph $\widehat{G}$ that exactly optimizes some score S. Typically, some combinatorial optimization techniques and heuristics, such as dynamic programming (Koivisto and Sood, 2004; Parviainen and Koivisto, 2011) or integer linear programming (Bartlett and Cussens, 2017; Cussens, 2020; Chen et al., 2021). GES and its extensions are located at the middle of the trade-off continuum since they are no longer exact but still consistent. On the other end of the trade-off dwell gradient-based routines where the search space over DAGs is relaxed to be continuous rather than discrete. This allows highly performant continuous optimization techniques including deep learning architectures to be used for causal structure learning (Zheng et al., 2018; Lachapelle et al., 2020; Ng et al., 2020). Due to the non-convex nature of the involved optimization problem consistency guarantees are challenging to establish (Deng et al., 2024).

4.1.3 Functional form restricting methods

When imposing suitable constraints on the functional form of the generating SEM, one can leverage induced asymmetries in the joint observational distribution to identify the true underlying DAG. This central idea is most clearly illustrated in the bivariate scalar case, described by the SEM

$$X = N_X, \quad Y = f_Y(X, N_Y), \tag{4.1}$$

with corresponding DAG G given by $X \to Y$. We previously noted that the DAG obtained by reversing the direction of the edge, i.e., $X \leftarrow Y$, lies within the same MEC as G. Consequently, *identifying* the true DAG G solely from observational data is impossible without imposing additional assumptions.

This can be more elegantly understood through the following reasoning: if the reversed causal direction were valid, one could represent X as a function of Y and some exogenous noise N_X, satisfying $Y \perp\!\!\!\perp N_X$. In fact, without restrictions on the functional class $\mathcal{F}$, Hyvärinen and Pajunen (1999) show that there always exists a suitable function $f \in \mathcal{F}$ ensuring $Y \perp\!\!\!\perp N_X$. Thus, the absence of constraints on $\mathcal{F}$ renders the SEM symmetric with respect to variables X and Y.

Subsequently, we discuss the most important restrictions imposed on the function class $\mathcal{F}$ that facilitate the *identification* of the causal direction between X and Y. For the purposes of the present Chapter, *structure*

identifiability refers to the ability to determine the causal direction between X and Y from observational data alone. In Section 5.1, we provide a more detailed discussion of the notion of identifiability in the context of bivariate SEMs. Throughout, the term *identifiability* is used to refer specifically to structure identifiability. Although our exposition focuses on the bivariate scenario for simplicity and clarity, the identifiability results we outline readily extend to the general multivariate case (Peters et al., 2011b).

Identifiability in linear SEMs The first restriction discussed in this context is to require the function class $\mathcal{F}$ to contain only linear functions, i.e.,

$$Y = bX + N_Y. \tag{4.2}$$

In this context, Shimizu et al. (2006) demonstrated that provided at most one of the variables X and N_Y is Gaussian, the reversed causal direction does not produce independence between Y and N_X. Independent component analysis (ICA) (Comon, 1994) can be employed to estimate this model, referred to as the linear non-Gaussian acyclic model (LiNGAM). Subsequently, Shimizu et al. (2011) introduced a direct estimation procedure specifically tailored for the LiNGAM framework. Although linear ICA indicates that a linear Gaussian SEM generally does not allow ruling out the reverse causal direction, Peters and Bühlmann (2013) showed that under the special condition of Gaussian noise variables with equal variances the reverse causal direction can be ruled out.

Identifiability in nonlinear SEMs A further restriction on the function class $\mathcal{F}$ is introduced by the nonlinear additive noise model (ANM) (Hoyer et al., 2008), wherein—as the name indicates—the noise enters the effect variable Y additively:

$$Y = f(X) + N_Y, \tag{4.3}$$

with f being a three-times differentiable nonlinear function. Hoyer et al. (2008) demonstrate that, except for certain specific combinations of the function f, the distribution P_X, and the noise distribution P_{N_Y}, the causal direction is identifiable. Notably, the introduction of nonlinearity breaks the symmetry between the observed variables, thus enabling inference of the causal direction. Chapter 5 provides a detailed discussion of the ANM framework. Several causal discovery algorithms have emerged from the ANM approach, including the regression with subsequent independence test (RESIT) algorithm (Peters et al., 2014), the SCORE algorithm (Rolland et al., 2022) based on score matching, and the Gradient-Based Neural DAG learning (GraN-DAG) algorithm (Lachapelle et al., 2020).

A more general identifiable model class is provided by the post-nonlinear causal model (PNL) (Zhang and Hyvärinen, 2009), where the effect Y undergoes an additional nonlinear distortion beyond the ANM framework. Specifically, the PNL model is defined as

$$Y = g(f(X) + N_Y), \tag{4.4}$$

where g is a three-times differentiable and invertible nonlinear function. The additional transformation g can be interpreted as representing nonlinear sensor or measurement errors, phenomena frequently encountered in real-world applications. Due to this increased complexity, the PNL model generally poses greater challenges in estimation compared to the ANM. In the bivariate case, Zhang and Hyvärinen (2009) propose estimating the

Table 4.1 Overview of properties, assumptions, and inferred graphs for the causal discovery classes examined in this chapter, with representative algorithms.

	Class	Suff.	Faith.	Further constr.	Output	Identif.	Consist.
PC (Spirtes et al., 1993)	*CB*	✓	✓	✗	CPDAG	CPDAG	✓
rankPC (Harris and Drton, 2013)		✓	✓	✗	CPDAG	CPDAG	✓
stablePC (Colombo and Maathuis, 2014)		✓	✓	✗	CPDAG	CPDAG	✓
ExactBN (Koivisto and Sood, 2004)	*SB*	✓	✓	✗	DAG	CPDAG	✓
GES (Chickering, 2003)		✓	✓	✗	CPDAG	CPDAG	✓
NOTEARS (Zheng et al., 2018)		✓	✓	✗	DAG	CPDAG	✗
DirectLiNGAM (Shimizu et al., 2011)	*FFR*	✓	✓	✓	DAG	DAG	✓
EqualVar (Chen et al., 2019)		✓	✓	✓	DAG	DAG	✓
RESIT (Peters et al., 2014)		✓	✓	✓	DAG	DAG	✓
SCORE (Rolland et al., 2022)		✓	✓	✓	DAG	DAG	✓
GraN-DAG (Lachapelle et al., 2020)		✓	✓	✓	DAG	DAG	✗

Suff.: causal sufficiency, Faith.: causal faithfulness, Further constr.: further constraints of the functional form
Identif.: Identifiable given the set of assumptions, Consist.: consistency.
CB: constraint-based, *SB*: score-based, *FFR*: functional form restricting

PNL by minimizing the mutual information between the estimated noise and the potential cause, employing multi-layer perceptrons (MLPs) to approximate the nonlinear functions f and g in both causal directions. The correct orientation is subsequently identified using kernel-based independence tests (Gretton et al., 2005) between the estimated noise terms and the candidate causes. More recently, Uemura et al. (2022) have extended this estimation approach to the multivariate setting.

In multivariate settings, most functional form restricting methods belong to the class of *order-based* methods (Teyssier and Koller, 2005). First, a (causal) topological ordering is obtained, where each node can only have parents that precede it in the topological ordering. Second, a super-DAG with all edges being present that obey the causal ordering is constructed. Then, superfluous edges are pruned for instance by sparse regression techniques (Peters et al., 2014). In Chapter 5, we discuss sparse regression for edge-pruning in greater detail.

In multivariate settings, most approaches based on restricting the functional form belong to the class of *order-based* methods (Teyssier and Koller, 2005). These methods typically proceed in two steps: first, a (causal) topological ordering is determined, ensuring each node can have only parents that precede it in the ordering. Second, a super-DAG is constructed, initially containing all edges consistent with this ordering. Subsequently, superfluous edges are pruned, e.g., using sparse regression techniques (Peters et al., 2014). Sparse regression as an edge-pruning strategy is discussed in greater detail in Chapter 5. Table 4.1 provides an overview of the causal discovery methods discussed above, their properties, and references to representative algorithms for each class.

In the second core article (Göbler et al., 2024b), a valid causal ordering is derived from expert knowledge combined with the layered structure inherent to the assembly line. Given that domain knowledge precludes edges across layers, the corresponding super-DAG is pruned using sparse additive models (SpAM) (Liu et al., 2007). The resulting DAG is then considered as the ground truth, serving as the basis for generating semisynthetic data. Leveraging this semisynthetic data generator, we benchmark several prominent constraint-based, score-based, and functional form restricting causal discovery algorithms, including the PC algorithm (Spirtes et al., 1993), NOTEARS (Zheng et al., 2018), GraN-DAG (Lachapelle et al., 2020), DirectLiNGAM (Shimizu et al., 2011), and SCORE along with its scalable extension DAS (Rolland et al., 2022; Montagna et al., 2023).

5 Nonlinear causal discovery for grouped data

The third core article of this thesis (Göbler et al., 2025) focuses on continuous additive noise models (ANMs) in a setting where the causal entities of interest are random vectors (groups) rather than scalar random variables. The treatment of groups of variables among which causal relations are to be determined arises in many real-world settings. Examples include neuroscience where causal relations among brain regions may constitute the research objective (Panzeri et al., 2017; Kohn et al., 2020). Another example is psychology where measurements are taken of certain psychological or societal traits based on several proxy variables that need to be treated jointly (Cronbach and Meehl, 1955; Antonoplis, 2022).

5.1 Identifiability

In Section 4.1, we briefly reviewed identifiability results for scalar, continuous, nonlinear additive noise models (ANMs). In this section, we revisit these results in greater detail, laying the groundwork for their extension to the group setting. Throughout this chapter we focus on the continuous case; for a treatment of the scalar discrete case, see Peters et al. (2011a).

Consider the following ANM over the two random variables X and Y,

$$X = N_X, \quad Y = f_Y(X) + N_Y, \tag{5.1}$$

where N_X and N_Y have strictly positive densities with respect to the Lebesgue measure, and f_Y is a three-times differentiable, nonlinear, and non-constant function. In this case, Peters et al. (2014) show that the requirement that f_Y is non-constant is equivalent to assuming *causal minimality*, a condition that is similar to but weaker than *causal faithfulness*. A distribution satisfies causal minimality with respect to a graph G if the Markov condition holds for G but fails for every proper subgraph $\widetilde{G} = (V, \widetilde{E})$ of G where $\widetilde{E} \subset E$. Given samples from a distribution generated by such an ANM, Hoyer et al. (2008) demonstrate that the causal direction is *generically identifiable*.

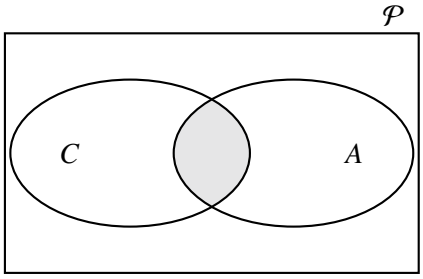


Figure 5.1 Depiction of the joint distributions that may be generated from the causal SEM (C) and the anticausal SEM (A) inside the set of all possible joint distributions $\mathcal{P}$.

We now make the notion of identifiability more precise. Consider Figure 5.1, which depicts the set of joint distributions that can be generated from the SEM with the *causal* graph $X \to Y$ (denoted C) and from the SEM with the *anti-causal* direction $X \leftarrow Y$ (denoted A), both contained within the set of all possible joint distributions $\mathcal{P}$. When we speak of identifiability, we are essentially reasoning about the size of the

intersection $C \cap A$. If C and A were to contain almost the same set of joint distributions, we would regard the model class as non-identifiable. Conversely, if the intersection is very small, we would regard the model class as identifiable.

For ANMs, we already know that this intersection cannot be empty, due to the well-known linear Gaussian case, which admits both directions. We illustrate this in the following example.

Example 5.1. Suppose we have

$$Y = \beta X + N_Y, \quad N_Y \perp\!\!\!\perp X, \tag{5.2}$$

with linear coefficient $\beta \in \mathbb{R}$, and X and N_Y normally distributed (w.l.o.g. with zero mean) and variances σ_X^2 and $\sigma_{N_Y}^2$, respectively. The joint distribution generated by this SEM is bivariate normal with covariance

$$\mathrm{Cov}(X, Y) = \beta \sigma_X^2. \tag{5.3}$$

If both directions are admissible, we can construct a new noise term N_X such that

$$X = \gamma Y + N_X, \quad N_X \perp\!\!\!\perp Y. \tag{5.4}$$

From the geometry of linear regression, it follows that

$$N_X = X - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} Y, \tag{5.5}$$

and therefore

$$X = \gamma Y + N_X, \quad \gamma = \frac{\beta \sigma_X^2}{\beta^2 \sigma_X^2 + \sigma_{N_Y}^2}. \tag{5.6}$$

By construction, N_X and Y are uncorrelated:

$$\mathrm{Cov}(N_X, Y) = \mathrm{Cov}\left(X - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)} Y,\, Y\right) = \mathrm{Cov}(X, Y) - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}\, \mathrm{Var}(Y) = 0. \tag{5.7}$$

Since all variables are Gaussian, uncorrelatedness implies independence, so $N_X \perp\!\!\!\perp Y$. Lastly, observe that

$$\gamma = \frac{\beta \sigma_X^2}{\beta^2 \sigma_X^2 + \sigma_{N_Y}^2} = \frac{\beta}{\beta^2 + \sigma_{N_Y}^2 / \sigma_X^2} \neq \frac{1}{\beta} \tag{5.8}$$

whenever $\sigma_{N_Y}^2 > 0$.

From linear ICA (Comon, 1994; Shimizu et al., 2006) it is known that the statement also holds in reverse: if the reverse causal direction is admissible and the SEM is linear, then X and N_Y must be Gaussian. Consequently, Hoyer et al. (2008) addressed the question of whether it is possible to characterize the intersection $C \cap A$ for nonlinear ANMs.

To answer this question, suppose that the graph is not identifiable, so that the joint densities factorizing according to the DAGs implied by the causal and anti-causal SEMs are equal, i.e.,

$$p_{N_Y}(y - f(x))\, p_X(x) = p_{XY}(x, y) = p_{N_X}(x - g(y))\, p_Y(y). \tag{5.9}$$

Define

$$\pi_1(x, y) := \log\big[p_{N_Y}(y - f(x))\, p_X(x)\big] = \nu\big(y - f(x)\big) + \xi(x), \tag{5.10}$$

and

$$\pi_2(x, y) \coloneqq \log\big[p_{N_X}(x - g(y))\, p_Y(y)\big] = \mu\big(x - g(y)\big) + \eta(y), \tag{5.11}$$

for $\nu \coloneqq \log p_{N_Y}$, $\xi \coloneqq \log p_X$, $\mu \coloneqq \log p_{N_X}$, and $\eta \coloneqq \log p_Y$. It is straightforward to verify that

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 \pi_2(x, y)/\partial x^2}{\partial^2 \pi_2(x, y)/\partial x\, \partial y}\right) = 0. \tag{5.12}$$

Since $\pi_1(x, y) = \pi_2(x, y)$, we also have

$$Q_1 \coloneqq \frac{\partial}{\partial x}\left(\frac{\partial^2 \pi_1(x, y)/\partial x^2}{\partial^2 \pi_1(x, y)/\partial x\, \partial y}\right) = 0, \tag{5.13}$$

which yields the following expression (arguments omitted for readability):

$$Q_1 = -2f'' + \frac{\nu' f'''}{\nu'' f'} - \frac{\xi'''}{\nu'' f'} + \frac{\nu' \nu''' f''}{(\nu'')^2} - \frac{\nu'(f'')^2}{\nu''(f')^2} - \xi'' \frac{\nu'''}{(\nu'')^2} + \xi'' \frac{f''}{\nu''(f')^2}. \tag{5.14}$$

Setting $Q_1 = 0$ and rearranging terms gives the differential equation

$$\xi''' = \xi''\left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu''' \nu' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}. \tag{5.15}$$

Thus, the size of the intersection $C \cap A$ for nonlinear ANMs is fully characterized by all cases in which the triple (f, P_X, P_{N_Y}) satisfies Eq. (5.15). Broadly speaking, if we fix (f, P_{N_Y}), the set of all distributions P_X that satisfy Eq. (5.15) lies in a three-dimensional space. Since the space of continuous distributions is infinite-dimensional in general, the backwards model is inadmissible for *most* distributions P_X. Consequently, Hoyer et al. (2008) conclude that for *generic* triples (f, P_X, P_{N_Y}), the graph is *generically identifiable*.

Although Eq. (5.15) demonstrates that $C \cap A$ is indeed a small set, we still lack a concrete understanding of the form of (f, P_X, P_{N_Y}) that satisfies Eq. (5.15). In this context, Zhang and Hyvärinen (2009) derive five classes of triples (f, P_X, P_{N_Y}) for which the graph is non-identifiable:

1. X and N_Y are Gaussian, f is linear.
2. X and N_Y are log-mix-lin-exp, f is linear.
3. X is log-mix-lin-exp, N_Y is one-sided asymptotically exponential, and f is strictly monotonic with $f'(x) \to 0$ as $x \to \pm\infty$.
4. X is log-mix-lin-exp, N_Y is a generalized mixture of two exponentials, and f is strictly monotonic with $f'(x) \to 0$ as $x \to \pm\infty$.
5. X is a generalized mixture of two exponentials, N_Y is two-sided asymptotically exponential, and f is strictly monotonic with $f'(x) \to 0$ as $x \to \pm\infty$.

We refer to Zhang and Hyvärinen (2009) for precise definitions of these distributional families. Suppose we are given the distribution generated from an identifiable bivariate ANM, in other words, the triple (f, P_X, P_{N_Y}) does not belong to either of the five classes. Then, the regression in the causal direction: $Y - \mathbb{E}[Y \mid X] = Y - f(X) = N_Y$ implies $N_Y \perp\!\!\!\perp X$. However, the residuals obtained from the regression in the anti-causal direction $X - \mathbb{E}[X \mid Y]$ are not independent of Y. Figure 5.2 illustrates this concept. The panel on the left plots

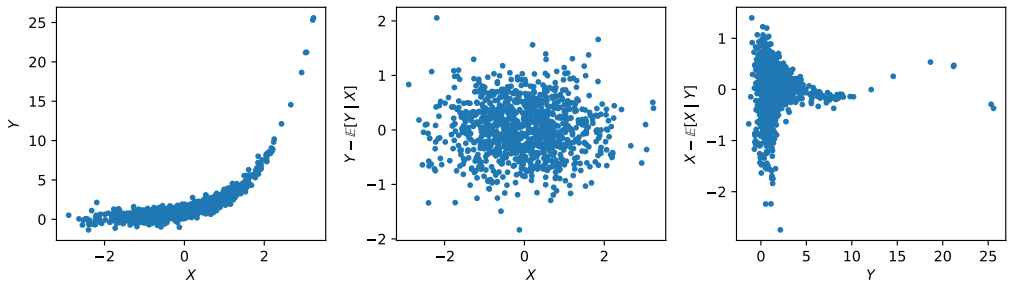


Figure 5.2 Illustration of bivariate identifiability in nonlinear additive noise models. The triple (f, P_X, P_{N_Y}) is chosen to be as follows: $(f = \exp, P_X = \mathcal{N}(0,1), P_{N_Y} = \mathcal{N}(0,1))$. The sample size is $n = 1,000$. The left panel demonstrates the relationship between X and Y. The middle panel demonstrates independence between cause and residual when regressed in the *correct* direction, while the right panel shows the lack of independence in the *anti-causal* direction.

X against Y where f is chosen to be the exponential function. Further, X and N_Y are generated from a standard Normal and $n = 1,000$ samples are drawn. The middle and right panel of Figure 5.2 plot the residuals from the two regressions against the supposed cause. Clearly, the regression in the causal direction (middle panel) leads to independent residuals while the regression in the anti-causal direction (right panel) does not.

Multivariate case So far, we have treated only the bivariate case. The reason for this is that Peters et al. (2014) demonstrate that an identification result analogous to Eq. (5.15) is sufficient to extend identifiability to the multivariate case. Suppose we are given a SEM on p variables, i.e.,

$$X_j = f_j(\mathbf{X}_{pa(j)}) + N_j, \quad j = 1, \dots, p. \tag{5.16}$$

If we fix all arguments of the function f_j except for one, we again obtain a bivariate model. However, the triple

$$\big(f_j(\mathbf{x}_{pa(j)\setminus\{i\}}, \underbrace{\cdot}_{X_i}),\ P_{X_i},\ P_{N_j}\big),$$

where the underbrace indicates the one parent variable that is not fixed, does not in general yield a condition analogous to Eq. (5.15) (Peters et al., 2014, Example 26). Instead, we require that there exists some set S among the ancestors of j, i.e.,

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

for each $j \in [p]$ such that the triple from such a *restricted ANM*

$$\big(f_j(\mathbf{x}_{S\setminus\{i\}}, \underbrace{\cdot}_{X_i}),\ P_{X_i|X_S},\ P_{N_j}\big),$$

does not satisfy the same differential equation as Eq. (5.15). Under this condition, the orientation between X_i and X_j is generically identifiable for all $\{i, j\} \in [p]$. In the third core article of the thesis (Göbler et al., 2025), we extend these identifiability results to the case of groups of variables.

Beyond the bivariate case, the theoretical identifiability results do not directly yield a finite-sample method for estimating the DAG under a *restricted ANM*. For instance, Rolland et al. (2022) propose the SCORE algorithm which utilizes score matching to estimate a causal ordering in the context of multivariate ANMs. In particular,

they show that, for multivariate ANMs with Gaussian noise, sink nodes can be identified by analyzing their entailed observational score. An estimate of the causal order is then obtained by iteratively removing sink nodes from the dataset. In similar spirit, Peters et al. (2014) exploit the fact that, for any node i in a SEM, its corresponding noise term is (unconditionally) independent of all other variables. This property allows the recursive identification of sink nodes, as they are independent of all remaining variables. The corresponding RESIT algorithm (see Section 4.1) first regresses each node on all others to obtain estimates of the noise terms, and then identifies a sink node by selecting the variable with the *most independent* residual. This node is prepended to the causal order and subsequently removed from the analysis. Repeating these steps $p - 1$ times yields an estimate of the full causal ordering.

Since RESIT is agnostic to the dimensionality of the involved random quantities, in our third core article of this thesis (Göbler et al., 2025) we apply the same procedure in the group case. Specifically, we repeatedly solve the multi-task regression problems by training multi-output deep neural networks whose output layers are sized to match the number of response variables in each group. We then compute the residuals and identify the group whose residuals exhibit the weakest dependence on the remaining variables. This group is designated as a *sink node* and subsequently removed from the set of variables. Dependence is measured via the HSIC (Gretton et al., 2005). Note that nonparametric independence tests are necessary, since the residuals are uncorrelated with the predictors by construction. Surprisingly, this procedure requires only $O(p^2)$ regressions and subsequent independence tests, making it polynomial in the number of variables. While most problems in causal structure learning are NP-hard, the scalability here ultimately depends on how well both the nonparametric regression method and the independence test perform as the number of nodes increases.

5.2 Pruning

For order-based methods such as SCORE and RESIT, causal structure learning is carried out in two phases. In the first phase, an estimate of the causal order π is obtained. In the second phase, a graph G^π is constructed by inserting all edges that are consistent with the causal order found in the first phase. If the estimated causal order agrees with a true causal order, then clearly $G^\pi \supseteq G$. Subsequently, superfluous edges are pruned to obtain the final graph estimate. Conceptually, the pruning phase amounts to solving several feature selection problems: for each node j in G^π, select only the active parents, i.e., relevant features among nodes that precede j in the estimated causal order π (see Bühlmann et al., 2014). These feature selection problems are typically solved either by hypothesis testing (Marra and Wood, 2011; Peters et al., 2014) or by sparse regression techniques (Liu et al., 2007; Ravikumar et al., 2009; Meier et al., 2009). In what follows, we focus on sparse regression techniques, as hypothesis testing performed consecutively may lead to order dependence (see Peters et al., 2014, Algorithm 1).

In the context of variable selection, sparse additive models (SpAM) have been shown to recover the correct sparsity pattern asymptotically (Ravikumar et al., 2009). More specifically, consider the regression function for node j in the super-DAG G^π, i.e.,

$$h_j(x) = \mathbb{E}[X_j \mid \mathbf{X}_{pa(j)} = \mathbf{x}_{pa(j)}]. \tag{5.17}$$

SpAM approximates this regression function by sums of the form

$$h_j(x) \approx \sum_{k \in S} h_{jk}(x_k), \quad h_{jk} \in \mathcal{H}_k, \text{ for } k \in S, \tag{5.18}$$

where the index set of nonzero components $S = \{ k : h_{jk} \neq 0 \}$ is a subset of $pa(j)$. Here, $\mathcal{H}_k$ denotes the Hilbert subspace of $L_2(P)$ consisting of P-measurable functions h_{jk} of the variable X_k. Since a direct sparse approximation to the regression function is both non-convex and computationally intractable, SpAM instead defines an alternative type of best sparse approximation via

$$\min_{h_{jk} \in \mathcal{H}_k,\, k \in pa(j)} \left\{ \mathbb{E}\left[X_j - \sum_{k \in pa(j)} h_{jk}(X_k) \right]^2 + \lambda \sum_{k \in pa(j)} \|h_{jk}\|_2 \right\}, \tag{5.19}$$

for some $\lambda \geq 0$ and $\|h_{jk}\|_2 = \sqrt{\mathbb{E}\big(h_{jk}(X_k)^2\big)}$. In terms of edge pruning, the final edge set of the pruned DAG is obtained by

$$\widehat{E} = \left\{ (k, j) : \widehat{h}_{jk} \neq 0 \right\}, \quad \forall j \in V,\ k \in pa(j). \tag{5.20}$$

In the second core article of this thesis (Göbler et al., 2024b), we apply SpAM directly after obtaining a valid causal order via expert domain knowledge, thereby estimating the DAG that we consider as ground truth. In the third core article (Göbler et al., 2025), the pruning phase amounts to variable selection among *groups* of nodes — a problem that has not yet been addressed in the literature, as both the response variables and all predictor variables are grouped. The case of scalar responses with grouped predictors has been addressed by Yin et al. (2012) with their group SpAM. If only the response variable is grouped and the remaining predictors are singletons, then the resulting model corresponds to the multi-response SpAM discussed by Liu et al. (2008). In the third core article, we combine these two approaches into the class of *multi-response group sparse additive models*.

6 Conclusion

This dissertation examines structure learning in both undirected and directed graphical models, with a focus on high-dimensional settings. High-dimensional undirected graphical models are a powerful tool for capturing complex dependencies in data. However, most methodological developments have focused on fully continuous or fully discrete settings, leaving a gap in the literature for mixed data scenarios. This dissertation addresses this gap by introducing methodologies for structure learning that can handle arbitrary mixed data types, thereby broadening the applicability of undirected graphical models in real-world high-dimensional contexts.

Directed probabilistic graphical models, particularly when framed through structural equation models, have also been an active area of research due to their ability to represent causal relationships. However, the number of possible directed acyclic graphs (DAGs) grows super-exponentially with the number of nodes, making causal structure learning with many variables especially challenging. This dissertation investigates important obstacles that practitioners frequently encounter in this setting. One such obstacle is the scarcity of high-quality benchmark datasets for evaluating causal discovery methods. Leveraging detailed domain knowledge from an industrial assembly line, we construct plausible ground-truth causal relationships. To ensure that the resulting observational distribution is Markov with respect to this ground-truth graph, we fit conditional distributions to real data and use them to generate semisynthetic data. This yields a flexible data generator that can produce samples of varying dimensions and sizes, enabling systematic benchmarking of causal discovery methods.

Another common scenario in real-world applications involves settings where the causal entities of interest are random vectors (groups) rather than scalar variables. We extend the additive noise model (ANM) framework to this grouped setting, establish structure identifiability results, and adapt the RESIT algorithm accordingly. As part of this adaptation, we introduce multi-response group sparse additive models for the pruning step common to order-based grouped causal discovery methods, which reduce to SpAM (Liu et al., 2007; Ravikumar et al., 2009) when all groups are singletons.

The proposed methodologies for undirected and directed graphical models serve as building blocks for advancing structure learning in real-world problems involving complex data types, such as mixed or grouped variables. For our first core contribution, a natural extension is to handle settings that combine continuous, latent continuous, and truly discrete variables, and to explore the integration of polyserial and polychoric correlation measures into directed graphical models. Regarding our second core contribution, we hope that the proposed semisynthetic pipeline will encourage the release of additional datasets, even under privacy constraints, thereby enabling the community to evaluate methods on a broader and more diverse set of benchmarks. Finally, for the grouped ANM framework, an important direction for future work is to obtain a full characterization of the non-identifiable cases, analogous to the scalar setting. This also includes studying the effect of group dimension on identifiability, which we suspect may help break symmetries in the observational distribution.

7 Summary of Core Publications

7.1 High-dimensional undirected graphical models for arbitrary mixed data

In the first core article, we address the pervasive challenge posed by datasets comprising both continuous and discrete variables in the context of high-dimensional undirected graph learning. Traditional graphical modeling approaches typically assume homogeneous data, either entirely continuous (Gaussian graphical models) or discrete (Ising-type models). However, practical scenarios often involve heterogeneous data types, rendering standard tools inapplicable.

To address this challenge, this article proposes a comprehensive framework grounded in latent Gaussian models and latent Gaussian copula models, where observed mixed data arise from underlying continuous latent variables. Specifically, we build upon classical statistical concepts such as polychoric and polyserial correlations, integrating these within the high-dimensional graphical modeling context. The resulting methodology is versatile and capable of handling arbitrary mixtures of continuous and discrete data, while maintaining scalability and interpretability.

We develop covariance matrix estimators suitable for latent Gaussian and nonparanormal models that can subsequently be integrated into existing sparse precision matrix estimation routines. The main contribution of this article is the careful integration of polyserial and polychoric correlations into the high-dimensional graphical modeling framework, accompanied by established theoretical guarantees, including concentration inequalities that characterize the statistical properties and reliability of our estimators.

Empirical validation is conducted using both simulated datasets and practical examples drawn from real-world biomedical applications, including phenotyping data from the UK Biobank. These experiments demonstrate that our methods perform comparably to oracle models with access to the true underlying latent structure. Notably, our approach simplifies model specification and implementation, avoiding complex preprocessing steps or manual selection of specialized bridge functions required by competing methods.

Overall, the first core article contributes to the field by providing an easy-to-use computationally efficient, and theoretically justified framework for high-dimensional graphical modeling of arbitrarily mixed data types.

Individual contributions As the lead author of this article, I formulated the necessary concepts, developed all proofs, wrote software implementations in the form of an R-package, conducted simulation studies, and drafted the manuscript. Mathias Drton convceived the idea to investigate polychoric and polyserial correlations in the context of undirected graphical models for discrete and continuous data. Sach Mukherjee provided biostatistical expertise and Anne Miloschewski applied the R-package to real data from the UK Biobank.

7.2 `causalAssembly`: Generating semi-synthetic real data for benchmarking causal discovery

The second core article of this thesis addresses the challenge of benchmarking causal discovery methods when real-world data lacks known ground truth. Causal discovery — the process of inferring causal relationships from observational data — has seen significant theoretical and algorithmic advances in recent years. However, empirical validation remains a bottleneck, as most real datasets lack confirmed causal structures, and privacy constraints often prevent the release of high-quality data. Consequently, evaluations frequently rely on synthetic data that, while informative, fail to capture the complex dependencies and heterogeneity found in practice.

To help address these issues, this article introduces `causalAssembly`, a semisynthetic data generation framework based on real production data from a highly automated manufacturing assembly line. This industrial setting naturally embeds complex, nonlinear, and sequential dependencies across production stations, which are monitored through detailed sensor measurements. By collaborating with domain experts and examining the underlying physical processes at selected stations, we are able to identify a set of causal relationships within individual stations and use the line structure to infer a consistent causal ordering.

This domain-informed structure is then formalized into a ground truth causal directed acyclic graph, which serves as the backbone for data generation. To guarantee that the generated data are Markov with respect to this graph, the authors apply distributional random forests to estimate the conditional distributions implied by the graph and compose them into a joint distribution according. This allows to sample repeatedly from the joint distribution allowing users to conduct proper simulation studies with known ground truth.

The `causalAssembly` pipeline is general-purpose and modular: it can be applied to any real dataset where partial or full domain knowledge is available. If set up accordingly, it allows the generation of datasets with adjustable dimensionality and complexity while preserving privacy. Benchmarking experiments are provided for several widely used causal discovery algorithms.

In summary, this second core article contributes a practical and extensible framework for semisynthetic data generation, bridging the gap between synthetic and real-world benchmarks in causal discovery. It enables reproducible, interpretable evaluations of causal learning algorithms under realistic conditions, where known causal ground truth is otherwise unavailable.

Individual contributions I am the main author of this article. I formulated the specific research problem, developed all proofs, wrote software implementations, conducted simulation studies, and drafted the manuscript. The idea of estimating conditional distributions based on partial domain knowledge in order to be able to sample from the joint distribution stems from a discussion among Mathias Drton and me. Tobias Windisch provided useful suggestions and feedback as well as coding support. Tim Pychynski provided domain knowledge regarding the production line. Steffen Sonntag aided in setting up the code base and Martin Roth provided useful feedback throughout the project.

7.3 Nonlinear causal discovery for grouped data

The third core article of this thesis focuses on causal discovery in settings where the fundamental units of analysis are groups of variables, rather than individual scalar measurements. While classical causal discovery methods typically operate on scalar variables, many real-world applications — such as those in neuroscience, climate science, psychology, and industrial manufacturing — involve groups of related measurements that collectively represent the causal entities of interest.

Instead of working at the level of individual variables followed by aggregation to the group level, this work aims to infer causal relationships directly between variable groups. To achieve this, we extend the class of nonlinear additive noise models (ANMs) to the group setting and establish identifiability results through a differential equation that generalizes the scalar case. Building on these results, we propose an extension of the RESIT algorithm of Peters et al. (2014), which we refer to as *GroupRESIT*.

GroupRESIT consists of two phases. In the first phase, we infer a causal ordering by fitting deep learning models to obtain residuals from a multi-response regression, followed by nonparametric unconditional independence tests to identify sink nodes. In the second phase, given the causal ordering, we prune the corresponding super-graph to estimate the final group-level DAG. For this purpose, we introduce MURGS (Multi-response Group Sparse Additive Models), a novel method that promotes sparsity at the group level. MURGS features an efficient block coordinate descent algorithm with a closed-form backfitting update.

We evaluate the proposed method on synthetic datasets, demonstrating improved performance compared to existing causal discovery approaches in grouped settings. Additionally, we apply GroupRESIT to real-world manufacturing data, where partial ground truth is available, further supporting the practical relevance and utility of our approach.

In summary, this third core article contributes both theoretical advances and practical tools for nonlinear causal discovery in grouped data contexts. The framework is designed so that regression and testing components can be easily replaced with future state-of-the-art methods, supporting broader applicability of causal discovery in diverse scientific and industrial settings.

Individual contributions I am the lead author of this article. I formulated the specific research problem, developed all theoretical results and proofs, implemented the software, conducted the simulation studies and real-world application, and drafted the manuscript. The original idea of vector causal discovery in the context of additive noise models was conceived by Mathias Drton. Tobias Windisch provided extensive feedback throughout the project, contributed to manuscript preparation, and assisted in developing parts of the code. In particular, he offered substantial support in executing and managing the simulation study.

Bibliography

Abegaz, F. and Wit, E. (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica*, 69(4):419–441.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541.

Antonoplis, S. (2022). Studying socioeconomic status: Conceptual problems and an alternative path forward. *Perspectives on Psychological Science*, 18(2):275–292.

Bartlett, M. and Cussens, J. (2017). Integer linear programming for the Bayesian network structure learning problem. *Artificial Intelligence*, 244:258–271.

Behrouzi, P. and Wit, E. C. (2019). Detecting epistatic selection with partially observed genotype data by using copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 68(1):141–160.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B*, 26:211–252.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

Brenner, E. and Sontag, D. (2013). SparsityBoost: a new scoring function for learning Bayesian network structure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 112–121, Arlington, Virginia, USA. AUAI Press.

Buhl, S. r. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Statist.*, 20(3):263–270.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, 42(6):2526–2556.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Ćevid, D., Michel, L., Näf, J., Bühlmann, P., and Meinshausen, N. (2022). Distributional random forests: heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.*, 23:Paper No. [333], 79.

Chen, R., Dash, S., and Gao, T. (2021). Integer programming for causal structure learning in the presence of latent variables. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1550–1560. PMLR.

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64.

Chen, W., Drton, M., and Wang, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980.

Cheng, J., Li, T., Levina, E., and Zhu, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.*, 26(2):367–378.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In *Learning from data (Fort Lauderdale, FL, 1995)*, volume 112 of *Lect. Notes Stat.*, pages 121–130. Springer, New York.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.

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

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314. Higher Order Statistics.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Cussens, J. (2020). GOBNILP: Learning Bayesian network structure with integer programming. In Jaeger, M. and Nielsen, T. D., editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 605–608. PMLR.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30(1):56–66.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Deng, C., Bello, K., Ravikumar, P., and Aragam, B. (2024). Markov equivalence and consistency in differentiable structure learning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 91756–91797. Curran Associates, Inc.

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.*, 5(2A):969–993.

Dressler, J., Storz, M. A., Müller, C., Kandil, F. I., Kessler, C. S., Michalsen, A., and Jeitler, M. (2022). Does a plant-based diet stand out for its favorable composition for heart health? dietary intake data from a randomized controlled trial. *Nutrients*, 14(21).

Drton, M. and Maathuis, M. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393. Copyright © 2017 by Annual Reviews.

Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.*, 22(3):430–449.

Druzdzel, M. J. and Van Leijen, H. (2001). Causal reversibility in Bayesian networks. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(1):45–62.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(2):405–421.

Feng, H. and Ning, Y. (2019). High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, N. and Nachman, I. (2000). Gaussian process networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 211–219, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, 98(2):227–255.

Göbler, K., Drton, M., Mukherjee, S., and Miloschewski, A. (2024a). High-dimensional undirected graphical models for arbitrary mixed data. *Electron. J. Stat.*, 18(1):2339–2404.

Göbler, K., Windisch, T., and Drton, M. (2025). Nonlinear causal discovery for grouped data. In Chiappa, S. and Magliacane, S., editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1453–1475. PMLR.

Göbler, K., Windisch, T., Drton, M., Pychynski, T., Roth, M., and Sonntag, S. (2024b). `causalAssembly`: Generating realistic production data for benchmarking causal discovery. In Locatello, F. and Didelez, V.,

editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, volume 3734 of *Lecture Notes in Comput. Sci.*, pages 63–77. Springer, Berlin.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14:3365–3383.

Hermes, S., van Heerwaarden, J., and Behrouzi, P. (2024). Copula graphical models for heterogeneous mixed data. *J. Comput. Graph. Statist.*, 33(3):991–1005.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81.

Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Kohn, A., Jasper, A. I., Semedo, J. D., Gokcen, E., Machens, C. K., and Yu, B. M. (2020). Principles of corticocortical communication: Proposed schemes and design considerations. *Trends in Neurosciences*, 43(9):725–737.

Koivisto, M. and Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *J. Mach. Learn. Res.*, 5:549–573.

Kruskal, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.*, 53:814–861.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2020). Gradient-based neural DAG learning. In *International Conference on Learning Representations*.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278.

Langley, P. (2011). The changing science of machine learning. *Machine Learning*, 82(3):275–279.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411.

Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Statist.*, 24(1):230–253.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328.

Liu, H., Wasserman, L., and Lafferty, J. (2008). Nonparametric regression and classification with joint sparsity constraints. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Liu, H., Wasserman, L., Lafferty, J., and Ravikumar, P. (2007). SpAM: Sparse additive models. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Longjohn, R., Kelly, M., Singh, S., and Smyth, P. (2024). Benchmark data repositories for better benchmarking. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 86435–86457. Curran Associates, Inc.

