The Annals of Statistics
2018, Vol. 46, No. 6A, 2904–2938
https://doi.org/10.1214/17-AOS1643

CAUSAL INFERENCE IN PARTIALLY LINEAR STRUCTURAL EQUATION MODELS

BY DOMINIK ROTHENHÄUSLER[1], JAN ERNEST[1,2] AND PETER BÜHLMANN

ETH Zürich

We consider identifiability of partially linear additive structural equation models with Gaussian noise (PLSEMs) and estimation of distributionally equivalent models to a given PLSEM. Thereby, we also include robustness results for errors in the neighborhood of Gaussian distributions. Existing identifiability results in the framework of additive SEMs with Gaussian noise are limited to linear and nonlinear SEMs, which can be considered as special cases of PLSEMs with vanishing nonparametric or parametric part, respectively. We close the wide gap between these two special cases by providing a comprehensive theory of the identifiability of PLSEMs by means of (A) a graphical, (B) a transformational, (C) a functional and (D) a causal ordering characterization of PLSEMs that generate a given distribution $\mathbb{P}$. In particular, the characterizations (C) and (D) answer the fundamental question to which extent nonlinear functions in additive SEMs with Gaussian noise restrict the set of potential causal models, and hence influence the identifiability.

On the basis of the transformational characterization (B) we provide a score-based estimation procedure that outputs the graphical representation (A) of the distribution equivalence class of a given PLSEM. We derive its (high-dimensional) consistency and demonstrate its performance on simulated datasets.

1. Introduction. Causal inference is fundamental in many scientific disciplines. Examples include the identification of causal molecular mechanisms in genomics [24, 25], the investigation of causal relations among activity in brain regions from fMRI data [19] or the search for causal associations in public health [7].

A major research topic in causal inference aims at establishing causal dependencies based on purely observational data. The notion "observational" commonly refers to the fact that one obtains the data from the system of variables under consideration without subjecting it to external manipulations. Typically, one then assumes that the observed data has been generated by an underlying causal model and tries to draw conclusions about its structure.

Received July 2016; revised October 2017.

[1]These authors contributed equally to this work.

[2]Supported in part by the Max Planck ETH Center for Learning Systems and by the Swiss National Science Foundation Grant 2-77991-14.

MSC2010 subject classifications. Primary 62G99, 62H99; secondary 68T99.

Key words and phrases. Causal inference, distribution equivalence class, graphical model, high-dimensional consistency, partially linear structural equation model.

Two main research tasks in this setting are identifiability and estimation of the underlying causal model. We consider identifiability of partially linear additive structural equation models with Gaussian noise (PLSEMs) and estimation of distributionally equivalent models to a given PLSEM. Thereby, we also include robustness results for errors in the neighborhood of Gaussian distributions.

So far, there exists a wide "identifiability gap" for PLSEMs, as their identifiability has only been characterized for the two special cases where all the functions are linear or all the functions are nonlinear. We close this "identifiability gap" by providing comprehensive characterizations of the identifiability of the general class of PLSEMs from various perspectives.

Unlike in regression where partially linear models are mainly studied because of efficiency gains in estimation, the use of partially linear models has a deeper meaning in causal inference. In fact, as we will show, it is closely connected to identifiability. The functional form of an additive component directly influences the identifiability of the corresponding (and also other) causal relations. For this reason, we strongly believe that the understanding of the identifiability of PLSEMs is important. First and foremost, it raises the awareness of potentially limited (or increased) identifiability in the presence of linear (or nonlinear) relations in the data. Second, by not restricting the functions to be either all linear or all nonlinear, PLSEMs allow for a flexible modeling approach.

We start by reviewing and introducing important concepts in Section 1.1. We then provide a brief overview of related work in Section 1.2 and explicitly state the main contributions of this paper in Section 1.3.

