Partial Homoscedasticity in Causal Discovery With Linear Models

Jun Wu and Mathias Drton

Abstract—**Recursive linear structural equation models and the associated directed acyclic graphs (DAGs) play an important role in causal discovery. The classic identifiability result for this class of models states that when only observational data is available, each DAG can be identified only up to a Markov equivalence class. In contrast, recent work has shown that the DAG can be uniquely identified if the errors in the model are homoscedastic, i.e., all have the same variance. This equal variance assumption yields methods that, if appropriate, are highly scalable and also sheds light on fundamental information-theoretic limits and optimality in causal discovery. In this paper, we fill the gap that exists between the two previously considered cases, which assume the error variances to be either arbitrary or all equal. Specifically, we formulate a framework of partial homoscedasticity, in which the variables are partitioned into blocks and each block shares the same error variance. For any such groupwise equal variances assumption, we characterize when two DAGs give rise to identical Gaussian linear structural equation models. Furthermore, we show how the resulting distributional equivalence classes may be represented using a completed partially directed acyclic graph (CPDAG), and we give an algorithm to efficiently construct this CPDAG. In a simulation study, we demonstrate that greedy search provides an effective way to learn the CPDAG and exploit partial knowledge about homoscedasticity of errors in structural equation models.**

Index Terms—**Causal discovery, covariance matrix, equal variance, graphical model, structural equation model.**

I. Introduction

A STRUCTURAL equation model (SEM) describes the stochastic dependence among a group of random variables in terms of noisy functional relationships between causes and their effects. In their interpretation as causal models, SEMs furnish models of the variables' joint distribution not only in observational studies but also under experimental interventions, providing a rigorous definition of the causal effect of an intervention [1], [2]. In this realm, the problem of causal discovery, also known as structure learning, is the problem of learning a causal model from available data by determining the direct causes of each variable with the help of statistical methods [3]. For this purpose, it is convenient to take a point of view of probabilistic graphical modeling and represent each SEM with the help of a directed graph, with vertices corresponding to the random variables at hand, and directed edges linking the functionally related variables [4]. We tacitly assume all considered SEMs to be recursive, i.e., the underlying graph is a directed acyclic graph (DAG).

Each DAG uniquely encodes a causal model, but when only observational data are available, different DAGs (i.e., different systems of structural equations) may be equivalent in the sense of defining the same statistical model for the observations. The target of inference then becomes an equivalence class of empirically indistinguishable SEMs. Hallmark theory for graphical models provide a characterization of SEMs in terms of conditional independence, show how this characterization leads to efficient decision criteria for the observational equivalence of two SEMs given by two different DAGs, and finally develop a graphical representation of the resulting Markov equivalence class by means of a completed partially directed acyclic graph (CPDAG). This theory is summarized, for instance, in [5].

The theory just mentioned holds for fully nonparametric models. Somewhat amazingly, the same arguments and results also apply to the widely considered special case of linear SEMs with Gaussian errors with arbitrary variances, where again only a Markov equivalence class may be inferred from observational data; see, e.g., the review in [6] or the recent work of [7], [8]. However, other modeling assumptions generally behave differently and lead to a setting where every DAG defines a unique model for observational data. This has been shown, for example, for linear models with non-Gaussian errors [9], [10] and nonlinear additive noise models [11], [12]. Similarly, additional assumptions may render linear Gaussian SEMs identifiable. The most prominent example in this direction are linear SEMs with Gaussian errors that all share a common variance [13]. Despite its restrictive nature, the equal variance setting plays a useful role as a test bed for developing scalable causal discovery algorithms [14], [15], [16], [17]. It also allows for a derivation of fundamental information-theoretic limits in the form of lower bounds on the sample size needed for consistent estimation of the DAG, which opens the door for optimality results on learning algorithms [18]. Further recent work related to equal variance models includes [19], [20], [21].

In this paper, we consider the realm of linear Gaussian SEMs that fall between the classical case with arbitrary error variances and the case with all error variances equal. To this

Manuscript received 26 April 2023; revised 21 September 2023; accepted 24 October 2023. Date of publication 1 November 2023; date of current version 5 December 2023. This work was supported by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 883818. *(Corresponding author: Mathias Drton.)*

The authors are with the Department of Mathematics, TUM School of Computation, Information and Technology, and the Munich Data Science Institute, Technical University of Munich, 80939 Munich, Germany (e-mail: jun1.wu@tum.de; mathias.drton@tum.de).

Digital Object Identifier 10.1109/JSAIT.2023.3328476

end we formulate a novel framework: partial homoscedasticity of the errors. Specifically, we encode the modeling assumptions in a partition over the variables, with the interpretation that the errors associated to the variables in the same partition block have equal variances. In this framework, the minimal and maximal partitions recover the previously studied cases. The minimal partition in which every variable forms a block of size one recovers the classical case of arbitrary error variances. The maximal partition in which all variables are in one block gives the equal variance case. Weakening the all-equal assumption to groupwise equal variances is natural for data sets in which only certain subgroups of variables are measured on similar scales (e.g., in multiomics data in biology). We will show that this weakening continues to be very favorable for identifiability of causal structure; e.g., direct causes are identifiable for any variable that appears in some equal variance constraint.

Our results are developed by first providing an implicit description of partially homoscedastic linear Gaussian SEMs, with a partition specifying groupwise equal variances. This description is based on conditional independence constraints as well as constraints we deduce from the equalities of error variances. We then characterize when two DAGs define the same partially homoscedastic linear Gaussian SEM. This characterization shows that because of the equal variance constraints, the existence of a pair of variables in the same block of the considered partition determines the orientation of the edges that have these two nodes as endpoints. Moreover, we generalize the concept of a CPDAG to the partially homoscedastic case and provide an algorithm for efficient construction of the CPDAG. This algorithm is an adjusted version of the general algorithm for constructing an equivalence class under background knowledge [22].

The remainder of this paper is organized as follows: In Section II, we introduce basic notation and necessary background for linear SEMs and their representation using DAGs. In Section III, we discuss the partially homoscedastic setup with groupwise equal error variances and derive the equal variance constraints that are needed in an implicit description of the models. In Section IV, we characterize the equivalence classes in our setup and give an algorithm to construct the CPDAG. In Section V we propose a greedy search scheme for model selection based on information criteria, which is seen to be effective in a simulation study. We conclude with a discussion in Section VI.

II. Background

A. Linear Structural Equation Models

A linear structural equation model (SEM) is specified via a linear system consisting of equations among variables $\{X_i : i \in V\}$ and random errors $\{\varepsilon_i : i \in V\}$, where V is an index set of size $|V| = p$. In terms of the random vectors $X := (X_i)_{i\in V}$ and $\epsilon := (\varepsilon_i)_{i\in V}$, the linear system can be written as

$$X = \Lambda^T X + \epsilon, \tag{1}$$

where $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{V\times V}$ is a matrix of coefficients representing the causal structure among the variables. Assuming that the matrix $I - \Lambda$ is invertible (I is the identity matrix), the linear equation system (1) is solved uniquely by $X = (I - \Lambda)^{-T}\epsilon$, with covariance matrix

$$\mathrm{Var}\,[X] = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}, \tag{2}$$

where Ω is the covariance matrix of the random vector of stochastic errors ϵ. We assume that the errors are independent such that $\Omega = \mathrm{diag}(\boldsymbol{\omega})$ is a diagonal matrix with all diagonal entries positive. Any specific SEM makes the assumption that a subset of the entries of Λ is zero. Such an assumption is naturally encoded in a directed graph. We will detail the connection between a linear SEM and its directed graph in Section II-C.

B. Graph Terminology

A directed graph $G = (V, E)$ is a tuple that pairs a set of nodes V with a set of edges $E \in V \times V$. The elements in E are ordered pairs of the form (i, j), $i \neq j$, encoding the edge $i \to j$ in the graph. We say that the edge (i, j) is an outgoing edge from the arrow tail i and an incoming edge to the arrow head j. The node i is a *parent* of j and the node j is a *child* of i. We denote the sets of all parents and children of a node i by $\mathrm{pa}(i)$ and $\mathrm{ch}(i)$, respectively. Similarly, the notation $\mathrm{an}(i)$ denotes the set of *ancestors* of i and $\mathrm{de}(i)$ denotes the set of *descendants* of i. For simplicity, we adopt the convention that $i \notin \mathrm{an}(i)$ and $i \in \mathrm{de}(i)$. If $\mathrm{an}(i) \cap \mathrm{de}(i) = \emptyset$, then node i does not lie on any directed cycle. If $\mathrm{an}(i) \cap \mathrm{de}(i) = \emptyset$ for every node i, then the graph is a *directed acyclic graph* (DAG). To distinguish those sets in different graphs, we use graph index as the subscript: pa_G, etc.

A *collider triple* in a directed graph $G = (V, E)$ is a triple of vertices (i, j, k) such that there are edges between i and j and between j and k, with j being a head on both these edges ($i \to j \leftarrow k$). If there exists an edge between i and k, i.e., $(i, k) \in E$ or $(k, i) \in E$, then the middle node j is a *shielded collider* in this collider triple. Otherwise, the node j is an *unshielded collider*.

In a directed graph $G = (V, E)$, a *path* is an alternating sequence of nodes from V and edges from E, such that each edge in the sequence is an edge between the nodes that precede and succeed it. Note that this definition allows a path to contain a node more than once. Given a fixed set $S \subseteq V$, two nodes $i, j \notin S$ are *d-connected* by S if G contains a path from i to j that has all colliders in S and all non-colliders outside S. If there exists no such path, then i and j are *d-separated* by S, denoted as $i \perp_d j \mid S$. A *trek* is a path without collider triples, making its endpoints d-connected given $\emptyset$. A trek takes the form

$$i_l^L \leftarrow \cdots \leftarrow i_1^L \leftarrow i_0 \to i_1^R \to \cdots \to i_r^R,$$

where i_0 is called the top node. This top node is characterized by not being the head of any edge on the trek. Every trek has a left hand side and a right hand side, corresponding to the nodes with superscript L or R. By convention, the top node i_0 is included in both the left hand and the right hand side of the trek.