Maathuis, M. H. and Nandy, P. (2016). A review of some recent advances in causal inference. In *Handbook of big data*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 387–407. CRC Press, Boca Raton, FL.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Comput. Statist. Data Anal.*, 55(7):2372–2387.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 403–410, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Mohammadi, A., Abegaz, F., van den Heuvel, E., and Wit, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 66(3):629–645.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. (2023). Scalable causal discovery with score matching. In van der Schaar, M., Zhang, C., and Janzing, D., editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 752–771. PMLR.

Nandy, P., Hauser, A., and Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *Ann. Statist.*, 46(6A):3151–3183.

Ng, I., Ghassami, A., and Zhang, K. (2020). On the role of sparsity and DAG constraints for learning linear DAGs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ng, I., Huang, B., and Zhang, K. (2024). Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, pages 71–105. PMLR.

OEIS Foundation Inc. (2025). A003024: Number of acyclic digraphs (or DAGs) with n labeled nodes. `https://oeis.org/A003024`.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460.

Olsson, U., Drasgow, F., and Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, 47(3):337–347.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., and Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507.

Pappe, C. L., Lutzenberger, S., Göbler, K., Meier, S., Jeitler, M., Michalsen, A., and Dommisch, H. (2025). Effect of a whole-food plant-based diet on periodontal parameters in patients with cardiovascular risk factors: A secondary sub-analysis of a randomized clinical trial. *Journal of Clinical Periodontology*, 52(1):125–136.

Parviainen, P. and Koivisto, M. (2011). Ancestor relations in the presence of unobserved variables. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 581–596. Springer Berlin Heidelberg.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, reasoning, and inference.

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195(262-273):1–47.

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika*, 9(1/2):116–139.

Peters, J. and Bühlmann, P. (2013). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Scholkopf, B. (2011a). Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2011b). Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 589–598.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053.

Quan, X., Booth, J. G., and Wells, M. T. (2018). Rank-based approach for estimating correlations in mixed ordinal data. arXiv: 1809.06255.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. (2021). AI and the everything in the whole wide world benchmark. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Ramsey, J. D. and Andrews, B. (2018). FASK with interventional knowledge recovers edges from the sachs model. *CoRR*, abs/1805.03108.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):1009–1030.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

Reisach, A., Seiler, C., and Weichwald, S. (2021). 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.

Robinson, R. W. (1973). Counting labeled acyclic digraphs. In *New directions in the theory of graphs (Proc. Third Ann Arbor Conf., Univ. Michigan, Ann Arbor, Mich., 1971)*, pages 239–273. Academic Press, New York-London.

Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. (2022). Score matching enables causal discovery of nonlinear additive noise models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753. PMLR.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, 48(3):1514–1538.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). Directlingam: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction, and search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

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

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1).

Squires, C. and Uhler, C. (2022). Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815.

Teyssier, M. and Koller, D. (2005). Ordering-based search: a simple and effective algorithm for learning Bayesian networks. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pages 584–590, Arlington, Virginia, USA. AUAI Press.

Tian, J. and Pearl, J. (2001). Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI'01, pages 512–521, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, volume 2 of *Stanford Studies in Mathematics and Statistics*, pages 448–485. Stanford Univ. Press, Stanford, CA.

Uemura, K., Takagi, T., Takayuki, K., Yoshida, H., and Shimizu, S. (2022). A multivariate causal discovery based on post-nonlinear model. In Schölkopf, B., Uhler, C., and Zhang, K., editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 826–839. PMLR.

Verma, T. and Pearl, J. (1990a). Causal networks: semantics and expressiveness. In *Uncertainty in artificial intelligence, 4*, volume 9 of *Mach. Intelligence Pattern Recogn.*, pages 69–76. North-Holland, Amsterdam.

Verma, T. and Pearl, J. (1990b). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI'90, pages 255–270, USA. Elsevier Science Inc.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571.

Yang, E., Baker, Y., Ravikumar, P., Allen, G., and Liu, Z. (2014). Mixed graphical models via exponential families. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1042–1050, Reykjavik, Iceland. PMLR.

Yang, Z., Ning, Y., and Liu, H. (2018). On semiparametric exponential family graphical models. *Journal of Machine Learning Research*, 19(57):1–59.

Yin, J., Chen, X., and Xing, E. P. (2012). Group sparse additive models. *Proc. Int. Conf. Mach. Learn.*, 2012:871–878.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, K. and Hyvärinen, A. (2009). 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, Arlington, Virginia, USA. AUAI Press.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, USA. AUAI Press.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020). Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR.

A Core Publications

A.1 High-dimensional undirected graphical models for arbitrary mixed data

Electronic Journal of Statistics
Vol. 18 (2024) 2339–2404
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2254

High-dimensional undirected graphical models for arbitrary mixed data

Konstantin Göbler

Technical University of Munich, Robert Bosch GmbH,
e-mail: konstantin.goebler@tum.de

Mathias Drton

Munich Center for Machine Learning, Technical University of Munich,
e-mail: mathias.drton@tum.de

Sach Mukherjee

German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany,
e-mail: sach.mukherjee@dzne.de

University of Bonn, Bonn, Germany

MRC Biostatistics Unit, University of Cambridge, UK

Anne Miloschewski

German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany,
e-mail: anne.miloschwski@dzne.de

Abstract: Graphical models are an important tool in exploring relationships between variables in complex, multivariate data. Methods for learning such graphical models are well-developed in the case where all variables are either continuous or discrete, including in high dimensions. However, in many applications, data span variables of different types (e.g., continuous, count, binary, ordinal, etc.), whose principled joint analysis is nontrivial. Latent Gaussian copula models, in which all variables are modeled as transformations of underlying jointly Gaussian variables, represent a useful approach. Recent advances have shown how the binary-continuous case can be tackled, but the general mixed variable type regime remains challenging. In this work, we make the simple but useful observation that classical ideas concerning polychoric and polyserial correlations can be leveraged in a latent Gaussian copula framework. Building on this observation, we propose a flexible and scalable methodology for data with variables of entirely general mixed type. We study the key properties of the approaches theoretically and empirically.

Keywords and phrases: Generalized correlation, high-dimensional statistics, latent Gaussian copula, mixed data, polychoric/polyserial correlation, undirected graphical models.

Received September 2022.

1. Introduction

Graphical models are widely used in the analysis of multivariate data, providing a convenient and interpretable way to study relationships among potentially large numbers of variables. They are key tools in modern statistics and machine learning and play an important role in diverse applications. Undirected graphical models are used in a wide range of settings, including, among others, systems biology, omics, deep phenotyping [see, e.g. 11, 14, 29] and as a component within other analyses, including two-sample testing, unsupervised learning, hidden Markov modeling, and more [examples include 44, 42, 38, 39, 34].

A significant portion of the literature on graphical models has concentrated on scenarios where either only continuous variables or only discrete variables are present. Regarding the former case, Gaussian graphical models have been extensively studied, including in the high-dimensional regime [see among others 27, 17, 2, 21, 49, 37, 7]. In such models, it is assumed that the observed random vector follows a multivariate Gaussian distribution, and the graph structure of the model is given by the zero pattern in the inverse covariance matrix. Generalizations for continuous, non-Gaussian data have also been studied [28, 24, 14]. In the latter case, discrete graphical models – related to Ising-type models in statistical physics – have also been extensively studied [see, e.g. 43, 36].

However, in many applications, it is common to encounter data that entail *mixed* variable types, i.e., where the data vector includes components of different types (e.g., continuous-Gaussian, continuous-non-Gaussian, count, binary, etc.). Such "column heterogeneity" (from the usual convention of samples in rows and variables in columns) is the rule rather than the exception. For instance, in biomedical applications, the construction of gene regulatory networks using expression profiling of genes may involve jointly analyzing gene expression levels alongside categorical phenotypes. Similarly, diagnostic data in many medical applications may contain continuous measurements such as blood pressure and discrete information about disease status or pain levels.

In analyzing such data, estimating a joint multivariate graphical model spanning the various variable types is often of interest. In practice, this is sometimes done using *ad hoc* pipelines and data transformations. However, in graphical modeling, since the model output is intended to be scientifically interpretable and involves statements about properties such as conditional independence between variables, the use of *ad hoc* workflows without an understanding of the resulting estimation properties is arguably problematic.

There have been three main lines of work that tackle high-dimensional graphical modeling for mixed data. The earliest approach is conditional Gaussian modeling of a mix of categorical and continuous data [22] as treated by Cheng et al. [9], Lee and Hastie [23]. A second approach is to employ neighborhood selection, which amounts to separate modeling of conditional distributions for each variable given all others [see, e.g. 8, 47, 41]. A third approach uses latent Gaussian models, with a key recent reference being the paper of Fan et al. [12], who proposed a latent Gaussian copula model for mixed data. The generative structure in their work posits that the discrete data is obtained from latent

continuous variables thresholded at certain (unknown) levels. However, in [12], only a mix of binary and continuous data is considered. Their setting does not allow for more general combinations (including counts or ordinal variables) as found in many real-world applications.

This third approach will be the focus of this paper, which aims to provide a simple framework for working with latent Gaussian copula models to analyze general mixed data. To do so, we combine classical ideas concerning polychoric and polyserial correlations with approaches from the high-dimensional graphical models and copula literature. As we discuss below, this provides an overall framework that is scalable, general, and straightforward from the user's point of view.

Already in the early 1900s, Pearson [32, 33] worked on the foundations of these ideas in the form of the tetrachoric and biserial correlation coefficients. From these arose the maximum likelihood estimators (MLEs) for the general version of these early ideas, namely the polychoric and the polyserial correlation coefficients. One drawback of these original measures is that they have been proposed in the context of latent Gaussian variables. A richer distributional family is the nonparanormal proposed by Liu, Lafferty and Wasserman [24] as a nonparametric extension to the Gaussian family. A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is a member of the nonparanormal family when $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))^T$ is Gaussian, where $\{f_k\}_{k=1}^d$ is a set of univariate monotone transformation functions. Moreover, if the f_j's are monotone and differentiable, the nonparanormal family is equivalent to the Gaussian copula family. As the polychoric and polyserial correlation assume that observed discrete data are generated from latent continuous variables, they adhere to a latent copula approach.

We propose two estimators of the latent correlation matrix, which can subsequently be plugged into existing precision matrix estimation routines, such as the graphical lasso (glasso) [17], CLIME [7], or the graphical Dantzig selector [49]. The first is appropriate under a latent Gaussian model and unifies the aforementioned MLEs. The second is more general and is applicable under the latent Gaussian copula model. Both approaches can deal with discrete variables with arbitrarily many levels. We show that both estimators exhibit favorable theoretical properties and include empirical results based on real and simulated data. The main contributions of the paper are as follows:

- We posit that integrating polychoric and polyserial correlations into the latent Gaussian copula framework offers an elegant, straightforward, and highly effective approach to graphical modeling for comprehensively diverse mixed data sets.
- We present theoretical findings on the performance of the proposed estimators, encompassing their behavior in high-dimensional scenarios. The concentration results underscore the statistical validity of the introduced procedures.
- We empirically examine the estimators through a series of simulations and a practical example involving real phenotyping data of mixed types sourced from the UK Biobank. Our findings illustrate the practical util-

ity of the proposed methods, demonstrating that their performance often closely aligns with an oracle model granted access to true latent data.

Our proposed procedure provides users with a method for conducting statistically sound graphical modeling of mixed data that is both straightforward to implement and carries no more overhead than conventional high-dimensional Gaussian graphical modeling approaches. Our procedure requires no manual specification of variable-type-specific model components, such as bridge functions.

The remainder of this paper is organized as follows. In Sections 2 and 3, we present the estimators based on polychoric and polyserial correlations, including theoretical guarantees in terms of concentration inequalities. In Section 4, we describe the experimental setup used to test the proposed approaches on simulated data together with the results themselves. We conclude with a summary of our findings in Section 5 and point towards our `R` package **hume**, providing users with a convenient implementation of the methods developed in this study.

2. Background and model set-up

The objective of this paper is to learn the structure of undirected graphical models applicable to a wide range of mixed and high-dimensional data. To achieve this, we extend the Gaussian copula model [24, 25, 46], enabling the incorporation of both discrete and continuous data of any nature.

Definition 2.1 (The nonparanormal model)**.** *A random vector of continuous variables* $\mathbf{X} = (X_1, \ldots, X_d)$ *follows a* d*-dimensional nonparanormal distribution if there exists a set of monotone and differentiable univariate functions* $f = \{f_1, \ldots, f_d\}$ *such that the transformed vector* $f(\mathbf{X}) = (f_1(X)_1, \ldots, f_d(X)_d)$ *is multivariate Gaussian with mean* 0 *and covariance matrix* Σ*, i.e.* $f(\mathbf{X}) \sim N(0, \Sigma)$*. We write*

$$\mathbf{X} \sim NPN(0, \Sigma, f), \tag{1}$$

where without loss of generality, the diagonal entries in Σ *are equal to one.*

As demonstrated by Liu, Lafferty and Wasserman [24], the model in Eq. (1) is a semiparametric Gaussian copula model. The following definition indicates how to extend this model to the presence of general mixed data.

Definition 2.2 (latent Gaussian copula model for general mixed data)**.** *Let* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *be a* d*-dimensional random vector with* $\mathbf{X}_1$ *a* d_1*-dimensional vector of possibly ordered discrete variables, and* $\mathbf{X}_2$ *a* d_2*-dimensional vector of continuous variables with* $d = d_1 + d_2$*. Suppose there exists a* d_1*-dimensional random vector of latent continuous variables* $\mathbf{Z}_1 = (Z_1, \ldots, Z_{d_1})^T$ *such that the following relation holds:*

$$X_j = x_j^r \quad \textit{if} \quad \gamma_j^{r-1} \leq Z_j < \gamma_j^r \quad \textit{for all } j = 1, \ldots d_1 \textit{ and } r = 1, \ldots, l_j + 1, \tag{2}$$

where γ_j^r *represents some unknown thresholds with* $\gamma_j^0 = -\infty$ *and* $\gamma_j^{l_j+1} = +\infty$*,* $x_j^r \in \mathbb{N}_0$ *and* $l_j + 1$ *the number of discrete levels of* X_j *for all* $j \in 1, \ldots, d_1$*.*

Then, $\mathbf{X}$ *satisfies the latent Gaussian copula model if* $\mathbf{Z} := (\mathbf{Z}_1, \mathbf{X}_2) \sim NPN(0, \boldsymbol{\Sigma}, f)$. *We write*

$$\mathbf{X} \sim LNPN(0, \boldsymbol{\Sigma}, f, \boldsymbol{\gamma}), \tag{3}$$

where $\boldsymbol{\gamma} = \cup_{j=1}^{d_1}\{\gamma_j^r, r = 0, \ldots, l_j + 1\}$.

Note that Definition 2.2 entails the class of Gaussian copula models if no discrete variables are present and the class of latent Gaussian models if $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2) \sim \mathrm{N}(0, \boldsymbol{\Sigma})$. As shown by Fan et al. [12], the latent Gaussian copula model (LGCM) is invariant concerning any re-ordering of the discrete variables.

We denote $[d] = \{1, \ldots, d\}$, $[d_1] = \{1, \ldots, d_1\}$, and $[d_2] = \{d_1 + 1, \ldots, d_2\}$, respectively. Several identifiability issues arise in the latent Gaussian copula class. First, the mean and the variances are not identifiable unless the monotone transformations f were restricted to preserve them. Note that this only affects the diagonal entries in $\boldsymbol{\Sigma}$, not the full covariance matrix. Therefore, without loss of generality, we assume the mean to be the zero vector and $\Sigma_{jj} = 1$ for all $j \in [d]$. Another identifiability issue relates to the unknown threshold parameters. To ease notation, let $\Gamma_j^r \equiv f_j(\gamma_j^r)$ and $\Gamma_j \equiv \{f_j(\gamma_j^r)\}_{r=0}^{l_j+1}$. In the LGCM, only the transformed thresholds Γ_j rather than the original thresholds are identifiable from the discrete variables. We assume, without loss of generality, that the transformed thresholds retain the limiting behavior of the original thresholds, i.e., $\Gamma_j^0 = -\infty$ and $\Gamma_j^{l_j+1} = \infty$.

Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ denote the latent precision matrix. Then, the zero-pattern of $\boldsymbol{\Omega}$ under the LGCM still encodes the conditional independencies of the latent continuous variables [24]. Thus, the underlying undirected graph is represented by $\boldsymbol{\Omega}$ just as for the parametric normal. Note that the LGCM for general mixed data in Definition 2.2 agrees with that of Quan, Booth and Wells [35] and of Feng and Ning [13]. The problem phrased by Fan et al. [12] is a special case of Definition 2.2. A more detailed comparison between both approaches can be found in Section 3. Nominal discrete variables need to be transformed into a dummy system.

For the remainder of the paper, assume we observe an independent n-sample of the d-dimensional vector $\mathbf{X}$ which is assumed to follow an LGCM of the form $\mathrm{LNPN}(0, \boldsymbol{\Sigma}, f, \Gamma)$, where $\Gamma = \cup_{j=1}^{d_1}\Gamma_j$. We estimate $\boldsymbol{\Sigma}$ by considering the corresponding entries separately i.e. the couples (X_j, X_k) for $j, k \in [d]$. Consequently, we have to keep in view three possible cases depending on the couple's variable types, respectively:

***Case I*:** Both X_j and X_k are continuous, i.e. $j, k \in [d_2]$.
***Case II*:** X_j is discrete and X_k is continuous, i.e. $j \in [d_1], k \in [d_2]$ and vice versa.
***Case III*:** Both X_j and X_k are discrete, i.e. $j, k \in [d_1]$.

2.1. Maximum-likelihood estimation under the latent Gaussian model

At the outset, we examine each of the three cases under the latent Gaussian model, a special case of the LGCM where all transformations are identity functions. Consider *Case I*, where both X_j and X_k are continuous. This corresponds to the regular Gaussian graphical model set-up discussed thoroughly, for instance, in [37]. Hence, the estimator for $\mathbf{\Sigma}$ when both X_j and X_k are continuous is:

Definition 2.3 (MLE $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case I*)**.** *Let $\bar{x}_j$ denote the sample mean of X_j. The estimator $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{d_1<j<k\leq d_2}$ of the correlation matrix $\mathbf{\Sigma}$ is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \tag{4}$$

for all $d_1 < j < k \leq d_2$.

This is the Pearson product-moment correlation coefficient, which, of course, coincides with the maximum likelihood estimator (MLE) for the bivariate normal couple $\{(X_j, X_k)\}_{i=1}^{n}$.

Turning to *Case II*, let X_j be ordinal and X_k be continuous. We are interested in the product-moment correlation Σ_{jk} between two jointly Gaussian variables, where X_j is not directly observed but only the ordered categories (see Eq. (2)). This is called the *polyserial* correlation [31]. The likelihood and log-likelihood of the n-sample are defined by:

$$\begin{aligned} L^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) &= \prod_{i=1}^{n} p(x^r_{ij}, x_{ik}, \Sigma_{jk}) = \prod_{i=1}^{n} p(x_{ik})p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}) \\ \ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) &= \sum_{i=1}^{n} \big[\log(p(x_{ik})) + \log(p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}))\big], \end{aligned} \tag{5}$$

where $p(x^r_{ij}, x_{ik}, \Sigma_{jk})$ denotes the joint probability of X_j and X_k and $p(x_{ik})$ the marginal density of the Gaussian variable X_k. MLEs are obtained by differentiating the log of the likelihood in Eq. (5) with respect to the unknown parameters, setting the partial derivatives to zero, and solving the system of equations for $\Sigma_{jk}, \mu, \sigma^2$, and Γ^r_j for $r \in [l_j]$. Under the latent Gaussian model, we have the special case that the thresholds are identifiable from the observed data as $\Gamma^r_j = \gamma^r_j$.

Definition 2.4 (MLE $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case II*)**.** *Recall the log-likelihood in Eq. (5). The estimator $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1<j\leq d_1<k\leq d_2}$ is defined by:*

$$\begin{aligned} \hat{\Sigma}^{(n)}_{jk} &= \operatorname*{arg\,max}_{|\Sigma_{jk}|\leq 1} \ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) \\ &= \operatorname*{arg\,max}_{|\Sigma_{jk}|\leq 1} \frac{1}{n}\ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) \end{aligned} \tag{6}$$

for all $1 < j \leq d_1 < k \leq d_2$.

Regularity conditions ensuring consistency and asymptotic efficiency, as well as asymptotic normality, can be verified to hold here [10].

Lastly, consider *Case III*, where both X_j and X_k are ordinal. The probability of an observation with $X_j = x_j^r$ and $X_k = x_k^s$ is given by

$$\begin{aligned} \pi_{rs} &:= p(X_j = x_j^r, X_k = x_k^s) \\ &= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r, \Gamma_k^{s-1} \leq Z_k < \Gamma_k^s) \\ &= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} \int_{\Gamma_k^{s-1}}^{\Gamma_k^s} \phi(z_j, z_k, \Sigma_{jk}) dz_j dz_k, \end{aligned} \tag{7}$$

where $r = 1, \ldots, l_j$ and $s = 1, \ldots, l_k$ and $\phi(x, y, \rho)$ denotes the standard bivariate density with correlation ρ. Then, as outlined by Olsson [30] the likelihood and log-likelihood of the n-sample are defined as:

$$\begin{aligned} L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= C \prod_{r=1}^{l_j} \prod_{s=1}^{l_k} \pi_{rs}^{n_{rs}}, \\ \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= \log(C) + \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs} \log(\pi_{rs}), \end{aligned} \tag{8}$$

where C is a constant and n_{rs} denotes the observed frequency of $X_j = x_j^r$ and $X_k = x_k^s$ in a sample of size $n = \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs}$. Differentiating the log-likelihood, setting it to zero, and solving for the unknown parameters yields the estimator for Σ for *Case III*:

Definition 2.5 (MLE $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case III*)**.** *Recall the log-likelihood in Eq.* (8)*. The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1 \leq j < k \leq d_1}$ *of* $\mathbf{\Sigma}$ *is defined by:*

$$\begin{aligned} \hat{\Sigma}_{jk}^{(n)} &= \underset{|\Sigma_{jk}| \leq 1}{\arg\max}\, \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) \\ &= \underset{|\Sigma_{jk}| \leq 1}{\arg\max}\, \frac{1}{n} \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s), \end{aligned} \tag{9}$$

for all $1 < j < k \leq d_1$.

Regularity conditions ensuring consistency and asymptotic efficiency, as well as asymptotic normality, can again be verified to hold here [20].

Summing up, under the latent Gaussian model, a special case of the LGCM, $\hat{\mathbf{\Sigma}}^{(n)}$ is a consistent and asymptotically efficient estimator for the underlying latent correlation matrix $\mathbf{\Sigma}$. Corresponding concentration results are derived in Section 3.5.

3. Latent Gaussian Copula models

Fan et al. [12] propose the binary LGCM, a special case of the LGCM allowing for the presence of binary and continuous variables. Following the approach of the nonparanormal SKEPTIC [25], they circumvent the direct estimation of monotone transformation functions $\{f_j\}_{j=1}^d$ by employing rank correlation measures, such as Kendall's tau or Spearman's rho. These measures remain invariant under monotone transformations. Notably, for *Case I*, a well-known mapping exists between Kendall's tau, Spearman's rho, and the underlying Pearson correlation coefficient Σ_{jk}. As a result, the primary contribution of Fan et al. [12] lies in deriving corresponding bridge functions for cases II and III. To reduce computational burden Yoon, Müller and Gaynanova [48] propose a hybrid multilinear interpolation and optimization scheme of the underlying latent correlation.

When considering the general mixed case, Fan et al. [12] advocate for binarizing all ordinal variables. This concept has been embraced by Feng and Ning [13], who suggest an initial step of binarizing all ordinal variables to create preliminary estimators. Subsequently, these estimators are meaningfully combined using a weighted aggregate. To extend the binary latent Gaussian copula model and explore generalizations regarding bridge functions, Quan, Booth and Wells [35] ventured into scenarios where a combination of continuous, binary, and ternary variables is present. However, a notable drawback of this approach becomes evident. Dealing with a mix of binary and continuous variables requires three bridge functions – one for each case. The complexity grows as discrete variables introduce distinct state spaces. In fact, a combination of continuous variables and discrete variables with k different state spaces necessitates $\binom{k+2}{2}$ bridge functions.

For this reason, we adopt an alternative approach to the latent Gaussian copula model when dealing with general mixed data, allowing discrete variables to possess any number of states. In this strategy, the number of cases to be considered remains consistent at three, as already introduced in the preceding section.

3.1. Nonparanormal Case I

For *Case I*, the mapping between Σ_{jk} and the population versions of Spearman's rho and Kendall's tau is well known [24]. Here we make use of Spearman's rho $\rho_{jk}^{Sp} = corr(F_j(X_j), F_k(X_k))$ with F_j and F_k denoting the cumulative distribution functions (CDFs) of X_j and X_k, respectively. Then $\Sigma_{jk} = 2\sin\frac{\pi}{6}\rho_{jk}^{Sp}$ for $d_1 < j < k \leq d_2$. In practice, we use the sample estimate

$$\hat{\rho}_{jk}^{Sp} = \frac{\sum_{i=1}^n (R_{ij} - \bar{R}_j)(R_{ik} - \bar{R}_k)}{\sqrt{\sum_{i=1}^n (R_{ij} - \bar{R}_j)^2 \sum_{i=1}^n (R_{ik} - \bar{R}_k)^2}},$$

with R_{ij} corresponding to the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$ and $\bar{R}_j = 1/n \sum_{i=1}^n R_{ij} = (n+1)/2$; compare [25]. From this, we obtain the following estimator:

Definition 3.1 (Nonparanormal estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case I*)**.** *The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{d_1<j<k\leq d_2}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = 2\sin\frac{\pi}{6}\hat{\rho}^{Sp}_{jk}, \tag{10}$$

for all $d_1 < j < k \leq d_2$.

3.2. Nonparanormal Case II

In *Case II*, the complexity increases. Employing a rank-based approach for the nonparanormal model makes direct application of the ML procedure unfeasible, given that the continuous variable is not observed in its Gaussian form. Nevertheless, a two-step approach remains viable. First, an estimate of f_j must be formulated and subsequently employed in Definition 2.4. Yet, scrutinizing convergence rates for this procedure poses challenges, as the estimated transformation appears in multiple instances within the first-order condition of the MLE. Section A.3 provides further details and compares the two-step likelihood approach to the one we propose below.

Instead, we will proceed by suitably modifying other approaches that address the Gaussian case through a more direct *ad hoc* examination of the relationship between Σ_{jk} and the point polyserial correlation [3, 4]. Section A.3 compares the nonparanormal *Case II* estimation strategies.

In what follows, in the interest of readability, we omit the index in the monotone transformation functions but explicitly allow them to vary among the $\mathbf{Z}$. According to Definition 2.3, we have the following Gaussian conditional expectation

$$E[f(X_k) \mid f(Z_j)] = \mu_{f(X_k)} + \Sigma_{jk}\sigma_{f(X_k)}f(Z_j), \quad \text{for } 1 \leq j \leq d_1 < k \leq d_2, \tag{11}$$

where we can assume w.l.o.g. that $\mu_{f(X_k)} = 0$. After multiplying both sides with the discrete variable X_j, we move it into the expectation on the left-hand side of the equation. This is permissible as X_j is a function of $f(Z_j)$, i.e.

$$E[f(X_k)X_j \mid f(Z_j)] = \Sigma_{jk}\sigma_{f(X_k)}f(Z_j)X_j.$$

Now let us take again the expectation on both sides, rearrange and expand by σ_{X_j}, yielding

$$\Sigma_{jk} = \frac{E[f(X_k)X_j]}{\sigma_{f(X_k)}E[f(Z_j)X_j]} = \frac{r_{f(X_k)X_j}\sigma_{X_j}}{E[f(Z_j)X_j]}, \tag{12}$$

where $r_{f(X_k)X_j}$ is the product-moment correlation between the Gaussian (unobserved) variable $f(X_k)$ and the observed discretized variable X_j.

All that remains is to find sample versions of each of the three components in Eq. (12). Let us start with the expectation in the denominator $E[f(Z_j)X_j]$.

By assumption $f(\mathbf{Z}) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ and therefore w.l.o.g. $f(Z_j) \sim \mathrm{N}(0,1)$ for all $j \in 1, \ldots, d_1$. Consequently, we have:

$$\begin{aligned} E[f(Z_j)X_j] &= \sum_{r=1}^{l_{j+1}} x_j^r \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} f(z_j) dF(f(z_j)) = \sum_{r=1}^{l_{j+1}} x_j^r \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} f(z_j)\phi(f(z_j))dz_j \\ &= \sum_{r=1}^{l_{j+1}} x_j^r \Big(\phi(\Gamma_j^r) - \phi(\Gamma_j^{r-1}) \Big) = \sum_{r=1}^{l_j} (x_j^{r+1} - x_j^r)\phi(\Gamma_j^r), \end{aligned} \tag{13}$$

where $\phi(t)$ denotes the standard normal density. Whenever the ordinal states are consecutive integers we have $\sum_{r=1}^{l_j}(x_j^{r+1} - x_j^r)\phi(\Gamma_j^r) = \sum_{r=1}^{l_j}\phi(\Gamma_j^r)$. Based on this derivation, it is straightforward to give an estimate of $E[f(Z_j)X_j]$ once estimates of the thresholds Γ_j have been formed (see Section 3.4 for more details). Let us turn to the numerator of Eq. (12). The standard deviation of X_j does not require any special treatment, and we simply use $\sigma_{X_j}^{(n)} = \sqrt{1/n \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}$ to be able to treat discrete variables with a general number of states. However, the product-moment correlation $r_{f(X_k),X_j}$ is inherently more challenging as it involves the (unobserved) transformed version of the continuous variables. Therefore, we proceed to estimate the transformation.

To this end, consider the marginal distribution function of X_k, namely

$$F_{X_k}(x) = P(X_k \leq x) = P(f(X_k) \leq f(x)) = \Phi(f(x)),$$

such that $f(x) = \Phi^{-1}(F_{X_k}(x))$. In this setting, Liu, Lafferty and Wasserman [24] propose to evaluate the quantile function of the standard normal at a Winsorized version of the empirical distribution function. This is necessary as the standard Gaussian quantile function $\Phi^{-1}(\cdot)$ diverges when evaluated at the boundaries of the $[0,1]$ interval. More precisely, consider $\hat{f}(u) = \Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(u)])$,where W_{δ_n} is a Winsorization operator, i.e.

$$W_{\delta_n}(u) \equiv \delta_n I(u < \delta_n) + uI(\delta_n \leq u \leq (1-\delta_n)) + (1-\delta_n)I(u > (1-\delta_n)).$$

The truncation constant δ_n can be chosen in several ways. Liu, Lafferty and Wasserman [24] propose to use $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$ in order to control the bias-variance trade-off. Thus, equipped with an estimator for the transformation functions, the product-moment correlation is obtained the usual way, i.e.

$$r_{\hat{f}(X_k),X_j}^{(n)} = \frac{\sum_{i=1}^n (\hat{f}(X_{ik}) - \mu(\hat{f}))(X_{ij} - \mu(X_j)}{\sqrt{\sum_{i=1}^n \Big(\hat{f}(X_{ik}) - \mu(\hat{f})\Big)^2} \sqrt{\sum_{i=1}^n \Big(X_{ij} - \mu(X_j)\Big)^2}},$$

where $\mu(\hat{f}) \equiv 1/n \sum_{i=1}^n \hat{f}(X_{ik})$ and $\mu(X_j) \equiv 1/n \sum_{i=1}^n X_{ij}$. The resulting estimator is a double-two-step estimator of the mixed couple X_j and X_k.

Definition 3.2 (Estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case II* nonparanormal)**.** *The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1<j\leq d_1<k\leq d_2}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \frac{r^{(n)}_{\hat{f}(X_k),X_j}\,\sigma^{(n)}_{X_j}}{\sum_{r=1}^{l_j}\phi(\hat{\Gamma}^r_j)(x_j^{r+1}-x_j^r)} \tag{14}$$

for all $1 < j \leq d_1 < k \leq d_2$.

3.3. Nonparanormal Case III

Lastly, let us turn to *Case III* where both X_j and X_k are discrete, but they might differ in their respective state spaces. In the previous section, the ML procedure could no longer be applied directly because we do not observe the continuous variable in its Gaussian form. In *Case III*, however, we only observe the discrete variables generated by the latent scheme outlined in Definition 2.3. Due to the monotonicity of the transformation functions, the ML procedure for *Case III* from Section 2.1 can still be applied, i.e.

Definition 3.3 (Nonparanormal estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case III*)**.** *The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1\leq j<k\leq d_1}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \operatorname*{arg\,max}_{|\Sigma_{jk}|\leq 1} \frac{1}{n}\ell^{(n)}_{jk}(\Sigma_{jk}, x_j^r, x_k^s) \tag{15}$$

for all $1 < j < k \leq d_1$.

In summary, the estimator $\hat{\mathbf{\Sigma}}^{(n)}$ under the latent Gaussian copula model is a simple but important tool for flexible mixed graph learning. By using ideas from polyserial and polychoric correlation measures, we not only have an easy-to-calculate estimator but also overcome the issue of finding bridge functions between all different kinds of discrete variables.

3.4. Threshold estimation

The unknown threshold parameters Γ_j for $j \in [d_1]$ play a key role in linking the observed discrete to the latent continuous variables. Therefore, being able to form accurate estimates of the Γ_j is crucial for both the likelihood-based procedures and the nonparanormal estimators outlined above.

We start by highlighting that we set the LGCM model up such that for each Γ_j, there exists a constant G such that $|\Gamma_j^r| \leq G$ for all $r \in [l_j]$, i.e., the estimable thresholds are bounded away from infinity. Let us define the cumulative probability vector $\pi_j = (\pi_j^1, \ldots, \pi_j^{l_j})$. Then, by Eq. (2), it is easy to see that

$$\begin{aligned}\pi_j^r &= \sum_{i=1}^{r} P(X_j = x_j^i) = P(X_j \leq x_j^r)\\ &= P(Z_j \leq \gamma_j^r) = P(f_j(Z_j) \leq f_j(\gamma_j^r)) = \Phi(\Gamma_j^r).\end{aligned} \tag{16}$$

From this equation, it is immediately clear that the thresholds satisfy $\Gamma_j^r = \Phi^{-1}(\pi_j^r)$. Consequently, when forming sample estimates of the unknown thresholds, we replace the cumulative probability vector with its sample equivalent, namely

$$\hat{\pi}_j^r = \sum_{k=1}^{r} \Big[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} = x_j^k) \Big] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} \leq x_j^r), \tag{17}$$

and plug it into the identity, i.e. $\hat{\Gamma}_j^r = \Phi^{-1}\big(\hat{\pi}_j^r\big)$ for $j \in [d_1]$. The following lemma assures that these threshold estimates can be formed with high accuracy.

Lemma 3.1. *Suppose the estimated thresholds are bounded away from infinity, i.e.,* $|\hat{\Gamma}_j^r| \leq G$ *for all* $j \in [d_1]$ *and* $r = 1, \ldots, l_j$ *and some* G*. The following bound holds for all* $t > 0$ *with Lipschitz constant* $L_1 = 1/(\sqrt{\frac{2}{\pi}} \min\{\hat{\pi}_j^r, 1 - \hat{\pi}_j^r\})$*:*

$$P\Big(|\hat{\Gamma}_j^r - \Gamma_j^r| \geq t\Big) \leq 2 \exp\Big(- \frac{2t^2 n}{L_1^2} \Big).$$

The proof of Lemma 3.1 is given in Section B.3. The requirement that the estimated thresholds are bounded away from infinity typically does not pose any restriction in finite samples. All herein-developed methods are applied in a two-step fashion. In the ensuing theoretical results, we stress this by denoting the estimated thresholds as $\bar{\Gamma}_j^r$.

3.5. Concentration results

Define $\mathbf{\Sigma}^*$ and $\mathbf{\Omega}^*$ as the true covariance matrix and its inverse, respectively. We start by stating the following assumptions:

Assumption 3.1. *For all* $1 \leq j < k \leq d$, $|\Sigma_{jk}^*| \neq 1$*. In other words, there exists a constant* $\delta > 0$ *such that* $|\Sigma_{jk}^*| \leq 1 - \delta$*.*

Assumption 3.2. *For any* Γ_j^r *with* $j \in [d_1]$ *and* $r \in [l_j]$ *there exists a constant* G *such that* $|\Gamma_j^r| \leq G$*.*

Assumption 3.3. *Let* $j < k$ *and consider the log-likelihood functions in Definition 2.4 and in Definition 2.5. We assume that with probability one,*

- $\{-1 + \delta, 1 - \delta\}$ *are not critical points of the respective log-likelihood functions.*
- *The log-likelihood functions have a finite number of critical points.*
- *Every critical point that is different from* Σ_{jk}^* *is non-degenerate.*
- *All joint and conditional states of the discrete variables have positive probability.*

Assumptions 3.1 and 3.2 ensure that $f(X_j)$ and $f(X_k)$ are not perfectly linearly dependent and that the thresholds are bounded away from infinity, respectively. Importantly, these constraints impose minimal restrictions in practice. Assumption 3.3 guarantees that the likelihood functions in Section 2.2 exhibit a "nice" behavior, representing a mild technical requirement.

Convergence results for latent Gaussian models The subsequent theorem, drawing on Mei, Bai and Montanari [26], hinges on four conditions, all substantiated in Section B.1. This concentration result specifically pertains to the MLEs introduced in Section 2.1 within the framework of the latent Gaussian model. We remark that related methodology has been applied by Anne, Aurélie and Clémence [1] in addressing zero-inflated Gaussian data under double truncation.

Theorem 3.2. *Suppose that Assumptions 3.1–3.3 hold, and let* $j \in [d_1]$ *and* $k \in [d_2]$ *for Case II and* $j, k \in [d_1]$ *for Case III. Let* $\alpha \in (0,1)$*, and let* $n \geq 4C\log(n)\log\left(\frac{B}{\alpha}\right)$ *with some known constants* B*,* C*, and* D *depending on cases II and III but independent of* (n,d)*. Then, it holds that*

$$P\left(\max_{j,k}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\right| \geq D\sqrt{\frac{\log(n)}{n}\log\left(\frac{B}{\alpha}\right)}\right) \leq \frac{d(d-1)}{2}\alpha. \tag{18}$$

Case I of the latent Gaussian model addresses the well-understood scenario involving observed Gaussian variables, with concentration results and rates of convergence readily available −see, for example, Lemma 1 in Ravikumar et al. [37]. Consequently, the MLEs converge to $\mathbf{\Sigma}^*$ at the optimal rate of $n^{-1/2}$, mirroring the convergence rate as if the underlying latent variables were directly observed.

Convergence of nonparanormal estimators Recall the three cases, which, in principle, will have to be considered again.

***Case I*:** When both random variables are continuous, concentration results follow immediately from Liu et al. [25] who make use of Hoeffding's inequalities for U-statistics.

***Case II*:** For the case where one variable is discrete and the other one continuous, we present concentration results below.

***Case III*:** When both variables are discrete, we make an important observation that Theorem 3.2 above still applies and needs not to be altered. We do not observe the continuous variables directly but only their discretized versions. Consequently, the threshold estimates remain valid under the monotone transformation functions, and so does the polychoric correlation.

The following theorem provides concentration properties for *Case II* under the LGCM.