1.1. *Problem description and important concepts.* We consider p random variables $X = (X_1, \dots, X_p)$ with joint distribution $\mathbb{P}$, which is assumed to be Markov with respect to an underlying directed acyclic graph (DAG). A DAG $D = (V, E)$ is an ordered pair consisting of a set of vertices $V = \{1, \dots, p\}$ associated with the variables $\{X_1, \dots, X_p\}$, and a set of directed edges $E \subset V^2$ such that there are no directed cycles. A directed edge between the nodes i and j in D is denoted by $i \to j$. Node i is called a *parent* of node j and j is called a *child* of i. Moreover, the edge is said to be oriented *out of* i and *into* j. If $i \to j$ or $i \leftarrow j$, i and j are called *adjacent* and the edge is *incident* to i and j. The *degree* of a node i, denoted by $\deg_D(i)$, counts the number of edges incident to node i in DAG D. A node k that can be reached from i by following directed edges is called *descendant* of i. We use the convention that any node is a descendant of itself. The set $\mathrm{pa}_D(j) = \{i \mid i \to j \text{ in } D\}$ consists of all parents of node j. The multi-index notation $X_{\mathrm{pa}_D(j)}$ denotes the set of variables $\{X_i\}_{i \in \mathrm{pa}_D(j)}$. An edge $i \to j$ is said to be *covered* in D, if $\mathrm{pa}_D(i) = \mathrm{pa}_D(j) \setminus \{i\}$. In that case, $\mathrm{pa}_D(i)$ is a *cover* for edge $i \to j$. The process of changing the orientation of a covered edge from $i \to j$ to $i \leftarrow j$ is referred to as a *covered edge reversal*. A triple (i, j, k) is called a *v-structure*, if $\{i, j\} \subseteq \mathrm{pa}_D(k)$ and i and j are not adjacent. The graph obtained by replacing all directed edges $i \to j$ by undirected edges $i - j$ is called *skeleton*

of D. The *pattern* of a DAG D is the graph with the same skeleton as D and $i \to j$ is directed if and only if it is part of a v-structure in D. A permutation $\sigma : V \to V$ is a *causal ordering* of D if $\sigma(i) < \sigma(j)$ for all $i \to j$ in D. DAGs may be used as underlying structures for structural equation models (SEMs). A SEM relates the distribution of every random variable $\{X_1, \dots, X_p\}$ to the distribution of its direct causes (the parents in the corresponding DAG D) and random noise. In its most general form,

$$X_j = f_j(X_{\mathrm{pa}_D(j)}, \varepsilon_j), \qquad j = 1, \dots, p, \tag{1.1}$$

where $\{f_j\}_{j=1,\dots,p}$ are functions from $\mathbb{R}^{|\mathrm{pa}_D(j)|+1} \to \mathbb{R}$ and $\{\varepsilon_j\}_{j=1,\dots,p}$ are mutually independent noise variables. Lastly, for a function $F : \mathbb{R}^p \to \mathbb{R}^p$, we write $\mathrm{D}F$ for the Jacobian of F.

1.1.1. *Main focus*: *PLSEMs*. In this paper we study the restriction of the general SEM in equation (1.1) to *partially linear additive SEMs with Gaussian noise (PLSEMs)* of the form

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) + \varepsilon_j, \tag{1.2}$$

where $\mu_j \in \mathbb{R}$, $f_{j,i} \in C^2(\mathbb{R})$, $f_{j,i} \not\equiv 0$, with $\mathbb{E}[f_{j,i}(X_i)] = 0$, and $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j^2 > 0$ for $j = 1, \dots, p$. Likewise, we may write

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D^{\mathrm{L}}(j)} \alpha_{j,i} X_i + \sum_{i \in \mathrm{pa}_D^{\mathrm{NL}}(j)} f_{j,i}(X_i) + \varepsilon_j,$$

with $\alpha_{j,i} \in \mathbb{R} \setminus \{0\}$, μ_j, $f_{j,i}$, ε_j as above, $\mathrm{pa}_D^{\mathrm{L}}(j) \cup \mathrm{pa}_D^{\mathrm{NL}}(j) = \mathrm{pa}_D(j)$ and $\mathrm{pa}_D^{\mathrm{L}}(j) \cap \mathrm{pa}_D^{\mathrm{NL}}(j) = \varnothing$. Note that we do not *a priori* fix the sets $\mathrm{pa}_D^{\mathrm{L}}(j)$ and $\mathrm{pa}_D^{\mathrm{NL}}(j)$. For $\mathbb{P}$ generated by a PLSEM with DAG D, the PLSEM corresponding to D is unique (cf. Lemma B.2 in the supplement [20]). Therefrom, we call an edge $i \to j$ in D a *(non)linear edge*, if $f_{j,i}$ in the PLSEM corresponding to D is (non)linear. Note that the concept of (non)linearity of an edge is defined with respect to a specific DAG D. Depending on the orientations of other edges, the status of an edge $i \to j$ may change from linear to nonlinear. An example is given in Figure 1.

The restriction to additive SEMs is interesting from both a statistical and computational perspective as the estimation of additive functions is well understood and one largely avoids the curse of dimensionality. The assumption of Gaussian noise is necessary for our characterization results in Section 2. In fact, identifiability properties may deteriorate in partially linear models with arbitrary noise distributions; see Section 1.2.4. We therefore consider PLSEMs to be among the most general SEMs with reasonable estimation properties. For an extension to error distributions in the neighborhood of the Gaussian distribution, see Section 4.