C. Graphical Models, Trek Rule and Conditional Independence

As noted earlier, any specific linear SEM is obtained by requiring a subset of the entries in the parameter matrix $\Lambda = (\lambda_{ij})$ in (1) to vanish. This requirement may be represented by drawing an edge $i \to j$ whenever the matrix entry λ_{ij} is *not* constrained to be zero.

Let $G = (V, E)$ be a DAG. Let $\mathbb{R}^E$ be the set of real $V \times V$-matrices with support in E, that is,

$$\mathbb{R}^E = \left\{\Lambda = \left(\lambda_{ij}\right) \in \mathbb{R}^{V\times V} : \lambda_{ij} = 0 \text{ if } i \to j \notin E\right\}. \quad (3)$$

Then the DAG G encodes the linear SEM with independent errors whose coefficient matrix is constrained to have a zero pattern given by $\mathbb{R}^E$. In the resulting linear SEM the covariance matrix is parametrized through the map

$$\phi_G : \mathbb{R}^E \times (0, \infty)^V \mapsto PD,$$
$$(\Lambda, \boldsymbol{\omega}) \mapsto (I - \Lambda)^{-T} \operatorname{diag}(\boldsymbol{\omega})(I - \Lambda)^{-1},$$

where PD denotes the cone of positive definite matrices. Note that for a DAG, the matrix $I - \Lambda$ is invertible for all $\Lambda \in \mathbb{R}^E$ because the row and columns of Λ can be permuted to bring the matrix in lower-triangular form.

Definition 1: The *linear Gaussian model* given by a DAG $G = (V, E)$ is the family of all multivariate normal distributions on $\mathbb{R}^V$ with covariance matrix in the set

$$M_G = \left\{\Sigma : \Sigma = \phi_G(\Lambda, \boldsymbol{\omega}),\ \Lambda \in \mathbb{R}^E,\ \boldsymbol{\omega} \in (0, \infty)^V\right\}.$$

The map ϕ_G computes the covariance matrix of the random vector X from the coefficient matrix Λ and the vector of error variances $\boldsymbol{\omega}$. A classical result known as the *trek rule* provides a combinatorial description of the coordinates of ϕ_G, i.e., of the individual covariances between entries of X; see, e.g., Theorem 4.1 in the review [6].

Theorem 1 (Trek rule): Let $G = (V, E)$ be a DAG, and let $\Lambda = (\lambda_{ij}) \in \mathbb{R}^E$ and $\boldsymbol{\omega} = (\omega_i) \in (0, \infty)^V$. For $i, j \in V$, let $\mathcal{T}(i, j)$ be the set of all treks between i and j. For a trek τ with top node i_0, we define the trek monomial

$$\tau(\Lambda, \boldsymbol{\omega}) = \omega_{i_0} \prod_{k \to l \in \tau} \lambda_{kl}.$$

Then the covariance between X_i and X_j equals the sum of all trek monomials for treks between i and j, i.e.,

$$\phi_G(\Lambda, \boldsymbol{\omega})_{ij} = \sum_{\tau \in \mathcal{T}(i,j)} \tau(\Lambda, \boldsymbol{\omega}), \quad i, j \in V.$$

SEMs naturally lead to conditional independence constraints and these may be read off from an underlying DAG by inspecting d-separation relations, which we recalled in Section II-B. Furthermore, in linear Gaussian SEMs, conditional independence corresponds to an algebraic constraint on the covariance matrix of the distribution. We recall these facts in the following theorem; see [6, Sec. 10] or also [23, Sec. 8] for further discussion and pointers to the original literature.

Theorem 2: Let $G = (V, E)$ be a DAG. Let i, j be two distinct nodes, and let $S \subseteq V \backslash \{i, j\}$.

(i) If X is a multivariate normal random vector with covariance matrix Σ, then the conditional independence $X_i \perp\!\!\!\perp X_j \mid X_S$ holds if and only if $\det(\Sigma_{iS,jS}) = 0$. Here, $\Sigma_{iS,jS}$ is the submatrix of Σ obtained by selecting the rows indexed by $\{i\} \cup S$ and the columns indexed by $\{j\} \cup S$.

(ii) The conditional independence constraint $\det(\Sigma_{iS,jS}) = 0$ holds for all covariance matrices $\Sigma \in M_G$ if and only if the d-separation $i \perp_d j \mid S$ holds in G.

(iii) A matrix $\Sigma \in PD$ is in M_G if and only if $\det(\Sigma_{iS,jS}) = 0$ for all triples (i, j, S) with $i \perp_d j \mid S$ in G.

Similar submatrix notation is used throughout. So $\Sigma_{A,B}$ is the $A \times B$ submatrix of a matrix Σ, and to have compact notation for subsets of index set V, we define $iS := \{i\} \cup S$ and $ijS = \{i\} \cup \{j\} \cup S$ for $i, j \in V$ and $S \subset V$.

III. Partial Homoscedasticity

A. Setup

We now introduce a class of models that incorporate partial knowledge about equality among the variances in $\boldsymbol{\omega} = (\omega_i)_{i\in V}$ of the errors in the linear SEM given by a DAG $G = (V, E)$. We encode this partial knowledge in a partition of the vertex set V.

Definition 2: Let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be a family of non-empty subsets of V. Then Π is a partition of V if $\pi_1, \ldots, \pi_K$ are pairwise disjoint and $\cup_{k=1}^K \pi_k = V$. The sets $\pi_1, \ldots, \pi_K$ are the *blocks* of the partition. Corresponding to Π is the equivalence relation that has $i, j \in V$ equivalent if i, j are in the same block of Π; we then write $i \sim_\Pi j$.

In order to model a priori assumptions about the equality of some pairs of variances in $\boldsymbol{\omega} = (\omega_i)_{i\in V}$, we are led to consider the partition Π such that $i \sim_\Pi j$ precisely when the equality $\omega_i = \omega_j$ is implied by our a priori knowledge. Each one of the proposed partially homoscedastic linear SEMs is thus associated to a pair (G, Π), where $G = (V, E)$ is a DAG and Π is a partition of V. In this model, each block of Π groups a set of nodes that index error variances that are constrained to be equal to each other. In other words, we have homoscedasticity of the errors within each partition block, but possibly different variances between the blocks.

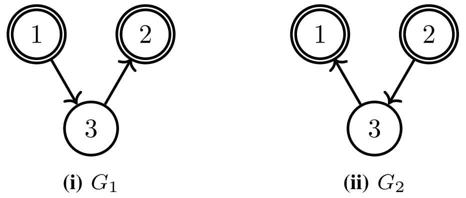
Definition 3: Let $G = (V, E)$ be a DAG, and let Π be a partition of V. The *partially homoscedastic linear Gaussian model* given by the pair (G, Π) is the family of all multivariate normal distributions on $\mathbb{R}^V$ with covariance matrix in the set

$$M_{G,\Pi} = \left\{\Sigma : \Sigma = \phi_G(\Lambda, \boldsymbol{\omega}),\ \Lambda \in \mathbb{R}^E,\right.$$
$$\left.\boldsymbol{\omega} \in (0, \infty)^V \text{ with } \omega_i = \omega_j \text{if } i \sim_\Pi j\right\}.$$

Given a partition Π, we call two DAGs G_1 and G_2 model equivalent if they induce the same partially homoscedastic linear Gaussian model, i.e., if $M_{G_1,\Pi} = M_{G_2,\Pi}$.

From the point of view of causal discovery, assumptions of (partial) homoscedasticity are interesting as extra constraints on error variances lead to a refinement of the Markov equivalence classes that result from conditional independence relations only. We exemplify this point here.

Example 1: Let G_1 and G_2 be the two DAGs in Figure 1. Consider first the finest partition $\Pi_{\min} = \{\{1\}, \{2\}, \{3\}\}$, which encodes that the error variances may be arbitrary positive numbers. Then $M_{G_1,\Pi_{\min}} = M_{G_1} = M_{G_2} = M_{G_2,\Pi_{\min}}$. Indeed, the two DAGs G_1 and G_2 are Markov equivalent, meaning that



(i) G_1 **(ii)** G_2

Fig. 1. Under the constraint $\omega_1 = \omega_2$, the two Markov equivalent DAGs G_1 and G_2 generate different partially homoscedastic linear Gaussian models.

they encode the same conditional independence relations. Both graphs feature precisely one d-separation relation, namely, that nodes 1 and 2 are d-separated by 3. The model may thus be defined by imposing the conditional independence of X_1 and X_2 given X_3, which under Gaussianity is equivalent to the constraint that the covariance matrix $\Sigma = (\sigma_{ij})$ satisfies the polynomial equation $\sigma_{12}\sigma_{23} - \sigma_{12}\sigma_{33} = 0$ (recall Theorem 2).

Now let us change the partition to $\Pi = \{\{1, 2\}, \{3\}\}$, i.e., we assume that $\omega_1 = \omega_2$. One can then show that the model given by G_1 is the set of 3×3 covariance matrices

$$M_{G_1,\Pi} = \big\{\Sigma \in PD : \sigma_{11}\sigma_{33} = \sigma_{22}\sigma_{33} - \sigma_{23}^2,\\ \sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33} = 0\big\},$$

whereas G_2 defines

$$M_{G_2,\Pi} = \big\{\Sigma \in PD : \sigma_{22}\sigma_{33} = \sigma_{11}\sigma_{33} - \sigma_{13}^2,\\ \sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33} = 0\big\}.$$

Both $M_{G_1,\Pi}$ and $M_{G_2,\Pi}$ are semialgebraic sets of dimension 4, but they are different and their intersection is of lower dimension.