Theorem 3.3. *Suppose that Assumptions 3.1 and 3.2 hold and* $j \in [d_1]$ *and* $k \in [d_2]$*. Then for any* $\epsilon \in \left[C_M\sqrt{\frac{\log d \log^2 n}{\sqrt{n}}}, 8(1+4c^2)\right]$*, with sub-Gaussian parameter* c*, generic constants* $k_i, i = 1,2,3$ *and constant* $C_M = \frac{48}{\sqrt{\pi}}\left(\sqrt{2M}-1\right)(M+2)$ *for some* $M \geq 2\left(\frac{\log d_2}{\log n}+1\right)$ *with* $C_\Gamma = \sum_{r=1}^{l_j}\phi(\bar{\Gamma}^r_j)(x_j^{r+1} - x_j^r)$ *and Lipschitz constant* L *the following probability bound holds*

$$
\begin{aligned}
&P\left(\max_{jk}\left|\hat{\Sigma}_{jk}^{(n)}-\Sigma_{jk}^{*}\right|\geq\epsilon\right)\\
&\quad\leq 8\exp\left(2\log d-\frac{\sqrt{n}\epsilon^{2}}{(64\ L\ C_{\gamma}\ l_{\max}\ \pi)^{2}\log n}\right)\\
&\quad+8\exp\left(2\log d-\frac{n\epsilon^{2}}{(4L\ C_{\gamma})^{2}\ 128(1+4c^{2})^{2}}\right)\\
&\quad+8\exp\left(2\log d-\frac{\sqrt{n}}{8\pi\log n}\right)+4\exp\left(-\frac{k_{1}n^{3/4}\sqrt{\log n}}{k_{2}+k_{3}}\right)+\frac{2}{\sqrt{\pi\log(nd_{2})}}.
\end{aligned}
$$

The proof of the theorem is given in Section B.4. The first four terms in the probability bound stem from finding bounds to different regions of the support of the transformed continuous variable. The last term is a consequence of the fact that we estimate the transform directly.

Regarding the scaling of the dimension in terms of sample size, the ensuing corollary follows immediately.

Corollary 3.4. *For some known constant K_{Σ} independent of d and n we have*

$$
P\left(\max_{j,k}\left|\hat{\Sigma}_{jk}^{(n)}-\Sigma_{jk}^{*}\right|>K_{\Sigma}\sqrt{\frac{\log d\log n}{\sqrt{n}}}\right)=o(1). \tag{19}
$$

The nonparanormal estimator for *Case II* converges to Σ_{jk}^{*} at rate $n^{-1/4}$, which is slower than the optimal parametric rate of $n^{-1/2}$. This stems not from the presence of the discrete variable but from the direct estimation of the transformation function f_j and the corresponding truncation constant δ_n. Both Xue and Zou [46] and Liu et al. [25] discuss room for improvement of the estimator for f_j to get a rate closer to the optimal one. Improvements depend strongly on the choice of truncation constant, striking a bias-variance balance in high-dimensional settings. In all of the numerical experiments below, we find that Theorem 3.3 gives a worst-case rate that does not appear to negatively impact performance compared to estimators that attain the optimal rate.

3.6. Estimating the precision matrix

Similar to Fan et al. [12], we plug our estimate of the sample correlation matrix into existing routines for estimating $\mathbf{\Omega}^{*}$. In particular, we employ the graphical lasso (glasso) estimator [17], i.e.

$$
\hat{\mathbf{\Omega}}=\underset{\mathbf{\Omega}\succeq 0}{\arg\min}\left[\operatorname{tr}(\hat{\mathbf{\Sigma}}^{(n)}\mathbf{\Omega})-\log|\mathbf{\Omega}|+\lambda\sum_{j\neq k}|\Omega_{jk}|\right], \tag{20}
$$

where $\lambda>0$ is a regularization parameter. As $\hat{\mathbf{\Sigma}}^{(n)}$ exhibits at worst the same theoretical properties as established in Liu, Lafferty and Wasserman [24], convergence rate and graph selection results follow immediately.

We do not penalize diagonal entries of $\boldsymbol{\Omega}$ and therefore have to make sure that $\hat{\boldsymbol{\Sigma}}^{(n)}$ is at least positive semidefinite to establish convergence in Eq. (20). Hence, we need to project $\hat{\boldsymbol{\Sigma}}^{(n)}$ into the cone of positive semidefinite matrices; see also [25, 12]. In practice, we use an efficient implementation of the alternating projections method proposed by Higham [18].

To select the tuning parameter in Eq. (20) Foygel and Drton [16] introduce an extended BIC (eBIC) in particular for Gaussian graphical models establishing consistency in higher dimensions under mild asymptotic assumptions. We consider

$$eBIC_\theta = -2\ell^{(n)}(\hat{\boldsymbol{\Omega}}(E)) + |E|\log(n) + 4|E|\theta\log(d), \tag{21}$$

where $\theta \in [0,1]$ governs penalization of large graphs. Furthermore, $|E|$ represents the cardinality of the edge set of a candidate graph on d nodes and $\ell^{(n)}(\hat{\boldsymbol{\Omega}}(E))$ denotes the corresponding maximized log-likelihood which in turn depends on λ from Eq. (20). In practice, first, one retrieves a small set of models over a range of penalty parameters $\lambda > 0$ (called *glasso path*). Then, we calculate the eBIC for each model in the path and select the one with the minimal value.

4. Numerical results

To numerically assess the accuracy of our mixed graph estimation approach, we commence with a simulation study in which the estimators are rigorously evaluated in a gold-standard fashion and compared against oracles.

4.1. Simulation setup

We start by constructing the underlying precision matrix $\boldsymbol{\Omega}^*$ whose zero pattern encodes the undirected graph. We set $\boldsymbol{\Omega}^*_{jj} = 1$ and $\boldsymbol{\Omega}^*_{jk} = s \cdot b_{jk}$ if $j \neq k$, where s is a constant signal strength chosen to assure positive definiteness. Furthermore, b_{jk} are realizations of a Bernoulli random variable with corresponding success probability $p_{jk} = (2\pi)^{-1/2}\exp\left[\|v_j - v_k\|_2/(2c)\right]$. In particular, $v_j = (v_j^{(1)}, v_j^{(2)})$ are independent realizations of a bivariate uniform $[0,1]$ distribution and c controls the sparsity of the graph.

Throughout the simulation, we set $s = 0.15$ and incrementally increase the dimensionality s.t. $d \in \{50, 250, 750\}$, representing a transition from small to large-scale graphs. We let $\boldsymbol{\Sigma}^* = (\boldsymbol{\Omega}^*)^{-1}$ be rescaled such that all diagonal elements are equal to 1. Given $\boldsymbol{\Sigma}^*$, we first obtain the partially latent continuous data $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2)$ where $\mathbf{Z} \sim \mathrm{NPN}_d(\mathbf{0}, \boldsymbol{\Sigma}^*, f)$. In practice, we draw n i.i.d. samples from $\mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma}^*)$ and apply the back-transform f^{-1} to each individual variable.

To generate general mixed data $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ according to the LGCM we need to appropriately threshold $\mathbf{Z}_1$. Let $\mathbf{X}_1$ be partitioned into equally sized collections of binary, ordinal, and Poisson distributed random variables, i.e., $\mathbf{X}_1 = (\mathbf{X}_1^{\mathrm{bin}}, \mathbf{X}_1^{\mathrm{ord}}, \mathbf{X}_1^{\mathrm{pois}})$. We use the inverse probability integral transform (IPT) to generate random samples from the respective cumulative distribution

functions, corresponding to the relationship described in Eq. (2). For $\mathbf{X}_1^{\text{bin}}$ IPT is employed with success probability drawn from Uniform $[0.4, 0.6]$ for 80% of $\mathbf{X}_1^{\text{bin}}$. We assign unbalanced classes to the remaining 20%, where the success probability is drawn from Uniform $[0.05, 0.1]$. Regarding $\mathbf{X}_1^{\text{ord}}$, IPT is used to generate samples from the multinomial distribution. To that end, we draw the number of categories from Uniform $[3, 7]$ and round it to the nearest integer. We set the probability of falling into one of these categories to be proportional to their number. Lastly, $\mathbf{X}_1^{\text{pois}}$ is generated using IPT with the rate parameter set to 6. In case we only need a mix of binary and continuous data, we set $\mathbf{X}_1 = \mathbf{X}_1^{\text{bin}}$.

Throughout the experiments, $\hat{\Omega}$ is chosen by minimizing the eBIC according to the procedure outlined in Section 3.6 with $\theta = 0.1$ for the low and medium, $\theta = 0.5$ for the high dimensional graphs. The sample size n is set to 200 for $d \in \{50, 250\}$ and 300 for $d = 750$. We set the number of simulation runs to 100. Lastly, we choose the sparsity parameter c such that the number of edges aligns roughly with the dimension – except for $d = 50$, where we allow for 200 edges following Fan et al. [12].

Performance metrics To evaluate performance, we report the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ using the Frobenius norm. Additionally, we employ graph recovery metrics. For this purpose, we calculate the number of true positives $\text{TP}(\lambda)$ and false positives $\text{FP}(\lambda)$ based on the *glasso path*. $\text{TP}(\lambda)$ represents the count of non-zero lower off-diagonal elements that are consistent both in $\mathbf{\Omega}^*$ and $\hat{\mathbf{\Omega}}$, while $\text{FP}(\lambda)$ denotes the count of non-zero lower off-diagonal elements in $\hat{\mathbf{\Omega}}$ that are zero in $\mathbf{\Omega}^*$.

The true positive rate $\text{TPR}(\lambda)$ and false positive rate $\text{FPR}(\lambda)$ are defined as $\text{TPR} = \frac{\text{TP}(\lambda)}{|E|}$ and $\text{FPR} = \frac{\text{FP}(\lambda)}{d(d-1)/2 - |E|}$, respectively. Finally, we consider the area under the curve (AUC), where a value of 0.5 corresponds to random guessing of edge presence and a value of 1 indicates perfect error-free recovery of the underlying latent graph (in the rank sense of ROC analysis).

4.2. Simulation results

Binary-continuous data We start by considering a mix of binary and continuous variables generated as outlined in Section 4.1 to compare our methods against the bridge function approach of Fan et al. [12]. For this purpose, Figure 1 depicts the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ and the AUC for the different estimators under the different (d, n) regimes. We include the following estimators for $\mathbf{\Omega}^*$: (1) An oracle estimator (`oracle`) that corresponds to estimating $\hat{\mathbf{\Sigma}}^{(n)}$ using the mapping between Spearman's rho and $\hat{\Sigma}^*_{jk}$ (Eq. (10)) based on realization of the (partially) latent continuous data $(\mathbf{Z_1}, \mathbf{X_2})$. (2) The bridge function based estimator (`bridge`) proposed by Fan et al. [12]. (3) The polychhoric and polyserial MLE estimator (`mle`) proposed in Section 2.1. (4) The general mixed estimator (`poly`) proposed in Section 3. In the left column, we set $f_j(x) = x$

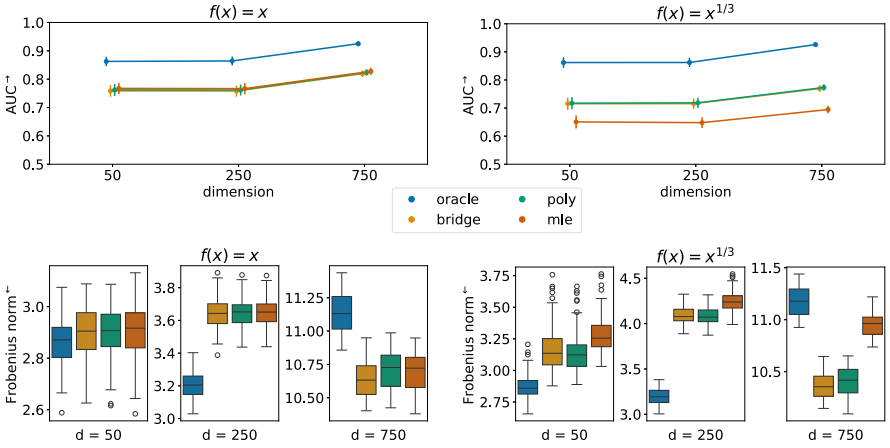


FIG 1. *Simulation results for the binary-continuous data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row reports mean and standard deviation of the AUC along simulation runs, and the bottom row depicts boxplots of the estimation error* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$*. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better, and* $\leftarrow$ *implies that smaller values are better.*

for all j, i.e., we recover the latent Gaussian model. In the right column, we set $f_j(x) = x^{1/3}$ for all j to recover the LGCM.

Figure 1 suggests that under the latent Gaussian model (left column), there are virtually no differences between all non-oracle estimators in graph recovery or estimation error. As expected, the `oracle` has the highest AUC and lowest estimation error across scenarios. The only exception is the estimation error when the dimension is $d = 750$. This surprising result stems from an increased FPR for `oracle` (see Figure 5 in the appendix) when minimizing the eBIC with the additional penalty set to $\theta = 0.5$. A higher penalty seems appropriate in this case. Meanwhile, the non-oracle estimators are more conservative, and the additional penalty appears to be chosen correctly in these cases.

In the right column of Figure 1, binary-continuous mixed data is generated from the LGCM. The *Case I* and *Case II* MLEs are misspecified in this case, which translates to lower AUC and higher estimation error. The remaining estimators are unaffected by the transformation. Encouragingly, we find no substantial performance difference between the bridge function approach and our procedure in any of the metrics considered, including the TPR and FPR results in Figure 5 in the appendix.

General mixed data Let us turn to the general mixed setting. While the bridge function approach by Fan et al. [12] does not extend beyond the binary-continuous mix, we can still compare our approach to the ensemble method developed by Feng and Ning [13]. Due to its close connection to the original method, we continue to denote the proposed ensemble estimator `bridge`.

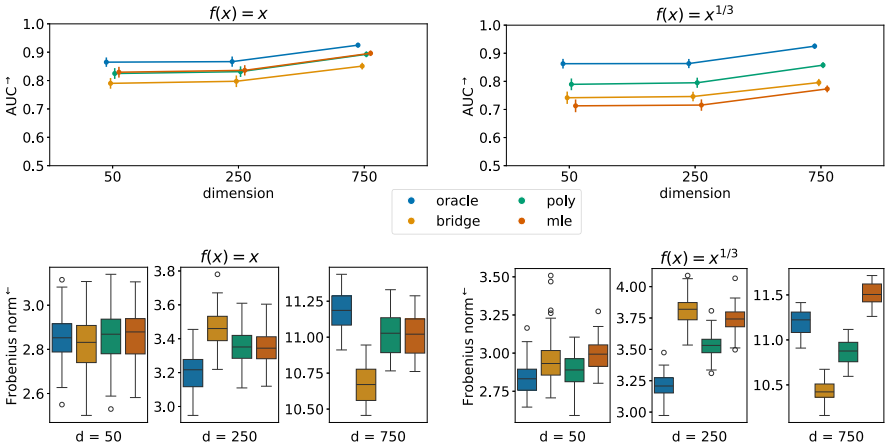


FIG 2. *Simulation results for the general mixed data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row reports mean and standard deviation of the AUC along simulation runs, and the bottom row depicts boxplots of the estimation error* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$*. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better, and* $\leftarrow$ *implies that smaller values are better.*

Similar to above, Figure 2 depicts the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ and the AUC and left and right columns correspond to latent Gaussian und LGCM settings, respectively. This time, given general mixed data, differences in terms of AUC between the estimators are noticeable. When the transformations are the identity, the MLE is correctly specified, and it is tied with `poly` in terms of AUC. The `bridge` estimator performs worse than the other two estimators. When the transformations are $f_j(x) = x^{1/3}$, the MLE is misspecified, and our `poly` estimator performs best among non-oracle estimators in terms of AUC. The `bridge` estimator performs only marginally better than the misspecified `mle`.

Turning to estimation error results, when $f_j(x) = x$, the `poly` and `mle` estimators perform similarly across dimensions. As the `oracle` estimator is formed on the latent continuous data, it is unaffected by any discretization and behaves the same as in the binary-continuous case above. The `bridge` ensemble estimator accounts for a slightly higher estimation error when $d = 250$ and a lower one when $d = 750$. This pattern can be explained by the FPR of the `bridge` estimator as illustrated in Figure 6 in the appendix. While the FPR is slightly higher in the $d = 250$ case, it is lower in the $d = 750$ case. Similar to the `oracle` results, this appears to be a consequence of the additional penalty term in the eBIC. Considering the estimation error when $f_j(x) = x^{1/3}$, the `poly` and `bridge` estimators retain their performance. As before, the `mle` estimator is misspecified and performs worse than the other two estimators.

Overall, the simulation results suggest that our proposed `poly` estimator performs similarly (binary-continuous data setting) or better (general mixed setting) than the current state-of-the-art. In particular, the `poly` estimator achieves

good performance scores regarding recovery of the graph structure in the general mixed setting. Estimation error results are more sensitive to the choice of the additional high-dimensional penalty. These empirical results suggest that despite the theoretically slower convergence rate, the `poly` estimator is competitive regarding graph recovery and estimation error.

5. Conclusion

Estimating high-dimensional undirected graphs from general mixed data is a challenging task. We propose an innovative approach that blends classical generalized correlation measures, specifically polychoric and polyserial correlations, with recent concepts from high-dimensional graphical modeling and copulas.

A pivotal insight guiding our approach is recognizing that polychoric and polyserial correlations can be effectively modeled through a latent Gaussian copula. Although adapting polyserial correlation to the nonparanormal case demands careful consideration, the polychoric correlation requires no adjustments. The resulting estimators exhibit favorable theoretical properties, even in high dimensions, and demonstrate robust empirical performance in our simulation study.

The framework we put forward builds on prior work extending the graphical lasso for Gaussian observations to nonparanormal models and subsequently to mixed data, as seen in the contributions of Fan et al. [12], followed by Quan, Booth and Wells [35] and Feng and Ning [13]. A key distinction in our approach is the absence of the need to specify bridge functions. Indeed, our method seamlessly handles various types of mixed data without requiring additional user effort.

6. Software

Software in the form of the `R` package **hume** is available on the corresponding author's GitHub page (`https://github.com/konstantingoe/hume`). The `R`-code to reproduce the simulation study conducted in the paper is available under `https://github.com/konstantingoe/mixed_hidim_graphs`.

Appendix A: Methodology

In the following sections, we derive the MLEs for the latent Gaussian model.

A.1. Case II MLE derivation

In *Case II*, we consider the instance where X_j is ordinal and X_k is continuous. Recall that in the latent Gaussian model, we take all $\{f_k\}_{k=1}^d$ to be the identity. Consequently, X_k is Gaussian and $\Gamma_j^r = f_j(\gamma_j^r) = \gamma_j^r$ for all $j \in [d_1]$ and $r \in [l_j]$. Recall that we are interested in the product-moment correlation Σ_{jk} between

two jointly Gaussian variables, where X_j is not directly observed, but only the ordered categories (Eq. (2)) are given. The likelihood of the n-sample is defined by:

$$\begin{aligned} L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k) &= \prod_{i=1}^{n} p(x_{ij}^r, x_{ik}, \Sigma_{jk}) \\ &= \prod_{i=1}^{n} p(x_{ik}) p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}), \end{aligned} \tag{22}$$

where $p(x_{ij}^r, x_{ik}, \Sigma_{jk})$ denotes the joint probability of X_j and X_k and $p(x_{ik})$ the marginal density of the Gaussian variable X_k, i.e.

$$p(x_{ik}) = \left(2\pi\sigma\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{x_{ik} - \mu}{\sigma_{ik}}\right)^2 \right].$$

Furthermore, the conditional probability of X_j in Eq. (5) can be written as:

$$\begin{aligned} p(X_j = x_j^r \mid X_k, \Sigma_{jk}) &= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r \mid X_k, \Sigma_{jk}) \\ &= p(Z_j \leq \Gamma_j^r \mid X_k, \Sigma_{jk}) - p(Z_j \leq \Gamma_j^{r-1} \mid X_k, \Sigma_{jk}) \\ &\quad \Phi(\tilde{\Gamma}_j^r) - \Phi(\tilde{\Gamma}_j^{r-1}), \quad r = 1, \ldots, l_j, \end{aligned} \tag{23}$$

where

$$\tilde{\Gamma}_j^r = \frac{\Gamma_j^r - \Sigma_{jk}\tilde{X}_k}{\sqrt{1 - (\Sigma_{jk})^2}},$$

with $\tilde{X}_k = \frac{X_k - \mu_k}{\sigma_k}$ and $\Phi(t)$ denoting the standard normal distribution function. This follows straight from the fact that the conditional distribution of Z_j is Gaussian with mean $\Sigma_{jk}\tilde{X}_k$ and variance $(1 - (\Sigma_{jk})^2)$. The log-likelihood is then $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)$ with

$$\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k) = \sum_{i=1}^{n} \left[\log(p(x_{ik})) + \log(p(x_j^r \mid x_{ik}, \Sigma_{jk})) \right]. \tag{24}$$

Due to the heavy computational burden involved when estimating all parameters simultaneously, a *two-step estimator* has been proposed [31]. That is, in a first step μ_k, σ_k^2 are estimated by $\bar{X}_k$ and s_k^2, respectively. Moreover, the thresholds $\Gamma_j^r, r = 1, \ldots, l_j$ are estimated by the quantile function of the standard normal distribution evaluated at the cumulative marginal proportions of x_j^r just as described in Section 3.4.

In a second step, all that remains is obtaining the MLE for Σ_{jk} now with the readily computed estimates from the first step:

$$\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n} \frac{1}{p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})} \frac{\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}. \tag{25}$$

Let us take a closer look at the partial derivative of the conditional probability in Eq. (25):

$$\begin{aligned}
&\frac{\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}} \\
&= \frac{\partial \Phi(\tilde{\Gamma}_j^r)}{\partial \Sigma_{jk}} - \frac{\partial \Phi(\tilde{\Gamma}_j^{r-1})}{\partial \Sigma_{jk}} \\
&= \phi(\tilde{\Gamma}_j^r)\frac{\partial \tilde{\Gamma}_j^r}{\partial \Sigma_{jk}} - \phi(\tilde{\Gamma}_j^{r-1})\frac{\partial \tilde{\Gamma}_j^r}{\partial \Sigma_{jk}} \\
&= (1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\Big[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_{ik}) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_{ik})\Big],
\end{aligned} \tag{26}$$

where

$$\tilde{X}_k = \frac{X_k - \bar{X}_k}{\sqrt{s_k^2}} \quad \text{and} \quad \tilde{\Gamma}_j^r = \frac{\Gamma_j^r - \Sigma_{jk}\tilde{X}_k}{\sqrt{1-(\Sigma_{jk})^2}}.$$

The last equality in Eq. (26) follows from taking the derivative and applying the chain rule. Putting all the pieces together, we obtain

$$\begin{aligned}
\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n}\Bigg[&\frac{1}{p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \\
&\Big[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_{ik}) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_{ik})\Big]\Bigg].
\end{aligned} \tag{27}$$

Setting Eq. (27) to zero and solving for Σ_{jk} yields the *Case II* two-step MLE for Σ_{jk}. Note that this is a nonlinear optimization problem, which can efficiently be solved utilizing some quasi-Newton method.

A.2. Case III MLE derivation

Turning to *Case III*, where both X_j and X_k are ordinal variables, we argue in Section 3.3 in the manuscript that the MLE for the polychoric correlation remains valid for the LGCM. The probability of an observation taking values $X_j = x_j^r$ and $X_k = x_k^s$ is

$$\begin{aligned}
\pi_{rs} &\coloneqq p(X_j = x_j^r, X_k = x_k^s) \\
&= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r, \Gamma_k^{s-1} \leq Z_k < \Gamma_k^s) \\
&= p(\Gamma_j^{r-1} \leq f_j(Z_j) < \Gamma_j^r, \Gamma_k^{s-1} \leq f_k(Z_k) < \Gamma_k^s) \\
&= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r}\int_{\Gamma_k^{s-1}}^{\Gamma_s^k} \phi_2(z_j, z_k, \Sigma_{jk})dz_j dz_k,
\end{aligned} \tag{28}$$

where $r \in [l_j]$ and $s \in [l_k]$ and $\phi_2(x, y, \rho)$ denotes the standard bivariate density with correlation ρ. Then, as in the manuscript, the likelihood and log-likelihood of the n-sample are defined as:

$$\begin{aligned} L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= C \prod_{r=1}^{l_j} \prod_{s=1}^{l_k} \pi_{rs}^{n_{rs}}, \\ \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= \log(C) + \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs} \log(\pi_{rs}). \end{aligned} \tag{29}$$

where C is a constant and n_{rs} denotes the observed frequency of $X_j = x_j^r$ and $X_k = x_k^s$ in a sample of size $n = \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs}$. Similar to *Case II* above, we employ the *two-step estimator* for the polychoric correlation. Given the threshold estimates from the first step, let us state the derivative of $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)$ with respect to Σ_{jk} explicitly. First, recall that from Eq. (28)

$$\begin{aligned} \pi_{rs} &= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} \int_{\Gamma_k^{s-1}}^{\Gamma_s^k} \phi_2(z_j, z_k, \Sigma_{jk}) dz_j dz_k \\ &= \Phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \Phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ &\quad - \Phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \Phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}), \end{aligned} \tag{30}$$

where $\Phi_2(u, v, \rho)$ is the standard bivariate normal distribution function with correlation parameter ρ. Note also that we have $\frac{\partial \Phi_2(u,v,\rho)}{\partial \rho} = \phi_2(u, v, \rho)$; see [40]. Taking the derivative of $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)$ with respect to Σ_{jk} yields

$$\begin{aligned} \frac{\partial \ell^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}} &= \sum_{r=1}^{l_{X_j}} \sum_{s=1}^{l_{X_k}} \frac{n_{rs}}{\pi_{rs}} \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \\ &= \sum_{r=1}^{l_{X_j}} \sum_{s=1}^{l_{X_k}} \frac{n_{rs}}{\pi_{rs}} \Big[\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) - \\ &\qquad \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}) \Big]. \end{aligned}$$

Again, setting the derivative to zero and solving for Σ_{jk} using some quasi-Newton method yields the *Case III* two-step MLE for Σ_{jk}.

A.3. Comparison of Case II estimators under the LGCM

In this section, we conduct an empirical comparison of *Case II* estimators within the framework of the LGCM. Specifically, we examine the *Case II* MLE derived under the latent Gaussian model, as discussed in Section 2.1, which involves incorporating the estimated transformations $\hat{f}$ at appropriate locations. Furthermore, we investigate the ad hoc estimator presented in Section 3.2 in more detail.

We start by rewriting Eq. (27), where we replace occurrences of the X_k with the corresponding transformation $\hat{f}_k(X_k)$. The resulting transformation-based first-order condition (FOC) for the *Case II* MLE under the LGCM becomes:

$$\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k))}{\partial \Sigma_{jk}} \tag{31}$$

$$= \sum_{i=1}^{n} \Bigg[\frac{1}{\Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \tag{32}$$

$$\Big[\phi(\tilde{\Gamma}_j^r(\hat{f}_k))(\Gamma_j^r \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) - \phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))(\Gamma_j^{r-1} \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) \Big] \Bigg], \tag{33}$$

where

$$\hat{f}_k(\tilde{x}_{ik}) = \frac{\hat{f}_k(x_{ik}) - \hat{f}_k(\bar{x}_k)}{\sqrt{\hat{f}_k(s_k)^2}} \quad \text{and} \quad \tilde{\Gamma}_j^r(\hat{f}_k) = \frac{\Gamma_j^r - \Sigma_{jk} \hat{f}_k(\tilde{x}_{ik})}{\sqrt{1 - (\Sigma_{jk})^2}}.$$

In the ensuing empirical evaluation, the data generation scheme is as follows. First, we generate n data points $(z_{ij}, x_{ik})_{i=1}^n$ from a standard bivariate normal with correlation Σ_{jk}^*. Second, we apply the same transformation $f_t^{-1}(x) = 5x^5$ for $t \in \{j, k\}$ to all the data points. Third, we generate binary data x_{ij}^r by randomly choosing $f_j^{-1}(z_{ij})$-thresholds (guaranteeing relatively balanced classes) and then applying inversion sampling.

Computing the transformation-based MLE for *Case II* can be achieved in several ways. Consider the plugged-in log-likelihood function, i.e.

$$\begin{aligned} \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k)) &= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \log(p(x_j^r \mid \hat{f}_k(x_{ik}), \Sigma_{jk})) \big] \\ &= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k)) \big]. \end{aligned} \tag{34}$$

One strategy to optimize Eq. (34) is direct maximization with a quasi-Newton optimization procedure to determine the optimal values for $\hat{\Sigma}_{jk}$. This strategy is used, for instance, in the R package `polycor` [15]. Alternatively, another approach involves utilizing the Eq. (31) and solving for $\hat{\Sigma}_{jk}$ through a nonlinear root-finding method. To do this, we employ Broyden's method [6].

In Figure 3, we generate data according to the scheme above for $n = 1000$ and a grid of true correlation values $\Sigma_{jk}^* \in [0, 0.98]$ with a step size of $s = 0.02$. Due to symmetry, taking only positive correlations is sufficient for comparison purposes. For each correlation value along the grid, we generate 100 mixed binary-continuous data pairs and compute the MLE (using the abovementioned strategies) and the ad hoc estimator from Section 3.2. We plot the true correlation values against the absolute difference between estimates and true correlation and the corresponding Monte-Carlo standard error for the MLE and the ad hoc estimator.

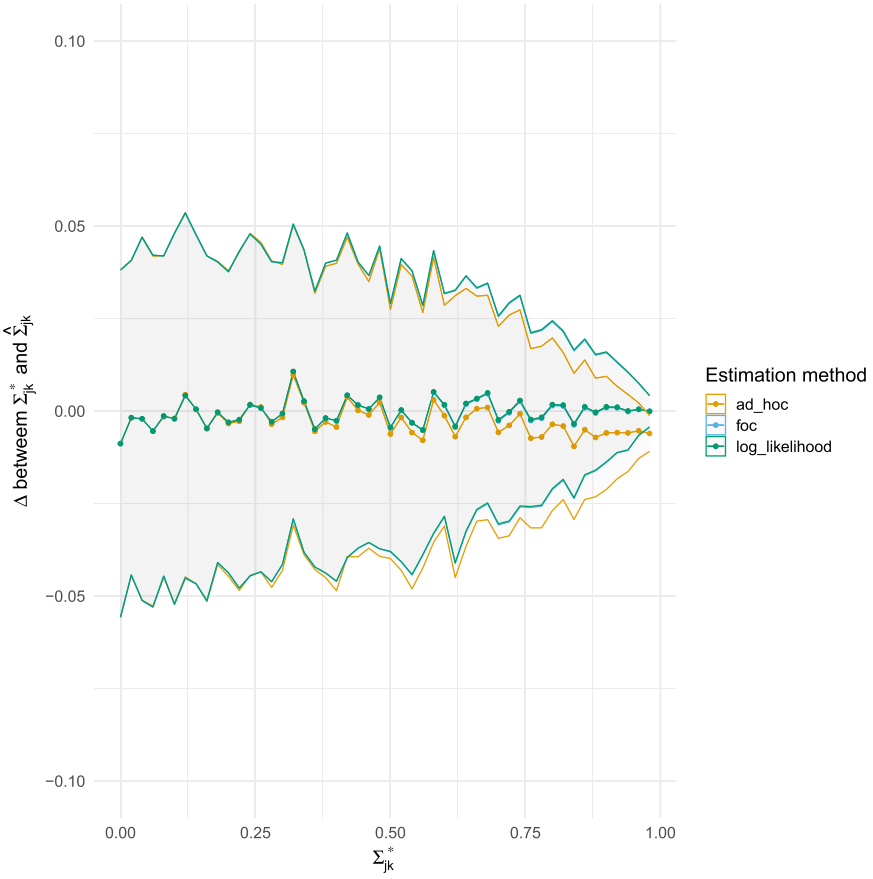


FIG 3. *Comparison of the Case II MLE under the LGCM and the ad hoc estimator.*

As expected, both strategies for attaining the MLE yield the same results. The ad hoc estimator's bias becomes noticeable only when the underlying correlation Σ^*_{jk} exceeds 0.75 and it remains at such a mild level that we consider it negligible [see 31, for a similar observation]. The strength of the ad hoc estimator lies in its simplicity and computational efficiency. The left panel of Figure 4 shows the median computation time surrounded by the first and third quartiles. We compare the two MLE optimization strategies and the ad hoc estimator for a grid of sample sizes $n \in [50, 10000]$ with a step size of $s_t = 50$. Here we fix $\Sigma^*_{jk} = .87$ and repeat each calculation 100 times recording the time elapsed.

The right panel of Figure 4 demonstrates computation time across a grid of length 200 of values for $\Sigma^*_{jk} \in [-.98, 98]$. The sample size is, in this case, fixed at $n = 1000$. The ad hoc estimator is consistently and considerably faster than the MLE, regardless of the strategy used. The difference in computation time is especially pronounced for large sample sizes and correlation values approaching

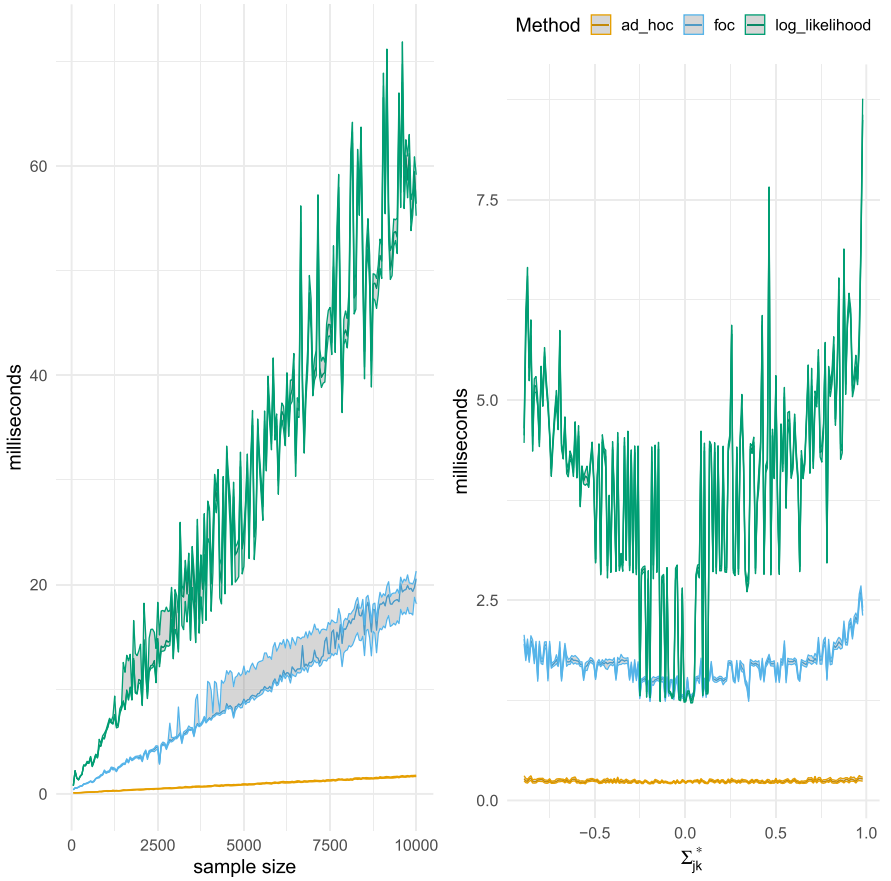


FIG 4. *Computation time in milliseconds for the Case II MLE and ad hoc estimators. We report the median (solid line) and the first and third quartile (shaded area) of recorded computation time. In the left panel, we compare computation time against a grid of sample sizes* $n \in [50, 10000]$ *with a step size of* $s_t = 50$. *In the right panel, we compare computation time against a grid of true correlation values* $\Sigma^*_{jk} \in [-.98, .98]$.

the endpoints of the $[-1, 1]$-interval. Setting the FOC to zero and solving for Σ_{jk} is computationally more efficient than directly maximizing the log-likelihood function. The time difference in MLE strategies is more pronounced at the endpoints of the $[-1, 1]$ interval. The ad hoc estimator is not affected by this issue. Therefore, in the high-dimensional setting we consider in this paper, the ad hoc estimator is preferable to the MLE due to (1) its computational efficiency, (2) its simplicity, which allows us to form concentration inequalities, and (3) its robustness to the underlying correlation value.

Appendix B: Proofs

B.1. Proof of Theorem 3.2

Condition B.1 (Gradient statistical noise)**.** *The gradient of the log-likelihood function is τ^2-sub-Gaussian. That is, for any $\lambda \in \mathbb{R}$ and for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ for j,k according to Case II or Case III.*

$$\mathbb{E}\left[\exp\left(\lambda\Big(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\Big)\right)\right] \leq \exp\Big(\frac{\tau^2\lambda^2}{2}\Big), \tag{35}$$

where ℓ_{jk} corresponds to the log-likelihood functions in Definitions 2.4 and 2.5, respectively.

Case II: *Recall that*

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \frac{1}{p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})} \frac{\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}.$$

Replacing these with the derivations made in Eq. (27), we write

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \frac{(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}}{\Phi(\tilde{\Gamma}_j^r) - \Phi(\tilde{\Gamma}_j^{r-1})}\left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_k)\right].$$

It is easy to see that $p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}) \in (0,1)$ almost surely. Assumption 3.3 makes sure that we exclude impossible events where $p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}) = 0$. Moreover, we require that $\Gamma_j^r > \Gamma_j^{r-1}, \forall j \in 1, \ldots, d_1$ this implies that $\Phi(\tilde{\Gamma}_j^r) > \Phi(\tilde{\Gamma}_j^{r-1})$. In other words, there exists a $\kappa > 0$ such that $p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}) \leq \frac{1}{\kappa}$.

Let us now turn to $\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})/\partial \Sigma_{jk}$. First, for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ we clearly have $1 \leq (1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \leq \varpi$ for $\varpi > 1$. What's more, the density of the standard normal is bounded, i.e., $\big|\phi(t)\big| \leq (2\pi)^{-\frac{1}{2}}$ for all $t \in \mathbb{R}$. Similarly,

$$\Big|\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_k)\Big| \leq \Big|\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k)\Big| \leq L_1,$$

due to Assumption 3.2. Therefore,

$$\left|\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}}\right| \leq \kappa L_1,$$

and $\Big(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\Big)$ is zero-mean and bounded. Then by Hoeffding's [19] lemma, the gradient of the log-likelihood function is τ^2-sub-Gaussian with $\tau = 2\kappa L_1$

Case III: *Recall that we have*

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}} = \frac{1}{\pi_{rs}} \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}, \quad \text{for some } j < k \in [d_1].$$

Considering

$$\begin{aligned}\pi_{rs} =& \Phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \Phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ & - \Phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \Phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}),\end{aligned}$$

we note that this again has to be in $(0,1)$ *due to Assumptions 3.1 and 3.2. such that* $\pi_{rs} \leq \frac{1}{\xi}$.