FIG. 1. *Two DAGs D_1 and D_2 with linear edges (dashed) and nonlinear edges (solid). Let us give a brief outlook: let $\mathbb{P}$ be generated by a PLSEM with DAG D_1. In this paper we prove that there exists a PLSEM with DAG D_2 that generates the same distribution $\mathbb{P}$. Moreover, we show that D_1 and D_2 are the only two DAGs with a corresponding PLSEM that generates $\mathbb{P}$. For now, simply note that $1 \to 3$ is linear in D_1, but nonlinear in D_2.*

1.1.2. *Main task: Characterization of all PLSEMs that generate $\mathbb{P}$.* The main task of this paper is the systematic characterization of all PLSEMs that generate a given distribution $\mathbb{P}$ under very general assumptions. In particular: how do edge functions in different PLSEMs relate to each other? How does changing a single linear edge to a nonlinear edge affect the set of potential underlying PLSEMs? Do causal orderings of different DAGs corresponding to PLSEMs that generate $\mathbb{P}$ share certain properties?

Under faithfulness, it may be natural to characterize all PLSEMs that generate $\mathbb{P}$ by their corresponding DAGs as they are restricted to a subset of the Markov equivalence class (see Section 1.2.1). For a distribution $\mathbb{P}$ that has been generated by a faithful PLSEM, we call the set of DAGs

$$\mathscr{D}(\mathbb{P}) := \left\{ D \,\middle|\, \begin{array}{l} \mathbb{P} \text{ is faithful to } D \text{ and there exists a} \\ \text{PLSEM with DAG } D \text{ that generates } \mathbb{P} \end{array} \right\}$$

the *(PLSEM) distribution equivalence class*. Can we build on characterizations of the Markov equivalence class to characterize $\mathscr{D}(\mathbb{P})$? For example, can $\mathscr{D}(\mathbb{P})$ also be graphically represented by a single PDAG? Is it possible to efficiently estimate $\mathscr{D}(\mathbb{P})$? Before we explain our approaches to answer these questions in Section 1.3, let us briefly summarize related work.

1.2. *Related work.* First, in Section 1.2.1, we discuss the identifiability of general SEMs. We then motivate why our theoretical results close a relevant "gap" by reviewing existing identifiability results for two special cases of PLSEMs where either all the functions $f_{j,i}$ are linear (Section 1.2.2) or nonlinear (Section 1.2.3). Finally, we briefly comment on the assumption of Gaussian noise in Section 1.2.4.

1.2.1. *Identifiability of general SEMs.* In the general SEM as defined in equation (1.1), one cannot draw any conclusions about D given $\mathbb{P}$ without making further assumptions. One such assumption commonly made is faithfulness (cf. Section 2.1). Under faithfulness, one can identify the Markov equivalence class of D (a set of DAGs that all entail the same conditional independences); see, for example, [15]. Markov equivalence classes are well characterized. In fact, the Markov

equivalence class of a DAG D consists of all DAGs with the same skeleton and v-structures as D [28] and can be graphically represented by a single partially directed graph (cf. Section 2.1). Moreover, any two Markov equivalent DAGs can be transformed into each other by a sequence of distinct covered edge reversals [6].

The estimation of the general SEM is difficult due to the curse of dimensionality in fully nonparametric estimation. In combination with the unidentifiability, this motivates the use of restricted SEMs, which have better estimation properties and for which it is possible to achieve (partial) identifiability of the SEM (even without assuming faithfulness); see Section 2.2 or [18] for an overview.

1.2.2. *Special case of PLSEM*: *Linear Gaussian SEM*. A widespread specification of PLSEMs are linear Gaussian SEMs, which have the same identifiability properties as the general SEMs: without additional assumptions they are unidentifiable, whereas under faithfulness, their distribution equivalence class equals the Markov equivalence class; see, for example, [23].

The estimation of the Markov equivalence class of linear Gaussian SEMs in the low-dimensional case has been addressed in, for example, [5, 22], whereas the high-dimensional scenario (requiring sparsity of the true underlying DAG) is discussed in, for example, [2, 10, 13, 27].

An exception of identifiability of linear Gaussian SEMs occurs if all ε_j have equal variances $\sigma_j^2 = \sigma^2 > 0, \forall j$. Under this assumption, the true underlying DAG D is identifiable [16]. Yet, the assumption of equal noise variances seems to be overly restrictive in many scenarios. In general, the linearity assumption may be rather restrictive if not implausible in some cases.