In the remainder of this section, we develop a more general algebraic description of partially homoscedastic linear Gaussian models. This description furnishes the basis for solving the problem of deciding model equivalence, as developed in Section IV.

B. Equal Variance Constraints

The key difference between partially homoscedastic models and the classical case of arbitrary Gaussian errors is the emergence of constraints due to the equalities among error variances. To exhibit these constraints we first review how an error variance can be identified from the covariance matrix of the observations.

Theorem 3: Let $G = (V, E)$ be a DAG, and let $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ for $\Lambda = (\lambda_{ij}) \in \mathbb{R}^E$ and $\boldsymbol{\omega} = (\omega_i) \in (0, \infty)^V$. Then for any $i \in V$, the error variance ω_i can be computed from the covariance matrix $\Sigma = (\sigma_{ij})$ as

$$\omega_i = \sigma_{ii} - \Sigma_{i,A}\big(\Sigma_{A,A}\big)^{-1}\Sigma_{A,i}, \tag{4}$$

where A may be any subset with $\text{pa}(i) \subseteq A \subseteq V \setminus \text{de}(i)$.

Proof: We adapt the proof of [6, Theorem 7.1], where $A = \text{pa}(i)$. If a trek from j to i ends at i with an edge of the form $k \leftarrow i$, then the trek is a directed path from i to j and $j \in \text{de}(i)$. Now since $A \subseteq V \setminus \text{de}(i)$, every trek from a node in A to i must end with an edge $k \to i$. In other words, such a trek must visit a parent of i as the last node before i. Theorem 1 thus implies that

$$\Sigma_{A,i} = \Sigma_{A,\text{pa}(i)}\Lambda_{\text{pa}(i),i} = \Sigma_{A,A}\Lambda_{A,i}. \tag{5}$$

To see the first equality, partition the concerned sets of treks according to which element of $\text{pa}(i)$ is visited right before i. The second equality is then due to $\text{pa}(i) \subseteq A$ and $\lambda_{ki} = 0$ for $k \notin \text{pa}(i)$.

A similar reasoning for treks from i to i gives that

$$\begin{aligned}\sigma_{ii} &= \omega_i + \Lambda_{\text{pa}(i),i}^T\Sigma_{\text{pa}(i),\text{pa}(i)}\Lambda_{\text{pa}(i),i}\\ &= \omega_i + \Lambda_{A,i}^T\Sigma_{A,A}\Lambda_{A,i}.\end{aligned} \tag{6}$$

The claim now follows by rewriting (6) to $\omega_i = \sigma_{ii} - \Lambda_{A,i}^T\Sigma_{A,A}\Lambda_{A,i}$ and substituting $\Lambda_{A,i}$ by $(\Sigma_{A,A})^{-1}\Sigma_{A,i}$, as justified by (5). ∎

We immediately obtain the following corollary for an equal variance assumption.

Corollary 1: If two random errors ϵ_i and ϵ_j have equal variances, i.e., i and j are in the same block of a considered partition Π, then all covariance matrices $\Sigma = (\sigma_{ij})$ in $M_{G,\Pi}$ satisfy that

$$\sigma_{ii} - \Sigma_{i,A_i}\big(\Sigma_{A_i,A_i}\big)^{-1}\Sigma_{A_i,i} = \sigma_{jj} - \Sigma_{j,A_j}\big(\Sigma_{A_j,A_j}\big)^{-1}\Sigma_{A_j,j} \tag{7}$$

for all subsets A_i and A_j such that $\text{pa}(i) \subseteq A_i \subseteq V \setminus \text{de}(i)$ and $\text{pa}(j) \subseteq A_j \subseteq V \setminus \text{de}(j)$.

The fact from Theorem 3 admits the following converse.

Theorem 4: Let $G = (V, E)$ be a DAG, and let $i \in V$ be one of its nodes. Let $A \subseteq V \setminus \{i\}$. Fix any vector of positive error variances $\boldsymbol{\omega} \in (0, \infty)^V$. If for all $\Lambda \in \mathbb{R}^E$ the matrix $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ satisfies equation (4), then it must hold that $\text{pa}(i) \subseteq A \subseteq V \setminus \text{de}(i)$.

Proof: Suppose there exists a node $k \in \text{pa}(i) \setminus A$. Choose Λ to have all entries zero except for λ_{ki}. For this choice, the trek rule in Theorem 1 implies that $\Sigma_{i,A} = 0$ and, thus, the right hand side of (4) is equal to σ_{ii}. But the trek rule also yields that $\sigma_{ii} = \omega_i + \lambda_{ki}^2\omega_k > \omega_i$, which contradicts the assumption that (4) holds. We conclude that $\text{pa}(i) \subseteq A$.

Next, suppose that there exists a node $k \in A \setminus (V \setminus \text{de}(i)) = \text{de}(i) \cap A$. Then G contains a (non-trivial) directed path from i to k. Without loss of generality, we may assume that all interior nodes on the path between i and k are not in A. Indeed, we can always pick k to be the first node in A that lies on the path. So the path is of the form $i \to m_1 \to \cdots \to m_t \to k$ with $m_1, \ldots, m_t \notin A$. Now, take Λ with all entries zero except $\lambda_{im_1}, \lambda_{m_1m_2}, \ldots, \lambda_{m_{t-1}m_t}, \lambda_{m_tk}$. The trek rule in Theorem 1 asserts that $\sigma_{ii} = \omega_i$ under this parameterization (every trek between i and i has at least one edge with zero edge weight). But then equation (4) becomes

$$\begin{aligned}\omega_i &= \sigma_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2\Big[\big(\Sigma_{A,A}\big)^{-1}\Big]_{kk}\\ &= \sigma_{ii} - \Big(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\Big)^2\frac{1}{\sigma_{kk}} < \sigma_{ii} = \omega_i,\end{aligned}$$

which is again a contradiction. We conclude that $A \subseteq V \setminus \text{de}(i)$. ∎

Combining Theorems 3 and 4, we can characterize equal variance constraints that hold in a partially homoscedastic linear model. For the proof, see Appendix A.

Theorem 5: Let $G = (V, E)$ be a DAG, and let Π be a partition of the vertex set V. Suppose $i \sim_\Pi j$ are two distinct nodes that lie in the same block of Π, and let $A_i \subseteq V \backslash \{i\}$ and $A_j \subseteq V \backslash \{j\}$. Then the equation (7) holds for all matrices $\Sigma \in M_{G,\Pi}$ if and only if $\mathrm{pa}(i) \subseteq A_i \subseteq V \backslash \mathrm{de}(i)$ and $\mathrm{pa}(j) \subseteq A_j \subseteq V \backslash \mathrm{de}(j)$.

Theorem 5 yields a full algebraic characterization of partially homoscedastic linear Gaussian models. Every equal variance condition corresponds to a collection of equations between conditional variances, in which conditioning sets may be taken from a range of sets. The different conditioning sets will ultimately lead to equivalent constraints once the equal variance constraints are combined with conditional independence constraints.

We now record an observation that will be important for later considerations of model equivalence. It refers to the smallest and largest conditioning sets, where we partially order sets by set inclusion and extend the ordering lexicographically to pairs of sets; i.e., $(A_i, A_j) \le (B_i, B_j)$ if $A_i \subsetneq B_i$ or if $A_i = B_i$ and $A_j \subseteq B_j$.

Corollary 2: Let G be a DAG, and let Π be a partition of V such that $i \sim_\Pi j$ are in the same block of the partition. Let $\mathcal{A}_{ij}$ be the family of all pairs (A_i, A_j) with $A_i \subseteq V \backslash \{i\}$ and $A_j \subseteq V \backslash \{j\}$ for which equation (7), i.e.,

$$\sigma_{ii} - \Sigma_{i,A_i}\left(\Sigma_{A_i,A_i}\right)^{-1}\Sigma_{A_i,i} = \sigma_{jj} - \Sigma_{j,A_j}\left(\Sigma_{A_j,A_j}\right)^{-1}\Sigma_{A_j,j},$$

holds for all covariance matrices $\Sigma \in M_{G,\Pi}$. Then

(i) $\mathcal{A}_{ij}$ contains a unique minimal pair, namely, $A_i = \mathrm{pa}(i)$ and $A_j = \mathrm{pa}(j)$, and

(ii) $\mathcal{A}_{ij}$ contains a unique maximal pair, namely, $B_i = V \backslash \mathrm{de}(i)$ and $B_j = V \backslash \mathrm{de}(j)$.

C. Characterization of the Models

In Section III-B we considered a class of algebraic constraints that require equality of conditional variances, and we characterized which of these constraints hold in a given partially homoscedastic linear model. As we show in this section, combining the variance constraints with conditional independence constraints from d-separation relations yields an implicit algebraic description of partially homoscedastic linear models. We begin by revisiting the original proof in [24, Ths. 1 and 3] for soundness and completeness of d-separation in SEMs, which via a slight modification also applies to our specialized setting. In other words, we clarify in the following proposition that equal variance assumptions do not alter the set of conditional independence relations in a linear SEM.

Proposition 1: Let $G = (V, E)$ be a DAG, and let Π be a partition of V. Let i, j be two distinct nodes, and let $S \subseteq V \backslash \{i, j\}$. Then the conditional independence $X_i \perp\!\!\!\perp X_j \mid X_S$ holds for all multivariate normal random vectors X with covariance matrix in $M_{G,\Pi}$ if and only if the d-separation $i \perp_d j \mid S$ holds in G.