Now let us show that

$$\begin{aligned}\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} = \Big[&\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ &- \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk})\Big]\end{aligned}$$

is bounded. Indeed, the density of the standard bivariate normal random variable is of the form $\phi_2(x,y) = ce^{-q(x,y)}$. *Since* $q(x,y)$ *is a quadratic function of* x, y *it follows that* $\big|\phi_2(x,y)\big| \leq c$. *Therefore, every element in* $\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}$ *is bounded and thus* $\left|\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}\right| \leq K_1$. *By the same argument as for Case II* $\left(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\right)$ *is zero-mean and bounded and by Hoeffding's lemma the gradient of the log-likelihood function is* τ^2*-sub-Gaussian with* $\tau = 2\xi K_1$. *Based on these arguments, we can conclude that the condition for gradient statistical noise is satisfied.*

Condition B.2 (Hessian statistical noise)**.** *The Hessian of the log-likelihood function is* τ^2*-sub-exponential, i.e. for all* $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ *and for* j, k *according to Case II or Case III we have*

$$\left\| \frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} \right\|_{\psi_1} \leq \tau^2, \tag{36}$$

where $\|\cdot\|_{\psi_1}$ *denotes the Orlicz* ψ_1*-norm, defined as*

$$\|X\|_{\psi_1} := \sup_{p \geq 1} \frac{1}{p} \mathbb{E}\Big(\big|X - \mathbb{E}(X)\big|^p\Big)^{\frac{1}{p}}.$$

Note that ℓ_{jk} *corresponds to the respective log-likelihood functions in Definitions 2.4 and 2.5.*

Case II: *We have*

$$\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} = \frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk})} - \left(\frac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})} \right)^2.$$

$$\begin{aligned} \textit{Clearly, } \left| \frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} \right| &\leq \frac{\left| \partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2 \right|}{\left| p(x_j^r \mid x_k, \Sigma_{jk}) \right|} + \left(\frac{\left| \partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk} \right|}{\left| p(x_j^r \mid x_k, \Sigma_{jk}) \right|} \right)^2 \\ &\leq \kappa L_2 + \kappa^2 L_1^2, \end{aligned} \tag{37}$$

where it remains to show that $\left| \partial^2 p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})/\partial \Sigma_{jk}^2 \right| \leq L_2$. *Indeed, we can rewrite our objective as follows:*

$$\begin{aligned} &\frac{\partial}{\partial \Sigma_{jk}} \left(\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} \right) \\ &= \frac{\partial}{\partial \Sigma_{jk}} \left((1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k) \right] \right) \\ &= \frac{3\Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) + \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k)^2}{(1 - \Sigma_{jk}^2)^3} \\ &+ \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1 - \Sigma_{jk}^2)^{-\frac{3}{2}}} - \frac{3\Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k) \\ &- \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k)^2}{(1 - \Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1 - \Sigma_{jk}^2)^{-\frac{3}{2}}}. \end{aligned} \tag{38}$$

Thus,

$$\begin{aligned} \left| \frac{\partial}{\partial \Sigma_{jk}} \left(\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} \right) \right| \leq & \left| \frac{3\Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) \right| \\ & + \left| \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k)^2}{(1 - \Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1 - \Sigma_{jk}^2)^{-\frac{3}{2}}} \right| \\ & \leq L_2, \end{aligned}$$

due to Assumptions 3.1 and 3.2 and because both $\phi(t)$ *and* $\phi'(t)$ *are bounded for all* $t \in \mathbb{R}$. *Therefore, the inequality in Eq. (37) follows and* $\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} - \mathbb{E}\left(\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} \right)$ *is bounded by* $2(\kappa L_2 + \kappa^2 L_1^2)$. *Hence, for all* $p \geq 1$

$$\frac{1}{p} \mathbb{E} \left[\left| \partial^2 \ell_{jk}/\partial \Sigma_{jk}^2 - \mathbb{E}\big(\partial^2 \ell_{jk}/\partial \Sigma_{jk}^2\big) \right|^p \right]^{\frac{1}{p}} \leq \frac{2}{p} \big(\kappa L_2 + \kappa^2 L_1^2\big). \tag{39}$$

Finally, for $\tau = 2\kappa L_1$ *we can choose* L_1 *and* κ *such that*

$$2(\kappa L_2 + \kappa^2 L_1^2) \leq \tau^2 = 4\kappa^2 L_1^2.$$

Thus, the Hessian statistical noise-condition for Case II is satisfied.

Case III: *Let us start with the Hessian of* ℓ_{jk} *in the polychoric case:*

$$\frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}^2} = \frac{\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2}{\pi_{rs}} - \left(\frac{\partial \pi_{rs}/\partial \Sigma_{jk}}{\pi_{rs}} \right)^2$$

$$\text{Thus: } \left| \frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}^2} \right| \leq \frac{\left| \partial^2 \pi_{rs}/\partial \Sigma_{jk}^2 \right|}{|\pi_{rs}|} + \left(\frac{\left| \partial \pi_{rs}/\partial \Sigma_{jk} \right|}{|\pi_{rs}|} \right)^2 \tag{40}$$

$$\leq \xi K_2 + \xi^2 K_1^2.$$

Again it remains to show that $\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2 \leq K_2$. *Consider*

$$\begin{aligned}
&\left| \frac{\partial}{\partial \Sigma_{jk}} \left(\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \right) \right| \\
&= \Big| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\
&- \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}) \Big| \\
&\leq \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) \right| + \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \right| \\
&+ \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) \right| + \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}) \right| \\
&\leq K_2,
\end{aligned}$$

where each of the derivatives of the bivariate density is bounded, since we assume that the correlation is bounded away from one and minus one, i.e., $\Sigma_{jk} \in [-1 + \delta, 1 - \delta]$.

Similar to Case II, for $\tau = 2\xi K_1$ *we can choose* K_1 *and* ξ *such that*

$$2(\xi k_2 + \xi^2 k_1^2) \leq \tau^2 = 4\xi^2 K_1^2.$$

Consequently, the Hessian statistical noise-condition for Case III is satisfied, which concludes the proof of the Hessian statistical noise-condition.

Concerning the third condition, we introduce some additional notation. Denote the sample risk by $\hat{R}_{jk}^{(n)}(\Sigma_{jk})$ for j, k according to *Case II* and *Case III*, i.e.,

$$\textit{Case II}: \quad \hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \frac{1}{n} \sum_{i=1}^{n} \big[\log(p(x_{ik})) + \log(p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})) \big],$$

and

$$\textit{Case III:} \quad \hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \frac{1}{n}\sum_{r=1}^{l_{X_j}}\sum_{s=1}^{l_{X_k}} n_{rs}\log(\pi_{rs}).$$

Lastly, we define $R_{jk}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\hat{R}_{jk}^{(n)}(\Sigma_{jk})$ to be the population risk for each of the respective cases.

Condition B.3 (Hessian regularity)**.** *The Hessian regularity condition consists of three parts:*

1. *The second derivative of the population risk $R_{jk}(\Sigma_{jk})$ is bounded at one point. That is, there exists one $|\bar{\Sigma}_{jk}| \leq 1-\delta$ and $H > 0$ such that $|R''_{jk}(\bar{\Sigma}_{jk})| \leq H$.*
2. *The second derivative of the log-likelihood with respect to Σ_{jk} is Lipschitz continuous with integrable Lipschitz constant, i.e. there exists a $M^* > 0$ such that $\mathbb{E}[M] \leq M^*$, where*

$$M = \sup_{\substack{|\Sigma_{jk}^{(1)}|,\ |\Sigma_{jk}^{(2)}|\leq 1-\delta,\\ \Sigma_{jk}^{(1)}\neq\Sigma_{jk}^{(2)}}} \frac{\left|\ell''_{jk}(\Sigma_{jk}^{(1)}) - \ell''_{jk}(\Sigma_{jk}^{(2)})\right|}{\left|\Sigma_{jk}^{(1)} - \Sigma_{jk}^{(2)}\right|}.$$

3. *The constants H and M^* are such that $H \leq \tau^2$ and $M^* \leq \tau^3$.*

We need some intermediate results that make dealing with $R_{jk}(\Sigma_{jk})$ easier. First, note that $\mathbb{E}_{\Sigma_{jk}^*}\hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\ell_{jk}(\Sigma_{jk})$. Second, for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ and $m \in \{1,2\}$

$$R_{jk}^m(\Sigma_{jk}) = \frac{\partial^m}{\partial\Sigma_{jk}^m}\mathbb{E}_{\Sigma_{jk}^*}\ell_{jk}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\frac{\partial^m}{\partial\Sigma_{jk}^m}\ell_{jk}(\Sigma_{jk}),$$

by Lemma B.7 and Corollary B.8.

1. *Recall the first part of the Hessian regularity condition, whereby Eq. (37) and Eq. (40) for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ we have*

$$\textit{Case II:} \quad \left|\frac{\partial^2}{\partial\Sigma_{jk}^2}\ell_{jk}(\Sigma_{jk})\right| \leq \kappa L_2 + \kappa^2 L_1^2$$

and

$$\textit{Case III:} \quad \left|\frac{\partial^2}{\partial\Sigma_{jk}^2}\ell_{jk}(\Sigma_{jk})\right| \leq \xi K_2 + \xi^2 K_1^2$$

for cases II and III, respectively. Consequently, the claim in the first part of the Hessian regularity condition holds for Case II and Case III for any $|\bar{\Sigma}_{jk}| \leq 1-\delta$ with $H_1 = \kappa L_2 + \kappa^2 L_1^2$ and $H_2 = \xi K_2 + \xi^2 K_1^2$. Moreover, we also have $H_1 \leq \tau^2 = 4\kappa^2 L_1^2$ and $H_2 \leq \tau^2 = 4\xi^2 K_1^2$.

2. *The second part of the Hessian regularity condition requires that the second derivative of the log-likelihood with respect to Σ_{jk} is Lipschitz continuous with integrable Lipschitz constant. By the mean-value-theorem, all we need to show is that we can find a bound on the third derivative of the log-likelihood function.*

Case II: *Note that we have*

$$\begin{aligned}
&\frac{\partial^3}{\partial\Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk}) \\
&\quad= \frac{\partial}{\partial\Sigma_{jk}}\left[\frac{\partial^2\ell_{jk}}{\partial\Sigma_{jk}^2}\right] \\
&\quad= \frac{\partial}{\partial\Sigma_{jk}}\left[\frac{\partial^2 p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}^2}{p(x_j^r \mid x_k,\Sigma_{jk})} - \left(\frac{\partial p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}}{p(x_j^r \mid x_k,\Sigma_{jk})}\right)^2\right] \\
&\quad= \frac{\partial^3 p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}^3}{p(x_j^r \mid x_k,\Sigma_{jk})} \\
&\quad\quad- 3\frac{\big(\partial p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}\big)\big(\partial^2 p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}^2\big)}{(p(x_j^r \mid x_k,\Sigma_{jk}))^2} \\
&\quad\quad+ 2\left(\frac{\partial p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}}{p(x_j^r \mid x_k,\Sigma_{jk})}\right)^3.
\end{aligned}$$

Hence

$$\left|\frac{\partial^3}{\partial\Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk})\right| \le \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3.$$

It remains to show therefore, that

$$\left|\partial^3 p(x_j^r \mid x_k,\Sigma_{jk})/\partial\Sigma_{jk}^3\right| \le L_3.$$

When taking the derivative of Eq. (38), it is obvious that the resulting statement is bounded due to Assumptions 3.1 and 3.2 and the fact that $\phi(t), \phi'(t), \phi''(t)$ are all bounded for all $t \in \mathbb{R}$.
Therefore, by applying the mean-value-theorem we get

$$M_1 \le \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3,$$

and the natural choice for

$$M_1^* = \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3,$$

where it follows that

$$M_1^* \le \tau^3 = 8\kappa^3 L_1^3.$$

Case III: *We proceed similarly and first consider*

$$\begin{aligned}
&\frac{\partial^3}{\partial \Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk}) \\
&\quad= \frac{\partial}{\partial \Sigma_{jk}}\left[\frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}^2}\right] \\
&\quad= \frac{\partial}{\partial \Sigma_{jk}}\left[\frac{\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2}{\pi_{rs}} - \left(\frac{\partial \pi_{rs}/\partial \Sigma_{jk}}{\pi_{rs}}\right)^2\right] \\
&\quad= \frac{\partial^3 \pi_{rs}/\partial \Sigma_{jk}^3}{\pi_{rs}} - 3\frac{\left(\partial \pi_{rs}/\partial \Sigma_{jk}\right)\left(\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2\right)}{(\pi_{rs})^2} \\
&\quad+ 2\left(\frac{\partial \pi_{rs}/\partial \Sigma_{jk}}{\pi_{rs}}\right)^3.
\end{aligned} \tag{41}$$

Hence

$$\left|\frac{\partial^3}{\partial \Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk})\right| \leq \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3.$$

Taking a closer look at $\left|\partial^3 \pi_{rs}/\partial \Sigma_{jk}^3\right|$*, boundedness again follows from the fact that the quadratic function in the exponential of the bivariate normal density does not vanish. Thus, we have*

$$M_2 \leq \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3,$$

and the natural choice for

$$M_2^* = \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3,$$

where we have

$$M_2^* \leq \tau^3 = 8\xi^3 K_1^3.$$

These steps validate the Hessian regularity condition.

Condition B.4 (Population risk is strongly Morse)**.** *There exist* $\epsilon > 0$ *and* $\eta > 0$ *such that* $R_{jk}(\Sigma_{jk})$ *is* (ϵ, η)*-strongly Morse, i.e.*

1. *For all* Σ_{jk} *such that* $|\Sigma_{jk}| = 1 - \delta$ *we have that* $|R'_{jk}(\Sigma_{jk})| > \epsilon$.
2. *For all* Σ_{jk} *such that* $|\Sigma_{jk}| \leq 1 - \delta$*:*

$$|R'_{jk}(\Sigma_{jk})| \leq \epsilon \implies |R''_{jk}(\Sigma_{jk})| \geq \eta.$$

Put differently, $R_{jk}(\Sigma_{jk})$ *is* (ϵ, η)*-strongly Morse if the boundaries* $\{-1+\delta, 1-\delta\}$ *are not critical points of* $R_{jk}(\Sigma_{jk})$ *and if* $R_{jk}(\Sigma_{jk})$ *only has finitely many critical points that are all non-degenerate.*

Let us verify that $R''(\Sigma_{jk}) \neq 0$ *for cases II and III. Indeed by Lemma B.7 and Corollary B.8 we can rewrite* $R''_{jk}(\Sigma_{jk})$ *and obtain*

$$
\begin{aligned}
&R''(\Sigma_{jk}) \\
&\quad = \mathbb{E}_{\Sigma_{jk}}\left[\frac{\partial^2 \ell_{jk}(\Sigma_{jk})}{\partial \Sigma_{jk}^2}\right] \\
&\quad = \mathbb{E}_{\Sigma_{jk}^*}\begin{cases} \dfrac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk})} - \left(\dfrac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})}\right)^2 & \textit{Case II,} \\ \dfrac{\partial^2 \pi(\Sigma_{jk})_{rs}/\partial \Sigma_{jk}^2}{\pi(\Sigma_{jk})_{rs}} - \left(\dfrac{\partial \pi(\Sigma_{jk})_{rs}/\partial \Sigma_{jk}}{\pi(\Sigma_{jk})_{rs}}\right)^2 & \textit{Case III,} \end{cases}
\end{aligned}
$$

where in $\pi(\Sigma_{jk})_{rs}$ *we made the dependence on* Σ_{jk} *explicit.*

Case II: *Recall that we have*

$$
\begin{aligned}
&\mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk}^*)}\right] \\
&\quad = \int_{-\infty}^{\infty} \sum_{r=1}^{l_j+1} \frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk}^*)} p(x_j^r, x_k; \Sigma_{jk}^*) dx_k \\
&\quad = \int_{-\infty}^{\infty} \sum_{r=1}^{l_j+1} \frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk}^*)} p(x_j^r \mid x_k, \Sigma_{jk}^*) p(x_k) dx_k \\
&\quad = \sum_{r=1}^{l_j+1} \partial^2 p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial \Sigma_{jk}^2,
\end{aligned}
$$

with

$$
\begin{aligned}
&\sum_{r=1}^{l_j+1} \partial^2 p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial \Sigma_{jk}^2 \\
&\qquad = \sum_{r=1}^{l_j+1} \left[\frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) \right. \\
&\qquad + \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \\
&\qquad - \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k) \\
&\qquad \left. - \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \right] \\
&\qquad = 0,
\end{aligned}
$$

since all terms except the ones involving $\phi(\tilde{\Gamma}_j^0)$ *and* $\phi(\tilde{\Gamma}_{l_j+1}^j)$ *cancel and furthermore* $\lim_{t \to \pm\infty} \phi(t) = \lim_{t \to \pm\infty} \phi'(t) = 0$.

Case III: *Similarly, consider*

$$
\begin{aligned}
&\mathbb{E}_{\Sigma^*_{jk}}\left[\frac{\partial^2\pi(x_j^r, x_s^k; \Sigma^*_{jk})/\partial\Sigma^2_{jk}}{\pi(x_j^r, x_s^k; \Sigma^*_{jk})}\right] \\
&\quad= \sum_r \sum_s \left[\frac{\partial^2\pi(x_j^r, x_s^k; \Sigma^*_{jk})/\partial\Sigma^2_{jk}}{\pi(x_j^r, x_s^k; \Sigma^*_{jk})} P(X_j = x_j^r, X_k = x_k^s)\right] \\
&\quad= \sum_r \sum_s \left[\partial^2\pi(x_j^r, x_s^k; \Sigma^*_{jk})/\partial\Sigma^2_{jk}\right] \\
&\quad= \sum_r \sum_s \Big[q(\Gamma_j^r, \Gamma_k^s, \Sigma^*_{jk})\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma^*_{jk}) \\
&\quad- q(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma^*_{jk})\phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma^*_{jk}) \\
&\quad- q(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma^*_{jk})\phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma^*_{jk}) \\
&\quad+ q(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma^*_{jk})\phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma^*_{jk})\Big] \\
&\quad= q(\Gamma_j^{l_j+1}, \Gamma_k^{l_k+1}, \Sigma^*_{jk})\phi_2(\Gamma_j^{l_j+1}, \Gamma_k^{l_k+1}, \Sigma^*_{jk}) \\
&\quad- q(\Gamma_j^{l_j+1}, \Gamma_k^0, \Sigma^*_{jk})\phi_2(\Gamma_j^{l_j+1}, \Gamma_k^0, \Sigma^*_{jk}) \\
&\quad- q(\Gamma_j^0, \Gamma_k^{l_k+1}, \Sigma^*_{jk})\phi_2(\Gamma_j^0, \Gamma_k^{l_k+1}, \Sigma^*_{jk}) \\
&\quad+ q(\Gamma_j^0, \Gamma_k^0, \Sigma^*_{jk})\phi_2(\Gamma_j^0, \Gamma_k^0, \Sigma^*_{jk}) \\
&\quad= 0,
\end{aligned}
$$

with $q(s, t, \Sigma^*_{jk}))$ denoting the corresponding quadratic function from the derivative of the bivariate normal density. As above, $\Gamma_k^{l_k+1} = \infty$ and $\Gamma_k^0 = -\infty$ for all $k \in 1, \dots d_1$. This, together with the fact that all other terms cancel when summing over r, s, $\phi(\cdot)$ is zero in all points containing $\Gamma_k^{l_k+1}, \Gamma_k^0$ and so the last equality follows.

From this it follows, that $R''_{jk}(\Sigma^*_{jk})$ can only be zero if $\partial p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial\Sigma_{jk}$ for Case II and $\partial\pi(\Sigma^*_{jk})_{rs}/\partial\Sigma_{jk}$ for Case III are zero. However, this is not possible due to Assumptions 3.2 and 3.3. To see this note that in Eq. (26) $\partial p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial\Sigma_{jk}$ can only be zero if either $\Gamma_j^r = \Gamma_j^{r-1}$ which we ruled out in the definition of the LGCM, or if $|\Gamma_j^r| = |\Gamma_j^{r-1}| = \infty$ which is ruled out by Assumption 3.2. We would not observe any discrete states in the first place if we had $r = \{0, l_j + 1\}$. Assumption 3.3 rules this case out. Consequently, there exist $\epsilon > 0$ and $\eta > 0$ such that $R_{jk}(\Sigma_{jk})$ is (ϵ, η)-strongly Morse

With these considerations, we have verified the required four conditions to hold such that Theorem 2 in Mei, Bai and Montanari [26] holds for each couple (j, k) according to cases II and III. More precisely, let $\alpha \in (0, 1)$. Now, letting $n \geq 4C\log(n)\log(\frac{B}{\alpha})$ where $C = C_0\left(\frac{\tau^2}{\epsilon^2} \vee \frac{\tau^4}{\eta^2} \vee \frac{\tau^2 L^2}{\eta^4}\right)$ and $B = \tau(1 - \delta)$ with $\tau = 2[\kappa L_1 \vee \xi K_1]$ and C_0 denoting a universal constant. Letting further $L =$

$\sup_{\Sigma_{jk}:|\Sigma_{jk}|\leq 1-\delta}|R'''_{jk}(\Sigma_{jk})|$ we obtain

$$\mathbb{P}\left(\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{*}_{jk}\right|\geq\frac{2\tau}{\eta}\sqrt{C_0\frac{\log(n)}{n}\left[\log\left(\frac{B}{\alpha}\right)\vee 1\right]}\right)\leq\alpha, \tag{42}$$

and consequently, the result in Theorem 3.2 follows.

B.2. Proof of Lemmas B.5 to B.8

Lemma B.5. *For all $|\Sigma_{jk}|\leq 1-\delta$ and all $j\in[d_1],k\in[d_2]$ we have*

$$\int_S\frac{\partial}{\partial\Sigma_{jk}}p(x_j^r\mid x_k,\Sigma_{jk})d\mu(x_j^r)=\frac{\partial}{\partial\Sigma_{jk}}\int_S p(x_j^r\mid x_k,\Sigma_{jk})d\mu(x_j^r),$$

where μ is the counting measure on S, the corresponding discrete space.

Proof. Clearly, from Eq. (26) we have

$$\begin{aligned}&\frac{\partial}{\partial\Sigma_{jk}}p(x_j^r\mid x_k,\Sigma_{jk})\\&\qquad=(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)-\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)\right],\end{aligned}$$

and therefore

$$\begin{aligned}&\int_S(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)-\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)\right]d\mu(x_j^r)\\&\quad=(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\sum_{r=1}^{l_j}\left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)-\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)\right]\\&\quad=0\\&\quad=\frac{\partial}{\partial\Sigma_{jk}}\sum_{r=1}^{l_j}p(x_j^r\mid x_k,\Sigma_{jk})\\&\quad=\frac{\partial}{\partial\Sigma_{jk}}1,\end{aligned}$$

since all terms except the ones involving $\phi(\tilde{\Gamma}_j^0)$ and $\phi(\tilde{\Gamma}_j^{l_j+1})$ cancel and

$$\lim_{t\to\pm\infty}\phi(t)=\lim_{t\to\pm\infty}\phi'(t)=0,$$

and probabilities associated with all possible values must sum up to one. □

Corollary B.6. *For all $|\Sigma_{jk}|\leq 1-\delta$ and all $j\in[d_1],k\in[d_2]$ we have*

$$\int_S\frac{\partial^2}{\partial\Sigma_{jk}^2}p(x_j^r\mid x_k,\Sigma_{jk})d\mu(x_j^r)=\frac{\partial^2}{\partial\Sigma_{jk}^2}\int_S p(x_j^r\mid x_k,\Sigma_{jk})d\mu(x_j^r),$$

where μ is the counting measure on S, the corresponding discrete space.

Proof. From Eq. (38) we obtain

$$\begin{aligned}
&\frac{\partial^2}{\partial\Sigma_{jk}^2} p(x_j^r \mid x_k, \Sigma_{jk}) \\
&\quad = \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k) \\
&\quad + \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \\
&\quad - \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k) \\
&\quad - \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}}.
\end{aligned}$$

By similar arguments to Lemma B.5, when taking the sum over all possible states, all terms except the ones involving $\phi(\tilde{\Gamma}_j^0)$ and $\phi(\tilde{\Gamma}_j^{l_j+1})$ still cancel as they appear in every additive term in the above equation – recall that $\phi'(t) = -t\phi(t)$ – and equality then follows immediately. □

Lemma B.7. *For all $|\Sigma_{jk}| \leq 1-\delta$ we have*

1. $$\frac{\partial}{\partial\Sigma_{jk}}\mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial}{\partial\Sigma_{jk}}\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right],$$

   i.e.

   $$\begin{aligned}
   &\frac{\partial}{\partial\Sigma_{jk}}\int_{S\times\mathbb{R}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\
   &\quad = \int_{S\times\mathbb{R}} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k)
   \end{aligned}$$

   where ε is the product measure on $S\times\mathbb{R}$ defined by

   $$\varepsilon := \mu \otimes \lambda$$

   with μ denoting the counting measure on the corresponding discrete space S and λ the Lebesgue measure on the corresponding Euclidean space.

2. $$\frac{\partial}{\partial\Sigma_{jk}}\mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial}{\partial\Sigma_{jk}}\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right],$$

   i.e.

   $$\begin{aligned}
   &\frac{\partial}{\partial\Sigma_{jk}}\int_{S\times S'} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k^s) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k^s) d\varpi(x_j^r, x_k^s) \\
   &\quad = \int_{S\times S'} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k^s) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k^s) d\varpi(x_j^r, x_k^s)
   \end{aligned}$$

where ϖ is the product measure on $S \times S'$ defined by

$$\varpi \coloneqq \mu \otimes \mu'$$

with μ and μ' denoting the counting measure on the corresponding discrete space S and S', respectively.

Proof. Let us start with 1. and rewrite the right-hand side:

$$\begin{aligned}\int_{S\times\mathbb{R}} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\ = \int_{\mathbb{R}} \sum_{r=1}^{l_j} \frac{\partial}{\partial\Sigma_{jk}} \log p(x_j^r, x_k, \Sigma_{jk}) p(x_j^r, x_k, \Sigma_{jk}^*) dx_k.\end{aligned}$$

The left-hand side corresponds to

$$\begin{aligned}\frac{\partial}{\partial\Sigma_{jk}} \int_{S\times\mathbb{R}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\ = \frac{\partial}{\partial\Sigma_{jk}} \int_{\mathbb{R}} \sum_{r=1}^{l_j} \log p(x_j^r, x_k, \Sigma_{jk}) p(x_j^r, x_k, \Sigma_{jk}^*) dx_k.\end{aligned}$$

By Lebesgue's Dominated Convergence Theorem, we can interchange integration and differentiation as $\log p(x_j^r, x_k, \Sigma_{jk})$ is absolutely continuous s.t. its derivative exists almost everywhere and

$$\left| \frac{\partial \log p(x_j^r, x_k, \Sigma_{jk})}{\partial\Sigma_{jk}} \right|$$

is upper bounded by some integrable function. Indeed, the latter requirement has already been shown in Condition B.1. The second point follows by the same arguments where $\log p(x_j^r, x_k^s, \Sigma_{jk}) = \log(C) + \log(\pi_{rs})$ is absolutely continuous and bounded as shown in Condition B.1. This concludes the proof. □

Corollary B.8. *For all $|\Sigma_{jk}| \leq 1 - \delta$ we have*

$$\frac{\partial^2}{\partial\Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2}{\partial\Sigma_{jk}^2} \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right], \text{ for Case II and}$$

$$\frac{\partial^2}{\partial\Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2}{\partial\Sigma_{jk}^2} \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right], \text{ for case III.}$$

Proof. The claim follows immediately by the same arguments as in Lemma B.7 and the bound on the second derivative of the log-likelihood functions in Condition B.2, respectively. □

B.3. Proof of Lemma 3.1

First, note that $\Phi^{-1}(\cdot)$ is Lipschitz on $[\Phi(-G), \Phi(G)]$ with a Lipschitz constant $L_1 = \nabla\Phi^{-1}(\hat{\pi}_j^r) = 1/(\sqrt{\frac{2}{\pi}}\min\{\hat{\pi}_j^r, 1-\hat{\pi}_j^r\})$. Then we have

$$\begin{aligned}|\hat{\Gamma}_j^r - \Gamma_j^r| &= \Big|\Phi^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r)\big) - \Phi^{-1}\big(\Phi(\Gamma_j^r)\big)\Big| \\ &\le L_1\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r) - \Phi(\Gamma_j^r)\Big|.\end{aligned}$$

Applying Hoeffding's inequality, we obtain for some $t > 0$

$$\begin{aligned}P\Big(|\hat{\Gamma}_j^r - \Gamma_j^r| \ge t\Big) &\le P\Big(L_1\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r) - \Phi(\Gamma_j^r)\Big| \ge t\Big) \\ &\le 2\exp\Big(-\frac{2t^2 n}{L_1^2}\Big).\end{aligned}$$

This concludes the proof.

B.4. Proof of Theorem 3.3

In what follows, the proof of Theorem 3.3 revolves largely around the Winsorized estimator introduced in Section 3.2. Recall that

$$\hat{f}(x) = \Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(x)])$$

where

$$W_{\delta_n}(u) \equiv \delta_n I(u < \delta_n) + uI(\delta_n \le u \le (1-\delta_n)) + (1-\delta_n)I(u > (1-\delta_n)),$$

with truncation constant $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$. Recall $f(x) = \Phi^{-1}(F_{X_k}(x))$, and let $g = f^{-1}$.

Assume w.l.o.g. that we have consecutive integer scoring in our discrete variable X_j such that the polyserial ad hoc estimator simplifies as

$$\hat{\Sigma}_{jk}^{(n)} = \frac{r_{\hat{f}(X_k),X_j}^{(n)}\sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)(x_j^{r+1} - x_j^r)} = \frac{r_{\hat{f}(X_k),X_j}^{(n)}\sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)} = \frac{S_{\hat{f}(X_k)X_j}}{\sigma_{\hat{f}(X_k)}^{(n)}\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)}, \quad (43)$$

for all $j \in [d_1], k \in [d_2]$. $S_{\hat{f}(X_k)X_j}$ denotes the sample covariance between the $\hat{f}(X_k)$ and the X_j, i.e.

$$S_{\hat{f}(X_k)X_j} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}(X_{ik}) - \mu_n(\hat{f})\Big)\Big(X_{ij} - \mu_n(X_j)\Big),$$

where $\mu_n(\hat{f}) = 1/n \sum_{i=1}^n \hat{f}(X_{ik})$ and $\mu_n(X_j) = 1/n \sum_{i=1}^n X_{ij}$. Moreover, $\sigma^{(n)}_{\hat{f}(X_k)}$ denotes the sample standard deviation of the Winsorized estimator, i.e.

$$\sigma^{(n)}_{\hat{f}(X_k)} \equiv \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(X_{ik}) - \mu_n(\hat{f})\right)^2}.$$

Recall that we treat the threshold estimates as given. In particular, we have here $\phi(\bar{\Gamma}^r_j)$, therefore note further that $\phi(\cdot)$, namely the density function of the standard normal is Lipschitz with Lipschitz constant $L_0 = (2\pi)^{-1/2}\exp(-1/2)$, s.t.

$$\left|\phi(\bar{\Gamma}^r_j) - \phi(\Gamma^r_j)\right| \leq L_0\left|\bar{\Gamma}^r_j - \Gamma^r_j\right| \leq \left|\bar{\Gamma}^r_j - \Gamma^r_j\right|,$$

as $L_0 < 1$. Consequently, the statements regarding the accuracy of the threshold estimates in Section 3.4 still hold here.

The outline of the proof is as follows: We start by forming concentration bounds for the sample covariance and the sample standard deviation separately. Then, we argue that the quotient of the two will be accurate in terms of a Lipschitz condition on the corresponding compactum. Let us start with the sample covariance. To study the Winsorized estimator, we consider the interval $[g(-\sqrt{M\log n}), g(\sqrt{M\log n})]$ for a choice of $M > 2$. As the behavior of the estimator is different for the endpoints, we further split this interval into a middle and an end part, respectively, i.e.,

$$\begin{aligned}\mathbb{M}_n &\equiv (g(-\sqrt{\beta\log n}), g(\sqrt{\beta\log n}))\\ \mathbb{E}_n &\equiv [g(-\sqrt{M\log n}), g(-\sqrt{\beta\log n})) \cup (g(\sqrt{\beta\log n}), g(\sqrt{M\log n})].\end{aligned}$$

This is only necessary for $\hat{f}(X_k)$ since $X_j \in 1,\ldots,l_j$ is discrete and can therefore only take finitely many values. Now consider the sample covariance, where we have for any $t > 0$ that

$$\begin{aligned}&P\left(\max_{j,k}\left|S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j}\right| > 2t\right)\\ &\quad= P\Bigg(\max_{j,k}\Bigg|\frac{1}{n}\sum_{i=1}^n\Big[\hat{f}(X_{ik})X_{ij} - f(X_{ik})X_{ij}\\ &\qquad- \mu_n(\hat{f})\mu_n(X_j) + \mu_n(f)\mu_n(X_j)\Big]\Bigg| > 2t\Bigg)\\ &\quad\leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n\left[(\hat{f}(X_{ik}) - f(X_{ik}))X_{ij}\right]\right| > t\right)\\ &\qquad+ P\left(\max_{j,k}\left|(\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j)\right| > t\right).\end{aligned}$$

Let us take a closer look at the second term

$$
\begin{aligned}
&P\left(\max_{j,k} \left| (\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j) \right| > t \right) \\
&\qquad = P\left(\max_{j,k} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ik}) - f(X_{ik}) \right) \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right| > t \right) \\
&\qquad = P\left(\max_{k} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ik}) - f(X_{ik}) \right) \right| \max_{j} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right| > t \right) \\
&\qquad \leq P\left(\max_{k} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ik}) - f(X_{ik}) \right) \right| > \frac{t}{l_{\max}} \right),
\end{aligned}
$$

where X_j is a discrete random variable with finite level set and $l_{\max} \equiv \max_j l_j > 0$.

Now, define

$$\triangle_i(j,k) \equiv (\hat{f}(X_{ik}) - f(X_{ik}))X_{ij}$$

and

$$\tilde{\triangle}_{r,s} \equiv (\hat{f}(s) - f(s))r,$$

for $r = 1, \ldots, l_j$. Furthermore, consider the event $\mathcal{A}_n$, where

$$\mathcal{A}_n \equiv \{ g(-\sqrt{M \log n}) \leq X_{1k}, \ldots, X_{nk} \leq g(\sqrt{M \log n}), k = d_1 + 1, \ldots, d \}.$$

The bound for the complement arises from the Gaussian maximal inequality Liu, Lafferty and Wasserman [24, Lemma 13], i.e.,

$$P(\mathcal{A}_n^c) \leq P\left(\max_{i,k \in \{1,\ldots,n\} \times \{d_1+1,\ldots,d\}} \left| f(X_{ik}) \right| > \sqrt{2 \log(n d_2)} \right) \leq \frac{1}{2\sqrt{\pi \log(n d_2)}}.$$

The following lemma gives insight into the behavior of the Winsorized estimator along the end region.

Lemma B.9. *On the event $\mathcal{A}_n$, consider $\beta = \frac{1}{2}$, $t \geq C_M \sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$ and $A = \sqrt{\frac{2}{\pi}}(\sqrt{M} - \sqrt{\beta})$, then*

$$P\left(\max_{j,k} \frac{1}{n} \sum_{X_k \in \mathbb{E}_n} \left| (\hat{f}(X_{ik}) - f(X_{ik}))X_{ij} \right| > \frac{t}{2} \right) \leq \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big),$$

and

$$P\left(\max_{k} \frac{1}{n} \sum_{X_k \in \mathbb{E}_n} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| > \frac{t}{2} \right) \leq \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big),$$

where $k_i, i \in \{1, 2, 3\}$ are generic constants independent of sample size and dimension.