1.2.3. *Special case of PLSEM*: *Nonlinear additive SEM with Gaussian noise*. Interestingly, the assumption of exclusively nonlinear functions $f_{j,i}$ in equation (1.2) greatly improves the identifiability properties; see [9] for the bivariate case and [18] for a general treatment. In fact, if all $f_{j,i}$ are nonlinear and three times differentiable, $\mathscr{D}(\mathbb{P})$ only consists of the single true underlying DAG D [18], Corollary 31(ii). The nonlinearity assumption is crucial, though. The authors provide an example where two DAGs are distribution equivalent if one of the nonlinear functions is replaced by a linear function [18], Example 26.

Various estimation methods have been introduced for additive nonlinear SEMs to infer the underlying DAG [14, 18, 26]. In particular, a restricted maximum likelihood estimation method called CAM, which is consistent in the low- and high-dimensional setting (assuming a sparse underlying DAG), has been proposed specifically for nonlinear additive SEMs with Gaussian noise [3].

1.2.4. *Identifiability of PLSEMs with non-Gaussian or arbitrary noise*. The identifiability properties of linear SEMs generally improve if one allows for non-Gaussian noise distributions. In fact, if all but one ε_j are assumed to be non-Gaussian (commonly referred to as LiNGAM setting), the underlying DAG D is

identifiable [21]. A general theory for linear SEMs with arbitrary noise distributions is presented in [8]. Both papers also propose estimation procedures for the respective model classes.

Unfortunately, the situation is different for PLSEMs: identifiability can be lost if one considers PLSEMs with non-Gaussian (or arbitrary) noise distributions. This can be seen from a specific example of a bivariate linear SEM with Gumbel-distributed noise, which is identifiable in the LiNGAM framework, but for which there exists a nonlinear additive backward model [9]. Still, this example seems to be rather particular. In fact, for bivariate additive SEMs, all unidentifiable cases of additive models can be classified into five categories; see [18, 31]. Based on bivariate identifiability, it has been shown that one can conclude multivariate identifiability under an additional assumption referred to as IFMOC assumption [17]. For instance, this approach allows to conclude identifiability of the multivariate LiNGAM and CAM settings and as such covers settings with both, Gaussian or non-Gaussian noise and all linear or all nonlinear functions. However, it is less explicit than the results presented in Section 2. In particular, it does not allow for a characterization of the distribution equivalence class of a PLSEM with Gaussian noise where some of the edge functions are linear and some are nonlinear.

1.3. *Our contribution*. As discussed in Section 1.2, there exists a wide "identifiability gap" for PLSEMs. Their identifiability has only been studied for the two special cases of linear SEMs and entirely nonlinear additive SEMs. Moreover, to the best of our knowledge, it has not yet been understood to what extent (single) nonlinear functions in additive SEMs with Gaussian noise restrict the underlying causal model. We close the "identifiability gap" for PLSEMs and answer the questions raised in Section 1.1.2 with the following theoretical results:

(A) A graphical representation of $\mathscr{D}(\mathbb{P})$ with a single partially directed graph $G_{\mathscr{D}(\mathbb{P})}$ in Section 2.1.1 (analogous to the use of CPDAGs to represent Markov equivalence classes).

(B) A transformational characterization of $\mathscr{D}(\mathbb{P})$ through sequences of covered *linear* edge reversals in Section 2.1.2 (analogous to the characterization of Markov equivalence classes via sequences of covered edge reversals in [6]).

(C) A functional characterization of PLSEMs in Section 2.2.1: all PLSEMs that generate the same distribution $\mathbb{P}$ are constant rotations of each other.

(D) A causal orderings characterization of PLSEMs in Section 2.2.2. In particular, it precisely specifies to what extent nonlinear functions in PLSEMs restrict the set of potential causal orderings.

The first two characterizations hold only under faithfulness, the third and fourth are general. We will give details on the precise interplay between nonlinearity and faithfulness in Section 2.3. Building on the transformational characterization result in (B), we provide an efficient score-based estimation procedure that outputs

the graphical representation $G_{\mathscr{D}(\mathbb{P})}$ in (A) given $\mathbb{P}$ and one DAG $D \in \mathscr{D}(\mathbb{P})$. The proposed algorithm only relies on sequences of local transformations and score computations, and hence is feasible for large graphs with numbers of variables in the thousands (assuming reasonable sparsity). We demonstrate its performance on simulated data. Moreover, we provide some robustness results for identifiability in the neighborhood of Gaussian noise and we derive (high-dimensional) consistency based on the consistency proof of the CAM methodology in [3].