Proof: The "if" follows from Theorem 2 because $M_{G,\Pi} \subseteq M_G$. For the "only if", suppose that i and j are not d-separated by S. We then have to construct an example of $\Sigma \in M_{G,\Pi}$ in which the conditional independence does not hold, i.e., $\det(\Sigma_{iS,jS}) \neq 0$. To this end, we may slightly modify an example in [24]. The modification uses equal error variances to ensure Σ is in $M_{G,\Pi}$ and not merely in M_G. We provide the details in Appendix B. ■

All ingredients in place, we can now fully describe a partially homoscedastic linear Gaussian model in terms of conditional independence relations and equal variance constraints. Evidently, all constraints are algebraic, i.e., can be expressed in terms of polynomials.

Theorem 6: Let $G = (V, E)$ be a DAG, and let Π be a partition of V. Then a covariance matrix $\Sigma \in PD$ is in the partially homoscedastic linear model $M_{G,\Pi}$ if and only if Σ satisfies all conditional independence constraints given by d-separations and all equal variance constraints from Corollary 1.

Proof: The "if" follows from Proposition 1 and Corollary 1. For the "only if", let Σ satisfy all conditional independence and equal variance constraints associated to G. By Theorem 2(iii), a covariance matrix that satisfies all conditional independence constraints given by d-separation has to be an element of M_G. Hence, there exist $\Lambda \in \mathbb{R}^E$ and $\boldsymbol{\omega} \in (0, \infty)^V$ such that $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega}) \in M_G$. But then, by Theorem 3, the equalities among conditional variances imply that $\omega_i = \omega_j$ for $i \sim_\Pi j$. Therefore, $\Sigma \in M_{G,\Pi}$. ■

IV. Equivalence Classes

A. Model Equivalence of DAGs

Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs with the same given vertex set. An important problem for causal discovery is to decide whether the two DAGs are equivalent in the sense of defining the same statistical model for the observations at hand.

Definition 4: Let Π be a fixed partition of the index set V. Two DAGs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are *Π-model equivalent* if $M_{G_1,\Pi} = M_{G_2,\Pi}$. We denote this by $G_1 \approx_\Pi G_2$.

The classic case of linear SEMs with arbitrary error variances corresponds to the partition $\Pi_{\min} = \{\{i\} : i \in V\}$. In this case we have no equal variance constraints, and G_1 and G_2 are $\Pi_{\min}$-model equivalent if and only if G_1 and G_2 are Markov equivalent, meaning they induce the same conditional independence relations or, equivalently, have the same d-separation relations. Graphical models theory further tells us that G_1 and G_2 are Markov equivalent if and only if they have the same skeleton and unshielded collider triples [5].

Based on the algebraic characterization in Theorem 6, we are able to give the following extension of the Markov equivalence theory to the setting of partially homoscedastic models.

Theorem 7: Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs, and let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be a partition of the index set V. Then G_1 and G_2 are Π-model equivalent if and only if the following two conditions hold:

(i) G_1 and G_2 have the same skeleton and unshielded colliders, and

(ii) $\mathrm{pa}_{G_1}(i) = \mathrm{pa}_{G_2}(i)$ for all nodes i that belong to a partition block π_k of size $|\pi_k| \geq 2$.

Before proving the theorem, we would like to give some intuition for condition (ii), which is the result of the assumed equality among error variances. In the usual Markov equivalence theory, a DAG may contain directed edges that can be reversed while leaving the associated model unchanged. In the linear Gaussian case, such a reversal generally leads to a change of conditional variances (i.e., error variances) for the nodes incident to the reversed edge. Condition (ii) in the above theorem reflects the fact that an equality among error variances can generally not be preserved in this edge reversal step. Thus, if a node i is subject to an equal variance constraint, then its parent set $\mathrm{pa}(i)$ must be preserved across model equivalent graphs.

Proof of Theorem 7: For the "if" direction, suppose that conditions (i) and (ii) hold. By the standard Markov equivalence theory, condition (i) implies that G_1 and G_2 have the same d-separation relations and, thus, $M_{G_1} = M_{G_2}$. Now, let Σ be an arbitrary element of $M_{G_1,\Pi}$. Since $M_{G_1,\Pi} \subseteq M_{G_1} = M_{G_2}$, there is a (unique) choice of $\Lambda^{(2)} \in \mathbb{R}^{E_2}$ and $\boldsymbol{\omega}^{(2)} \in (0,\infty)^V$ such that $\Sigma = \phi_{G_2}(\Lambda^{(2)}, \boldsymbol{\omega}^{(2)})$. Let $i \neq j$ be any two nodes with $i \sim_\Pi j$, i.e., there is a partition block π_k of size $|\pi_k| \geq 2$ that contains both i, j. By Equation (7) of Corollary 1, since $\Sigma \in M_{G_1,\Pi}$, we have

$$\sigma_{ii} - \Sigma_{i,\mathrm{pa}_{G_1}(i)}\left(\Sigma_{\mathrm{pa}_{G_1}(i),\mathrm{pa}_{G_1}(i)}\right)^{-1}\Sigma_{\mathrm{pa}_{G_1}(i),i}$$
$$= \sigma_{jj} - \Sigma_{j,\mathrm{pa}_{G_1}(j)}\left(\Sigma_{\mathrm{pa}_{G_1}(j),\mathrm{pa}_{G_1}(j)}\right)^{-1}\Sigma_{\mathrm{pa}_{G_1}(j),j}.$$

By condition (ii), $\mathrm{pa}_{G_1}(i) = \mathrm{pa}_{G_2}(i)$ and $\mathrm{pa}_{G_1}(j) = \mathrm{pa}_{G_2}(j)$. Therefore, we have

$$\omega_i^{(2)} = \sigma_{ii} - \Sigma_{i,\mathrm{pa}_{G_2}(i)}\left(\Sigma_{\mathrm{pa}_{G_2}(i),\mathrm{pa}_{G_2}(i)}\right)^{-1}\Sigma_{\mathrm{pa}_{G_2}(i),i}$$
$$= \sigma_{jj} - \Sigma_{j,\mathrm{pa}_{G_2}(j)}\left(\Sigma_{\mathrm{pa}_{G_2}(j),\mathrm{pa}_{G_2}(j)}\right)^{-1}\Sigma_{\mathrm{pa}_{G_2}(j),j} = \omega_j^{(2)}.$$

We conclude that $\Sigma \in M_{G_2,\Pi}$ and, thus, $M_{G_1,\Pi} \subseteq M_{G_2,\Pi}$. Swapping the role of G_1 and G_2, we conclude that $M_{G_1,\Pi} = M_{G_2,\Pi}$ and $G_1 \approx_\Pi G_2$.

For the "only if" direction, suppose $M_{G_1,\Pi} = M_{G_2,\Pi}$. Theorem 6 implies that G_1 and G_2 induce the same conditional independence constraints and the same set of equal variance constraints (as specified in Corollary 1). We deduce that G_1 and G_2 have the same d-separation relations and, thus, condition (i) holds. Let i, j be any two distinct nodes in the same partition block π_k. Since G_1 and G_2 induce the same set of equal variance constraints, the set $\mathcal{A}_{ij}$ defined in Corollary 2 is the same for G_1 as for G_2. Corollary 2 now implies that the unique minimal element of $\mathcal{A}_{ij}$ must be comprised of the parent sets of node i and j in both G_1 and G_2. But this means that $\mathrm{pa}_{G_1}(i) = \mathrm{pa}_{G_2}(i)$ and $\mathrm{pa}_{G_1}(j) = \mathrm{pa}_{G_2}(j)$. Therefore, condition (ii) holds. ■

Remark 1: The two extreme cases of our setup are the classic setting in which all variances are freely varying ($|\Pi| = |V|$ or in other words $\Pi = \Pi_{\min} = \{\{i\} : i \in V\}$) and the previously studied case with all variances equal ($|\Pi| = 1$ or in other words $\Pi = \Pi_{\max} = \{V\}$). When $\Pi = \Pi_{\min}$, condition (ii) in Theorem 7 never applies and the theorem is just the classic Markov equivalence theorem. When $\Pi = \Pi_{\max}$, condition (ii) applies to all nodes, and Theorem 7 thus recovers the fact that under an equal variance assumption no two DAGs define the same model.

Remark 2: Another interesting special case arises in the context of two-sample problems, in which we observe each one of d variables under two different experimental conditions. In this setting, it is of interest to estimate the difference between the two DAGs for the two samples. This problem is greatly simplified by assuming equality of the two error variances that arise in the structural equations for the two independent copies of the kth random variables, $k = 1, \dots, d$ [25]. We can accommodate the two-sample problem in our framework by grouping all $2d$ random variables together. We then have a single combined graph of even size $|V| = 2d$ that consists of the disjoint union of the two DAGs for the two samples. The equal variance assumption in [25] corresponds to a partition $\Pi = \{\pi_1, \dots, \pi_d\}$ with $|\pi_1| = \cdots = |\pi_d| = 2$. Each partition block contains the two copies of one variable as it is observed in the two samples. Theorem 7 implies that under this partition the combined DAG is uniquely determined by the joint distribution of the observations from the two samples.

B. Completed Partially Directed Acyclic Graph (CPDAG)

Beyond characterizing equivalence of two DAGs as in Theorem 7, it is also of interest to provide a representation of each equivalence class. Similar to the classic heteroscedastic setup, we can represent the equivalence class by a *completed partially directed acyclic graph* (CPDAG) [26].