Proof. Let $\theta_1 \equiv \frac{n^{\beta/2}t}{4A\sqrt{\log n}}$ and let us first consider the bound for the first inequality.

$$
\begin{aligned}
&P\bigg(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\bigg)\\
&\qquad=P\Big(\max_{j,k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\\
&\qquad\qquad\cap\max_{jk}\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(t)-f(t)\Big||r|>\theta_1\Big)\\
&\qquad+P\Big(\max_{j,k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\\
&\qquad\qquad\cap\max_{jk}\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r|\le\theta_1\Big)\\
&\le P\Big(\max_{jk}\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r|>\theta_1\Big)+P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}}>\frac{t}{2\theta_1}\Big).
\end{aligned}
$$

Similarly, for the bound of the second inequality, we have

$$
\begin{aligned}
&P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\Big)\\
&=P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\cap\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)\\
&+P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\cap\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|\le\theta_1\Big)\\
&\qquad\le P\Big(\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)+P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}}>\frac{t}{2\theta_1}\Big).
\end{aligned}
$$

Recall that $\sup\{1,\ldots,l_j\}=l_j>0$. Furthermore, Lemma 16 in [24] states that for all n

$$
\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|<\sqrt{2(M+2)\log n} \tag{44}
$$

With this in mind, we have

$$
P\Big(\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)\le d_2P\Big(\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big).
$$

Recall that $C_M=8/\sqrt{\pi}(\sqrt{2M}-1)(M+2)$ and since $t\ge C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$, we have

$$\theta_1 = \frac{n^{\beta/2}t}{4A\sqrt{\log n}} \geq \frac{C_M\sqrt{\log d_2 log^2 n}}{4A\sqrt{\log n}} = 2(M+2)\log n.$$

Consequently, we have

$$\theta_1 \geq 2(M+2)\log n \geq \sqrt{2(M+2)\log n},$$

as well as

$$\frac{\theta_1}{l_j} \geq \sqrt{2(M+2)\log n},$$

such that

$$P\Big(\sup_{t\in\mathbb{E}_n}\Big|\hat{f}(t)-f(t)\Big| > \theta_1\Big) = P\Big(\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r| > \theta_1\Big) = 0.$$

In both cases, the second term is equivalent. We have

$$\begin{aligned}
P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{t}{2\theta_1}\Big) &= P\Big(\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{nt}{2\theta_1}\Big)\\
&= P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > \frac{nt}{2\theta_1} - P(X_{1k}\in n\mathbb{E}_n)\Big)\\
&\leq P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > \frac{nt}{2\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}}\Big).
\end{aligned}$$

Choosing θ_1 this way guarantees that

$$\frac{nt}{2\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}} = nA\sqrt{\frac{\log n}{n^\beta}} > 0.$$

Then, using Bernstein's inequality, we get

$$\begin{aligned}
&P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{t}{2\theta_1}\Big)\\
&\qquad\leq P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > nA\sqrt{\frac{\log n}{n^\beta}}\Big)\\
&\qquad\leq \exp\Big(-\frac{k_1 n^{2-\beta}\log n}{k_2 n^{1-\beta/2}\sqrt{\log n} + k_3 n^{1-\beta/2}\sqrt{\log n}}\Big),
\end{aligned}$$

where $k_1, k_2, k_3 > 0$ are generic constants independent of n and d_2. Collecting terms finishes the proof. □

Turning back to the first decomposition of the sample covariance, we have

$$P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\left[(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\right]\right|>t\right)$$
$$\leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t,\mathcal{A}_n\right)+P(\mathcal{A}_n^c)$$
$$\leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t\cap\mathcal{A}_n\right)+\frac{1}{2\sqrt{\pi\log(nd_2)}}.$$

Further, we have

$$P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t\cap\mathcal{A}_n\right)$$
$$\leq P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\left|\triangle_i(j,k)\right|>\frac{t}{2}\right)+P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\left|\triangle_i(j,k)\right|>\frac{t}{2}\right)$$
$$+\frac{1}{2\sqrt{\pi\log(nd_2)}}$$
$$\leq P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\left|\triangle_i(j,k)\right|>\frac{t}{2}\right)+\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)$$
$$+\frac{1}{2\sqrt{\pi\log(nd_2)}},$$

where $X_k\in\mathbb{M}_n$ is shorthand notation for $i:X_{ik}\in\mathbb{M}_n$. We derive the bound of the second term in Lemma B.9. Thus, let us continue with the first term, where

$$P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\left|\triangle_i(j,k)\right|>\frac{t}{2}\right)$$
$$\leq d^2P\left(\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{M}_n}\left|\tilde{\triangle}_{r,s}\right|>\frac{t}{2}\right)$$
$$=d^2P\left(\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{M}_n}\left|(\hat{f}(s)-f(s))\right||r|>\frac{t}{2}\right)$$
$$=d^2P\left(\sup_{s\in\mathbb{M}_n}\left|(\hat{f}(s)-f(s))\right|>\frac{t}{2l_j}\right),$$

where clearly $\sup\{1,\ldots,l_j\}=l_j>0$. Define the event

$$\mathbb{B}_n\equiv\{\delta_n\leq\hat{F}_{X_k}(g_j(s))\leq1-\delta_n,\quad k\in[d_2]\}.$$

Now, from the definition of the Winsorized estimator, we observe that

$$
\begin{aligned}
&d^2P\left(\sup_{s\in\mathbb{M}_n}\left|(\hat{f}(s)-f(s))\right|>\frac{t}{2l_j}\right)\\
&\quad\le d^2P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(s)])-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\cap\mathbb{B}_n\right)+P(\mathbb{B}_n^c)\\
&\quad\le d^2P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(\hat{F}_{X_k}(s))-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\right)\\
&\quad+2\exp\left(2\log d-\frac{\sqrt{n}}{8\pi\log n}\right),
\end{aligned}
$$

where the expression for $P(\mathbb{B}_n^c))$ follows directly from Lemma 19 in [24]. Now, define

$$
\begin{aligned}
T_{1n}&\equiv\max\left\{F_{X_k}(g(\sqrt{\beta\log n})),1-\delta_n\right\}\\
T_{2n}&\equiv 1-\min\left\{F_{X_k}(g(-\sqrt{\beta\log n})),\delta_n\right\},
\end{aligned}
$$

where it follows directly that $T_{1n}=T_{1n}=1-\delta_n$. Consequently, we apply the mean value theorem and get

$$
\begin{aligned}
&P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(\hat{F}_{X_k}(s))-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\right)\\
&\qquad\le P\left((\Phi^{-1})'(\max(T_{1n},T_{2n}))\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{2l_j}\right)\\
&\qquad= P\left((\Phi^{-1})'(1-\delta_n)\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{2l_j}\right)\\
&\qquad\le P\left(\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{(\Phi^{-1})'(1-\delta_n)2l_j}\right)\\
&\qquad\le 2\exp\left(-2\frac{nt^2}{4l_j^2[(\Phi^{-1})'(1-\delta_n)]^2}\right),
\end{aligned}
$$

where the last inequality arises from applying the Dvoretzky-Kiefer-Wolfowitz inequality. Now, we have that

$$
\begin{aligned}
(\Phi^{-1})'(1-\delta_n)&=\frac{1}{\phi\big(\Phi^{-1}(1-\delta_n)\big)}\\
&\le\frac{1}{\phi\Big(\sqrt{2\log\frac{1}{\delta_n}}\Big)}=\sqrt{2\pi}\Big(\frac{1}{\delta_n}\Big)=8\pi n^{\beta/2}\sqrt{\beta\log n}.
\end{aligned}
$$

Therefore,

$$d^2 P\bigg(\sup_{s\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat{F}_{X_k}(s))-\Phi^{-1}(F_{X_k}(s))\Big|>\frac{t}{2l_j}\bigg) \leq 2\exp\bigg(2\log d-\frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\bigg).$$

Collecting the remaining terms we have

$$P\bigg(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\bigg) \leq 2\exp\bigg(2\log d-\frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\bigg)+2\exp\Big(2\log d-\frac{\sqrt{n}}{8\pi\log n}\Big).$$

Thus, we have for the first term in the covariance matrix decomposition

$$\begin{aligned}P\bigg(\max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^{n}\Big[(\hat{f}(X_{ik})-f(X_{ik}))X_{ij})\Big]\bigg|>t\bigg) &\leq P\bigg(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\bigg)\\ &+\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)+\frac{1}{2\sqrt{\pi\log(nd_2)}}\\ &\leq 2\exp\bigg(2\log d-\frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\bigg)+2\exp\Big(2\log d-\frac{\sqrt{n}}{8\pi\log n}\Big)\\ &+\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)+\frac{1}{2\sqrt{\pi\log(nd_2)}}.\end{aligned}$$

Let us turn back to the second term of the first sample covariance decomposition, i.e.

$$\begin{aligned}P\bigg(\max_{j,k}\Big|(\mu_n(\hat{f})-\mu_n(f))\mu_n(X_j)\Big|>t\bigg) &\leq P\bigg(\max_{k}\bigg|\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}(X_{ik})-f(X_{ik})\Big)\bigg|>\frac{t}{l_{\max}}\cap\mathcal{A}_n\bigg)+\frac{1}{2\sqrt{\pi\log(nd_2)}}.\end{aligned}$$

Now, analogous to before, we find

$$P\Big(\max_{k}\frac{1}{n}\sum_{i=1}^{n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{l_{\max}}\cap\mathcal{A}_n\Big)$$

$$
\begin{aligned}
&\le P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\Big|\hat f(X_{ik})-f(X_{ik})\Big|>\frac{t}{2l_{\max}}\Big)\\
&+P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\Big|\hat f(X_{ik})-f(X_{ik})\Big|>\frac{t}{2l_{\max}}\Big)\\
&\le P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\Big|\hat f(X_{ik})-f(X_{ik})\Big|>\frac{t}{2l_{\max}}\Big)\\
&+\exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)\\
&\le d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat f(t)-f(t)\Big|>\frac{t}{2l_{\max}}\Big)\\
&+\exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big).
\end{aligned}
\tag{45}
$$

Let us take a closer look at

$$
\begin{aligned}
&d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat f(t)-f(t)\Big|>\frac{t}{2l_{\max}}\Big)\\
&\qquad\le d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(W_{\delta_n}[\hat F_{X_k}(t)])-\Phi^{-1}(F_{X_k}(t))\Big|>\frac{t}{2l_{\max}}\cap\mathbb{B}_n\right)\\
&\qquad+d_2 P(\mathbb{B}_n^c)\\
&\qquad\le d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat F_{X_k}(t))-\Phi^{-1}(F_{X_k}(t))\Big|>\frac{t}{2l_{\max}}\right)\\
&\qquad+2\exp\Big(\log d_2-\frac{\sqrt{n}}{8\pi\log n}\Big).
\end{aligned}
$$

The definition of the event $\mathbb{B}_n$ is the same as above. Then applying once more the Dvoretzky–Kiefer–Wolfowitz inequality we end up with the following upper bound:

$$
\begin{aligned}
&d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat F_{X_k}(t))-\Phi^{-1}(F_{X_k}(t))\Big|>\frac{t}{2l_{\max}}\right)\\
&\qquad\qquad\le 2\exp\Big(\log d_2-\frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2\log n}\Big).
\end{aligned}
$$

Collecting terms and simplifying yields

$$
P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\Big[(\hat f(X_{ik})-f(X_{ik}))X_{ij}\Big)\Big]\right|>t\right)
$$

$$
\begin{aligned}
&+ P\bigg(\max_{j,k} \Big| (\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j) \Big| > t \bigg) \\
&\leq 2\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_j^2\ \pi^2 \log n} \right) + 2\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\quad + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{2\sqrt{\pi \log(nd_2)}} \\
&\quad + 2\exp\left(\log d_2 - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\quad + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{2\sqrt{\pi \log(nd_2)}} \\
&\leq 4\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 4\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\quad + 2\exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{\sqrt{\pi \log(nd_2)}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
&P\bigg(\max_{j,k} \Big| S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j} \Big| > 2t \bigg) \\
&\qquad \leq 4\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 4\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\qquad\quad + 2\exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{\sqrt{\pi \log(nd_2)}},
\end{aligned}
\tag{46}
$$

which completes the considerations regarding the sample covariance.

As a next step, we need to bound the error of the sample standard deviation of the Winsorized estimator

$$
\sigma_{\hat{f}(X_k)}^{(n)} \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Big(\hat{f}(X_{ik}) - \mu_n(\hat{f}) \Big)^2}.
$$

Consider the following decomposition of the standard deviation of the Winsorized estimator,

$$
\Big| \sigma_{\hat{f}(X_k)}^{(n)} - \sigma_{f(X_k)}^{(n)} \Big|
$$

$$
\begin{aligned}
&= \left| \sqrt{1/n \sum_{i=1}^{n} \left(\hat{f}(X_{ik}) - \mu_n(\hat{f}) \right)^2} - \sqrt{1/n \sum_{i=1}^{n} \left(f(X_{ik}) - \mu_n(f) \right)^2} \right| \\
&= \frac{1}{\sqrt{n}} \left| \left(\|\hat{f}(X_k) - \mu_n(\hat{f})\|_2 - \|f(X_k) - \mu_n(f)\|_2 \right) \right| \\
&\leq \frac{1}{\sqrt{n}} \|\hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f)\|_2 \\
&\leq \frac{1}{\sqrt{n}} \sqrt{n} \|\hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f)\|_\infty \\
&= \sup_{i:X_{ik} \in \{1,\dots,n\}} \left| \hat{f}(X_{ik}) - f(X_{ik}) + \mu_n(f) - \mu_n(\hat{f}) \right| \\
&= \sup_{i:X_{ik} \in \{1,\dots,n\}} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| + \left| \mu_n(\hat{f}) - \mu_n(f) \right| \\
&\leq \sup_{i:X_{ik} \in \{1,\dots,n\}} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| + \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right|.
\end{aligned}
$$

The first inequality is due to the reverse triangle inequality. The ensuing inequalities arise from applying standard norm equivalences. As before, we analyze both terms separately since we have to take care of the behavior of the Winsorized estimator, taking values at the end and the middle interval. We have for any $t > 0$,

$$
\begin{aligned}
&P\left(\max_k \left| \sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)} \right| > 2t \right) \\
&\qquad \leq P\Big(\max_k \sup_{i:X_{ik} \in \{1,\dots,n\}} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| > t, \mathcal{A}_n \Big) \\
&\qquad + P\Big(\max_k \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| > t, \mathcal{A}_n \Big) + P(\mathcal{A}_n^c).
\end{aligned}
$$

Note that the second term is, in effect, equivalent to Eq. (45) above such that we can immediately conclude that

$$
\begin{aligned}
&P\Big(\max_k \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ik}) - f(X_{ik}) \right| > t \cap \mathcal{A}_n \Big) \\
&\qquad \leq d_2 P\Big(\sup_{t \in \mathbb{M}_n} \left| \hat{f}(t) - f(t) \right| > \frac{t}{2} \Big) + \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big) \\
&\qquad \leq 2 \exp\Big(\log d_2 - \frac{\sqrt{n} t^2}{64 \pi^2 \log n} \Big) + 2 \exp\Big(\log d_2 - \frac{\sqrt{n}}{8 \pi \log n} \Big) \\
&\qquad + \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big).
\end{aligned}
$$

The bound for the end region follows again from Lemma B.9.

Similarly, we find that

$$\begin{aligned}
&P\Big(\max_k \sup_{i:X_{ik}\in\{1,\dots,n\}} \Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > t \cap \mathcal{A}_n\Big) \\
&\qquad \le d_2 P\Big(\sup_{t\in\mathbb{M}_n} \Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) + d_2 P\Big(\sup_{t\in\mathbb{E}_n} \Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) \\
&\qquad = d_2 P\Big(\sup_{t\in\mathbb{M}_n} \Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big),
\end{aligned}$$

where the bound over the end region has been shown in Lemma B.9. Thus, we only have to take care of

$$\begin{aligned}
&P\Big(\sup_{t\in\mathbb{M}_n} \Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) \\
&\le P\left(\sup_{t\in\mathbb{M}_n} \Big|\Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(t)]) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2} \cap \mathbb{B}_n\right) + P(\mathbb{B}_n^c) \\
&\le P\left(\sup_{t\in\mathbb{M}_n} \Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2}\right) + 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi\log n}\Big).
\end{aligned}$$

The definition of the event $\mathbb{B}_n$ is the same as above. Then again, by the Dvoretzky-Kiefer-Wolfowitz inequality, we end up with the following upper bound:

$$P\left(\sup_{t\in\mathbb{M}_n} \Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2}\right) \le 2\exp\Big(-\frac{\sqrt{n}t^2}{64\pi^2\log n}\Big).$$

Collecting terms, we arrive at the concentration bound for the sample standard deviation:

$$\begin{aligned}
&P\left(\max_k \Big|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)}\Big| > 2t\right) \\
&\qquad \le 4\exp\Big(\log d_2 - \frac{\sqrt{n}t^2}{64\pi^2\log n}\Big) + 4\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi\log n}\Big) \\
&\qquad + \exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3}\Big) + \frac{1}{2\sqrt{\pi\log(nd_2)}}.
\end{aligned}$$

With these intermediate results, we have shown that the sample covariance (numerator) and the sample standard deviation (denominator) can be estimated accurately.

The following lemma provides us with the means to form a probability bound for

$$\max_{jk}\Big|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\Big|.$$

Lemma B.10. *Consider the polyserial ad hoc estimator* $\hat{\Sigma}^{(n)}_{jk}$ *for* $1 \leq j \leq d_1 < k \leq d_2$ *and let* $\epsilon \in \left[C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}, 8(1+4c^2)\right]$, *where* c *is the corresponding sub-Gaussian parameter of the discrete variable. Both the numerator and the denominator are bounded, i.e.*

$$S_{\hat{f}(X_k)X_j} \in [-(1+\epsilon), 1+\epsilon],$$

and

$$\sigma^{(n)}_{\hat{f}(X_k)} \in [1-\epsilon, 1+\epsilon].$$

Consequently, $\hat{\Sigma}^{(n)}_{jk}$ *is Lipschitz with constant* L. *The following decomposition holds*

$$\begin{aligned}
\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^{*}_{jk}\right| &= \max_{jk}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^{(n)}_{jk} + \Sigma^{(n)}_{jk} - \Sigma^{*}_{jk}\right| \\
&\leq \max_{jk}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^{(n)}_{jk}\right| + \max_{jk}\left|\Sigma^{(n)}_{jk} - \Sigma^{*}_{jk}\right| \\
&\leq L\Big(\max_{jk}\left|S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j}\right| + C_\Gamma \max_{k}\left|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{\hat{f}(X_k)}\right| \\
&+ \max_{jk}\left|S_{f(X_k)X_j} - S^{*}_{f(X_k)X_j}\right| + C_\Gamma \max_{k}\left|\sigma^{(n)}_{f(X_k)} - 1\right|\Big),
\end{aligned}$$

where $C_\Gamma \equiv \sum_{r=1}^{l_j} \phi(\bar{\Gamma}^r_j)(x_j^{r+1} - x_j^r)$.

Proof. Let us assume w.l.o.g. that X_j – the discrete variable – has mean zero and variance one. By the Cauchy-Schwarz inequality, the true covariance of the pair is bounded from above by 1, i.e.

$$\left|S^{*}_{f(X_k)X_j}\right| \leq \sigma^2_{f(X_k)}\sigma^2_{X_j} = 1.$$

Earlier we have shown that for some $t \geq C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$ we have

$$\begin{aligned}
&P\left(\max_{j,k}\left|S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j}\right| > 2t\right) \\
&\quad\leq 4\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l^2_{\max}\ \pi^2 \log n}\right) + 4\exp\left(2\log d - \frac{\sqrt{n}}{8\pi\log n}\right) \\
&\quad + 2\exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3}\Big) + \frac{1}{\sqrt{\pi\log(nd_2)}}.
\end{aligned} \tag{47}$$

According to Lemma 1 in [37] with sub-Gaussian parameter $c > 0$, $f(X_k)$ is standard Gaussian and thus sub-Gaussian and X_j is discrete and bounded and therefore also sub-Gaussian we have the following tail bound

$$P\left(\max_{jk}\left|S_{f(X_k)X_j} - S^{*}_{f(X_k)X_j}\right| \geq t\right) \leq 4d^2\exp\left\{-\frac{nt^2}{128(1+4c^2)^2}\right\},$$

for all $t \in (0, 8(1+4c^2))$. Therefore with high probability for $j \in [d_1], k \in [d_2]$, we have

$$S_{\hat{f}(X_k)X_j} \in [S_{f(X_k)X_j} - 2t, S_{f(X_k)X_j} + 2t],$$

and since

$$S_{f(X_k)X_j} \in [-1-t, 1+t],$$

with high probability, we have

$$S_{\hat{f}(X_k)X_j} \in [-(1+3t), 1+3t].$$

Similar considerations hold for the sample standard deviation. We have already shown that

$$\begin{aligned} P\left(\max_k \left|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)}\right| > 2t\right) \\ \leq 4\exp\left(\log d_2 - \frac{\sqrt{n}t^2}{64\pi^2 \log n}\right) + 4\exp\left(\log d_2 - \frac{\sqrt{n}}{8\pi \log n}\right) \\ + \exp\left(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3}\right) + \frac{1}{2\sqrt{\pi \log(nd_2)}}. \end{aligned}$$

Furthermore, we use Lemma 1 again in [37] to form a bound for the variance. Since $f(X_k)$ is standard Gaussian and hence sub-Gaussian with parameter $c = 1$ we immediately get

$$P\left(\max_k \left|(\sigma^{(n)}_{f(X_k)})^2 - 1\right| \geq t\right) \leq 4d_2 \exp\left\{-\frac{nt^2}{128(1+4)^2}\right\}.$$

Put differently, with high probability

$$(\sigma^{(n)}_{f(X_k)})^2 \in [1-t, 1+t],$$

and consequently, we also have with high probability

$$\sigma^{(n)}_{f(X_k)} \in [\sqrt{1-t}, \sqrt{1+t}].$$

Since the interval $[1-t, 1+t]$ is always at least as wide as $[\sqrt{1-t}, \sqrt{1+t}]$, for all $t > 0$ with high probability we then also have

$$\sigma^{(n)}_{f(X_k)} \in [1-t, 1+t].$$

Putting these things together, we obtain that with high probability

$$\sigma^{(n)}_{\hat{f}(X_k)} \in [1-3t, 1+3t].$$

In order to finish the proof consider the following function $h : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ defined by

$$h(u, v) = \frac{u}{v},$$

with $\nabla h = (1/v, -u/v^2)^T$.

As we have just shown for the polyserial ad hoc estimator, $\nabla h = (1/v, -u/v^2)^T$ is bounded with

$$\sup\|\nabla h\|_2 = \sqrt{\left(\frac{1}{1-3t}\right)^2 + \left(-\frac{(1+3t)}{(1-3t)^2}\right)^2} \coloneqq L.$$

Consequently, h is Lipschitz, and we have the following decomposition

$$\begin{aligned}\big|h(u,v) - h(u',v')\big| &= \big|h(u,v) - h(\tilde{u},\tilde{v}) + h(\tilde{u},\tilde{v}) - h(u',v')\big| \\ &\leq \big|h(u,v) - h(\tilde{u},\tilde{v})\big| + \big|h(\tilde{u},\tilde{v}) - h(u',v')\big| \\ &\leq L\big(|u-\tilde{u}| + |v-\tilde{v}|\big) + L\big(\big|\tilde{u}-u'\big| + \big|\tilde{v}-v'\big|\big).\end{aligned}$$

Finally, taking $\epsilon = 3t$ finishes the proof. □

At last, collecting terms, we find that for $j \in [d_1]$ and $k[d_2]$ and any

$$\epsilon \in \left[C_M\sqrt{\frac{\log d \log^2 n}{\sqrt{n}}}, 8(1+4c^2)\right]$$

the following bound holds

$$\begin{aligned}&P\left(\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\right| \geq \epsilon\right) \\ &\quad\leq P\left(\max_{j,k}\left|S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j}\right| > \frac{\epsilon}{4L}\right) \\ &\qquad + P\left(\max_{k}\left|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)}\right| > \frac{\epsilon}{4LC_\Gamma}\right) \\ &\qquad + P\left(\max_{jk}\left|S_{f(X_k)X_j} - S^*_{f(X_k)X_j}\right| \geq \frac{\epsilon}{4L}\right) \\ &\qquad + P\left(\max_{k}\left|(\sigma^{(n)}_{f(X_k)})^2 - 1\right| \geq \frac{\epsilon}{4lC_\Gamma}\right)\end{aligned}$$

The conclusion of Theorem 3.3 follows by plugging in the corresponding concentration bounds and simplifying.

Appendix C: Additional simulation setup and results

C.1. FPR and TPR results for binary and general mixed data

This section provides additional simulation results for the binary-continuous and general mixed data setting. In particular, we report TPR and FPR for the latent Gaussian model and the LGCM with $f_j(x) = x^{1/3}$ for all j. The results are depicted in Figures 5 and 6 for the binary-continuous and general mixed data settings, respectively.

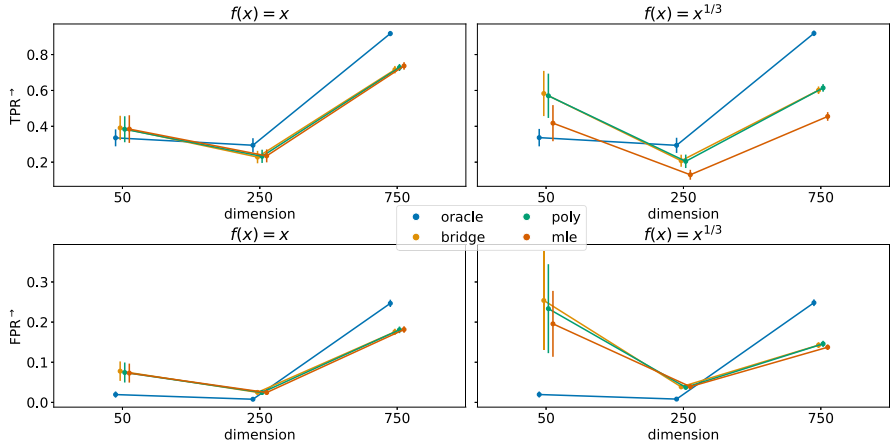


FIG 5. *Simulation results for the binary-continuous data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top and bottom rows report mean and standard deviation of the TPR and FPR along simulation runs, respectively. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better.*

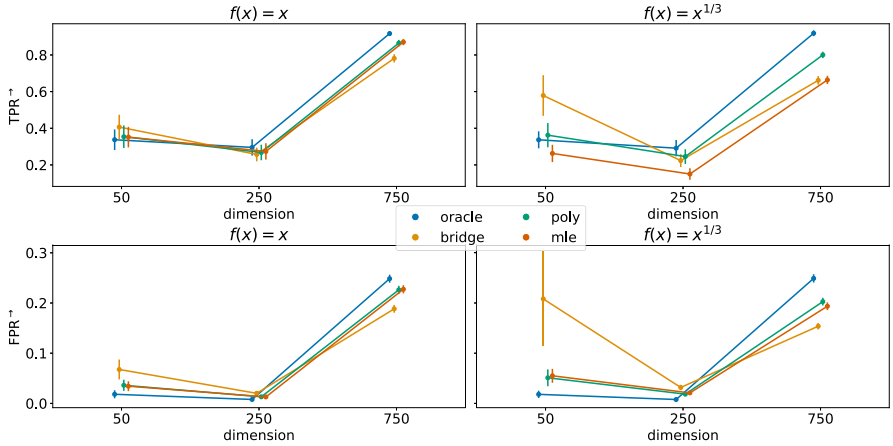


FIG 6. *Simulation results for the general mixed data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top and bottom rows report mean and standard deviation of the TPR and FPR along simulation runs, respectively. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better.*

Figure 5 illustrates that our `poly` estimator performs almost identical to the gold standard `bridge` estimator proposed by Fan et al. [12]. In the right column, the `mle` estimator is misspecified, and TPR is noticeably lower than

for the remaining estimators. Interestingly, it appears that the misspecified `mle` estimator does not infer additional edges, which is reflected in the FPR holding level with the other estimators.

The general mixed data setting results reported in Figure 6 are similar. The `poly` estimator performs almost identically to the `mle` estimator under the latent Gaussian model in terms of FPR and TPR. The `bridge` estimator has higher FPR and TPR in the $d = 50$ case and lower FPR and TPR in the $d = 750$ case. Under the LGCM, similar to the binary-continuous data setting, the `mle` estimator is misspecified and has a lower TPR than the other estimators while the FPR holds level. The `bridge` ensemble estimator has an even higher FPR and TPR in the $d = 50$ case and a slightly lower FPR and TPR in the $d = 750$ case.

C.2. Ternary mixed data results

We additionally compare our `poly` and `mle` estimators with a generalization of the bridge function approach proposed by Quan, Booth and Wells [35] given a mix of ternary, binary, and continuous data. In this case, $\binom{2+2}{2} = 6$ individual bridge functions are needed.

The (d, n)-setup is analogous to the binary-continuous data setting from the main text. Starting with `mle`, the pattern from the binary-continuous setting continues to show in the ternary-binary-continuous mix. Whenever $f(x) = x$, the `mle` estimator generally performs best, in particular concerning graph recovery.

Table 1: Ternary mixed data structure learning; Simulated data with 100 simulation runs. Standard errors in brackets.

$d, n, f(x)$		Oracle $\hat{\Omega}$	ternary $\hat{\Omega}_\tau$	$\hat{\Omega}_{\text{MLE}}$	$\hat{\Omega}_r$
$50, 200, x$	$\|\hat{\Omega} - \Omega\|_F$	2.860 (0.098)	2.936 (0.105)	2.935 (0.106)	2.930 (0.109)
	FPR	0.016 (0.005)	0.067 (0.017)	0.071 (0.021)	0.075 (0.023)
	TPR	0.340 (0.046)	0.370 (0.061)	0.381 (0.068)	0.389 (0.070)
	AUC	0.880 (0.013)	0.758 (0.019)	0.769 (0.019)	0.764 (0.020)
$50, 200, x^3$	$\|\hat{\Omega} - \Omega\|_F$	2.856 (0.116)	2.942 (0.102)	3.053 (0.098)	2.935 (0.108)
	FPR	0.016 (0.007)	0.068 (0.019)	0.076 (0.020)	0.075 (0.022)
	TPR	0.342 (0.051)	0.372 (0.059)	0.280 (0.051)	0.391 (0.066)
	AUC	0.882 (0.015)	0.759 (0.019)	0.691 (0.020)	0.768 (0.019)

$250, 200, x$	$\|\hat{\Omega} - \Omega\|_F$	3.185 (0.097)	3.742 (0.090)	3.709 (0.089)	3.711 (0.091)
	FPR	0.006 (0.001)	0.025 (0.003)	0.024 (0.003)	0.025 (0.003)
	TPR	0.308 (0.034)	0.238 (0.033)	0.237 (0.031)	0.235 (0.030)
	AUC	0.884 (0.014)	0.759 (0.018)	0.773 (0.018)	0.768 (0.018)
$250, 200, x^3$	$\|\hat{\Omega} - \Omega\|_F$	3.199 (0.096)	3.757 (0.096)	3.894 (0.096)	3.724 (0.087)
	FPR	0.006 (0.001)	0.025 (0.003)	0.026 (0.003)	0.025 (0.003)
	TPR	0.302 (0.034)	0.239 (0.032)	0.143 (0.027)	0.237 (0.032)
	AUC	0.882 (0.012)	0.759 (0.016)	0.691 (0.016)	0.767 (0.015)
$750, 300, x$	$\|\hat{\Omega} - \Omega\|_F$	11.181 (0.134)	10.830 (0.129)	10.640 (0.122)	10.659 (0.118)
	FPR	0.256 (0.006)	0.179 (0.006)	0.180 (0.006)	0.179 (0.005)
	TPR	0.937 (0.009)	0.723 (0.016)	0.744 (0.017)	0.736 (0.016)
	AUC	0.939 (0.006)	0.820 (0.009)	0.831 (0.009)	0.828 (0.009)
$750, 300, x^3$	$\|\hat{\Omega} - \Omega\|_F$	11.196 (0.130)	10.838 (0.129)	11.250 (0.130)	10.646 (0.137)
	FPR	0.256 (0.006)	0.180 (0.006)	0.173 (0.006)	0.179 (0.006)
	TPR	0.937 (0.009)	0.724 (0.016)	0.590 (0.020)	0.737 (0.016)
	AUC	0.939 (0.006)	0.820 (0.008)	0.743 (0.011)	0.828 (0.009)

However, when $f_j(x) = x^3$, performance drops notably, which is driven not by an increased FPR but by a decreased TPR. Again, results for `bridge` and `poly` behave similarly across metrics. No performance reduction, neither in estimation error nor in graph recovery, can be detected when comparing `poly` to the ternary `bridge` estimator.

Appendix D: Application to COVID-19 data

This section presents the results of an analysis of real-world health data (from the UK Biobank). We are interested in investigating associations between the

severity of a COVID-19 infection and various potential risk factors. This analysis is intended to illustrate the use of the proposed methods in a real-world, mixed variable type example.

TABLE 2. *Estimated partial correlations between COVID-19 severity and the listed variables for data sets A, B and C.*

Covid-19 severity assoc. Variables	Data set A	Data set B	Data set C
age	0.162	0.134	0.140
waist circ.	0.031	0.009	0.011
deprev. idx	0.016	–	–
sex	0.007	–	–
hypertension	0.075	0.035	0.037
heart attack	–	0.073	0.065
diabetes	–	0.062	0.055
chr. bronch.	–	–	0.012
wisd. teeth surg.	–	−0.003	–

D.1. Data set and variables

We first describe the data set used here, which is a part of the UK Biobank COVID-19 resource in which UK Biobank data were linked to clinical COVID data. To construct an indicator of COVID-19 severity, we consider subjects who tested positive for COVID-19 at some point in 2020. Based on that, we created an indicator variable (COVID severity) to capture whether each subject had a severe outcome within six weeks of infection (meaning either hospitalized, hospitalized, receiving critical care, or died). Around 14% experienced such a severe outcome. The analysis includes $n = 8672$ observations on $d = 712$ variables (risk factors and covariates with less than 40% missingness). Missing values were imputed using `missForest` R-package using default settings. Variables expressing more than 20 states were treated as continuous. The remaining data include 665 binary variables, 25 count variables, and 8 categorical variables. Many of the binary variables represent the status for relatively rare conditions. This means that the share of the minority class of these indicators (i.e., the fraction of samples with the least frequent value of the variable) can often be minimal. To understand the effects of such rare events on the analysis, we define three data sets (named A, B, and C) with inclusion rules requiring respectively at least a $25\%, 2\%, 1\%$ share of observations falling into the minority class.

D.2. Results

We present the results of a joint analysis of the variables using real UK Biobank data. We emphasize that the analysis aims to illustrate the proposed estimators' behavior and not fully understand the risk factors for severe COVID-19. There has been much work done on factors influencing the risk of severe COVID-19

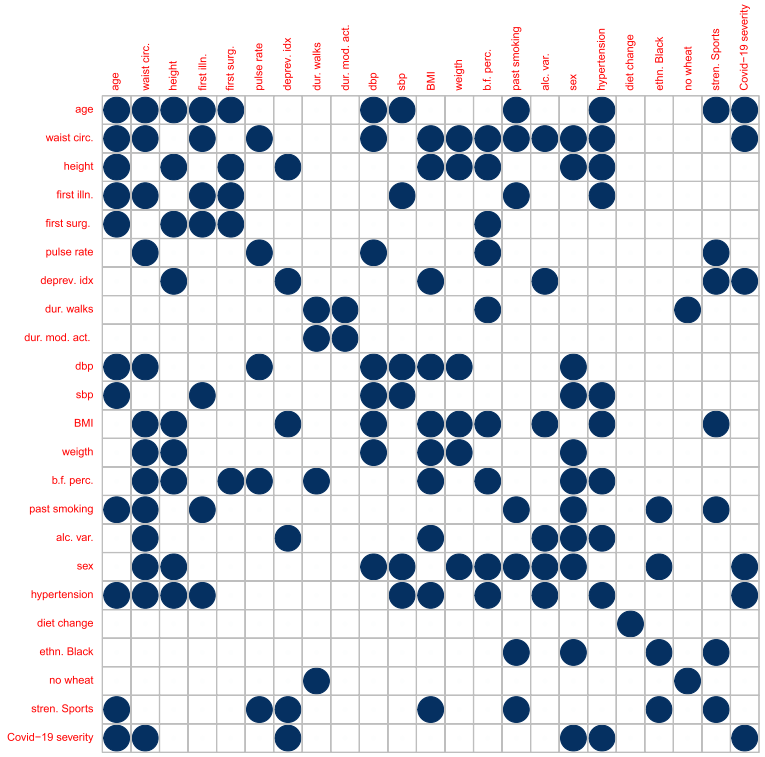


FIG 7. *Plot of the estimated adjacency matrix of data set A.*

and its treatment [see, among others, 45, 5] and we direct the interested reader to the references for further information.

Table 2 gives a summary of the estimated links (indicated as a visualization of the partial correlations) between the variables (including COVID-19 severity). Considering in particular links to COVID-19 severity, we see that `age`, `waist circ.`, `hypertension`, `heart attack` and `diabetes` are quite stable links throughout the different data sets. The effect sizes in terms of partial correlations are penalized and should be interpreted in relative terms. In particular, `age` retains a relatively large signal, which is in line with the known strong influence of age on COVID-19 severity [see, e.g. 45].

Finally, we present more detailed results of the analysis of data set A. Figure 7 shows the estimated adjacency matrix and Figure 8 depicts the estimated precision matrix $\hat{\mathbf{\Omega}}_A$. These results highlight the type of output, spanning different kinds of variables, that is readily available from the proposed method.

D.3. Variable description for real-world data application

Table 3 gives an overview of the variables in the UK Biobank data set.

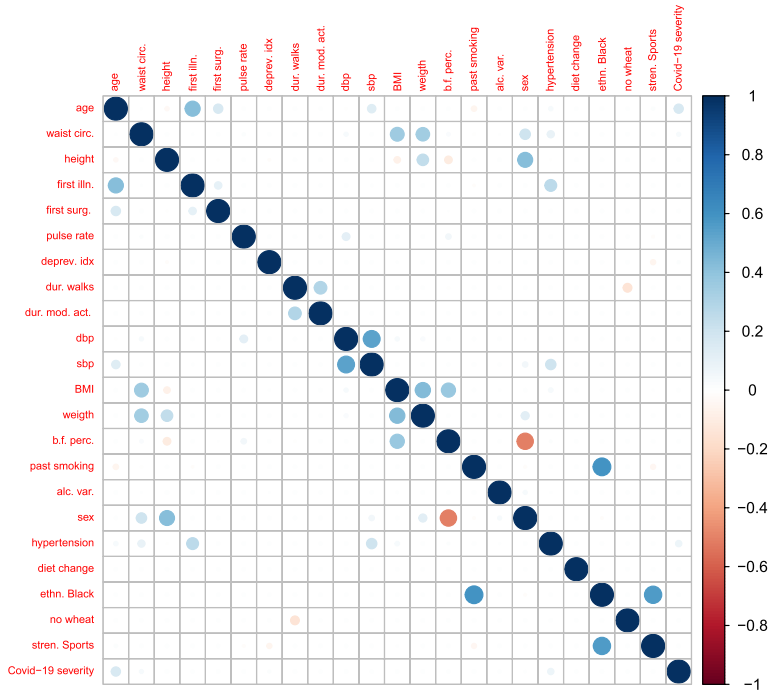


FIG 8. *Plot of the estimated precision matrix of data set A.*

Table 3: Variable description of the real world application.

Variable Name	Description
age	age in. years in 2020
waist circ.	waist circumference in cm
height	standing in height in cm
first illn.	age at which illness first occurred
first surg.	age at which operation was done first
pulse rate	pulse rate measured in bpm
deprev. idx	Townsend deprivation index at recruitment
dur. walks	duration of walks in minutes per day
dur. mod. act.	duration of moderate activity in minutes per day
dbp	diastolic blood pressure in mmHg
sbp	systolic blood pressure in mmHg
BMI	in kg/m2
weigth	in kg
b.f. perc.	body fat percentage in %
walking	number of days per week walked 10+ minutes

mod. phys. act.	number of days per week of moderate physical activity 10+ minutes
vig. phys. act.	number of days per week of vigorous physical activity 10+ minutes
cheese	answer to "How often do you eat cheese per week?"
stair climb.	answer to "At home, during the last 4 weeks, about how many times a DAY do you climb a flight of stairs? (approx 10 steps)"
curr. smoking	categorial, "Do you smoke tobacco now?" (yes, no, occasionally)
past smoking	categorial, "How often did you smoke tobacco?" (never, once/twice, occasionally, on most days)
diet var.	categorial, "Does your diet change?" (never, sometimes, often)
alc. freq.	categorial, "How often do you drink alcohol?" (never, special occasions only, 1-3 per month, 1-2 per week, 3-4 per week, almost daily)
alc. var.	categorial, "Compared to 10 years ago, do you drink?" (more, about the same, less)
sex	binary indicator with 0=female, 1=male
hypertension	hypertension, binary indicator with 0=no, 1=yes
angina	angina, binary indicator with 0=no, 1=yes
heart attack	heart attack, binary indicator with 0=no, 1=yes
stroke	stroke, binary indicator with 0=no, 1=yes
dvt	deep venous thrombosis, binary indicator with 0=no, 1=yes
asthma	asthma, binary indicator with 0=no, 1=yes
chr. bronch.	emphysema/chronic bronchitis, binary indicator with 0=no, 1=yes
gord	gastro-oesophageal reflux/gastric reflux, binary indicator with 0=no, 1=yes
ibs	irritable bowel syndrome, binary indicator with 0=no, 1=yes
gall stones	cholelithiasis/gall stones, binary indicator with 0=no, 1=yes
kidn./bladder stone	kidney stone/ureter stone/bladder stone, binary indicator with 0=no, 1=yes
diabetes	diabetes, binary indicator with 0=no, 1=yes
diabtes 2	type 2 diabetes, binary indicator with 0=no, 1=yes
myxoedema	hypothyroidism/myxoedema, binary indicator with 0=no, 1=yes
migraine	migraine, binary indicator with 0=no, 1=yes
glaucoma	glaucoma, binary indicator with 0=no, 1=yes
cataract	cataract, binary indicator with 0=no, 1=yes
depression	depression, binary indicator with 0=no, 1=yes

panic attacks	anxiety/panic attacks, binary indicator with 0=no, 1=yes
back probl.	back problems, binary indicator with 0=no, 1=yes
osteoporosis	osteoporosis, binary indicator with 0=no, 1=yes
spine arthr.	spine arthritis/spondylitis, binary indicator with 0=no, 1=yes
slipped disc	prolapsed disc/slipped disc, binary indicator with 0=no, 1=yes
anaemia	iron deficiency anaemia, binary indicator with 0=no, 1=yes
ut. fibroids	uterine fibroids, binary indicator with 0=no, 1=yes
allerg. rhinitis	heyfever/allergic rhinitis, binary indicator with 0=no, 1=yes
enlarged prost.	enlarged prostate, binary indicator with 0=no, 1=yes
pneumonia	pneumonia, binary indicator with 0=no, 1=yes
endometr.	endometriosis, binary indicator with 0=no, 1=yes
ear disor.	ear/vestibular disorder, binary indicator with 0=no, 1=yes
headaches	headaches (not migraine), binary indicator with 0=no, 1=yes
ecz./dermat.	eczema/dermatitis, binary indicator with 0=no, 1=yes
psoriasis	psoriasis, binary indicator with 0=no, 1=yes
div. disease	diverticular disease/diverticulitis, binary indicator with 0=no, 1=yes
osteoarthr.	osteoarthritis, binary indicator with 0=no, 1=yes
gout	gout, binary indicator with 0=no, 1=yes
high chol.	high cholesterol, binary indicator with 0=no, 1=yes
hiat. hern.	hiatus hernia, binary indicator with 0=no, 1=yes
sciatica	sciatica, binary indicator with 0=no, 1=yes
appendic.	appendicitis, binary indicator with 0=no, 1=yes
back pain	back pain, binary indicator with 0=no, 1=yes
arthritis	arthritis (nos), binary indicator with 0=no, 1=yes
measles	measles/morbillivirus, binary indicator with 0=no, 1=yes
chickpox	chickenpox, binary indicator with 0=no, 1=yes
tonsillitis	tonsillitis, binary indicator with 0=no, 1=yes
ptca	coronary angioplasty (ptca)+/-stent, binary indicator with 0=no, 1=yes
ear surg.	ear surgery, binary indicator with 0=no, 1=yes
sinus surg.	nasal/sinus,nose surgery, binary indicator with 0=no, 1=yes
vasectomy	vasectomy, binary indicator with 0=no, 1=yes
soft tiss. surg.	mucsle/soft tissue surgery, binary indicator with 0=no, 1=yes

hip repl.	hip replacement/revision, binary indicator with 0=no, 1=yes
knee repl.	knee replacement/revision, binary indicator with 0=no, 1=yes
spine surg.	spine or back surgery, binary indicator with 0=no, 1=yes
bil. ooph.	bilateral oophorectomy, binary indicator with 0=no, 1=yes
hysterect.	hysterectomy, binary indicator with 0=no, 1=yes
steril.	sterilisation, binary indicator with 0=no, 1=yes
lumpect.	lumpectomy, binary indicator with 0=no, 1=yes
ing. hernia rep.	inguinal/femoral hernia repair, binary indicator with 0=no, 1=yes
umb. hernia rep.	umbilical hernia repair, binary indicator with 0=no, 1=yes
cataract extr.	catarct extraction/lens implant, binary indicator with 0=no, 1=yes
red./fix. bone frac.	reduction or fixationof bone fracture, binary indicator with 0=no, 1=yes
cholecystect.	cholecystectomy/gall bladder removal, binary indicator with 0=no, 1=yes
appendicect.	appendicectomy, binary indicator with 0=no, 1=yes
c-sec.	caesarian section, binary indicator with 0=no, 1=yes
tonsillest.	tonsillectomy, binary indicator with 0=no, 1=yes
var. vein surg.	varicose vein surgery, binary indicator with 0=no, 1=yes
wisd. teeth surg.	wisdom teeth surgery, binary indicator with 0=no, 1=yes
piles surg.	haemorroidectomy/piles surgery/banding of piles, binary indicator with 0=no, 1=yes
male circ.	male circumcision, binary indicator with 0=no, 1=yes
squint corr.	squint correction, binary indicator with 0=no, 1=yes
arthrosc.	arthroscopy (nos), binary indicator with 0=no, 1=yes
foot surg.	foot surgery, binary indicator with 0=no, 1=yes
knee surg.	knee surgery (not replacement), binary indicator with 0=no, 1=yes
shoulder surg.	shoulder surgery, binary indicator with 0=no, 1=yes
car. tunn. surg.	carpal tunnel surgery, binary indicator with 0=no, 1=yes
valg. surg.	bunion/hallus valgus surgery, binary indicator with 0=no, 1=yes
rem. mole	removal of mole/skin lesion, binary indicator with 0=no, 1=yes

ov. cyst. rem.	ovarian cyst removal/surgery, binary indicator with 0=no, 1=yes
d+c	dilatation and curettage, binary indicator with 0=no, 1=yes
cone biops.	cone biopsy, binary indicator with 0=no, 1=yes
endosc.	endoscopy/gastroscopy, binary indicator with 0=no, 1=yes
colonosc.	colonoscopy/sigmoidoscopy, binary indicator with 0=no, 1=yes
laparosc.	laparoscopy, binary indicator with 0=no, 1=yes
rhinoplast.	rhinoplasty/nose surgery, binary indicator with 0=no, 1=yes
tonsil surg.	tonsillectomy/tonsil surgery, binary indicator with 0=no, 1=yes
ing. hern. rep.	inguinal hernia repair, binary indicator with 0=no, 1=yes
illn. ind. diet	Major dietary changes in the last 5 years because of illness, binary indicator with 0=no, 1=yes
diet change	Major dietary changes in the last 5 years because of other reason, binary indicator with 0=no, 1=yes
ethn. Mixed	Ethnicity – mixed, binary indicator with 0=no, 1=yes
ethn. Asian	Ethnicity – Asian, binary indicator with 0=no, 1=yes
ethn. Black	Ethnicity – Black, binary indicator with 0=no, 1=yes
no eggs	Never eat eggs or foods containing eggs, binary indicator with 0=no, 1=yes
no dairy	Never dairy products, binary indicator with 0=no, 1=yes
no wheat	Never eat wheat, binary indicator with 0=no, 1=yes
no sugar	Never eat sugar or foods/drinks containing sugar, binary indicator with 0=no, 1=yes
walk. f. pleas.	Types of physical activity in last 4 weeks – walking for pleasure, binary indicator with 0=no, 1=yes
exercises	Types of physical activity in last 4 weeks – other exercises (swimming, bowling etc.), binary indicator with 0=no, 1=yes
stren. Sports	Types of physical activity in last 4 weeks – strenuous sports, binary indicator with 0=no, 1=yes
Covid-19 severity	Covid-19 severity, binary indicator with 0=mild outcome and 1=severe outcome

Acknowledgments

We thank the anonymous reviewers whose insightful comments and constructive feedback significantly enhanced the quality and clarity of this paper. We further

want to thank Hongjian Shi for his helpful insights on the subject. Empirical results include work conducted using the UK Biobank Resource.