2. Comprehensive characterization of PLSEMs. In this section, we present our main theoretical results. They consist of characterizations of PLSEMs that generate a given distribution $\mathbb{P}$ from various perspectives. In Section 2.1, we assume that $\mathbb{P}$ is faithful to the underlying causal model and demonstrate that this leads to a transformational characterization and a graphical representation of $\mathscr{D}(\mathbb{P})$ very similar to the well-known counterparts characterizing a Markov equivalence class. Our main theoretical contributions, which hold under very general assumptions and, in particular, do not rely on the faithfulness assumption, are presented in Section 2.2. They fully characterize all PLSEMs that generate a given distribution $\mathbb{P}$ on a functional level. Moreover, they explain how nonlinear functions impose very specific restrictions on the set of potential causal orderings. Section 2.3 brings together the two previous sections by discussing the precise interplay of nonlinearity and faithfulness.

2.1. *Characterizations of $\mathscr{D}(\mathbb{P})$ under faithfulness.* Let $\mathbb{P}$ be generated by a PLSEM with DAG $D \in \mathscr{D}(\mathbb{P})$. The goal of this section is to characterize $\mathscr{D}(\mathbb{P})$. Recall that $\mathscr{D}(\mathbb{P})$ is the set of all DAGs D such that $\mathbb{P}$ is faithful to D and there exists a PLSEM with DAG D that generates $\mathbb{P}$. In words, faithfulness means that no conditional independence relations other than those entailed by the Markov property hold; see, for example, [22]. In particular, it implies that $\mathscr{D}(\mathbb{P})$ is a subset of the Markov equivalence class and all DAGs in $\mathscr{D}(\mathbb{P})$ have the same skeleton and v-structures [28]. Markov equivalence classes can be graphically represented with single graphs, known as CPDAGs (also referred to as essential graphs, maximally oriented graphs or completed patterns) [1, 6, 12, 28], where an edge is directed if and only if it is oriented the same way in all the DAGs in the Markov equivalence class, else it is undirected. The Markov equivalence class then equals the set of all DAGs that can be obtained from the CPDAG by orienting the undirected edges without creating new v-structures. In Section 2.1.1, we derive an analogous graphical representation of $\mathscr{D}(\mathbb{P})$.

Another useful (transformational) characterization result says that any two Markov equivalent DAGs can be transformed into each other by a sequence of distinct covered edge reversals [6]. We will demonstrate in Section 2.1.2 that a very similar principle transfers to $\mathscr{D}(\mathbb{P})$.

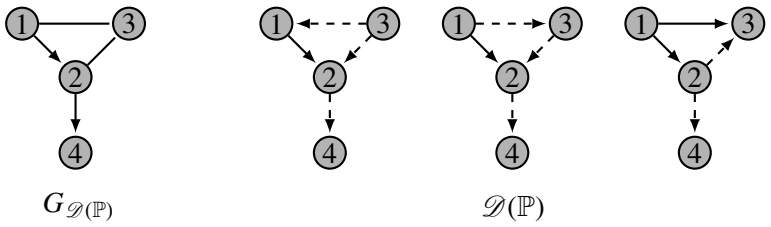

FIG. 2. *Graphical representation of $\mathscr{D}(\mathbb{P})$ with the single PDAG $G_{\mathscr{D}(\mathbb{P})}$. $\mathscr{D}(\mathbb{P})$ equals the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$. The graph with $2 \to 3 \to 1$ is not a consistent DAG extension of $G_{\mathscr{D}(\mathbb{P})}$ as it contains a cycle. Linear edges are dashed, nonlinear edges are solid.*

2.1.1. *Graphical representation of $\mathscr{D}(\mathbb{P})$.* The distribution equivalence class $\mathscr{D}(\mathbb{P})$ can be graphically represented by a single partially directed acyclic graph (PDAG). A PDAG is a graph with directed and undirected edges that does not contain any directed cycles. A *consistent DAG extension* of a PDAG is a DAG with the same skeleton, the same edge orientations on the directed subgraph of the PDAG and no additional v-structures.

DEFINITION 2.1. Let $\mathcal{E}$ be a set of Markov equivalent DAGs. We denote by $G_{\mathcal{E}}$ the PDAG that has the same skeleton as the DAGs in $\mathcal{E}$ and $i \to j$ in $G_{\mathcal{E}}$ if and only if $i \to j$ in all the DAGs in $\mathcal{E}$, else, $i - j$. We say that $G_{\mathcal{E}}$ is maximally oriented with respect to $\mathcal{E}$.