Definition 5: Let Π be a partition of the vertex set of a DAG $G = (V, E)$. The *completed partially directed acyclic graph* (CPDAG) of the DAG G *under partition* Π is the graph obtained by forming the union of all DAGs equivalent to G:

$$G_\Pi^* := \bigcup_{G' \approx_\Pi G} G'. \tag{8}$$

So, G_Π^* contains edge $i \to j$ if the edge is contained in some DAG $G' \approx_\Pi G$. It is customary to draw G_Π^* as a mixed graph with an undirected edge between nodes i and j for which both $i \to j$ and $j \to i$ are in G_Π^*.

We emphasize that an undirected edge in a CPDAG indicates that there exist two DAGs in the equivalence class in which the edge appears with opposite directions. Moreover, a CPDAG contains a directed edge $i \to j$ precisely when all DAGs in the equivalence class of G contain this edge.

In the standard heteroscedastic setting (i.e., $\Pi = \Pi_{\min} = \{\{i\} : i \in V\}$), the CPDAG may be constructed using an algorithm developed in [22]. In addition, Meek [22] shows how to construct a CPDAG in a setting where there is background knowledge about some of the edges. The background knowledge is of the form $\mathcal{K} = \langle \mathbf{F}, \mathbf{R} \rangle$, where $\mathbf{F}$ contains the edges not in the DAG and $\mathbf{R}$ contains the edges in the DAG. The algorithm first translates conditional independence statements into adjacencies and unshielded collider triples. Then the first 3 of the 4 orientation rules in [27] (Figure 2) are applied to obtain the CPDAG without background knowledge, which is exactly the CPDAG under Π_p. The last phase incorporates

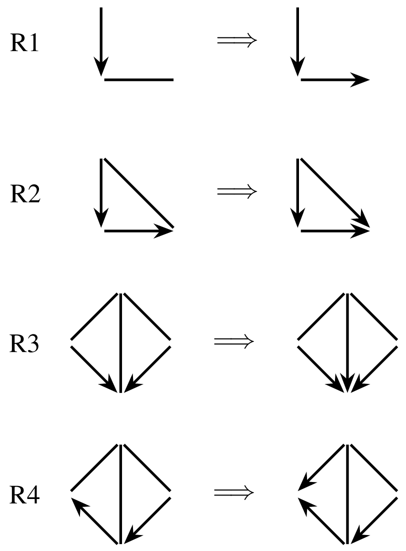


Fig. 2. The four orientation rules.

background knowledge and checks whether a compatible CPDAG exists or not. The following is the procedure, in which background knowledge is inserted edge by edge, and the CPDAG at the current step is denoted by G^*:

S1 If there is an edge $i \to j$ in $\mathbf{F}$ such that $i \to j$ in G^* then FAIL.

S1′ If there is an edge $i \to j$ in $\mathbf{R}$ such that $j \to i$ in G^* or i, j are not adjacent in G^* then FAIL.

S2 Randomly choose one edge $i \to j$ from $\mathbf{R}$, and let $\mathbf{R} = \mathbf{R} \backslash \{i \to j\}$.

S3 Orient $i \to j$ in G^* and close orientations under rules R1, R2, R3 and R4 in Figure 2.

S4 If $\mathbf{R} \neq \emptyset$, then go to step S1.

In our setup, we want to compute the equivalence class of a DAG under a partition, which restricts edge orientations in the DAG's equivalence class. These restrictions can be interpreted as providing background knowledge as considered by Meek. In our case, a CPDAG compatible to the background knowledge always exists, and we can use a simplified version of the general algorithm to construct the equivalence class.

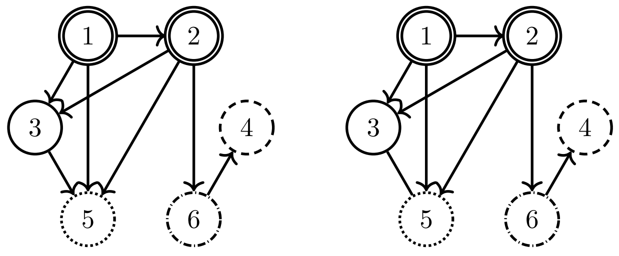
Given a DAG G and a partition Π, the equivalence class is obtained by the following algorithm (Algorithm 1). Theorem 8 below certifies the correctness of the algorithm. The proof of the theorem justifies the simplifications in the algorithm and is given in Appendix C.

Theorem 8: Given a DAG G and partition Π, Algorithm 1 outputs the CPDAG G^*_Π.

Example 2: Consider the DAG G in Figure 3 with node set $V = \{1, 2, 3, 4, 5, 6\}$ and the partition $\Pi = \{\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$, which is illustrated through different line styles when drawing nodes. In other words, the partition sequence is $(1, 1, 2, 3, 4, 5)$, where the i'th element of the sequence indicates the block that node i belongs to. To determine the equivalence class of G, we first keep the skeleton and unshielded colliders. Then those edges containing node 1 or 2 (partition block size ≥ 2) are oriented the same way as they are in G. Next, we propagate the edge orientation by rules R1 and R2, and we find that the edge between 4 and 6 is oriented as $4 \to 6$. Finally, the remaining edge $3 - 5$ can have both direction and is kept undirected in the final CPDAG that represents the equivalence class of G.

Algorithm 1 Constructing the Equivalence Class of a DAG, Given the Partition

```
Require: A DAG $G$, the partition $\Pi$
Create an empty graph $G'$
Copy the skeleton and all edge orientations with unshielded colliders of $G$ to $G'$
Apply rules R1, R2 and R3 on $G'$ until no more edges can be oriented
for $i \in V$ with $i \in \pi_k$ and $|\pi_k| \geq 2$ do
    Copy the orientation of edges in $G$ having one end-point at $i$ to $G'$
end for
Apply rules R1 and R2 on $G'$ until no more edges can be oriented
return $G^*_\Pi = G'$
```



(i) DAG G **(ii)** CPDAG G'

Fig. 3. A DAG and the corresponding CPDAG, under a fixed partition.

V. Greedy Search

A. Likelihood Inference

Let $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^{p \times n}$ be a data matrix comprised of n observations for each one of the $|V| = p$ considered variables. The columns of $\mathbf{X}$ are assumed to be generated as an i.i.d. sample from a joint multivariate normal distribution. Without loss of generality, we may assume the mean vector of the normal distribution to be zero. (Otherwise, we may estimate the means by sample means and recenter each row of the data matrix.) Define the sample covariance matrix $S = \mathbf{X}\mathbf{X}^T/n$. Then omitting a constant, the Gaussian log-likelihood function is $\frac{n}{2}(-\log\det(\Sigma) - \operatorname{tr}(\Sigma^{-1}S))$ with Σ being the covariance matrix.

For a fixed DAG $G = (V, E)$ and partition Π of V, the partially homoscedastic linear Gaussian model given by (G, Π) has the covariance matrix of the form $\Sigma = (I - \Lambda)^{-T}\operatorname{diag}(\boldsymbol{\omega})(I - \Lambda)^{-1}$. Thus, the model's log-likelihood function is

$$\ell_G(\Lambda, \boldsymbol{\omega}) = \frac{n}{2}\Big(- \log\det(\operatorname{diag}(\boldsymbol{\omega})) + \log\det(I - \Lambda)^2 - \operatorname{tr}\Big\{(I - \Lambda)\operatorname{diag}(\boldsymbol{\omega})^{-1}(I - \Lambda)^T S\Big\}\Big)$$

$$
\begin{aligned}
&= \frac{n}{2}\Big(-\log\det(\operatorname{diag}(\boldsymbol{\omega})) \\
&\qquad - \frac{1}{n}\operatorname{tr}\Big\{\mathbf{X}^T(I-\Lambda)\operatorname{diag}(\boldsymbol{\omega})^{-1}(I-\Lambda)^T\mathbf{X}\Big\}\Big) \\
&= \frac{n}{2}\left(-\sum_{i=1}^{p}\log\omega_i - \frac{1}{n}\sum_{i=1}^{p}\frac{1}{\omega_i}\left\|X_i - \Lambda^T_{\mathrm{pa}(i),i}X_{\mathrm{pa}(i)}\right\|^2\right),
\end{aligned}
$$

where the acyclicity of the DAG ensures that $\det(I-\Lambda) = 1$. Let $\Pi = \{\pi_1, \dots, \pi_K\}$. Then

$$
\ell_G(\Lambda, \boldsymbol{\omega}) = \frac{n}{2}\sum_{k=1}^{K}\ell_{G,\pi_k}(\Lambda, \omega_k)
$$

is the sum of log-likelihood values of the K blocks, with

$$
\ell_{G,\pi_k}(\Lambda, \omega_k) = -|\pi_k|\log\omega_k - \frac{1}{n\omega_k}\sum_{i\in\pi_k}\left\|X_i - \Lambda^T_{\mathrm{pa}(i),i}X_{\mathrm{pa}(i)}\right\|^2.
$$

We observe that the maximum log-likelihood is achieved at the pair $(\widehat{\Lambda}, \widehat{\boldsymbol{\omega}})$ with

$$
\begin{aligned}
\widehat{\Lambda}_{\mathrm{pa}(i),i} &= \operatorname*{argmin}_{\Lambda_{\mathrm{pa}(i),i}\in\mathbb{R}^{|\mathrm{pa}(i)|}} \|X_i - \Lambda^T_{\mathrm{pa}(i),i}X_{\mathrm{pa}(i)}\|^2, \\
\widehat{\omega}_k &= \frac{\sum_{i\in\pi_k}\left\|X_i - \widehat{\Lambda}^T_{\mathrm{pa}(i),i}X_{\mathrm{pa}(i)}\right\|^2}{n|\pi_k|}.
\end{aligned}
$$

In order to solve the problem of selecting the DAG G, we appeal to information criteria. Plugging the maximum likelihood estimate $(\widehat{\Lambda}, \widehat{\boldsymbol{\omega}})$ back into the log-likelihood function, we can compute the Bayesian information criterion (BIC) score for the DAG G given the data $\mathbf{X}$. This score decomposes into the sum of scores of each block:

$$
\begin{aligned}
s_{\mathrm{BIC}}(G) &= \frac{1}{n}\left(\ell_G\big(\widehat{\Lambda}, \widehat{\boldsymbol{\omega}}\big) - \frac{\log(n)}{2}|E|\right) \\
&= \frac{1}{2}\sum_{k=1}^{K}\left(-|\pi_k|\log\hat{\omega}_k - |\pi_k| - \frac{\log(n)}{n}\sum_{i\in\pi_k}|\mathrm{pa}(i)|\right),
\end{aligned} \tag{9}
$$

where the simplification arises from the fact that $\frac{1}{n\widehat{\omega}_k}\sum_{i\in\pi_k}\|X_i - \widehat{\Lambda}^T_{i,\mathrm{pa}(i)}X_{\mathrm{pa}(i)}\|^2 = |\pi_k|$.