Conflict of interest

None declared.

Funding

This work was partly supported by the Helmholtz AI project "Scalable and Interpretable Models for Complex And Structured Data" (SIMCARD), the UK Medical Research Council (MC-UU-00002/17) and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust).

This project also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant agreement No 883818].

References

[1] Anne, G.-P., Aurélie, G.-M. and Clémence, K. (2019). Graph estimation for Gaussian data zero-inflated by double truncation. arXiv:1911.07694.

[2] Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[3] Bedrick, E. J. (1992). A comparison of generalized and modified sample biserial correlation estimators. *Psychometrika* **57** 183–201. https://doi.org/10.1007/BF02294504 MR1173589

[4] Bedrick, E. J. and Breslin, F. C. (1996). Estimating the polyserial correlation coefficient. *Psychometrika* **61** 427–443. https://doi.org/10.1007/BF02294548 MR1424910

[5] Berlin, D. A., Gulick, R. M. and Martinez, F. J. (2020). Severe Covid-19. *New England Journal of Medicine* **383** 2451–2460. https://doi.org/10.1056/nejmcp2009575

[6] Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Math. Comp.* **19** 577–593. https://doi.org/10.2307/2003941 MR198670

[7] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. https://doi.org/10.1198/jasa.2011.tm10155 MR2847973

[8] Chen, S., Witten, D. M. and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. https://doi.org/10.1093/biomet/asu051 MR3335095

[9] Cheng, J., Li, T., Levina, E. and Zhu, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. https://doi.org/10.1080/10618600.2016.1237362 MR3640193
[10] Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30** 171–178. https://doi.org/10.2307/2529626 MR334376
[11] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. https://doi.org/10.1016/j.jmva.2004.02.009 MR2064941
[12] Fan, J., Liu, H., Ning, Y. and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 405–421. https://doi.org/10.1111/rssb.12168 MR3611752
[13] Feng, H. and Ning, Y. (2019). High-dimensional Mixed Graphical Model with Ordinal Data: Parameter Estimation and Statistical Inference. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 654–663. PMLR.
[14] Finegold, M. and Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. https://doi.org/10.1214/10-AOAS410 MR2840186
[15] Fox, J. (2022). polycor: Polychoric and Polyserial Correlations R package version 0.8-1.
[16] Foygel, R. and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culotta, eds.) **23** 604–612. Curran Associates, Inc.
[17] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045
[18] Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. https://doi.org/10.1016/0024-3795(88)90223-6 MR943997
[19] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR144363
[20] Jin, S. and Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika* **82** 67–85. https://doi.org/10.1007/s11336-016-9512-2 MR3614808
[21] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. https://doi.org/10.1214/09-AOS720 MR2572459
[22] Lauritzen, S. L. (1996). *Graphical models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York Oxford Science Publications. MR1419991
[23] Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed

graphical models. *J. Comput. Graph. Statist.* **24** 230–253. https://doi.org/10.1080/10618600.2014.900500 MR3328255

[24] Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[25] Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. https://doi.org/10.1214/12-AOS1037 MR3059084

[26] Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. https://doi.org/10.1214/17-AOS1637 MR3851754

[27] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. https://doi.org/10.1214/009053606000000281 MR2278363

[28] Miyamura, M. and Kano, Y. (2006). Robust Gaussian graphical modeling. *J. Multivariate Anal.* **97** 1525–1550. https://doi.org/10.1016/j.jmva.2006.02.006 MR2275418

[29] Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C. and Montana, G. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage* **103** 427–443. https://doi.org/10.1016/j.neuroimage.2014.07.033

[30] Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460. https://doi.org/10.1007/BF02296207 MR554892

[31] Olsson, U., Drasgow, F. and Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika* **47** 337–347. https://doi.org/10.1007/BF02294164 MR678066

[32] Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **195** 1–47.

[33] Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* **9** 116–139.

[34] Perrakis, K., Lartigue, T., Dondelinger, F. and Mukherjee, S. (2019). Regularized joint mixture models. arXiv:1908.07869. MR4582441

[35] Quan, X., Booth, J. G. and Wells, M. T. (2018). Rank-based approach for estimating correlations in mixed ordinal data. arXiv:1809.06255.

[36] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. https://doi.org/10.1214/09-AOS691 MR2662343

[37] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. https://doi.org/10.1214/11-EJS631 MR2836766

[38] Städler, N. and Mukherjee, S. (2013). Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *Ann. Appl. Stat.* **7** 2157–2179. https://doi.org/10.1214/13-AOAS662 MR3161717

[39] Städler, N. and Mukherjee, S. (2015). Multivariate gene-set testing based on graphical models. *Biostatistics* **16** 47–59. https://doi.org/10.1093/biostatistics/kxu027 MR3365410

[40] Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18** 342–353. https://doi.org/10.2307/2527476 MR145613

[41] Yang, Z., Ning, Y. and Liu, H. (2018). On semiparametric exponential family graphical models. *J. Mach. Learn. Res.* **19** Paper No. 57, 59. MR3899759

[42] Verzelen, N. and Villers, F. (2009). Tests for Gaussian graphical models. *Comput. Statist. Data Anal.* **53** 1894–1905. https://doi.org/10.1016/j.csda.2008.09.022 MR2649554

[43] Wainwright, M. J. and Jordan, M. I. (2006). Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing* **54** 2099–2109. https://doi.org/10.1109/tsp.2006.874409

[44] Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544. https://doi.org/10.1093/bioinformatics/btm129

[45] Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H. I., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., Harrison, D., Forbes, H., Schultze, A., Croker, R., Parry, J., Hester, F., Harper, S., Perera, R., Evans, S. J. W., Smeeth, L. and Goldacre, B. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584** 430–436. https://doi.org/10.1038/s41586-020-2521-4

[46] Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. https://doi.org/10.1214/12-AOS1041 MR3097612

[47] Yang, E., Baker, Y., Ravikumar, P., Allen, G. and Liu, Z. (2014). Mixed Graphical Models via Exponential Families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. Kaski and J. Corander, eds.). *Proceedings of Machine Learning Research* **33** 1042–1050. PMLR, Reykjavik, Iceland.

[48] Yoon, G., Müller, C. L. and Gaynanova, I. (2021). Fast Computation of Latent Correlations. *Journal of Computational and Graphical Statistics* **30** 1249-1256. https://doi.org/10.1080/10618600.2021.1882468 MR4356618

[49] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

A.2 `causalAssembly`: Generating Semi-Synthetic Real Data for Benchmarking Causal Discovery

Proceedings of Machine Learning Research vol 236:609–642, 2024 3rd Conference on Causal Learning and Reasoning

`causalAssembly`: Generating Realistic Production Data for Benchmarking Causal Discovery

Konstantin Göbler KONSTANTIN.GOEBLER@TUM.DE
Technical University of Munich, Germany
Robert Bosch GmbH

Tobias Windisch TOBIAS.WINDISCH@HS-KEMPTEN.DE
University of Applied Sciences Kempten

Mathias Drton MATHIAS.DRTON@TUM.DE
Technical University of Munich, Germany
Munich Center for Machine Learning (MCML)

Tim Pychynski TIM.PYCHYNSKI@DE.BOSCH.COM
Steffen Sonntag STEFFEN.SONNTAG@DE.BOSCH.COM
Martin Roth MARTIN.ROTH2@DE.BOSCH.COM
Robert Bosch GmbH

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Algorithms for causal discovery have recently undergone rapid advances and increasingly draw on flexible nonparametric methods to process complex data. With these advances comes a need for adequate empirical validation of the causal relationships learned by different algorithms. However, for most real and complex data sources true causal relations remain unknown. This issue is further compounded by privacy concerns surrounding the release of suitable high-quality data. To tackle these challenges, we introduce `causalAssembly`, a semisynthetic data generator designed to facilitate the benchmarking of causal discovery methods. The tool is built using a complex real-world dataset comprised of measurements collected along an assembly line in a manufacturing setting. For these measurements, we establish a partial set of ground truth causal relationships through a detailed study of the physics underlying the processes carried out in the assembly line. The partial ground truth is sufficiently informative to allow for estimation of a full causal graph by mere nonparametric regression. To overcome potential confounding and privacy concerns, we use distributional random forests to estimate and represent conditional distributions implied by the ground truth causal graph. These conditionals are combined into a joint distribution that strictly adheres to a causal model over the observed variables. Sampling from this distribution, `causalAssembly` generates data that are guaranteed to be Markovian with respect to the ground truth. Using our tool, we showcase how to benchmark several well-known causal discovery algorithms.

Keywords: Causal discovery, benchmarking, production data, distributional random forest

1. Introduction

Causal discovery is the process of learning causal relations among variables of interest from data. In recent years, the area has seen numerous theoretical and algorithmic developments (for an overview see, e.g. Glymour et al., 2019; Heinze-Deml et al., 2018; Maathuis et al., 2019; Vowels et al., 2022; Zanga et al., 2022). Causal structure learning is also gaining traction in industrial domains (compare e.g. the review in Vuković and Thalmann, 2022), where knowledge about causal relations forms the basis for downstream operations such as root-cause analysis (Budhathoki et al., 2022).

Despite the recent advances and expanded application areas, it remains challenging to provide adequate empirical validation of the causal relationships inferred by causal discovery algorithms (Gentzel et al., 2019; Eigenmann et al., 2020). Ground truth causal relationships are unknown for most real-world examples that involve complex data sources. In addition, privacy concerns often hinder the release of high-quality data. As a result, research papers proposing novel procedures often work with simplistic simulation setups that tend to have limited generalizability to real-world settings (Huegle et al., 2021; Reisach et al., 2021). When applied to real data and assessed by domain experts, results are often sobering (Heinze-Deml et al., 2018; Constantinou et al., 2021).

To help address these challenges we introduce `causalAssembly`, a semisynthetic data generation tool that leverages extensive domain knowledge and real production data to form a ground truth. The assembly line that originates the data is composed of multiple production stations where individual components are joined together through automated manufacturing processes. Each individual manufacturing process highly depends on the state of the preceding processes as well as on the raw components added in earlier stations. Due to the intricate and often nonlinear nature of the physical processes—including, e.g., the interaction of machine state, process control settings and component state, complex causal relationships arise across the entire assembly line. In the manufacturing plant, processes are computer-monitored, and the resulting measurements are stored throughout production.

We select a subset of production stations from the assembly line, each of which contains individual processes with known causal relationships. We determine these relationships by consulting with domain experts and carefully studying the physical processes that take place at each station. However, our process knowledge currently does not extend to relationships between processes. We show that the extensive domain knowledge of the individual processes, together with the line structure, gives rise to a causal ordering that agrees with that of the unknown ground truth layered graph. Consequently, we apply established feature selection strategies (e.g. Bühlmann et al., 2014) to prune the corresponding complete graph implied by this causal ordering. The resulting causal graph is treated as the appropriate ground truth in `causalAssembly`.

Next, we need to enable sampling from a ground truth distribution that factorizes according to the ground truth causal graph. This is achieved with a synthetization step that learns conditional distributions indicated by the ground truth graph via distributional random forests (DRF) (Cevid et al., 2022; Gamella et al., 2022). Combining the conditionals yields a joint distribution that strictly obeys a causal model for the observed variables. Thus, the distribution of semisynthetic data sampled with `causalAssembly` is guaranteed to be Markov with respect to the ground truth graph. We want to note that strong-faithfulness (see Uhler et al., 2013) can not be guaranteed to hold in the sampled data which may affect consistency results of some causal discovery algorithms.

Our contributions. We summarize our main contributions below:

- Using domain knowledge and real production data, we propose a semisynthetic data generation pipeline `causalAssembly`. Implemented as a Python library[1], the tool allows one to generate data of varying dimension and complexity in order to benchmark causal discovery algorithms.

- The proposed procedure to synthesize real data is general and can in principle be applied to any real dataset with full or partial domain knowledge. The data generation procedure assures resulting data to be Markovian under causal sufficiency.

1. https://github.com/boschresearch/causalAssembly

- The proposed semisynthetization of real data offers a way to release high-quality data to the public that would otherwise be inaccessible due to privacy concerns. To employ this procedure, knowledge regarding a causal order will be a necessary requirement.

- We provide first benchmark results of popular causal discovery algorithms including the PC algorithm (Spirtes and Glymour, 1991), DirectLiNGAM (Shimizu et al., 2011), NOTEARS (Zheng et al., 2018), GraN-DAG (Lachapelle et al., 2020), and SCORE (and its scalable extension DAS) (Rolland et al., 2022; Montagna et al., 2023).

In the remainder of the paper, we first address identifiability concerns and provide an overview of related work (Section 2). Section 3 describes the data generation pipeline, and Section 4 showcases benchmarking results. Section 5 gives a concluding discussion of our work. Supplementary Sections A-D provide additional information regarding data, algorithms and metrics used, as well as additional benchmark results.

2. Identifiability and common benchmarking strategies

2.1. General background

For natural number $p \in \mathbb{N}$, let $[p] := \{1, \dots, p\}$. Defining the vertex set as $V = [p]$, let $G = (V, E)$ be a directed acyclic graph (DAG) with edge set $E \subset V \times V$. The vertex set V indexes a collection of jointly distributed random variables $(X_v)_{v \in V}$. For $S \subseteq V$, we write $X_S := (X_s)_{s \in S}$ for the subcollection of variables indexed by S. We often write $v \to w$ or $X_v \to X_w$ instead of (v, w) to indicate the existence of an edge from v to w in G. A node $v \in V$ is a *parent* of another node $w \in V$ if $v \to w \in E$. The set of all parents of w in G is denoted by $pa_G(w)$. A *directed path* from v to w in G is a sequence of nodes $v_1, \dots, v_k$ starting at $v_1 = v$ and ending at $v_k = w$ with subsequent nodes connected by edges $v_i \to v_{i+1} \in E$. If there exists a directed path from v to w, then v is an *ancestor* of w and w a *descendant* of v. The sets of all ancestors and all descendants of node w in G are denoted $an_G(w)$ and $de_G(w)$, respectively and $w \notin an_G(w) \cup de_G(w)$. When it is clear which graph G is being considered, we simply write $pa(w)$, $de(w)$ or $an(w)$.

The DAG G permits causal relationships among the random variables, and we say that X_v is a direct cause of X_w if $v \to w$. Furthermore, X_v may be an indirect cause of X_w if there is a directed path from v to w; we refer to Pearl (2009); Peters et al. (2017) for more background. Each DAG G on V gives rise to a statistical model for the joint distribution P_X of the random vector $X = (X_v)_{v \in V}$. Here, the edges in G encode the permitted conditional dependence relations among the variables (see e.g. Drton and Maathuis, 2017; Lauritzen, 1996). Any distribution that satisfies the conditional independence constraints implied by G is said to be *Markov* with respect to G. In other words, the joint density of X factors into a product of conditional densities as $p(x) = \prod_{v \in V} p(x_v \mid x_{pa_G(v)})$ with $p(x_v \mid x_\emptyset) = p(x_v)$. The conditional independence relations holding in all such factorizing distributions can be obtained using the graphical criterion of *d-separation* (Pearl, 2009). A distribution P_X is *Markov* and *faithful* with respect to the underlying DAG G if the conditional independence relations in P_X correspond exactly to d-separation relations in G. Different DAGs may share the same set of conditional independence statements. Such DAGs are said to be Markov equivalent. All equivalent DAGs form a Markov equivalence class (MEC) (see e.g. Spirtes et al., 1993).

2.2. Layered DAGs

Let $G = (V, E)$ be a DAG, and let $\mathcal{L} := (V_1, \dots, V_K)$ be a partition of V into K ordered subsets. Then $L = (G, \mathcal{L})$ is a *layered DAG* if for every edge $(v, w) \in E$ with nodes $v \in V_s$, $w \in V_t$ with $s \neq t$, we have $s < t$ (see, e.g., Healy and Nikolov, 2002). In this case, the elements of $\mathcal{L}$ are the *layers* of L, and each layer V_s induces a subgraph $G[V_s]$ of G, which we denote by L_s and refer to as the layer-induced subgraph.

A permutation $\pi : [p] \to [p]$ is a causal ordering for the layered DAG $L = (G, \mathcal{L})$ if (i) $\pi(v) < \pi(w)$ for all (v, w) with $w \in de_G(v)$, and (ii) $\pi(v) < \pi(w)$ for all (v, w) such that $v \in V_s$ and $w \in V_t$ for layers with $s < t$. Such an ordering respects both the ordering given by directed paths in G and the ordering of the layers in $\mathcal{L}$. There always exist a causal ordering but it need not be unique. The set of all causal orderings of a layered DAG L is denoted by Π_L.

Proposition 1 *Let $L = (G, \mathcal{L})$ and $L' = (G', \mathcal{L})$ be two layered DAGs that share the same vertex set V and the same partition $\mathcal{L} = (V_1, \dots, V_K)$. If $L_s = L'_s$ for all $s \in [K]$, then $\Pi_L = \Pi_{L'}$.*

The proof can be found in Section B. Let $L = (G, \mathcal{L})$ be a layered DAG with vertex set V and edge set E. For any causal ordering $\pi \in \Pi_L$, we define a completed layered DAG $L^\pi = (G^\pi, \mathcal{L})$ by adding to L all edges $v \to w$ with $v \in V_s$ and $v \in V_t$ for $s < t$. So, the edge set of L^π is

$$E^\pi := E \cup \bigcup_{s<t} V_s \times V_t.$$

Clearly, L^π is a super-DAG of L in the sense of E being a subset of E^π. Often we have to consider all nodes up to a certain layer. Thus, we define $V_{1:s} := \cup_{i=1}^{s-1} V_i$ for $s \in [K]$. Note that $V_s \cap V_{1:s} = \varnothing$.

2.3. Identifiability of the causal structure

In general, the DAG G is not identifiable from P_X alone, and learning the graph or its MEC from data requires assumptions. Under the Markov and faithfulness conditions on P_X, G can be recovered up to its MEC. Prominent examples of algorithms that are based on this assumption are the PC algorithm (Spirtes et al., 1993) and Greedy Equivalence Search (GES) (Chickering, 2003). Under additional assumptions, we may not only recover the MEC but G itself. Note that the graphical model associated with G can also be expressed in terms of a structural causal model (SCM) (Peters et al., 2017). That is, in the absence of latent confounding, if P_X is Markov with respect to G then there exist independent noise variables $\epsilon = (\epsilon_1, \dots, \epsilon_p)$ and measurable functions g_v such that $X_v = g_v(X_{pa(v)}, \epsilon_v)$ for all $v \in [p]$. In this framework, if all the g_v are linear and all ϵ_v follow non-Gaussian distributions then the causal graph is identifiable from data (Shimizu, 2022). In a different vein, identifiability of G is also achieved under an additive noise model (ANM) (Hoyer et al., 2008). In other words, as long as all g_v are non-linear as defined in Hoyer et al. (2008), the noise ϵ_v may follow any distribution given that it is additive. A generalization to the post-nonlinear model is proposed by Zhang and Hyvärinen (2009). In contrast, linear SCMs with Gaussian noise terms are not identifiable. However, identifiability can be achieved in this setting under further assumptions such as equal noise variances (Peters and Bühlmann, 2014).

2.4. Additive noise models for synthetic data generation

Comprehensive assessment of methods for learning causal relations is challenging due to the scarcity of real data where true causal relations are known. Hence, the focus often lies on simulated synthetic

data. In order to accommodate non-linearities and also to ensure identifiability of the true DAG, ANMs are often employed for this purpose. However, synthetic data generation can leave unintended traces hinting at the true causal relations among the variables. In particular, Reisach et al. (2021) demonstrate that for commonly used simulation setups involving ANMs, irrespective of the choice of the functions g_v, ordering nodes by increasing marginal variance tends to agree with the causal ordering of the ANM. The authors introduce *varsortability*, a score that measures the degree of concordance between causal orderings of the ground truth DAG and the order of increasing marginal variance. When varsortability equals one then all directed paths lead from lower variance nodes to higher variance nodes. This is connected to the aforementioned identifiability results for linear SCMs with equal variance (Chen et al., 2019). Many recently proposed continuous optimization-based structure learning algorithms have been evaluated in simulations based on ANMs (e.g., Zheng et al., 2018; Ng et al., 2020; Lachapelle et al., 2020; Yu et al., 2021; Zhang et al., 2022, 2023). The resulting high varsortability on the raw data scale contributes to often staggering performances as the methods may inadvertently make use of variance order. In real systems, on the other hand, measurement scales are essentially irrelevant for the causal ordering. Thus, standardization would render data scale non-informative. Yet, remnants of high varsortability with distinctive covariance patterns may remain even after standardization (Reisach et al., 2021, 2023). Ultimately, all synthetic data generation processes are vulnerable to a lack of portraying common themes observable in practice.

Evaluating the performance of algorithms is a key factor that impacts which research directions the field is gravitating towards. This can be seen, for instance, in the case of continuous optimization-based structure learning. Following the development of NOTEARS (Zheng et al., 2018), numerous extensions have been proposed in short succession (see Vowels et al., 2022, for an overview). However, when evaluated outside of ANMs, performance often drops notably (Constantinou et al., 2021). Therefore, there is a clear necessity for more robust evaluation techniques in the field of causal discovery as outlined in Eigenmann et al. (2020); Gentzel et al. (2019). This motivates the present work which seeks to support benchmarking of causal discovery by providing tools for semisynthetic data generation that draw on partial ground truth and real data from a manufacturing plant.

2.5. Related work

Focusing on observational data, the available tools for empirical illustration and benchmarking of causal discovery algorithm can be broadly classified into three tiers: (1) Real datasets with expert-knowledge driven ground truth, such as the cause-effect pairs of Mooij et al. (2016) and data pertaining to already well understood systems in biology (Marbach et al., 2012; Replogle et al., 2022; Sachs et al., 2005). (2) Synthetic datasets generated based on well-established parameterizations (den Bulcke et al., 2006; Schaffter et al., 2011; Scutari, 2010). (3) Data generated using functional specifications; see, e.g., Chen and Fayek (2023); Huegle et al. (2021). Our proposed semisynthetic data generator takes a middle ground between (1) and (2), leveraging expert knowledge to establish a causal order while relying on real data to circumvent the need for explicit parameterizations.

3. Semisynthetic data generator

3.1. Process knowledge

Our real-world data come from an assembly line at a highly automated manufacturing plant. We focus on a selected subset of production stations at which press-in and staking processes are executed.

(a) Press-in *(b) Staking*

Figure 1: Illustration of the phases of a press-in and staking process. In press-fitting (a1) the tool moves downward axially until it contacts the valve (a2) and pushes it into a slightly smaller bore until its shoulder reaches the axial block position (a3).

In the following, we give a brief overview of these processes. A more in depth discussion of the relations among the central features of the physical procedures is given in supplementary Section A.1.

Press-in: In each fitting process, two components are joined in a pressing machine to form a mechanically strong and sometimes even leakage tight connection between the components. An inner cylindrical component (e.g., a valve, axle or pin) is pushed into a slightly smaller bore of a larger component. This is done by applying a mechanical force high enough to overcome the required friction force. After the friction force is exceeded, the inner component will move into the bore. The pressing machine stops as soon as a certain predefined press-in force is exceeded for the first time.

Staking: A staking process is an extension of a press-in process and has the objective to fixate the inner component in its final position by permanently deforming outer component materials. This is achieved by applying a large enough force to the tool to create a mechanical stress-strain state in the material resulting in permanent plastic material deformation.

Figure 1 depicts these two process steps in a schematic manner. During both steps the required forces and tool displacements are measured continuously by process control units. We extract all relevant numeric measurements for each process, and together with the mapping detailed in Section A.1 they build the foundation for our data generation scheme.

3.2. Modeling domain knowledge

Let $L^0 = (V, E^0, \mathcal{L})$ be the true and unknown layered DAG that encapsulates within as well as between process relations. The associated random variables $(X_v)_{v \in V}$ represent the features extracted from the $K = 10$ manufacturing processes that are carried out sequentially. These processes are lined up across five production stations. A user of `causalAssembly` will see which variables belong to each of these five stations. Each station accommodates two of the production processes (Fig. 2). In total, $p = |V| = 98$ features for $n = 15.581$ parts are collected. The features distribute across production stations as follows: *Station 1* with $|V_1 \cup V_2| = 6$, *Station 2* with $|V_3 \cup V_4| = 34$, *Station 3* with$|V_5 \cup V_6| = 16$, *Station 4* with $|V_7 \cup V_8| = 26$, and *Station 5* with $|V_9 \cup V_{10}| = 16$.

Each individual process we focus on is well understood, and we are able to state causal relations among the variables in V_s. These causal relations are considered to be complete and correct for the individual processes. We map the process knowledge into process DAGs $G_s^* = (V_s, E_s)$ for $s \in [K]$. The graph union $G_1^* \oplus \cdots \oplus G_K^*$ together with the ordered partitioning implied by the node sets of the process DAGs gives rise to the layered DAG $L^* = (G_1^* \oplus \cdots \oplus G_K^*, \mathcal{L})$ with $\mathcal{L} = (V_1, \ldots V_{10})$.

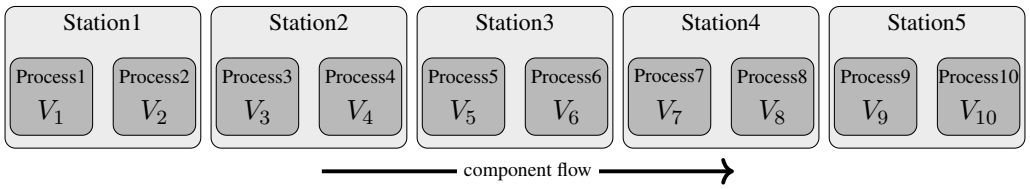


Figure 2: Illustration of the production line with five process stations each containing two successive processes. The first station prepares components while station two and four carry out staking tasks. The remaining stations perform press-in tasks.

Note that L^* and L^0 are both layered DAGs on the same vertex set V with the same layering $\mathcal{L}$. More importantly, the layer-induced subgraphs coincide, i.e., $L^*_s = L^0_s$ for all $s \in [10]$.

Unfortunately, due to their complex physical nature, available process knowledge does not extend to how the K processes influence each other. In other words, the edge sets of L^* and L^0 likely differ. However, Proposition 1 suggests that the process knowledge we gathered puts us in a favorable situation: Irrespective of whether there exist edges between layers, the set of all possible causal orderings is determined solely by the layers themselves, i.e. the process graphs. The complete layered DAG L^π for any $\pi \in \Pi_{L^*}$ is a super-DAG of the true layered DAG L^0. Consequently, we may use common variable selection techniques from regression to prune L^π and infer the layered ground truth DAG L^0 (Bühlmann et al., 2014, Section 2.5).

3.3. Sparse additive models (SpAM) for cross-process pruning

We perform the pruning step by regressing each variable X_v onto its parent variables X_k, $k \in pa_{L^\pi}(v)$. The resulting active sets of predictors determine the final edge set, forming the estimate $\hat{L}$. Unless stated otherwise, the parent sets considered below are obtained from L^π. When constructing L^π we add edges between the nodes from distinct processes V_s and V_t, and leave the layer-induced subgraphs as they are. Similarly, we only consider edges between processes for pruning. To prune, we use the flexible nonparametric framework of SpAMs (Ravikumar et al., 2009). In a SpAM, the regression function for a variable X_v with v in layer V_t, is modeled as

$$f_v(x_{pa(v)}) = \mathbb{E}(X_v \mid X_{pa(v)} = x_{pa(v)}) \approx \sum_{k \in pa(v)} f_{v,k}(x_k).$$

Sparsity is induced through a functional group-lasso penalty based on $\|f_{v,k}\|_2 := \sqrt{\mathbb{E}(f_{v,k}(X_k)^2)}$, the $L^2(P_{X_k})$ norm of component $f_{v,k}(X_k)$. For some $\lambda \geqslant 0$, the resulting population-level minimization problem can be stated as:

$$\min_{f_{v,k} \in \mathcal{F}_{v,k},\ k \in pa(v)} \left\{ \frac{1}{2}\mathbb{E}\big(X_v - \sum_{k \in pa(v)} f_{v,k}(X_k)\big) + \lambda \sum_{k \in V_{1:t}} \|f_{v,k}\|_2 \right\}.$$

Depending on the function classes $\mathcal{F}_{v,k}$, the minimization in SpAM becomes a convex program. Indeed, in practice, we express each function as $f_{v,k}(x_k) = \sum_{l=1}^{p_{vk}} \psi_{vkl}(x_k)\beta_{vkl}$, where $\{\psi_{vkl}\}_{l=1}^{p_{vk}}$ is a set of cubic splines. Let $\Psi_{vk}(x_k)$ be the $1 \times p_{vk}$ row vector of evaluations of the ψ_{vkl}, and gather the β_{vkl} in the coefficient vector $\theta_{vk} = (\beta_{vk1}, \ldots, \beta_{vkp_{vk}})^T$. Then $f_{v,k}(x_k) = \Psi_{vk}\theta_{vk}$. For estimation from data, we solve the empirical version of the SpAM problem, with cross-validation for

tuning parameter selection. The set $\{(k, v) : k \in V_{1:t}, \hat{\theta}_{vk} \neq 0\} \cup pa_{L_s^*}(k)$ then marks the indices of the parent variables of X_v remaining after pruning. Algorithm 1 outlines our procedure.

Algorithm 1 SpAM cross-process learning

Input: Data $\mathcal{D} \subseteq \mathbb{R}^p$, completed layered DAG L^π, regularization parameter $\lambda \geqslant 0$
Output: Edge set $E_{\text{across}} \subset V \times V$.

$E_{\text{across}} \leftarrow \{\}$;
for $t = 2, \ldots, K$ **do**
  Form a causal ordering π_t of the layer-induced subgraph L_t^π;
  for v *along* π_t **do**
    Obtain $(\hat{\theta}_{vk})_{V_{1:t}}$ through the following minimization problem:

$$\min_{\theta_{vk},\, k \in pa(v)} \left\{ \frac{1}{2|\mathcal{D}|} \sum_{x \in \mathcal{D}} \Big(x_v - \sum_{k \in pa(v)} \Psi_{vk}(x_k)\theta_{vk} \Big)^2 + \lambda \sum_{k \in V_{1:t}} \sqrt{\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (\Psi_{vk}(x_k)\theta_{vk})^2} \right\}$$

    $E_{\text{across}} \leftarrow E_{\text{across}} \cup \{(k, v) : k \in V_{1:t}, \hat{\theta}_{vk} \neq 0\}$
  end
end

For some variables X_v with v in a layer V_t, $t \in \{2, \ldots, K\}$, the process experts are able to state functional relations, up to noise. Denote $P_v^* := pa_{L^*}(v) \cap V_t$. Functional process knowledge takes the form $X_v \approx f_v^*(X_{P_v^*})$. To incorporate such knowledge, we alter Algorithm 1 by inserting the known function. The optimization problem is then

$$\min_{\theta_{vk},\, k \in pa(v)} \left\{ \frac{1}{2|\mathcal{D}|} \sum_{x \in \mathcal{D}} \Big(x_v - f_v^*(x_{P_v^*}) - \sum_{k \in V_{1:t}} \Psi_{vk}(x_k)\theta_{vk} \Big)^2 + \lambda \sum_{k \in V_{1:t}} \sqrt{\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} (\Psi_{vk}(x_k)\theta_{vk})^2} \right\}$$

Note that Algorithm 1 admits a natural way to also incorporate cross-process domain knowledge simply by adding the corresponding edge to E or dropping it from the corresponding parent set.

Figure 3 illustrates the process graph $\hat{L}$ after the pruning step using Algorithm 1. In the absence of latent confounding the sparsistency property of the SpAM procedure suggests that asymptotically we recover the true layered DAG L^0. Even in the presence of confounding the synthetization procedure illustrated in the next section assures that data will be "Markovian". From here on we consider the layered DAG $\hat{L}$ as ground truth in `causalAssembly`.

3.4. Synthetization

To ensure that the semisynthetic data come from a distribution that factorizes according to the ground truth DAG $\hat{L}$, we explicitly estimate the conditional distributions under the factorization implied by the independence constraints in $\hat{L}$. To that end, we employ distributional random forests (DRFs) (Cevid et al., 2022) to obtain estimates $\hat{P}_X^{\text{DRF}}(X_v \mid X_{pa(v)})$, $v \in V$, along a causal ordering of $\hat{L}$.

DRFs build on random forests (Breiman, 2001) where the original splitting criterion is replaced with a distributional metric (in our case the maximal mean discrepancy (MMD) statistic (Gretton et al., 2006)) in order to ensure homogeneity in each of the tree's leaf nodes. These give rise to neighborhoods of relevant training data points described by corresponding weighting functions.

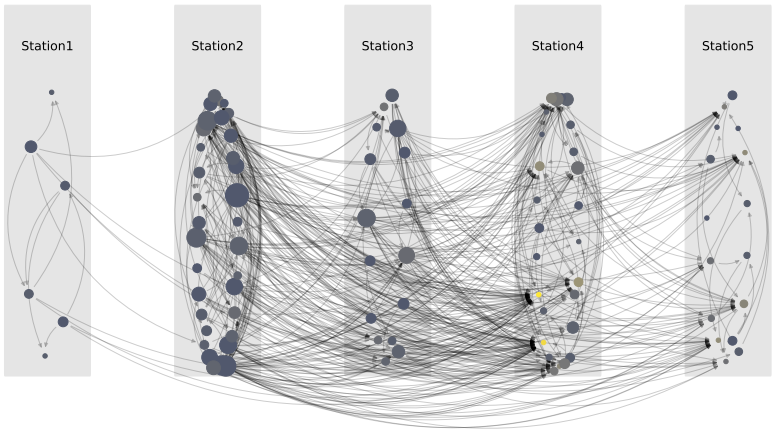


Figure 3: Assembly line ground truth after edges have been learned between processes using SpAM. We depict production stations, not the processes they are decomposed in. Node size increases with the number of out-edges. Node color gets brighter with the number of in-edges. In terms of sparsity, the assembly line ground truth graph accounts for around 10.2% of all possible connections.

The average of the tree-specific weighting functions can then be used to estimate the conditional distributions. The following paragraphs give a brief overview of the DRF procedure.

DRF growing Let L be a layered DAG. Assume we have grown N trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$ and denote $\mathcal{L}_j(x)$ the set of training data points that are assigned to the same leaf node as the test data point x in tree $\mathcal{T}_j$. Cevid et al. (2022) define the weighting function by the average of the tree-specific weighting functions i.e. $w_x(x_i) = 1/N \sum_{j=1}^{N} \mathbb{1}(x_i \in \mathcal{L}_j(x))/|\mathcal{L}_j(x)|$. Let n be the sample size of the training data. Estimating the conditional distribution $P(X_v \mid X_{pa_L(v)} = x)$ boils down to assigning weights to the point mass $\delta_{x_{iv}}$ at x_{iv} in the empirical distribution of X_v, i.e.

$$\hat{P}(X_v \mid X_{pa_L(v)} = x) = \sum_{i=1}^{n} w_x(x_i) \cdot \delta_{x_{iv}}.$$

Given $\hat{P}_X^{\text{DRF}}(X_v \mid X_{pa(v)})$ we sample a new dataset $\mathcal{D}_{\text{synth}} \subset \mathbb{R}^p$ from the production line as outlined in Algorithm 2. By construction, $\mathcal{D}_{\text{synth}}$ is drawn from the estimated joint distribution $\hat{P}_X^{\text{DRF}}$ and is guaranteed to be Markov with respect to the ground truth layered DAG $\hat{L}$.