For a given distribution equivalence class $\mathscr{D}(\mathbb{P})$, the corresponding PDAG $G_{\mathscr{D}(\mathbb{P})}$ is uniquely defined by Definition 2.1. Moreover, $G_{\mathscr{D}(\mathbb{P})}$ is a graphical representation of $\mathscr{D}(\mathbb{P})$ in the following sense.

THEOREM 2.1. *$\mathscr{D}(\mathbb{P})$ equals the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$.*

A proof can be found in Section A in the supplement. Theorem 2.1 states that one can represent $\mathscr{D}(\mathbb{P})$ with a single PDAG $G_{\mathscr{D}(\mathbb{P})}$ without loss of information, as $\mathscr{D}(\mathbb{P})$ can be reconstructed from $G_{\mathscr{D}(\mathbb{P})}$ by listing all consistent DAG extensions. An example is given in Figure 2. Note that $G_{\mathscr{D}(\mathbb{P})}$ can be interpreted as a maximally oriented graph with respect to some background knowledge as defined in [12]. For details, we refer to Section 3.2.

Conceptually, this is analogous to the use of CPDAGs to represent Markov equivalence classes. There are important differences, though: first of all, necessary and sufficient conditions have been derived for a graph to be a CPDAG of a Markov equivalence class [1], Theorem 4.1. These properties do not all transfer to $G_{\mathscr{D}(\mathbb{P})}$. For example, $G_{\mathscr{D}(\mathbb{P})}$ typically is not a chain graph; see Figure 2. Second, given a DAG D, the CPDAG (and hence a full characterization of the Markov equivalence

class) can be obtained by an iterative application of three purely graphical orientation rules (R1–R3 in Figure 6) applied to the pattern of D [12]. This is not true for $G_{\mathscr{D}(\mathbb{P})}$ and $\mathscr{D}(\mathbb{P})$. It is still feasible to obtain $G_{\mathscr{D}(\mathbb{P})}$ from a DAG $D \in \mathscr{D}(\mathbb{P})$, but it is crucial to know which of the functions in the (unique) corresponding PLSEM (cf. Lemma B.2 in the supplement) are linear and which are nonlinear. We will show in Section 3 that the transformational characterization in Theorem 2.2 gives rise to a consistent and efficient score-based procedure to estimate $G_{\mathscr{D}(\mathbb{P})}$ based on $D \in \mathscr{D}(\mathbb{P})$ and samples of $\mathbb{P}$.

2.1.2. *Transformational characterization of* $\mathscr{D}(\mathbb{P})$. Given $D \in \mathscr{D}(\mathbb{P})$, the distribution equivalence class $\mathscr{D}(\mathbb{P})$ can be comprehensively characterized via sequences of local transformations of DAGs.

THEOREM 2.2. *Assume that* $\mathbb{P}$ *has been generated by a PLSEM and that it is faithful to the underlying DAG. Then the following two results hold*:

(a) *Let* $D \in \mathscr{D}(\mathbb{P})$, $i \to j$ *covered in* D, *and* D' *be the DAG that differs from* D *only by the reversal of* $i \to j$. *Then* $D' \in \mathscr{D}(\mathbb{P})$ *if and only if* $i \to j$ *is linear in* D. *Furthermore, if* $i \to j$ *is covered and nonlinear in* D, *then* $i \to j$ *in all DAGs in* $\mathscr{D}(\mathbb{P})$.

(b) *Let* $D, D' \in \mathscr{D}(\mathbb{P})$. *Then there exists a sequence of distinct covered linear edge reversals that transforms* D *to* D'.

A proof can be found in Section B in the supplement and an illustration is provided in Figure 3. Note that the interesting part of this result is that $\mathscr{D}(\mathbb{P})$ is connected with respect to covered linear edge reversals. It will be of particular importance in the design of score-based estimation procedures for $\mathscr{D}(\mathbb{P})$ and $G_{\mathscr{D}(\mathbb{P})}$ in Section 3.

Theorem 2.2 covers the two special cases discussed in Section 1.2: if all the functions $f_{j,i}$ in equation (1.2) are linear, $\mathscr{D}(\mathbb{P})$ (which, in this setting, is equal to the Markov equivalence class) can be fully characterized by sequences of covered edge reversals of D (as all the edges are linear). If, on the contrary, all the functions $f_{j,i}$ in equation (1.2) are nonlinear, $\mathscr{D}(\mathbb{P})$ only consists of the DAG D as there is no covered linear edge in D.