B. Search Scheme

For model selection, we maximize the BIC score over the space of DAGs. The number of possible DAGs grows very quickly with the number of nodes p; e.g., we have 1.2×10^{15} for $p = 10$ [28]. We thus follow prior work and adopt a greedy search algorithm, which starts at some initial random or empty DAG and selects the DAG with highest BIC score in the local neighborhood at each step. The procedure terminates when the considered DAG has higher BIC score than all other DAGs in the local neighborhood. Here, we define the local neighborhood of a DAG G as the set of all DAGs that can be obtained from G by addition, removal or reversal of a single edge. Note that in this work we search only over DAGs, for a given fixed partition π.

An edge addition or removal always changes the equivalence class of a DAG. Whether an edge reversal creates a DAG in a different equivalence class is determined by parents and partition information, as specified in our previous results. We can thus search DAGs that are in the neighborhood and in different equivalence classes. To speed up the search we also restrict the local neighborhood to a random subset with size bound. To relieve issues of local optima, we perform the greedy search multiple times starting from different initial DAGs. In this work, for each realization of the greedy search, we restart the method 5 times with neighborhood size bound $k = 300$. We refer to this scheme as greedy search with Groupwise Equal Variances (GEV).

C. Simulation Study

We investigate the numerical performance of our algorithm (GEV) by comparing against greedy equivalence search (GES) [29] and the PC-algorithm [2]. The former tries to find the structure with maximum ℓ_0-penalized log-likelihood and the default penalty is $\log(n)/2n$, corresponding to the BIC score. The latter has a significance level α for conditional independence tests that determine edges. To make the score-based and the constraint-based methods comparable, we consider a grid of values for α from 10^{-5} to 0.8, increasing by the ratio 1.1 [30]. Then we can choose the value of α according to the maximum BIC score.

Both PC and GES algorithm do not incorporate any possible (partial) homoscedasticity and return a standard Markov equivalence class. In our partially homoscedastic model setup, the greedy search is performed among DAGs and returns the CPDAG of the final DAG, as described in Section V-B. Since the parental information (or edge directions) plays an important role in our idenfitiability result, we adopt the modified structural Hamming distance (SHD) in [13] as the error measurement. The classic SHD is the number of edge additions, deletions and reversals in order to transform one graph to another graph, i.e., all edge mistakes count as 1, while the modified version assigns a distance of 2 to each pair of reversed edges.

In the simulation we use 24 different configurations of (p, n, sp), where $p \in \{5, 10, 20, 40\}$ is the number of nodes, $n \in \{100, 500, 1000\}$ is the sample size and $sp \in \{sparse, dense\}$ controls the sparsity of randomly generated DAGs. In the sparse setting, each pair of nodes has the adjacency probability $prob = 3/(2p-2)$, while in the dense setting the probability is 0.3. For each (i, j) pair with $i < j$, we simulate independent uniform random variables $U_{ij} \sim U(0, 1)$. If $U_{ij} < prob$, the edge $i \to j$ is introduced. Every edge weight is uniformly drawn from $[-1, -0.3] \cup [0.3, 1]$, and the error variance of each partition block is uniformly drawn from $[0.3, 1]$. After traversing all node pairs, we randomly permute the node labels. For each configuration we run the simulation 100 times.

The box-plots that follow show the SHD between the true CPDAG and the estimated CPDAG obtained by the considered methods. We study the case of 2 partition blocks as well as a more subtle case with $\lceil p/3 \rceil + 1$ blocks. In the former case, the simulation experiments summarized in Figures 4 and 5 show that the greedy search algorithm for partially homoscedastic models is able to very effectively exploit the available homoscedasticity in both sparse and dense settings as well as across all three sample sizes and four

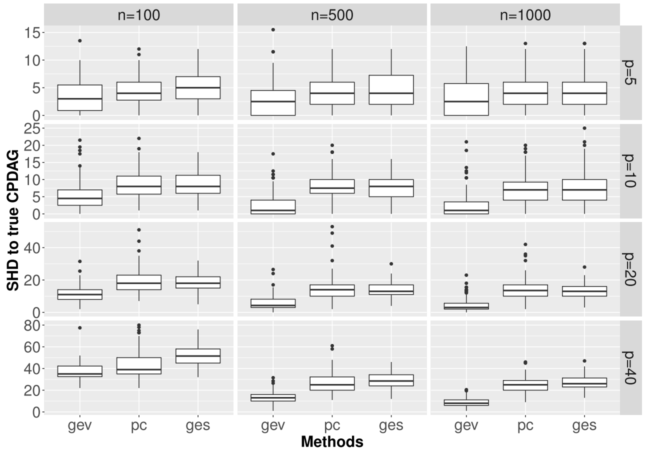


Fig. 4. Box-plots of SHD by groups of p and n, sparse graphs 2 blocks.

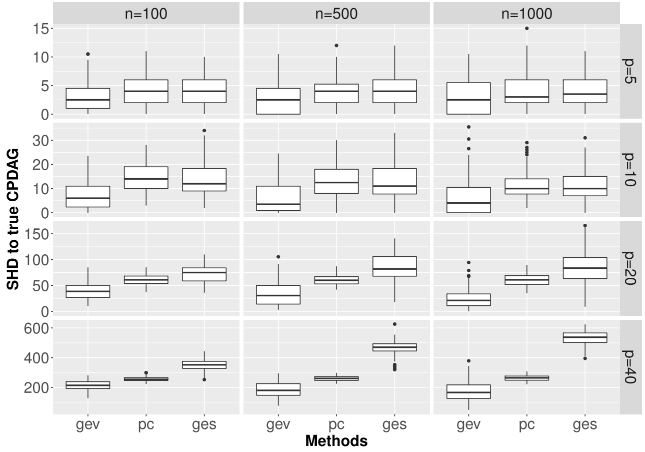


Fig. 5. Box-plots of SHD by groups of p and n, dense graphs, 2 blocks.

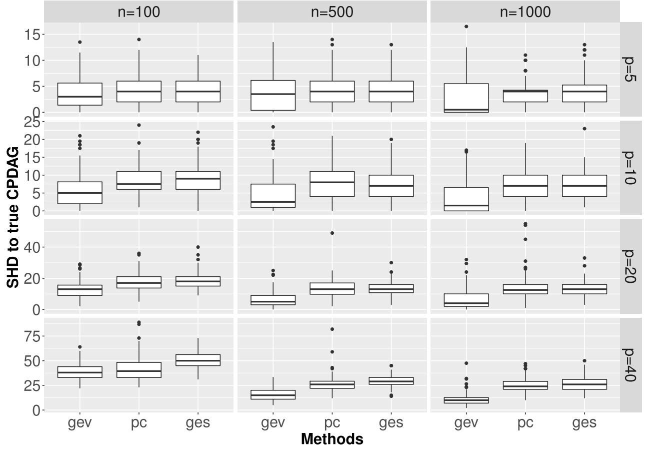


Fig. 6. Box-plots of SHD by groups of p and n, sparse graphs, $\lceil p/3 \rceil + 1$ blocks.

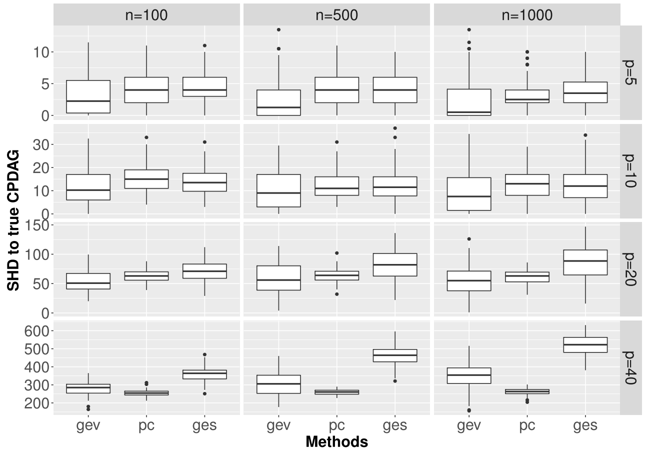


Fig. 7. Box-plots of SHD by groups of p and n, dense graphs, $\lceil p/3 \rceil + 1$ blocks.

dimensions. Its SHDs are consistently lower as for GES and PC. This said, the dense is clearly far more challenging than the sparse—as is to be expected. Moreover, the simulations confirm the intuition that the SHDs should be smaller if extra equal error variances information is utilized by the algorithm. For the larger number of blocks, Figure 6 shows the same pattern of clearly better performance. However, in the dense case depicted in Figure 7 one now sees the problem becoming difficult in the highest-dimensional case where the PC algorithm shows best performance.