Gamella et al. (2022) first proposed this approach using Sachs et al. (2005) data. By construction, it resolves confounding issues in the real data at the expense of $\hat{L}$ potentially containing more cross-process edges than L^0.

Remark 2 *We emphasize the crucial role of the acquired process knowledge. Although the synthetization procedure is technically capable of fitting DRFs to any configuration of conditional distributions given a sufficient sample size, domain knowledge plays a pivotal role in ensuring the generated data closely resembles the real data. To illustrate this point, we conduct experiments in Section C.2 where we assume no prior domain knowledge and instead rely solely on structure learning. Upon applying the semisynthetic procedure to a resulting process graph, we observe that in a majority of experiments, the causal discovery algorithm that initially generated the process graph*

Algorithm 2 Sampling from the causal DRFs

Input: Finite dataset $\mathcal{D}_{\text{original}} \subset \mathbb{R}^p$, (layered) DAG L, number of samples n to be generated
Output: A dataset $\mathcal{D}_{\text{synth}} \subset \mathbb{R}^p$ of size n

$\mathcal{D}_{synth} \leftarrow \varnothing$;
Form a causal ordering π of L;
while $|\mathcal{D}_{synth}| < n$ **do**
  for v *along* π **do**
    if $pa_L(v) = \varnothing$ **then**
      sample $\hat{x}_v \in \mathbb{R}$ via smooth bootstrapping from $\{x_v \in \mathbb{R} : x \in \mathcal{D}_{\text{original}}\}$;
    else
      sample $\hat{x}_v \in \mathbb{R}$ from $\hat{P}_X^{\text{DRF}}(X_v \mid X_{pa_L(v)} = \hat{x}_{pa_L(v)})$;
    end
    $\mathcal{D}_{synth} \leftarrow \mathcal{D}_{synth} \cup \{\hat{x}_v\}$;
  end
end

structure tends to exhibit superior performance. As we demonstrate in Section 4, this is not the case for our semisynthetic data generator. We investigate, whether learning the cross-process edges via SpAM makes learning these edges easier. However, Figure 12 in Section C.2 suggests that this is not the case. The combination of extensive process knowledge and highly flexible nonparametric techniques employed in `causalAssembly` *renders it virtually impossible for any causal discovery algorithm to unfairly "game" graph recovery results.*

To accommodate the analysis of data with varying dimensionality, we provide the option to isolate and generate data for any of the 5 production stations that form the natural subunits of the assembly line. To avoid creating unintended confounding we refit the DRFs and apply Algorithm 2 to each production station independently. Lastly, we want to stress that SpAMs are merely a pruning method. Their application in `causalAssembly` is not related to ANMs that have been discussed in Section 2.4.

3.5. Semisynthetic data description

To demonstrate that the semisynthetic "Markovian" data obtained from `causalAssembly` preserves key characteristics of the real data, we calculate the two-sample Kolmogorov-Smirnov (KS) statistic (Kolmogorov, 1933) between all variables in the real and a semisynthetic dataset. Figure 4 shows kernel density estimates of variables with largest and smallest KS statistics among all non-source nodes. Even in unfavorable cases where the absolute difference between the marginal distributions is largest, the semisynthetic data only shows moderate deviations from the real data. To further illustrate the similarity of the semisynthetic data to the real data, we depict pairwise correlation agreement in Figure 11 in the Appendix.

In Figure 5, we depict bivariate relationships between selected direct causes on the x-axis and their effects on the y-axis according to the ground truth DAG $\hat{L}$. Variables were selected within stations in (a) and (b) and between stations in (c). The size of the parent sets for the effect variables are $1, 3, 6$ for (a), (b), and (c), respectively. Naturally, with increasing parent set size bivariate

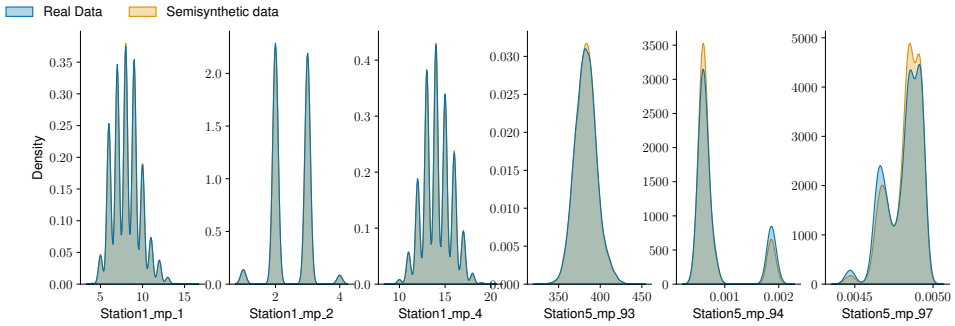


Figure 4: Kernel density plots (upper panel) of the same variables in the real (blue) and semisynthetic (yellow) data. Selection of those variables is based on comparing the highest (first three) and lowest (last three) agreement in terms of Kolmogorov-Smirnov statistic.

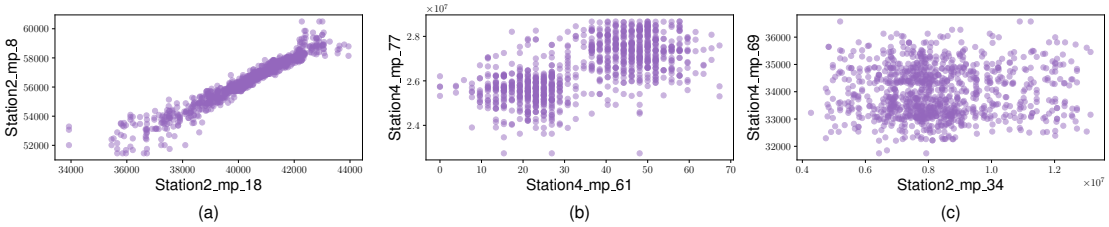


Figure 5: Bivariate scatter plots of selected nodes to showcase the different types of bivariate patterns to expect in data generated with `causalAssembly`. All node pairs are causally linked with x-axis variables being parents of y-axis variables. The number of parents of each y-axis variable varies from one parent in (a), three parents in (b) and six parents in (c).

relationships become harder to discern. While (a) depicts a clear linear relationship between the selected measurements, the relationships in (b) and (c) are masked by the remaining parent variables.

4. Benchmark results

As a proof of concept, we present first causal discovery benchmark results. In order to make sure that privacy concerns are respected, we first apply the semisynthetization procedure to the real data from the manufacturing plant. We then sample once from the joint distribution in `causalAssembly`. The resulting data is used for benchmarking as shown below. We present further benchmark results based on the individual production stations in Section D.2.

The purpose of the study we report on here is to demonstrate the versatility of the data generation scheme. A more elaborate set of experiments comprising varying sample size and adjusting hyperparameters will be taken up in future work. The PC algorithm does not output a DAG. Thus, to compare results additional harmonization steps are necessary. We implement two strategies *CPDAG-transform* and *average-dag* outlined in Section D.1. As the results following these strategies do not seem to differ substantially, we report the *average-dag* results and refer to Appendix D for the remaining experiments. A brief overview of the algorithms and metrics is provided in Section D.1.

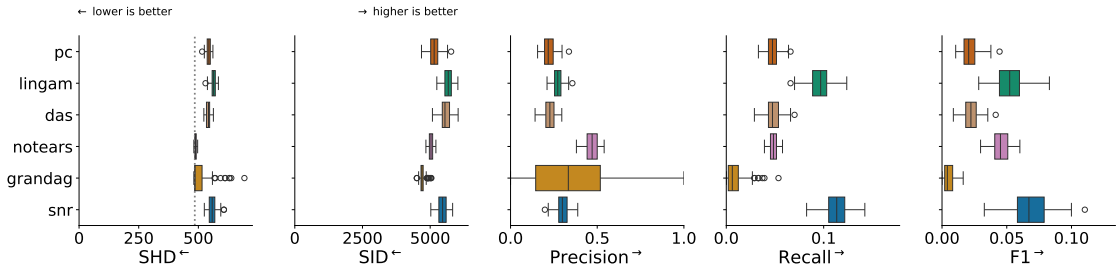


Figure 6: Chosen metrics for assessing causal discovery of the full assembly line ($|V| = 98$) using 100 simulation runs on a sample size of $n = 5000$. The dashed line corresponds to the SHD between the empty graph and ground truth indicating the number of edges in the ground truth graph. Note that we use DAS rather than SCORE on the full line due to its improved scaling to higher dimensions.

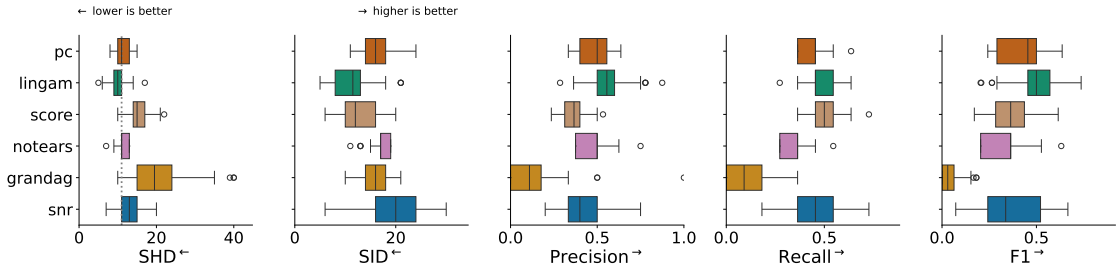


Figure 7: Chosen metrics for assessing causal discovery outcomes on Station 3 ($|V_5 \cup V_6| = 16$) using 100 simulation runs on a sample size of $n = 500$. The dashed line corresponds to the SHD between the empty graph and ground truth indicating the number of edges in the ground truth graph.

We standardize the data sampled in each of the 100 simulation runs before structure learning and report the average varsortability in Table 4 in Section D. The sortnregress (snr) routine serves as a reference point, as on standardized data it corresponds to choosing some ordering at random and regressing each node on its ancestors (*random regress*) (see also Reisach et al., 2021).

Figures 6 and 7 illustrate the benchmark results for the full line and Station 3, respectively. We focus on Station 3 because its results differ significantly from those of the full line. The benchmark results for the remaining four process stations are displayed in Section D.

Regarding the full line benchmark results in Figure 6, the default methods fail to beat an empty graph in terms of structural Hamming distance (SHD). Overall, NOTEARS performs best in terms of SHD and SID, but this is partly due to its tendency to output very sparse graphs, as evidenced by low Recall scores. Linear methods such as DirectLiNGAM and the PC algorithm with standard partial correlation tests perform similarly to *random regress*. GraN-DAG exhibits higher variability in results, with highly variable precision and very low recall and F1 scores. Inspecting the results from Station 3 depicted in Figure 7, we observe that most selected algorithms outperform random regress, especially in terms of structural intervention distance (SID) and F1 score. DirectLiNGAM exhibits the best performance across most metrics, while GraN-DAG again struggles with precision, recall, and F1. The PC algorithm and SCORE perform similarly, and NOTEARS fails to beat *random regress* in this case.

We emphasize that these results are based on out-of-the-box implementations of the algorithms, with no attempt to tune them to the data. Additionally, all the considered methods ignore the prior knowledge implied by the layered structure of the processes. We expect that causal discovery tools that can efficiently incorporate the layered structure will achieve better performance. Furthermore, tuning the hyperparameters of the algorithms and using statistical methods that are better equipped to handle complex nonlinear relationships is likely to yield better results.

While limited, this first benchmark study reveals that standard off-the-shelf methods struggle with data generated by `causalAssembly`. This largely corroborates findings from empirical studies, which show that the learning accuracy reported in the literature on synthetic data overestimates performance on more complex datasets (Constantinou et al., 2021; Rios et al., 2022).

5. Conclusion and Limitations

Empirical validation is key in understanding how causal discovery algorithms perform outside their laboratory settings. In this work, we leverage a complex dataset comprising measurements from a highly automated assembly line in a manufacturing plant, for which we are able to put together partial ground truth causal relationships on the basis of a detailed study of the underlying physics. Based on this dataset, we use the associated ground truth and develop a versatile synthetization pipeline based on DRFs that helps overcome common real world data problems such as lack of ground truth, confounding, and privacy concerns. At the same time, by preserving the real data characteristics as much as possible while making sure that the observational distribution is Markov with respect to the obtained ground truth, we overcome the issue of purely synthetic data often being too simplistic and not generalizable to real world settings. The resulting data generator can be used to benchmark causal discovery algorithms based on complex production data. We demonstrate this by providing initial benchmarking results for some well-known causal discovery algorithms.

An important limitation is that currently `causalAssembly` supports the generation of continuous data only. However, the methodology presented is more general and supports mixed data. We plan to include discrete features by selecting a wider range of processes in future releases of `causalAssembly`. Another limitation comes from the fact that ground truth relations are only available within and not across processes. In future releases we plan to address this by continuously validating cross-process edges together with process experts. Lastly, it would be interesting to take advantages of opportunities to involve interventional samples.

Acknowledgments

We thank the anonymous reviewers and Jens Ackermann for helpful comments and insights. The research for this project was partly conducted by authors employed at Robert Bosch GmbH.

References

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541, 1997. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1031833662. URL https://doi.org/10.1214/aos/1031833662.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/a:1010933404324. URL https://doi.org/10.1023/a:1010933404324.

Kailash Budhathoki, Lenon Minorics, Patrick Bloebaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/budhathoki22a.html.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, 42(6):2526–2556, 2014. ISSN 0090-5364,2168-8966. doi: 10.1214/14-AOS1260. URL https://doi.org/10.1214/14-AOS1260.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.

Jarry Chen and Haytham M. Fayek. The structurally complex with additive parent causality (scary) dataset. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz049. URL https://doi-org.eaccess.tum.edu/10.1093/biomet/asz049.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741–3782, 2014. ISSN 1532-4435,1533-7928.

Anthony C. Constantinou, Yang Liu, Kiattikun Chobtham, Zhigao Guo, and Neville K. Kitson. Large-scale empirical validation of Bayesian network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, 131:151–188, 2021. doi: 10.1016/j.ijar.2021.01.001. URL https://doi.org/10.1016/j.ijar.2021.01.001.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1), 2006. doi: 10.1186/1471-2105-7-43. URL https://doi.org/10.1186/1471-2105-7-43.

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.

Marco Eigenmann, Sach Mukherjee, and Marloes Maathuis. Evaluation of causal structure learning algorithms via risk estimation. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 2020. URL https://proceedings.mlr.press/v124/eigenmann20a.html.

Juan L. Gamella, Armeen Taeb, Christina Heinze-Deml, and Peter Bühlmann. Characterization and greedy learning of Gaussian structural causal models under unknown interventions. *arXiv preprint arXiv:2211.14897*, 2022.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a87c11b9100c608b7f8e98cfa316ff7b-Paper.pdf.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, June 2019. doi: 10.3389/fgene.2019.00524. URL https://doi.org/10.3389/fgene.2019.00524.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf.

Patrick Healy and Nikola S. Nikolov. A branch-and-cut approach to the directed acyclic graph layering problem. In *Graph drawing*, volume 2528 of *Lecture Notes in Comput. Sci.*, pages 98–109. Springer, Berlin, 2002. ISBN 3-540-00158-1. doi: 10.1007/3-540-36151-0\{_}10. URL https://doi.org/10.1007/3-540-36151-0_10.

Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018. doi: 10.1146/annurev-statistics-031017-100630. URL https://doi.org/10.1146/annurev-statistics-031017-100630.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *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.

Johannes Huegle, Christopher Hagedorn, Lukas Boehme, Mats Poerschke, Jonas Umland, and Rainer Schlosser. Manm-cs: Data generation for benchmarking causal structure learning from mixed discrete-continuous and nonlinear data. In *WHY-21 @ NeurIPS 2021*, 2021.

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.*, 14:111–152, 2013. ISSN 1532-4435,1533-7928.

A. N Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933. URL https://cir.nii.ac.jp/crid/1571135650766370304.

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.

Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019. ISBN 978-1-4987-8862-5.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012. doi: 10.1038/nmeth.2016. URL https://doi.org/10.1038/nmeth.2016.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 752–771. PMLR, 11–14 Apr 2023. URL https://proceedings.mlr.press/v213/montagna23b.html.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL http://jmlr.org/papers/v17/14-518.html.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d04d42cdf14579cd294e5079e0745411-Paper.pdf.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-89560-6; 0-521-77362-8. doi: 10.1017/CBO9780511803161. URL https://doi.org/10.1017/CBO9780511803161. Models, reasoning, and inference.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. ISSN 0006-3444,1464-3510. doi: 10.1093/biomet/ast043. URL https://doi.org/10.1093/biomet/ast043.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Comput.*, 27(3):771–799, 2015. ISSN 0899-7667,1530-888X. 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. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/books/elements-causal-inference.

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. ISSN 1369-7412,1467-9868. 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.

Alexander G. Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. Simple sorting criteria help find the causal order in additive noise models. *arXiv preprint arXiv:2303.18211*, 2023.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013. URL https://doi.org/10.1016/j.cell.2022.05.013.

Felix L. Rios, Giusi Moffa, and Jack Kuipers. Benchpress: A scalable and versatile workflow for benchmarking structure learning algorithms. *arXiv preprint arXiv:2107.03863*, 2022.

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, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809. URL https://www.science.org/doi/abs/10.1126/science.1105809.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr373. URL https://doi.org/10.1093/bioinformatics/btr373.

M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. URL http://www.jstatsoft.org/v35/i03/.

Shohei. Shimizu. *Statistical causal discovery: LiNGAM approach*. JSS Research Series in Statistics. Springer Japan, Tokyo, 1st ed. 2022. edition, 2022. ISBN 9784431557845.

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. ISSN 1532-4435,1533-7928.

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 and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991. doi: 10.1177/089443939100900106. URL https://doi.org/10.1177/089443939100900106.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. ISBN 0-387-97979-4. doi: 10.1007/978-1-4612-2748-9. URL https://doi.org/10.1007/978-1-4612-2748-9.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463, 2013. ISSN 0090-5364,2168-8966. doi: 10.1214/12-AOS1080. URL https://doi.org/10.1214/12-AOS1080.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022. doi: 10.1145/3527154. URL https://doi.org/10.1145/3527154.

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

Yue Yu, Tian Gao, Naiyu Yin, and Qiang Ji. DAGs with no curl: An efficient DAG structure learning approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12156–12166. PMLR, 2021. URL https://proceedings.mlr.press/v139/yu21a.html.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2022.09.004. URL https://www.sciencedirect.com/science/article/pii/S0888613X22001402.

An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. Boosting causal discovery via adaptive sample reweighting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LNpMtk15AS4.

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, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Ehsan M Abbasnejad, Mingming Gong, Kun Zhang, and Javen Qinfeng Shi. Truncated matrix power iteration for differentiable DAG learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=I4aSjFR7jOm.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *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.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR, 2020. URL https://proceedings.mlr.press/v108/zheng20a.html.

Appendix A. Assembly line dataset

We leverage a complex production dataset comprising real measurements from a highly automated assembly line in a manufacturing plant. The assembly line consists of multiple production stations along which individual components are joined together. Parts are processed in these production stations and are sequentially passed to the next station according to the line structure.

Often, stations perform the same process several times either in succession or in parallel to reduce overall cycle time. Raw components are added to the assembly at the respective stations. The quality of the assembly process depends on the respective machine state and settings (e.g. tool state, control parameters), the environmental conditions, and on the raw state of the components or intermediates. We carefully study the physical processes that take place in each of the stations.

A.1. Mapping process knowledge

Press-in In order for the inner cylindrical component to be pressed into the bore, mechanical force has to exceed friction force. The friction force depends on component properties, such as the geometrical dimensions (especially diameters and axial lengths), structural stiffness as well as the surface conditions (e.g. surface roughness). Also, machine properties such as misalignment of the pressing machine tool with respect to the work piece holder and cylindrical bore axis can affect the process. None of these properties is measured directly for each component. Depending on the product type, the nominal values of the component properties may differ.

During the fitting process the force acting on the pressing machine tool and its current vertical position are measured jointly using special sensors. In order to control and monitor the process both signals are analyzed during and after the process by the process control unit, for example using gradient detection algorithms. The change in measured position is the sum of both the elastic deformation of the pressing machine (driven by the force and structural machine stiffness) and the actual movement of the inner component relative to the bore. Figure 8 demonstrates how the machine tool force behaves along the vertical position. The most important features extracted from this procedure are stored in the manufacturing execution system. Table 1 depicts a selection of these and lays out dependencies.

Staking In staking the deformation process of the material is mainly driven by two factors. One factor is the material yield strength and hardening behavior. The other factor is the compression stress. The latter is imposed by the force acting on the tool and by the contact area between tool and component. If several staking processes are happening on the same component with several bore positions, the unknown material state is likely to be similar, confounding the observed force values. Again a typical force and position curve of a staking process as well as a description of extracted features including their interrelations are demonstrated in Figure 9 and Table 2, respectively.

Often, press-in and staking processes are run on the same machine. Unobserved machine properties such as structural stiffness, tool misalignment and force sensor positions may impact both processes. Furthermore, confounding can be caused by similarities of the component properties (same component with several bores) or due to batch effects.

Figure 10 depicts an exemplary press-in and staking process. Depending on the production cell and process step these might vary slightly in size and complexity along the production line. From Figure 10 we can see that the staking process is larger and more complex than the press-in process. This reflects the fact that a staking process is an extension of a press-in process.

Table 1: Description of relevant force and position features extracted from the measurements during the press-in process.

Parameter	Description
F_{offs}	Average force measured during closing of the gap between tool and work piece. This depends on the acceleration and mass of the tool as well as the force sensor measurement error.
F_1	Press-in force (friction force) required to reach the block position (S_1) as extracted from gradient detection algorithm.
F_2	Maximum force at the end of the process. This force depends on the structural stiffness and the difference between the positions S_2 and S_1.
S_0	Axial position of tool at first force increase as extracted from gradient detection algorithm. Position depends on geometrical tolerances of the machine and components as well as on the calibration of the position sensor.
S_1	Axial position of tool at reaching the block position. Position depends on the relative displacement S_0. S_1 is affected by the length of the bore as well as by the elastic deformation of the machine. Deformation is influenced by the required press-in force F_1 and by the structural machine and component stiffness. Once this gradient position is detected by the process control unit during process execution the command for reversing the tool movement is given.
S_2	Maximum tool position where the tool direction of movement is effectively reversed. This is affected by the position S_1, by the time required for computation of the gradient point, the tool speed, by the inertial mass of the tool and by the effective structural stiffness.

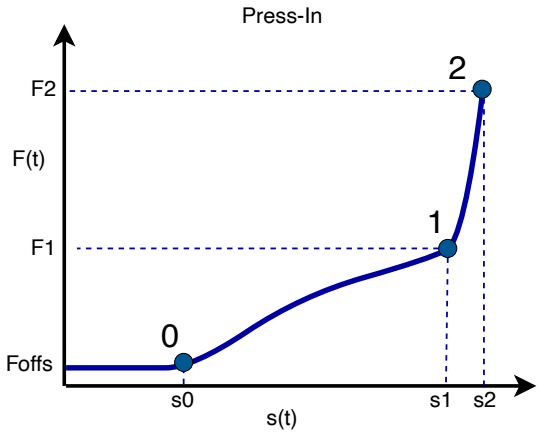


Figure 8: A typical Force (y-axis) and Position (x-axis) curve during press-in illustrating the characteristics of the process.

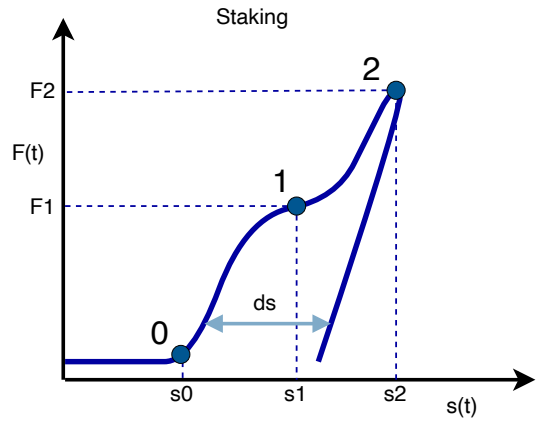


Figure 9: A typical Force (y-axis) and Position (x-axis) curve during staking illustrating the characteristics of the process.

Table 2: Description of relevant force and position features extracted from the measurements during the staking process.

Parameter	Description
F_1^*	Staking force required to reach the turning point of increasing structural stiffness as extracted from the detection algorithm. Once this turning point is detected by the process control unit during process execution the process command for reversing the tool movement is given after an additional, pre-defined force delta has been exceeded.
F_2^*	Maximum force at the end of the process. This force depends on the structural stiffness and the difference between the positions S_2^* and S_1^*.
S_0^*	Axial position of tool at first force increase as extracted from gradient detection algorithm. Position depends on geometrical tolerances of the machine and components as well as on the calibration of the position sensor.
S_1^*	Axial position of tool at reaching the turning point. Position depends on the relative displacement S_0. S_0 is affected by the length of the bore as well as by the elastic deformation of the machine. Deformation is influenced by the required press-in force F_1^* and by the structural machine and component stiffness.
S_2^*	Maximum tool position where the tool direction of movement is effectively reversed. This is affected by the position S_1^*, by the time required for computation of the turning point, the tool speed, by the inertial mass of the tool and by the effective structural stiffness.
ΔS^*	Difference in axial position of the tool after removing the pressing force measured at a pre-defined remaining force value. This indicates the true deformation length in the bore without the elastic machine deformation.

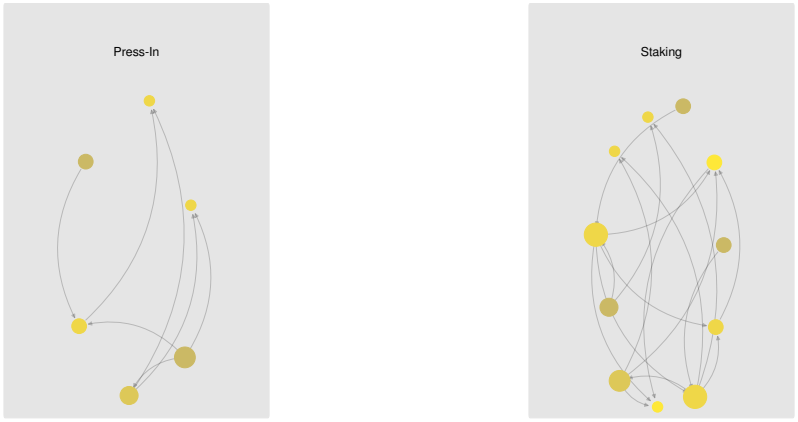


Figure 10: Process DAGs of an exemplary press-in and staking process. Node size increases with the number of out-edges. Node color gets brighter with the number of in-edges.

Appendix B. Proof of Proposition 1

Proof It suffices to show $\Pi_L \subseteq \Pi_{L'}$. So, let $\pi \in \Pi_L$ and consider any two nodes $v, w \in V$. If $v \in V_s$ and $w \in V_t$ with $s \neq t$, then we must have $s < t$ because L is a layered DAG. Since L' is based on the same partition $\mathcal{L}$, we have $\pi \in \Pi_{L'}$. If instead $v, w \in V_s$ for some $s \in [K]$, then $L_s = L'_s$. We conclude that v is a descendant of w in L if only if this is the case in L'. Thus, again $\pi \in \Pi_{L'}$. ■

Appendix C. Semisynthetic data generator

The semisynthetic data generation tool is available through the Python library `causalAssembly` (`https://github.com/boschresearch/causalAssembly`). In the library, we also provide the semisynthetic data itself (and for convenience one set of datasets with sample size $n = 500$) as well as the partial ground truth. All benchmark results presented both in the main text and the supplementary sections will also be accessible through the library.

C.1. Quality of the DRF procedure

To avoid overfitting when learning DRFs, we grow $N = 2000$ trees on a random subset of the training data for each conditional distribution. When forming each tree, the subsets are split again; one part is used to grow the tree, the other part is used to populate the leave nodes. We choose the Gaussian kernel for the MMD splitting criterion with bandwidth set according to the *median heuristic* (Gretton et al., 2012). Table 5 in Section D.3 gives an overview over the all relevant hyperparameters.

In the interest of quality control, we compare correlation patterns in the synthetic data largely agree with those of the real data as shown in Figure 11. Some divergence is to be expected as the

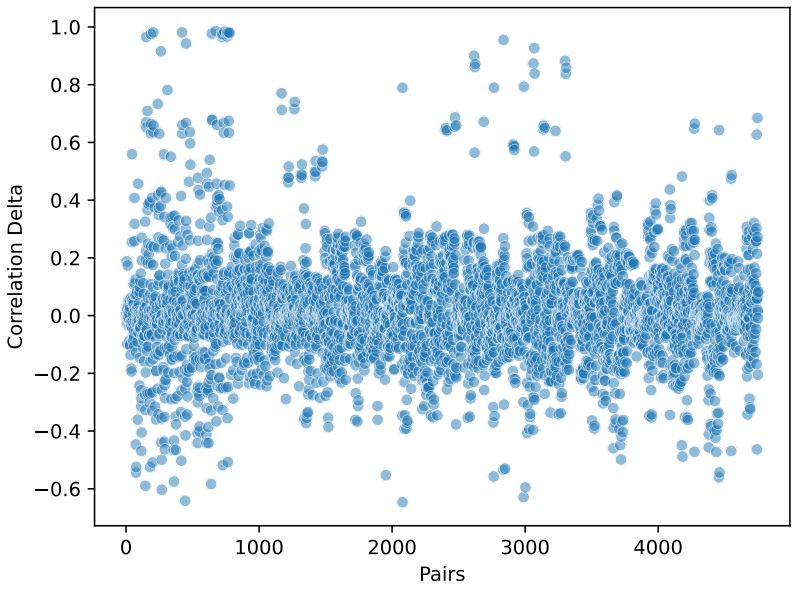


Figure 11: Pairwise correlation agreement between semisynthetic and real data. We depict the difference between the pearson correlation of the same pair in one draw of the semisynthetic data and the real production line data. Sample size is $n = 15.581$ for semisynthetic and real data.

real data comes with a wide range of confounding problems which our procedure deals with by employing the DRF procedure described in Section 3.4. Figure 11 confirms that for most of the pairs among the $p = 98$ variables, correlation patterns agree.

C.2. Influence of initial learning

Before investigating the influence of the absence of prior knowledge, we engage in a brief comparison of between and within process edge recovery. Figure 12 demonstrates that cross-process edges are indeed not recovered more easily as opposed to within-process edges. If any, the opposite is true and within-process edges achieve somewhat better precision and recall scores.

Turning to the absence of prior knowledge, learning edges within and across processes using causal discovery algorithms only without prior knowledge leaves distinct traces. Benchmarking outcomes are then tilted in favor of the routine employed in this initial step. We demonstrate this on processes that take place in Station 3 consisting of 16 nodes. We ignore existing process knowledge and learn from the data only. Figures 13 and 14 report SHD and F_1 score between the ground truth DAG learned by the initial causal discovery procedure and results based on corresponding semisynthetic data.

Next to overall patterns left by the initial learning step, we present sensitivity results with respect to measurement scale. The real data obtained from the process control unit carry a wide range of different measurement scales. These scales should ideally be irrelevant with respect to causal relationships. In a first step, we keep the raw data scale both for the initial learning and for the

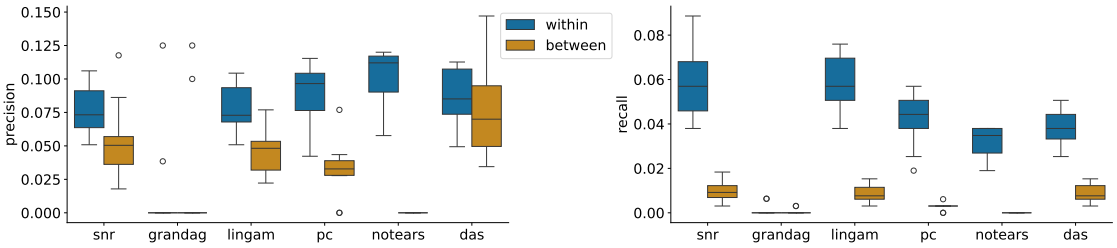


Figure 12: Comparison between within and cross process edges. Simulation run with 50 repetitions and sample size $n = 5000$. Precision and recall scores are relative to the amount of within and cross-process edges, respectively.

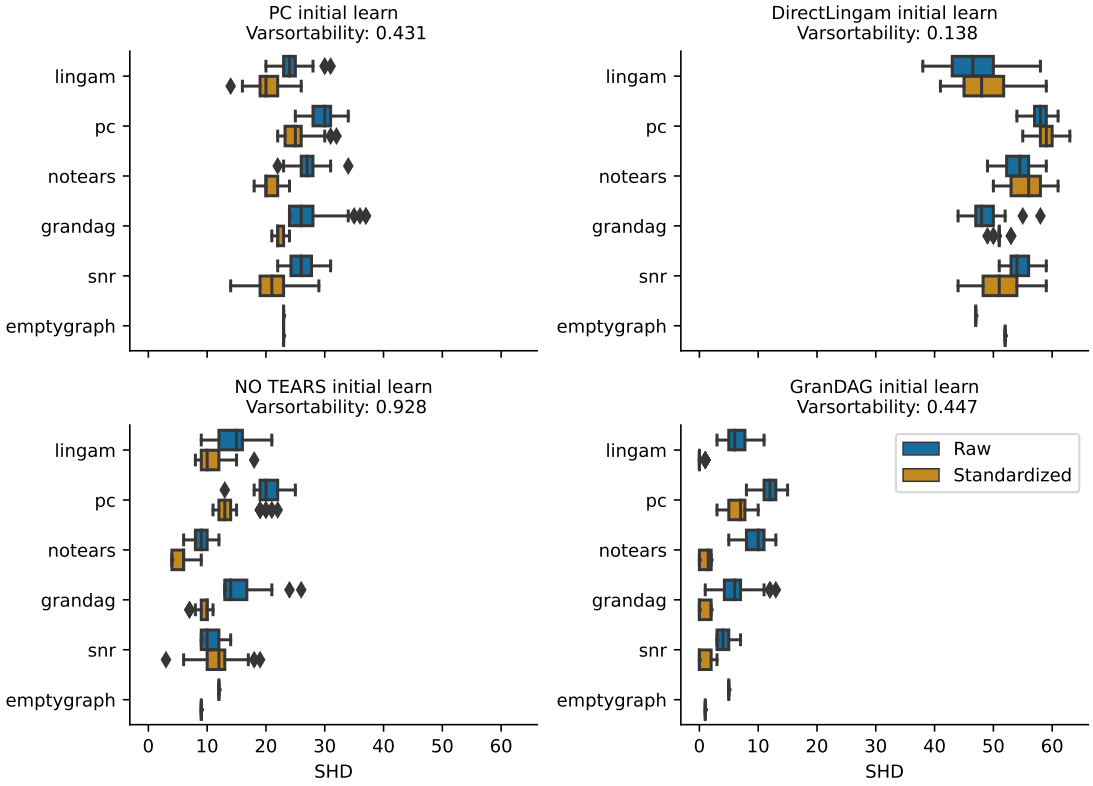


Figure 13: SHD for initial learning based on chosen causal discovery algorithms. Simulation run with 50 repetitions.

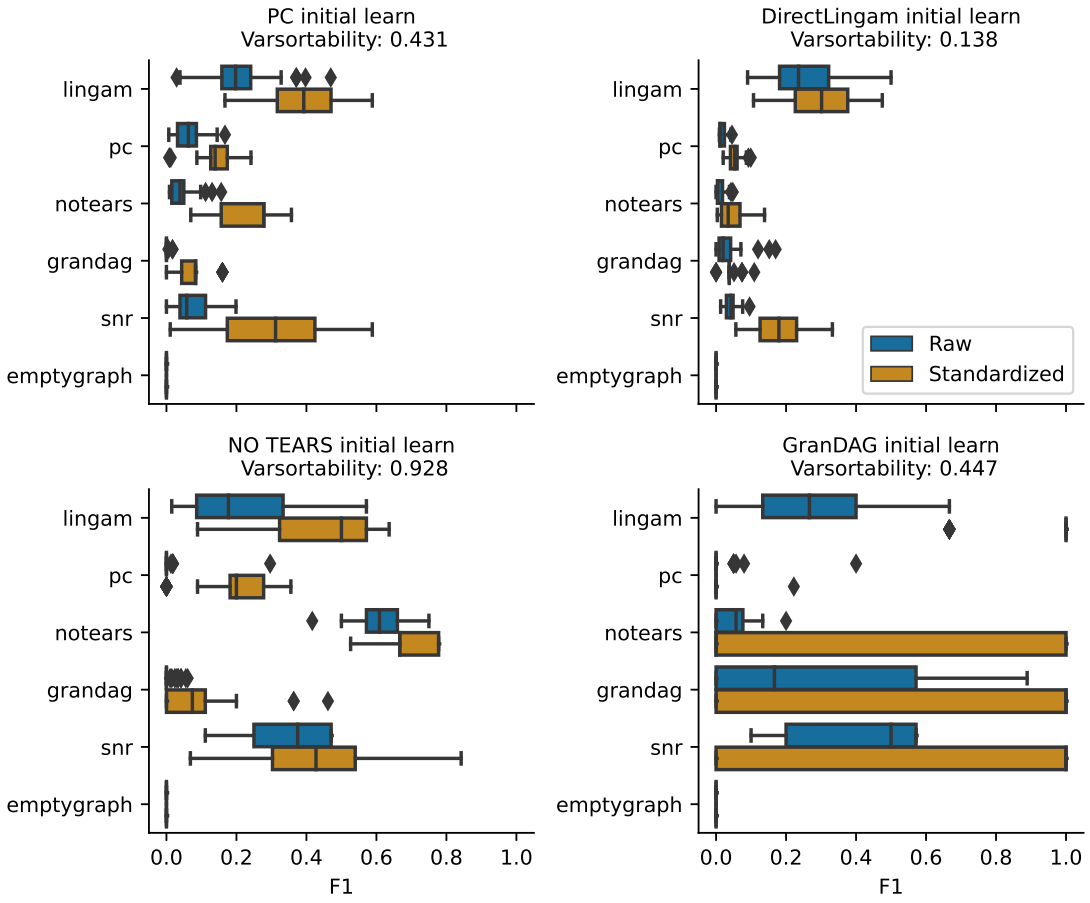


Figure 14: F_1 score for initial learning based on chosen causal discovery algorithms. Simulation run with 50 repetitions.

benchmark run based on the obtained ground truth. We report the average varsortability across simulation runs. In a second step, we standardize all data. We also report results for the empty graph to keep track of the number of edges inferred during the initial graph learning step. If the SHD of the empty graph changes under standardization then data scale has an effect on the initial ground truth.

DirectLiNGAM, NOTEARS, and GraN-DAG show a clear performance edge when the ground truth is originally obtained using the respective routine. This pattern is most clear in Figure 14. The PC algorithm does not seem to benefit too greatly which can to some degree be attributed to the random DAG completion. In contrast to the remaining methods, the PC algorithm is agnostic to measurement scale. DirectLiNGAM is in theory also scale agnostic, however scaling may influence the adjacency matrix estimation via adaptive lasso when predictors are not standardized. Regarding NOTEARS and GraN-DAG measurement units play a substantial role. NOTEARS obtains a ground truth with varsortability close to one on unstandardized data and recovers a very sparse DAG on standardized data. This suggests that NOTEARS essentially orders variables according to their variances when learned on raw data. GraN-DAG results are highly variable throughout all initial learning methods. The instability becomes particularly severe when applied to standardized data.