2.2. *General characterizations not assuming faithfulness*. In this section, we give general characterizations of PLSEMs that generate the same distribution $\mathbb{P}$, both, from the perspective of causal orderings and from a functional viewpoint. The functional characterization in Section 2.2.1 describes how the $f_{j,i}$ of different PLSEMs relate to each other. The characterization via causal orderings in Section 2.2.2 describes the set of causal orderings, such that there exists a corresponding PLSEM that generates the given distribution $\mathbb{P}$. It will show that nonlinear

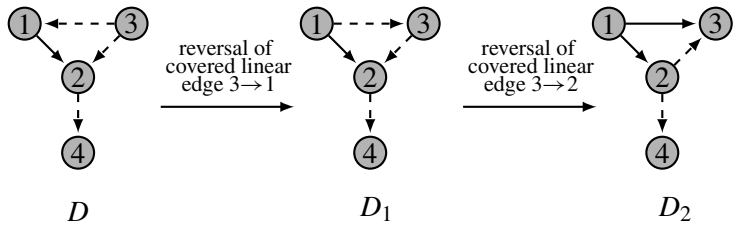

FIG. 3. *Transformational characterization of $\mathscr{D}(\mathbb{P})$ from Figure 2. Let $1 \to 2$ in D be nonlinear (solid) and all other edges in D be linear (dashed). Then D_1 and D_2 can be reached from D by the displayed sequence of covered linear edge reversals. Note that in D and D_2, $1 \to 2$ is covered but nonlinear, and hence cannot be reversed by Theorem 2.2(a). Moreover, $2 \to 4$ is not covered in any of D, D_1 and D_2, and hence cannot be reversed.*

functions impose a very specific structure on the model, which (perhaps surprisingly) is compatible with some of the previous theory on graphical models, as described in Section 1.2. Furthermore, it will help us understand in the general case how nonlinear functions restrict the set of PLSEMs that generate $\mathbb{P}$. Section 2.2.3 gives some intuition on the functional characterization in Section 2.2.1. Throughout this section, we assume that $\mathbb{P}$ is generated by a PLSEM as defined in equation (1.2).

2.2.1. *Functional characterization.* Let us first characterize the result on the level of SEMs. Consider a PLSEM that generates $\mathbb{P}$,

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) + \varepsilon_j,$$

where $f_{j,i}$, D, ε_j, μ_j, $\sigma_j^2 = \mathrm{Var}(\varepsilon_j)$ satisfy the assumptions from Section 1.1.1.

Let us define the function $F : \mathbb{R}^p \to \mathbb{R}^p$ by

$$F(x)_j := \frac{1}{\sigma_j}\Big(x_j - \mu_j - \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(x_i)\Big). \tag{2.1}$$

It turns out to be convenient to work with this function F. Notably, we do not lose any information by working with F instead of $f_{j,i}$, $\mathrm{pa}_D(j)$, μ_j and σ_j as these quantities can be recovered from F. Specifically, we can easily obtain the distribution of the errors from the function F as

$$\sigma_j := 1/\partial_j F_j. \tag{2.2}$$

By definition, $F(X) \sim \mathcal{N}(0, \mathrm{Id}_p)$. Note that F maps the observed random variable $X \in \mathbb{R}^p$ to the scaled residuals $\frac{\varepsilon_j}{\sigma_j}$. As for every $\varepsilon \in \mathbb{R}^p$, there exists exactly one $X \in \mathbb{R}^p$ that satisfies equation (1.2), F is invertible. Hence, if $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$, it

holds that $F^{-1}(Z) \sim X$. Using this, we obtain $\mu_j = \mathbb{E}_Z[F^{-1}(Z)_j]$ and we can recover the functions $f_{j,i}$ from the function F using the equations

$$f'_{j,i} = -\sigma_j \partial_i F_j \quad \text{and} \quad \mathbb{E}_Z f_{j,i}\big(F^{-1}(Z)_i\big) = 0. \tag{2.3}$$

Note that the equation on the left-hand side determines $f_{j,i}$ up to a constant, whereas the equation on the right-hand side determines the constant using only quantities that can be calculated from F. In the same spirit, $\mathrm{pa}_D(j)$ can be recovered from F via

$$\mathrm{pa}_D(j) = \{i \neq j : \partial_i F_j \not\equiv 0\}. \tag{2.4}$$

In this sense, instead of describing the PLSEM by $f_{j,i}, \mathrm{pa}_D(j), \mu_j$ and σ_j it can simply be described by the function $F : \mathbb{R}^p \to \mathbb{R}^p$. Now let us define

$$\mathcal{F}(\mathbb{P}) := \big\{F : \mathbb{R}^p \mapsto \mathbb{R}^p : F \text{ suffices (2.1) for a PLSEM that generates } \mathbb{P}\big\}.$$

We call the functions in this set *PLSEM-functions*. Let us define the set of orthonormal matrices $\mathcal{O}_n(\mathbb{R}) = \{O \in \mathbb{R}^{p\times p} : O O^t = \mathrm{Id}\}$. The following theorem follows from Lemma C.1 in the supplement. See also Remark C.4 in the supplement for details. It states that we can construct all PLSEMs that generate $\mathbb{P}$ by essentially *rotating* F.