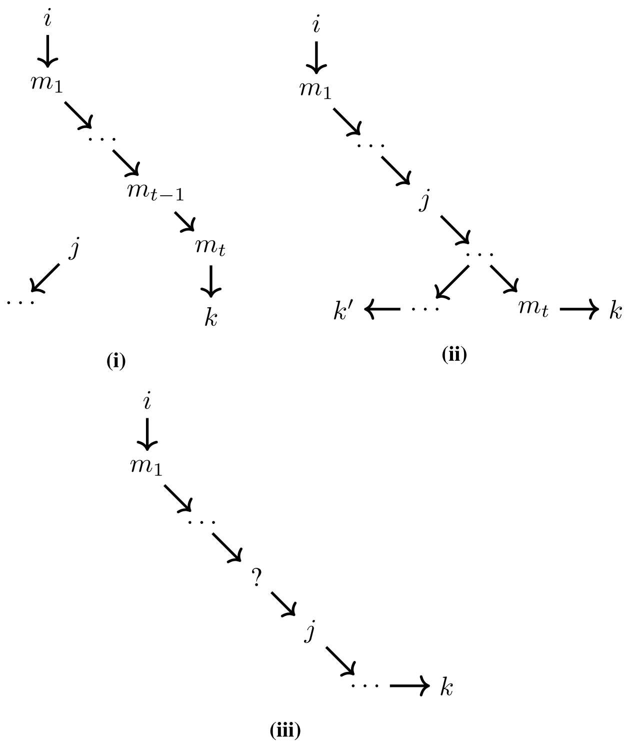
VI. Discussion

The framework of partially homoscedastic linear Gaussian models is a generalization of linear SEMs with equal error variances. It encodes equal variance assumption through a partition of the variables. The framework unifies the classical setting in which the error variances may be arbitrary and the equal error variance setup that has been studied in recent literature. These two cases are captured by the two extreme partitions, with a single block and all variables in separate blocks, respectively.

Each partially homoscedastic linear model can be characterized algebraically via conditional independence constraints and equal variance constraints. The former are well known from the classical graphical model perspective on linear SEMs, and we explicitly derived the latter in this paper. The equal variance constraints reveal the essence of how equal variance assumptions lead to identifiability of edge orientations. This perspective differs from previous work on the equal variance assumption and, in particular, work that considered ordering variances [e.g., [15]]. We also showed how equivalence classes in the partially homoscedastic setting are naturally represented by a refined CPDAG, which may be constructed efficiently with the help of existing results on CPDAGs in setting with background knowledge. For model selection, we demonstrated that greedy search provides an effective tool to exploit knowledge about partial homoscedasticity.

Appendix A
Proof of Theorem 5

The "if" direction is given by Corollary 1. For the "only if" direction, we distinguish several cases for the set A_i. The arguments for the corresponding different cases of A_j are analogous. In each case, we construct a set of parameters such that the considered rational equation in (7) does not hold.



(i) **(ii)** **(iii)**

Fig. 8. The three subcases when there exists a node $k \in \mathrm{de}(i) \cap A_i$.

a) $\exists\ k \in \mathrm{pa}(i)\setminus A_i$: We choose $\lambda_{ki} \neq 0$ and set all other edge weights equal to zero. Then since $k \notin A_i$, the trek rule implies that $\Sigma_{i,A_i} = 0$. Hence (7) yields that

$$\sigma_{ii} = \sigma_{jj} - \Sigma_{j,A_j}\left(\Sigma_{A_j,A_j}\right)^{-1}\Sigma_{A_j,j} \le \sigma_{jj}.$$

By the trek rule, it further holds that $\sigma_{jj} = \omega_j$ and $\sigma_{ii} = \omega_i + \lambda_{ki}^2\omega_k$. We arrive at the following contradiction:

$$\sigma_{ii} \le \sigma_{jj} = \omega_j = \omega_i < \omega_i + \lambda_{ki}^2\omega_k = \sigma_{ii}.$$

We conclude that $\mathrm{pa}(i) \subseteq A_i$.

b) $\exists\, k \in \mathrm{de}(i) \cap A_i$: There is then a directed path from i to k and as in the proof of Theorem 4, we assume that k was chosen such that this path is "minimal". In other words, the directed path is of the form $i \to m_1 \to \cdots \to m_t \to k$ with $m_1, \ldots, m_t \notin A_i$. We proceed by distinguishing three subcases (illustrated in Figure 8):

(i) Suppose k can be chosen such that there exists a directed path from i to k that is minimal in the above sense and does not intersect j. Then we can set all edge weights zero except those on the path. As in the proof of Theorem 4, we have $\sigma_{ii} = \omega_i = \omega_j = \sigma_{jj}$ and find a contradiction because under equation (7),

$$\begin{aligned}\omega_i &= \sigma_{ii} - \left(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\right)^2\frac{1}{\sigma_{kk}}\\ &< \sigma_{ii} = \sigma_{jj} = \omega_j.\end{aligned}$$

(ii) Next, consider the case where every minimal directed path from i to a node $k \in \mathrm{de}(i) \cap A_i$ contains the node j and where in addition $A_j \cap \mathrm{de}(j) \neq \emptyset$. Let $k' \in \mathrm{de}(j) \cap A_j$. Then there exists a directed path from j to k'. It follows that in this subcase j must be in $\mathrm{de}(i)$. Since the graph is a DAG, the considered directed path from j to k' may not contain i. Hence, we encounter exactly the situation of subcase (i), but with the role of i and j switched. Hence, also in this case we can construct a counterexample to equation (7).

(iii) The remaining subcase is that every minimal directed path from i to a node $k \in \mathrm{de}(i) \cap A_i$ contains the node j, and that these paths intersect A_j only after they have visited j. Select one such minimal directed path. If the node preceding j on the path is not in A_j, we can reduce the problem to case (a) by switching i and j ($\mathrm{pa}(j)\setminus A_j \neq \emptyset$). Otherwise, we set all edge weights zero except those on the considered minimal path. Let A_j' be the intersection of A_j and the nodes on the path. In the new DAG with only edges in the directed path, the set A_j' satisfies that $\mathrm{pa}(j) \subseteq A_j' \subseteq V\setminus \mathrm{de}(j)$, and thus

$$\begin{aligned}\omega_j &= \sigma_{jj} - \Sigma_{j,A_j'}\left(\Sigma_{A_j',A_j'}\right)^{-1}\Sigma_{A_j',j}\\ &= \sigma_{jj} - \Sigma_{j,A_j}\left(\Sigma_{A_j,A_j}\right)^{-1}\Sigma_{A_j,j}.\end{aligned}$$

However, computing the left hand side of (7) leads to a strict inequality.

$$\begin{aligned}&\sigma_{ii} - \Sigma_{i,A_i}\left(\Sigma_{A_i,A_i}\right)^{-1}\Sigma_{A_i,i}\\ &\quad= \omega_i - \left(\lambda_{im_1}\lambda_{m_tk}\prod_{s=2}^{t}\lambda_{m_{s-1}m_s}\right)^2\frac{1}{\sigma_{kk}}\\ &\quad< \omega_i = \omega_j = \sigma_{jj} - \Sigma_{j,A_j}\left(\Sigma_{A_j,A_j}\right)^{-1}\Sigma_{A_j,j}. \qquad \blacksquare\end{aligned}$$

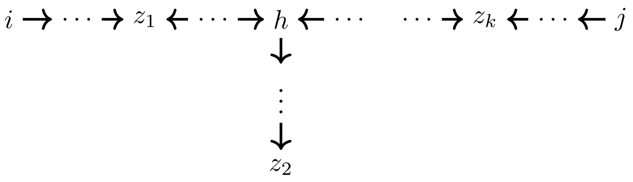


Fig. 9. An example of an active path q.

Appendix B
"Only If" Part of Proposition 1

Proof: If i and j are d-connected given S, then there exists a path q between i and j, on which every collider is in S (recall that our convention allows a path to visit the same node more than once). We denote the set of all these colliders by $S' = \{z_1, z_2, \ldots, z_k\} \subseteq S$; see Figure 9 for an illustration. In order to form a covariance matrix in $M_{G,\Pi}$, we assign the same weight $\rho \in (0, 1)$ to all edges of the path q and set all other edge weights zero. We set all error variances $\omega_i = 1$. Let Λ and $\boldsymbol{w}$ be the resulting choice of parameters, and let $\Sigma = \phi_G(\Lambda, \boldsymbol{\omega})$ the associated covariance matrix.