Overall, except for the PC algorithm there are clear traces left by initial learning algorithms applied to process cell structure learning discovery. In this context, the question arises whether to simply select a random DAG to represent process or station relationships. Following the proposed procedure and given sufficient sample size, the DRFs will be able to produce weights that allow drawing from such a ground truth graph. The choice of random DAG however influences strongly how close resulting synthetic data to the real world complexity will be. In the complete absence of ground truth knowledge, such a procedure might be an alternative to specifying a random SCM.

Appendix D. Benchmarking

D.1. Algorithms and metrics

Algorithms The following well-known structure learning algorithms are employed in our work.

PC algorithm (Spirtes et al., 1993) (`pc`)

A constraint based algorithm that in a first step performs conditional independence tests in a resource efficient way returning a skeleton and in a second step orients as many edges as possible. We use the stable version of the algorithm proposed by Colombo and Maathuis (2014). Conditional independence tests are performed using Fisher z-transformed (partial) Pearson correlation coefficients. Since the PC algorithm outputs a completed partially directed acyclic graph (CPDAG), we choose in each run a random DAG that is a member of the MEC implied by the CPDAG.

DirectLiNGAM (Shimizu et al., 2011; Hyvärinen and Smith, 2013) (`lingam`)

The direct method to learning the LiNGAM model (Shimizu et al., 2006) referred to in Section 2.3. First, the procedure detects source variables by employing entropy-based measures to find independence between the error variables. The *most independent* variable is placed first in the causal ordering and its effect removed from the other variables. As the ordering of the residuals resembles that of the variables themselves the algorithm continues to find the next variable in the ordering by finding the next *source*-residual.

NOTEARS (Zheng et al., 2018, 2020) (`notears`)

A score-based method that recasts the discrete DAG constraint as a continuous equality constraint. This allows for the application of a wide range of techniques for non-convex programs. In NOTEARS a linear SCM is assumed and for each variable, a feed-forward neural network is trained that gets all

other variables as input. The graph structure is then dynamically read-off from the networks during training by aggregating the weights of the input layer of each network. Acyclicity of the graph is enforced by training the networks simultaneously in an augmented Lagrangian schema such that the matrix exponential of the adjacency matrix has small trace.

GraN-DAG (Lachapelle et al., 2020) (`grandag`)
A score-based method and an extension of NOTEARS. In contrast to NOTEARS, the relation structure is read-off from the networks by aggregating all the weights till the output layer of the neural networks individually.

SCORE and *DAS* (Rolland et al., 2022; Montagna et al., 2023) (`score, das`)
A order-based method for learning nonlinear additive Gaussian noise models which has been shown to be identifiable from observational data (Peters et al., 2014). The method evolves around estimating the causal ordering based in approximating the score (the map $\nabla \log p(x)$ for a differentiable density $p(x)$) of the data distribution. Under this model, leaf nodes indicate a constant entry in the associated diagonal element of the score's Jacobian. This allows for sequential leaf node identification implying a causal ordering. The method concludes with a pruning step to remove spurious edges. Its extension *discovery at scale* (DAS) reduces the complexity of the pruning step.

sortnregress (Reisach et al., 2021) (`snr`)
A diagnostic tool to expose data scenarios with high varsortability. The procedure sorts variables by increasing marginal variance and then regresses each variable on their ancestors adding an ℓ-1 type penalty to induce sparsity.

Metrics The choice of evaluation metrics for assessing causal discovery outcomes matters (see Gentzel et al., 2019). We employ a selection of common structural performance metrics and note that in a full scale benchmark these should be extended to include a greater variety of evaluation measures.

Precision, Recall and F_1 score:
Precision in the graphical context refers to the fraction of correctly identified edges among all edges inferred. Recall, sets these 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. $F_1 = 2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$.

Structural hamming distance (SHD):
The SHD counts how many edge insertions, deletions, and reversals are necessary to turn the estimated DAG into the ground truth DAG.

Structural Interventional Distance (SID) (Peters and Bühlmann, 2015):
The SID counts the number of incorrectly inferred interventional distributions in the estimated DAG when compared to the ground truth DAG.

Harmonization We employ two harmonization strategies *CPDAG-transform* and *average-dag*.

(i) We transform all graphs, including the ground truth DAG, to their completed partially directed acyclic graph (CPDAG) representing the associated MEC (see Andersson et al., 1997, for details). Metrics will be calculated on CPDAGs only. Note that the SID is defined for DAGs and will not be part of the evaluation metrics following *CPDAG-transform*.

(ii) The second strategy, *average-dag*, involves choosing one member DAG from the implied MEC of algorithms that do not output a DAG. We then enumerate all DAGs in the MEC, calculate their structural Hamming distance (SHD) with respect to the ground truth DAG, and take the one whose SHD is closest to the average SHD. When running experiments on the full line, enumerating

Table 3: All benchmarks are run on a Linux machine with 32 GB Memory and 8 cores.

Result	Compute time
Full assembly line benchmark Fig 6 and 19 in the main text	16.90 hours
Station 1 benchmark Fig 15 and 20	1.07 hours
Station 2 benchmark Fig 16 and 21	4.39 hours
Station 3 benchmark Fig 7 and 22	2.22 hours
Station 4 benchmark Fig 17 and 23	3.54 hours
Station 5 benchmark Fig 18 and 24	2.33 hours

Table 4: Average varsortability over 100 draws using the semisynthetic data generation procedure for the full assembly line and for each production station.

	Varsortability
Full line	0.486122
Station1	0.669167
Station2	0.348319
Station3	0.330588
Station4	0.247204
Station5	0.210417

DAGs in their MEC becomes computationally infeasible, so we choose one random DAG from the associated MEC.

D.2. Additional benchmarking results

We present additional benchmarks for the CPDAG harmonization strategy and the remaining production stations from the MEC enumeration strategy. Average varsortability before standardization is reported in Table 4 and Table 3 states compute time for each of the benchmark scenarios. As expected, varsortability ranges between 0.7 in the first station and 0.22 in station five. The full assembly line exhibits varsortabiliy close to 0.5 indicating that ordering variables by their marginal variances does not give any insights regarding the causal ordering. In each simulation run, we sample from the DRF regrown on the corresponding layer-induced subgraphs.

We find that results vary substantially across stations. Figure 15 and 20 report results for the first station. DirectLiNGAM performs best while NOTEARS is not able to detect any edges. GraN-DAG results are very unstable throughout all benchmarks. The PC algorithm performs only slightly worse than DirectLiNGAM and SCORE's performance dropped when transformed to the corresponding MEC.

In Figure 16 and 21 Station 2 two benchmarks are depicted. In this case, NOTEARS performs best in the context of recovering the true DAG. When converted to CPDAGs, NOTEARS performance drops and PC algorithm performance increases drastically. Figure 22 reports Station 3 benchmarks for the CPDAG harmonization strategy. In this case the PC algorithms outperforms all other routines when compared to the ground truth CPDAG. In Figures 17 and 23 Station 4 benchmarks suggest again that the PC algorithm outperforms all other causal discovery algorithms. On station five (Figure

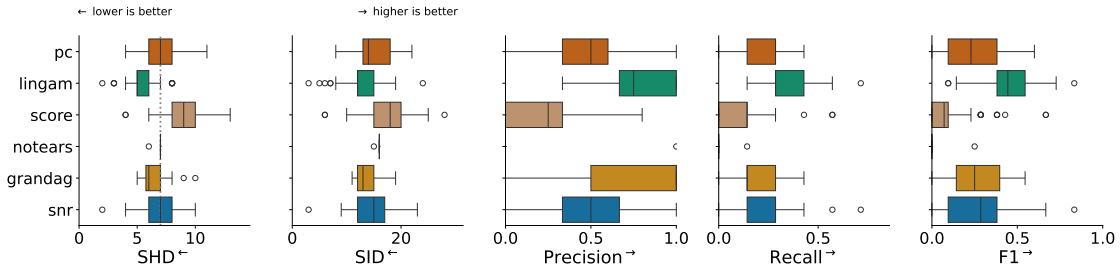


Figure 15: Chosen metrics for assessing causal discovery outcomes on Station 1 ($|V_1 \cup V_2| = 6$) using 100 simulation runs on a sample size of $n = 500$.

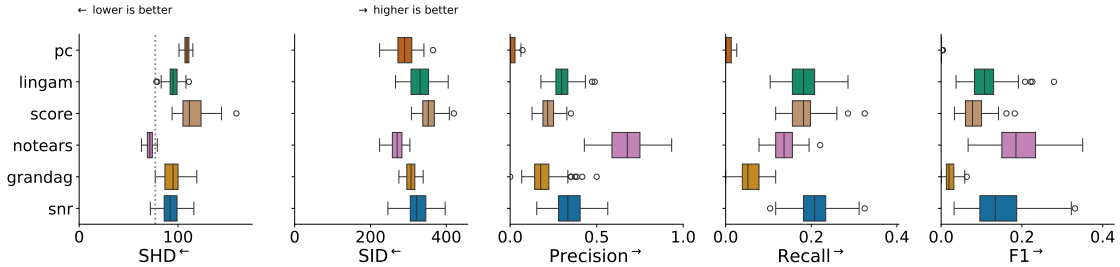


Figure 16: Chosen metrics for assessing causal discovery outcomes on Station 2 ($|V_3 \cup V_4| = 34$) using 100 simulation runs on a sample size of $n = 500$.

18 and 24) SCORE overall performs best on DAG level and the PC algorithm draws equal for the CPDAG harmonization strategy. Laslt, Figure 19 illustrates the full line benchmark results for the CPDAG harmonization strategy. As mentioned in the main text, NOTEARS performs best overall, GraN-DAGs behavior is very unreliable, and the PC algorithm is the runner-up. Yet, none of the procedures is able to perform significantly better than *random regress*.

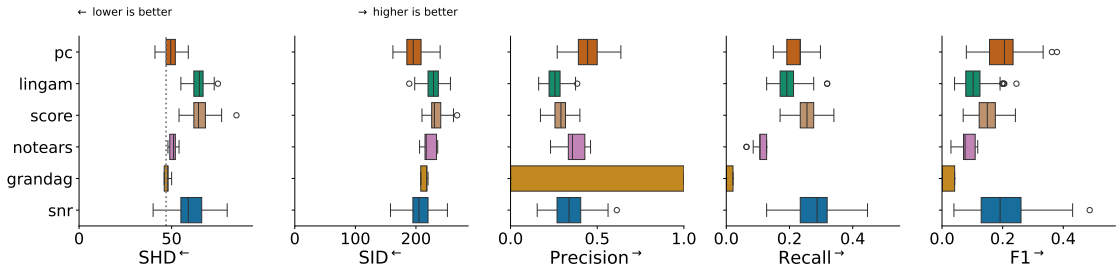


Figure 17: Chosen metrics for assessing causal discovery outcomes on Station 4 ($|V_7 \cup V_8| = 26$) using 100 simulation runs on a sample size of $n = 500$.

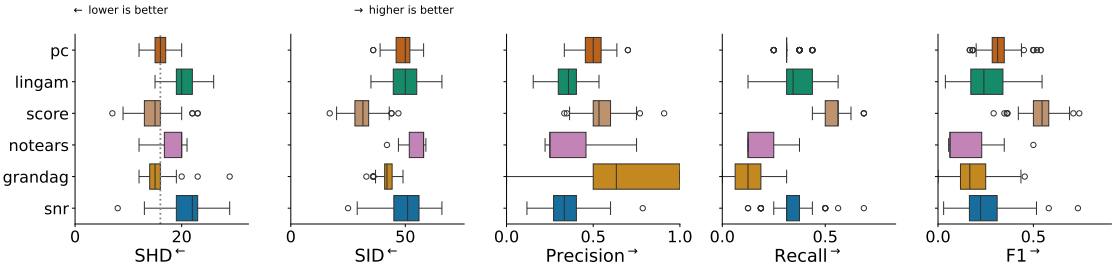


Figure 18: Chosen metrics for assessing causal discovery outcomes on Station 5 ($|V_9 \cup V_10| = 16$) using 100 simulation runs on a sample size of $n = 500$.

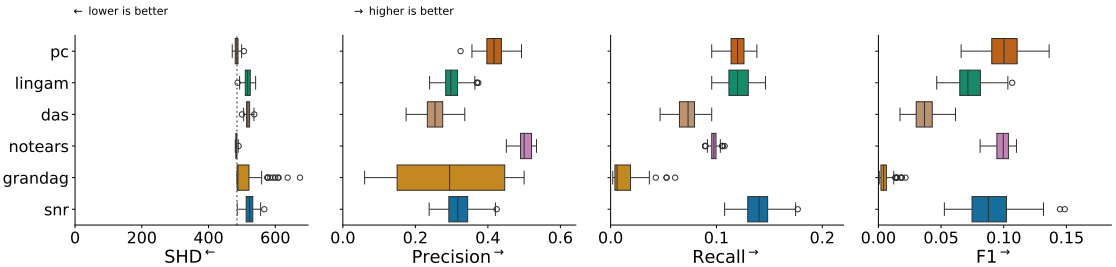


Figure 19: Chosen metrics for assessing causal discovery of the full assembly line based on CPDAG comparisons ($|V| = 98$) using 100 simulation runs on a sample size of $n = 5000$. The dashed line corresponds to the SHD between the empty graph and ground truth indicating the number of edges in the ground truth CPDAG.

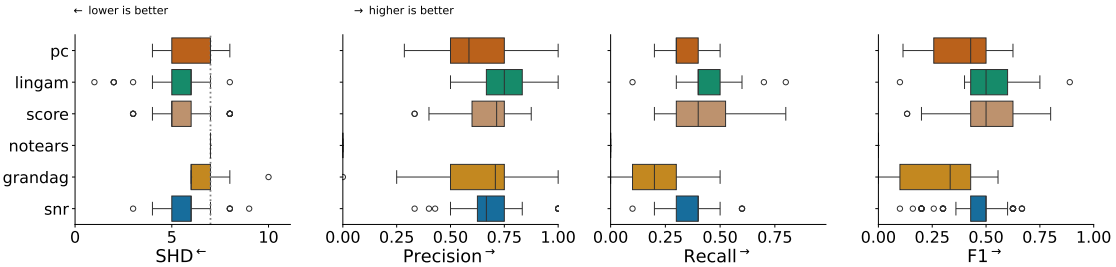


Figure 20: Chosen metrics for assessing causal discovery outcomes on Station 1 ($|V_1 \cup V_2| = 6$) after converting all DAGs to CPDAGs using 100 simulation runs on a sample size of $n = 500$.

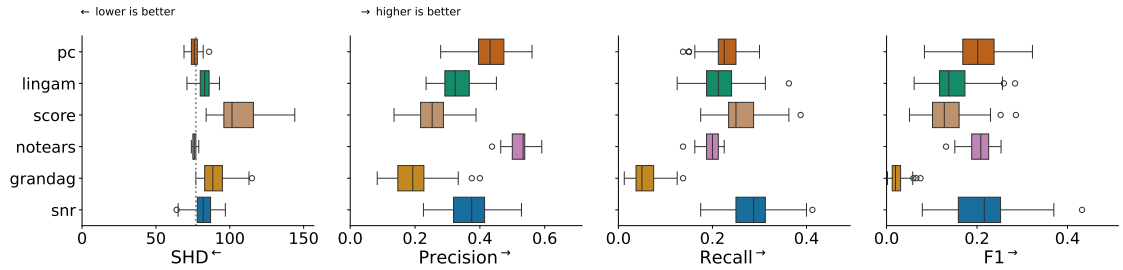


Figure 21: Chosen metrics for assessing causal discovery outcomes on Station 2 ($|V_3 \cup V_4| = 34$) after converting all DAGs to CPDAGs using 100 simulation runs on a sample size of $n = 500$.

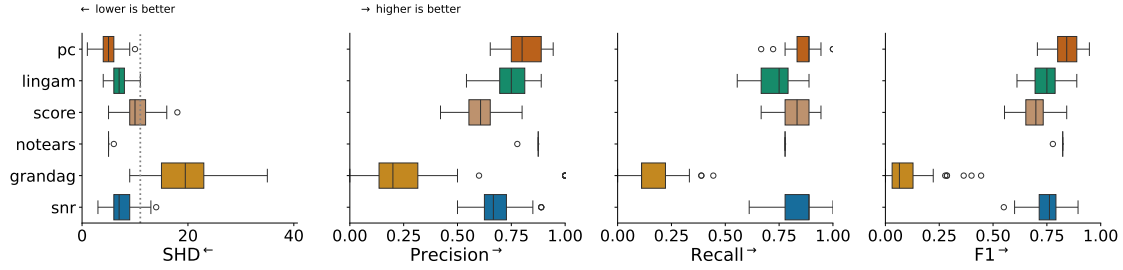


Figure 22: Chosen metrics for assessing causal discovery outcomes on Station 3 ($|V_5 \cup V_6| = 16$) after converting all DAGs to CPDAGs using 100 simulation runs on a sample size of $n = 500$.

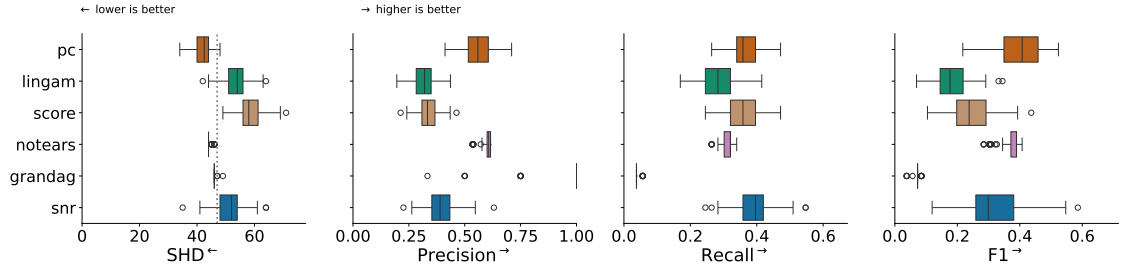


Figure 23: Chosen metrics for assessing causal discovery outcomes on Station 4 ($|V_7 \cup V_8| = 26$) after converting all DAGs to CPDAGs using 100 simulation runs on a sample size of $n = 500$.

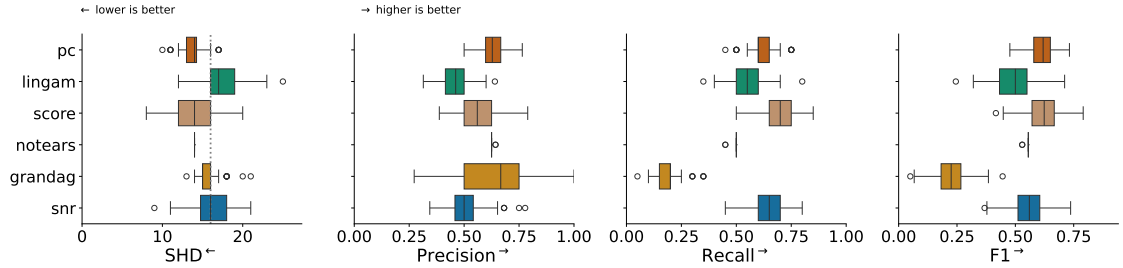


Figure 24: Chosen metrics for assessing causal discovery outcomes on Station 5 ($|V_9 \cup V_10| = 16$) after converting all DAGs to CPDAGs using 100 simulation runs on a sample size of $n = 500$.

D.3. Hyper and tuning parameters

Table 5: Hyper and tuning parameter choices for routines employed in this work.

	Parameters
SpAM	`num_folds` = 5 `cv_loss` = MSE `lambda_choice` = 1se `num_bsplines` = 6
DRF	`min_node_size` = 15 `num_trees` = 2000 `splitting_rule` = FourierMMD
PC Algorithm	`indep_test` : fisherz `alpha` = 0.05
NOTEARS	`lambda_1` = 0.1 `loss_type` = ℓ_2 `max_iter` = 100 `h_tol` = $1e-8$ `rho_max` = $1e+16$ `w_threshold` = 0.3
GraN-DAG	`input_dim` = V `hidden_num` = 2 `hidden_dim` = 10 `batch_size` =64 `lr` = 0.001 `iterations` = 10000 `model_name` = NonLinGaussANM `nonlinear` = leaky-relu `optimizer` = rmsprop `h_threshold` = $1e-8$ `lambda_init` = 0.0 `mu_init` = 0.001 `omega_lambda` = 0.0001 `omega_mu` = 0.9 `stop_crit_win` = 100

	`edge_clamp_range` = 0.0001
SCORE, DAS	`eta_G` = 0.001 `eta_H` = 0.001 `alpha` = 0.05 `n_splines` = 10 `splines_degree` = 3

Table 5 reports all hyperparameters and tuning parameter choices for benchmark results as well as for SpAM and DRF fitting. All benchmark algorithms were taken off-the-shelf to avoid providing any of the procedures with an unfair advantage.

A.3 Nonlinear Causal Discovery for Grouped Data

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.

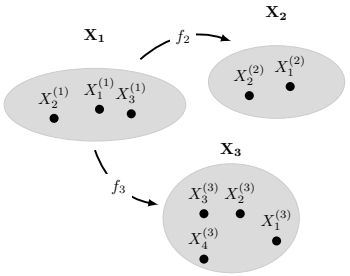


Figure 1: An example of a grouped additive noise model with three groups and varying group sizes.

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 $\mathbf{u} \coloneqq \mathbf{x}_g - f_g(\mathbf{x}_j)$ and further $\nu \coloneqq \log p_{\mathbf{N}_g}$, $\xi \coloneqq \log p_{\mathbf{X}_j}$, with arguments $\mathbf{u}, \mathbf{x}_j$, respectively. Additionally, $\pi_1(\mathbf{x}_j, \mathbf{x}_g) \coloneqq \log p_{\mathbf{X}_j, \mathbf{X}_g}(\mathbf{x}_j, \mathbf{x}_g)$ 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 : $S \coloneqq \{1, \ldots, p\}, \quad \pi \coloneqq [\cdot]$.
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
  $S \coloneqq S \setminus \{g^*\}$
  $\pi \coloneqq [g^*, \pi]$
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 $pa_{\mathcal{G}}(j) \coloneqq \{g \in pa_\pi(j) : f^{(k)}_{j,g,h} \neq 0\}$

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,

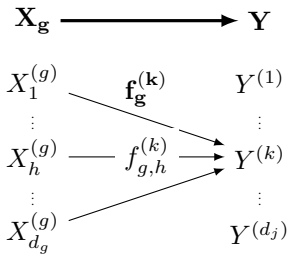


Figure 2: Overview of MURGS notation.

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 $\|f\|^2 \coloneqq \int_{\mathcal{Z}} f^2(z) dP_Z = \mathbb{E}[f^2]$. 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 $pa_j \coloneqq pa_{\mathcal{G}^\pi}(j)$. We define $Y^{(k)} \coloneqq X_k^{(j)}$ such that we may drop the corresponding subscripts, i.e. $f^{(k)}_{g,h} \coloneqq f^{(k)}_{j,g,h}$. 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 $f^{(k)}_{g,h} \coloneqq f^{(k)}_{g,h}(X_h^{(g)})$. 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 $m^* = \arg\max_m \frac{1}{m} \left(\sum_{l=1}^{m^*} \hat{s}_g^{(k_l)} - \sqrt{d_g}\lambda \right)$

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

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*

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

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

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

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 $S_{g,h}$.

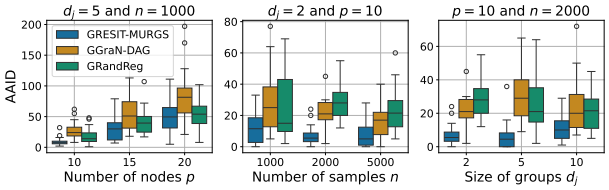


Figure 3: Performance of the algorithms under varying dimensions of the GANM.

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

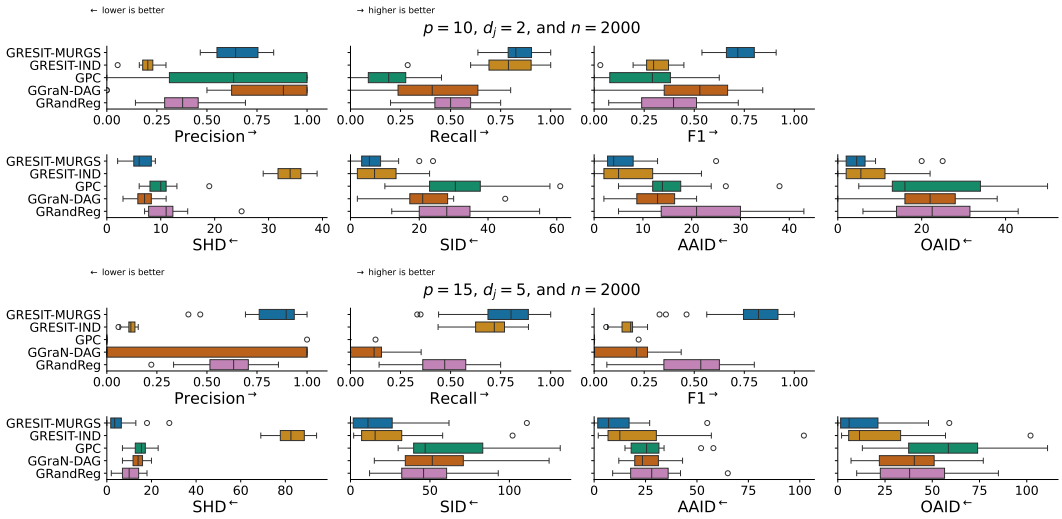


Figure 4: Simulation results based on 20 repetitions.

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.

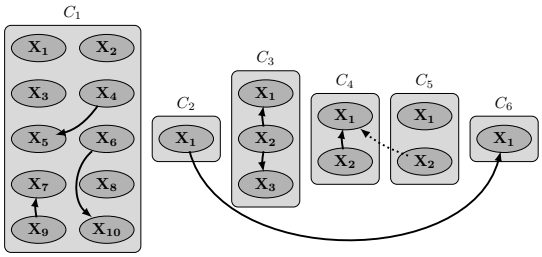


Figure 5: Learned causal edges from the real-world dataset using *GRESIT-MURGS*

.

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

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

and

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

where $\tilde{\nu} \coloneqq \log p_{\mathbf{N}_1}$ and $\eta \coloneqq \log p_{\mathbf{X}_2}$. 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 $\tilde{\mathbf{u}} \coloneqq \tilde{\mathbf{u}}(\mathbf{x}_1,\mathbf{x}_2) \coloneqq \mathbf{x}_1 - f_1(\mathbf{x}_2)$. 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

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

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.

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

We find

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

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

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

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.

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

Using the identity

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

where $\mathbf{A} \coloneqq \mathbf{A}(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{d_{x_1} \times d_{x_1}}$. 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., $\pi_1 \coloneqq \pi_1(\mathbf{x}_1, \mathbf{x}_2)$ 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.

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

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, $pa_Q \coloneqq pa_{\mathcal{G}_0}(Q) \setminus R$ and the set of parents without Q in $\mathcal{G}_0'$, i.e, $pa_R \coloneqq pa_{\mathcal{G}_0'}(R) \setminus Q$. Denote their union by $S \coloneqq pa_Q \cup pa_R$. 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 $(u_{h1},\ldots,u_{hd_j})^T \in \partial\|\cdot\|_\infty|_{(\|\mathbf{f}_g^{(1)}\|,\ldots,\|\mathbf{f}_g^{(d_j)}\|)^T}$ 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 k^{th} canonical unit vector in $\mathbb{R}^{d_j}$.

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

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

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

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 $\exists u^{(1)},\ldots,u^{(d_j)}$ such that $\sum_{k=1}^{d_j}|u|^{(k)} \le 1$ and $\exists v_1^{(k)},\ldots,v_{d_g}^{(k)}$ 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

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

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

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

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 $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ 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.

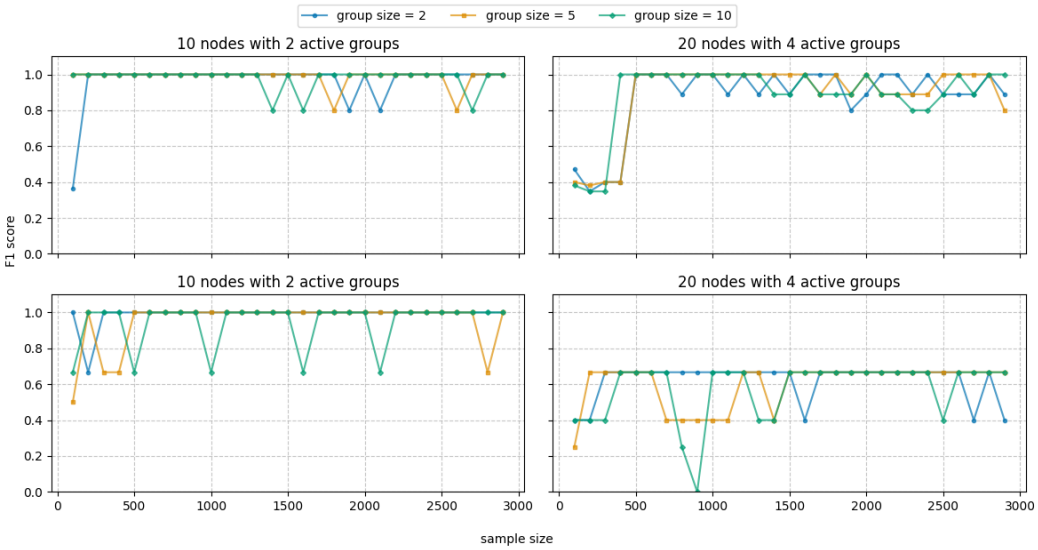


Figure 6: Feature selection capability of MURGS. Node size is $p = 10, 20$, with $2, 4$ active groups, respectively.

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.

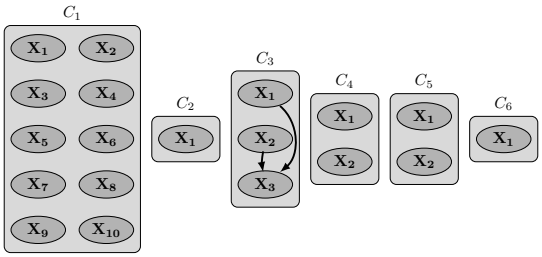


Figure 7: Learned causal edges from the real-world dataset using *GroupGraN-DAG*.

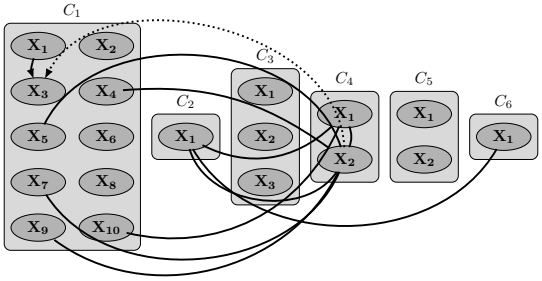


Figure 8: Learned causal edges from the real-world dataset using *GroupPC*.

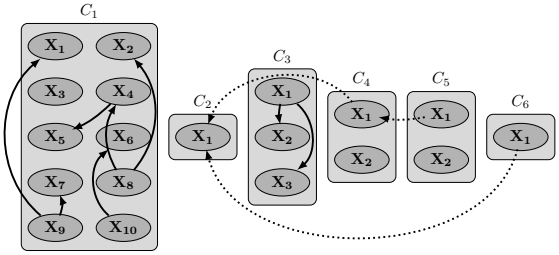
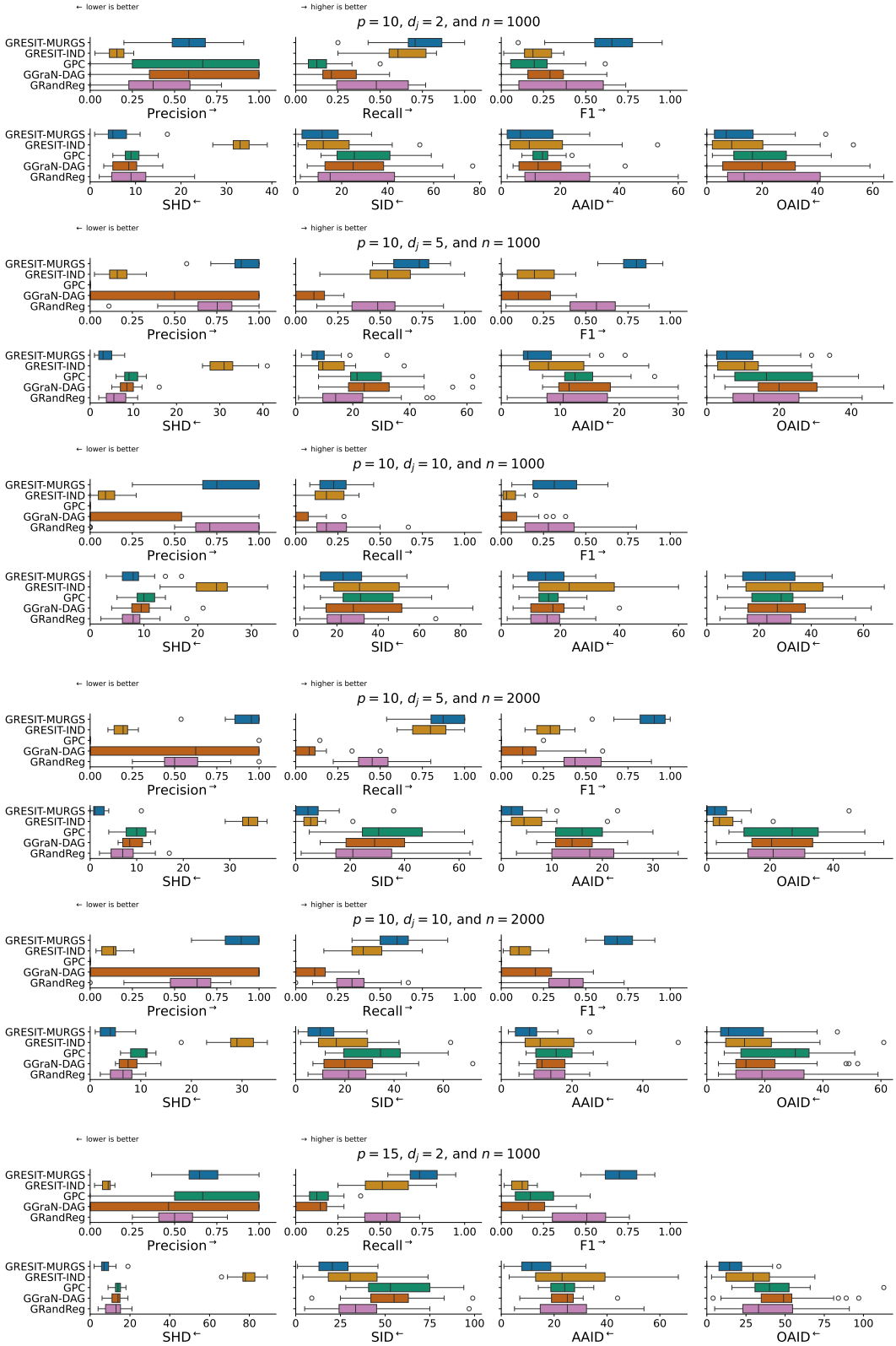
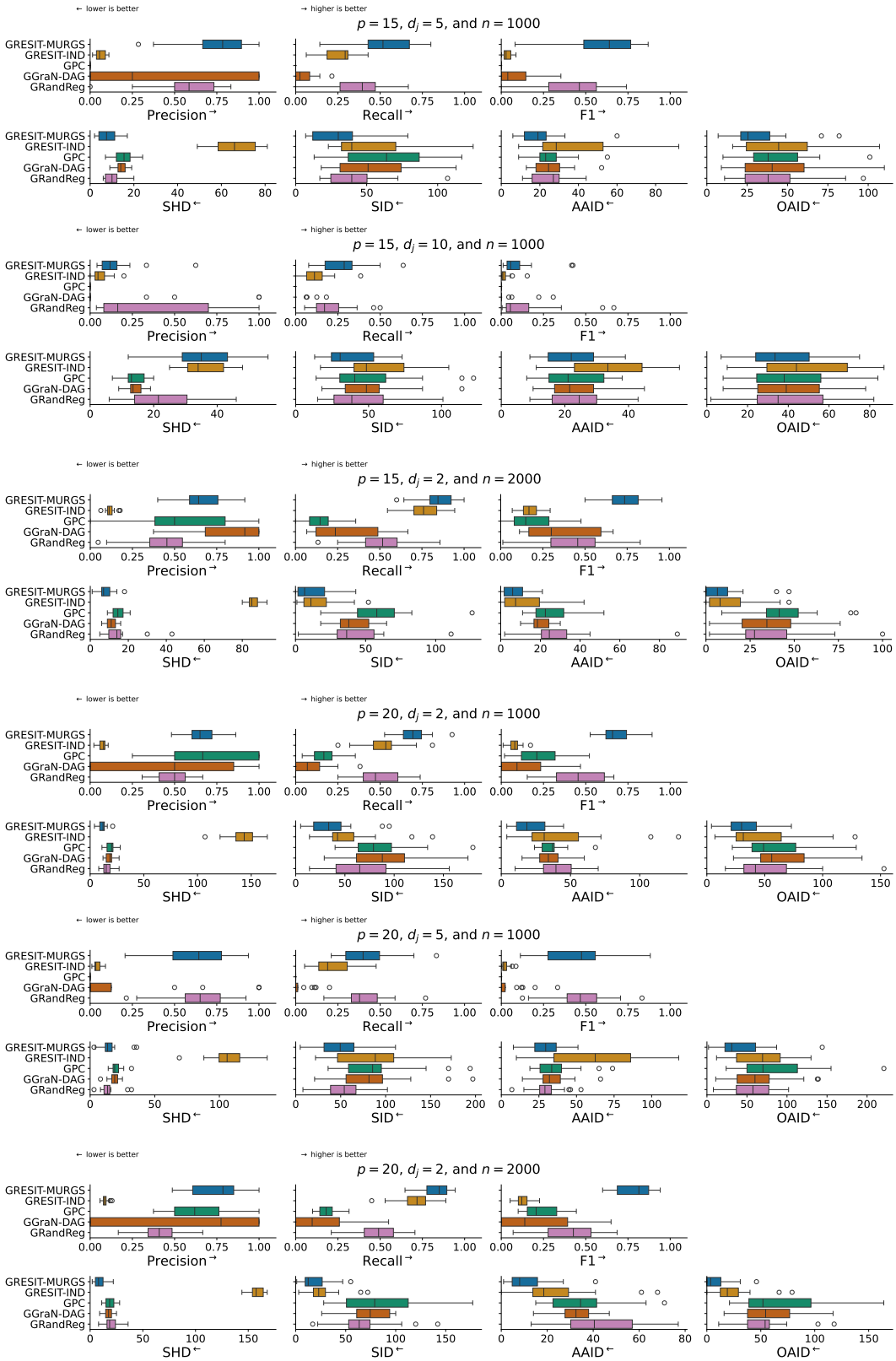


Figure 9: Learned causal edges from the real-world dataset using *GroupDirectLiNGAM*.


← 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 $n = 1000$
GRESIT-MURGS
GRESIT-IND
GPC
GGraN-DAG
GRandReg
Precision →
Recall →
F1 →
SHD ←
SID ←
AAID ←
OAID ←
$p = 15$, $d_j = 10$, 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 $n = 2000$