By the trek rule, the diagonal entries of $\Sigma = (\sigma_{kl})$ satisfy that

$$\sigma_{ii} = \sigma_{jj} = 1 \quad \text{and} \quad \sigma_{kk} = 1\ \forall\, k \notin S\setminus S',$$

because the fact that i and j are d-connected given S implies that the only nodes that are both in S and on the path q are the colliders in the set S'. Next, notice that there exists a unique nonzero trek between each pair of consecutive nodes in the sequence $i \equiv z_0, z_1, z_2, \ldots, z_k, z_{k+1} \equiv j$. Let r_t be the number of edges on the segment of q that goes from z_t to z_{t+1}. By the trek rule, for all $t = 0, \ldots, k$,

$$\sigma_{z_t,z_{t+1}} = \rho^{r_t}.$$

Ordering the nodes as $i, z_1, \ldots, z_k, j$ followed by the nodes in $S \backslash S'$, we obtain that

$$\Sigma_{ijS,ijS} = \left(\begin{array}{cccccc|c} 1 & \rho^{r_0} & 0 & \cdots & 0 & 0 & \\ \rho^{r_0} & \sigma_{z_1,z_1} & \rho^{r_1} & \cdots & 0 & 0 & \\ 0 & \rho^{r_1} & \sigma_{z_2,z_2} & \ddots & 0 & 0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & O \\ 0 & 0 & 0 & \ddots & \sigma_{z_k,z_k} & \rho^{r_k} & \\ 0 & 0 & 0 & \cdots & \rho^{r_k} & 1 & \\ \hline & & O & & & & I_{S\backslash S'} \end{array}\right). \tag{10}$$

Now observe that $\det(\Sigma_{iS,jS}) = \rho^{\sum_{t=0}^{k} r_t} \neq 0$. ■

Appendix C
Proof of Theorem 8

Algorithm 1 builds upon the work of Meek [22] who shows how to construct the CPDAG of an equivalence class when provided a set of conditional independence relations and arbitrary background knowledge about the edge orientations. His general algorithm first constructs the classical CPDAG by reading off unshielded colliders and propagating rules R1, R2, R3. Next, the general algorithm iteratively adds each edge from background knowledge and applies all rules R1, R2, R3, R4 to the 1-edge changes. Reference [22, Ths. 2–4] prove the correctness of the general algorithm.

The application of R1-R3 before inserting background knowledge creates the classical CPDAG for known conditional independence relations and without extra information (it is the CPDAG under partition $\Pi_{\min} = \{\{i\} : i \in V\}$). In our setup, we start with a DAG G in the equivalence class and determine directly the skeleton and unshielded colliders and the classical CPDAG via rules R1-R3.

The partial homoscedasticity encoded in the given partition Π now provides special 'background knowledge' that fixes the orientation of all the edges with one endpoint at special nodes. As we show in the remainder of this proof, when we insert this special knowledge into the classical CPDAG, the situations of R3 and R4 in [22] cannot arise. It thus suffices to apply only R1 and R2, and we can insert all the background knowledge simultaneously, because we know that all extra information is compatible and the desired CPDAG always exists.

For our proof of the correctness of the simplifications in Algorithm 1 over Meek's general procedure, recall that the equal variance constraints give the adjacency directions of all nodes whose block has size at least 2. The set **R** consists of edges incident to these nodes, and the set **F** consists of the reversal of the edges in **R**. We then argue as follows.

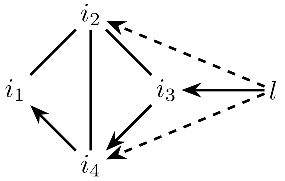


Fig. 10. $i_3 \to i_4$ from R1.

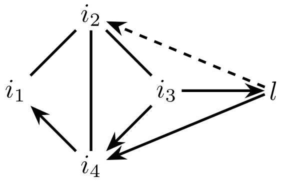


Fig. 11. $i_3 \to i_4$ from R2.

(i) First, we know there is at least one DAG in the equivalence class, so the general algorithm will not fail. That means the background knowledge check S1 and S1′ are redundant. We can just iteratively perform S2, S3 and S4 and obtain the same result.

(ii) Next, notice that we can add all edges in **R** simultaneously and close the orientations sequentially. Indeed, every newly oriented edge is dependent on some of the background knowledge. As long as all dependencies are added, the edge will be oriented without conflicts. Either adding edges sequentially or simultaneously would finally cover all dependencies of each orientable edge, and results in the same final output.

(iii) Finally, we claim that only the rules R1 and R2 become applicable in the orientation propagation step S3 of our algorithm. Indeed, there is an unshielded collider triple in R3, but the propagation with background knowledge does not make any new collider triples, otherwise the output CPDAG cannot have same conditional independence statements as the DAG itself. Hence, any pattern of R3 must have been obtained in the initial phase of constructing the classical CPDAG, and will not appear in the last propagation phase.

For R4, consider the first time that its pattern appears in the propagation phase. The orientation $i_3 \to i_4$ is not obtained in the classical CPDAG phase as otherwise $i_4 \to i_1$ would have also been oriented and the pattern of R4 appears in the classic CPDAG phase, which is a contradiction. If $i_3 \to i_4$ results from background knowledge directly, then we know the orientations of all adjacencies of either i_3 or i_4, which will orient $i_2 - i_3$ or $i_2 - i_4$. This is a contradiction. Figure 10 depicts the case of $i_3 \to i_4$ obtained from R1: unshielded triple $l \to i_3 - i_4$. The edge $l \to i_2$ must exist to keep $i_2 - i_3$ not oriented, consequently the undirected edge $i_2 - i_4$ implies the adjacency between l and i_4. The triple (l, i_3, i_4) is shielded, contradicting the pattern of R1. Figure 11 illustrates the case of $i_3 \to i_4$ obtained from

R2. To keep $i_2 - i_4$ not oriented, the edge $l \rightarrow i_2$ must exist. But then $i_2 - i_3$ can be oriented as $i_3 \rightarrow i_2$, which is again a contradiction.

In conclusion, we have proved that our modification to the general algorithm for equal variance constraints background knowledge is correct. ■

References

[1] J. Pearl, *Causality*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[3] M. Drton and M. H. Maathuis, "Structure learning in graphical modeling," *Annu. Rev. Stat. Appl.*, vol. 4, pp. 365–393, Mar. 2017.

[4] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, *Handbook of Graphical Models*. Boca Raton, FL, USA: CRC Press, 2019.

[5] M. Studený, "Conditional independence and basic Markov properties," in *Handbook of Graphical Models*. Boca Raton, FL, USA: CRC Press, 2019, pp. 3–38.

[6] M. Drton, "Algebraic problems in structural equation modeling," in *Proc. 50th Anniversary Gröbner Bases*, vol. 77, 2018, pp. 35–86.

[7] E. Mokhtarian, S. Akbari, A. Ghassami, and N. Kiyavash, "A recursive Markov boundary-based approach to causal structure learning," in *Proc. KDD Workshop Causal Disc.*, vol. 150, 2021, pp. 26–54.

[8] E. Mokhtarian, S. Akbari, F. Jamshidi, J. Etesami, and N. Kiyavash, "Learning Bayesian networks in the presence of structural side information," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 7814–7822.

[9] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, Oct. 2006.

[10] S. Shimizu, *Statistical Causal Discovery: LiNGAM Approach*. Tokyo, Japan: Springer 2022.

[11] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in Neural Information Processing Systems*, vol. 21. Red Hook, NY, USA: Curran Assoc., Inc., 2008, pp. 689–696.

[12] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Identifiability of causal graphs using functional models," in *Proc. 27th Conf. Uncertain. Artif. Intell.*, 2011, pp. 589–598.

[13] J. Peters and P. Bühlmann, "Identifiability of Gaussian structural equation models with equal error variances," *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[14] A. Ghoshal and J. Honorio, "Learning identifiable Gaussian Bayesian networks in polynomial time and sample complexity," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6460–6469.

[15] W. Chen, M. Drton, and Y. S. Wang, "On causal discovery with an equal-variance assumption," *Biometrika*, vol. 106, no. 4, pp. 973–980, 2019.

[16] A. Ghoshal and J. Honorio, "Learning linear structural equation models in polynomial time and sample complexity," in *Proc. 21st Int. Conf. Artif. Intell. Stat.*, vol. 84, 2018, pp. 1466–1475.

[17] M. Gao, Y. Ding, and B. Aragam, "A polynomial-time algorithm for learning nonparametric causal graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11599–11611.

[18] M. Gao, W. Ming Tai, and B. Aragam, "Optimal estimation of Gaussian DAG models," in *Proc. 25th Int. Conf. Artif. Intell. Stat.*, vol. 151, 2022, pp. 8738–8757.

[19] G. Park and Y. Kim, "Identifiability of Gaussian linear structural equation models with homogeneous and heterogeneous error variances," *J. Korean Stat. Soc.*, vol. 49, pp. 276–292, Jan. 2020.

[20] G. Park, "Identifiability of additive noise models using conditional variances," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 2896–2929, 2020.

[21] Y. Wang and A. Bhattacharyya, "Identifiability of linear AMP chain graph models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 10080–10089.

[22] C. Meek, "Causal inference and causal explanation with background knowledge," in *Proc. 11th Annu. Conf. Uncertain. Artif. Intell.*, 1995, pp. 403–410.

[23] T. Richardson and P. Spirtes, "Ancestral graph Markov models," *Ann. Stat.*, vol. 30, no. 4, pp. 962–1030, 2002.

[24] D. Geiger and J. Pearl, "On the logic of causal models," in *Proc. 4th Annu. Conf. Uncertain. Artif. Intell.*, 1988, pp. 3–14.

[25] Y. Wang, C. Squires, A. Belyaeva, and C. Uhler, "Direct estimation of differences in causal graphs," in *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY, USA: Curran Assoc., Inc., 2018.

[26] S. A. Andersson, D. Madigan, and M. D. Perlman, "A characterization of Markov equivalence classes for acyclic digraphs," *Ann. Stat.*, vol. 25, no. 2, pp. 505–541, 1997.

[27] T. Verma and J. Pearl, "An algorithm for deciding if a set of observed independencies has a causal explanation," in *Proc. 8th Annu. Conf. Uncertain. Artif. Intell.*, 1992, pp. 323–330.

[28] N. J. A. Sloane. "The on-line encyclopedia of integer sequences." 2022. [Online]. Available: http://oeis.org/A003024

[29] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 507–554, 2003.

[30] N. Harris and M. Drton, "PC algorithm for nonparanormal graphical models," *J. Mach. Learn. Res.*, vol. 14, pp. 3365–3383, Nov. 2013.