THEOREM 2.3 (Characterization of potential PLSEMs). *For a given* $F \in \mathcal{F}(\mathbb{P})$, *there exists a set of* (*constant*) *rotations* $\mathcal{O}_{\mathcal{F}(\mathbb{P})} \subset \mathcal{O}_n(\mathbb{R})$ *such that*

$$\mathcal{F}(\mathbb{P}) = \{O \cdot F : O \in \mathcal{O}_{\mathcal{F}(\mathbb{P})}\}.$$

A description and explicit formulae for each $O \in \mathcal{O}_{\mathcal{F}(\mathbb{P})}$ *can be found in Remark* C.4 *in the supplement*.

Astonishingly, in this sense, all PLSEMs that generate $\mathbb{P}$ are rotations of each other. The importance of this result lies in its simplicity: There are very simple linear relationships between the $f_{j,i}$ in one PLSEM and the $\tilde{f}_{j,i}$ in another PLSEM. The formulae in Section C in the supplement permit to fully characterize these matrices $\mathcal{O}_{\mathcal{F}(\mathbb{P})}$. In fact, the characterization in Lemma C.1 in the supplement is the first step toward all other characterizations.

2.2.2. *Characterization via causal orderings*. This section discusses a characterization of all potential causal orderings of a given PLSEM. Let us define the set of *potential causal orderings* as

$$\mathcal{S}(\mathbb{P}) := \left\{ \begin{array}{l} \sigma \text{ permutation on } \{1,\dots,p\} : \text{there is a PLSEM with DAG } D \\ \text{that generates } \mathbb{P} \text{ such that } \sigma(i) < \sigma(j) \text{ for all } i \to j \text{ in } D \end{array} \right\}.$$

Without assuming faithfulness, if all $f_{j,i}$ are linear, all permutations of $\{1,\dots,p\}$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. That

is, $\mathcal{S}(\mathbb{P})$ is equal to the set of all permutations of $\{1, \dots, p\}$. Roughly, the more nonlinear functions in the PLSEM, the smaller the resulting set $\mathcal{S}(\mathbb{P})$. The interesting point is that nonlinear edges restrict $\mathcal{S}(\mathbb{P})$ in a very specific way. Before we state the theorem, consider a PLSEM that generates $\mathbb{P}$, define the function $F : \mathbb{R}^p \to \mathbb{R}^p$ as in equation (2.1) and define the set

$$\mathcal{V} := \{(i, j) \in \{1, \dots, p\}^2 : e_j^t (\mathrm{D}F)^{-1} \partial_i^2 F \not\equiv 0\}, \tag{2.5}$$

where e_j, $j = 1, \dots, p$ is the standard basis of $\mathbb{R}^p$, t stands for the transpose and $\mathrm{D}F$ denotes the Jacobian of F. We will discuss the interpretation of the set $\mathcal{V}$ and the expression $e_j^t (\mathrm{D}F)^{-1} \partial_i^2 F$ in more detail later. For now, the potential causal orderings can be characterized as follows.

THEOREM 2.4 (Characterization of potential causal orderings).

$$\mathcal{S}(\mathbb{P}) = \{\sigma \text{ permutation on } \{1, \dots, p\} : \sigma(i) < \sigma(j) \text{ for all } (i, j) \in \mathcal{V}\}.$$

The proof of this theorem can be found in Section D in the supplement. In words, all permutations of the indices that do not swap any of the tuples in $\mathcal{V}$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. And for all permutations of indices for which one of the tuples in $\mathcal{V}$ is switched, there exists *no* PLSEM with this causal ordering that generates $\mathbb{P}$. Moreover, by Lemma E.1(b) in the supplement, if $(i, j) \in \mathcal{V}$, then j is a descendant of i in every PLSEM that generates $\mathbb{P}$.

Now, let us give some intuition on the set $\mathcal{V}$. For $e_j^t (\mathrm{D}F)^{-1} \partial_i^2 F$ to be nonzero, it is necessary that there is a directed path from node i to node j that begins with a nonlinear edge. However, the existence of such a path is not sufficient, due to potential cancellations. An example is given in Figure 4 where the causal ordering of nodes 1 and 3 is not fixed even though $\partial_1^2 F_3 \not\equiv 0$. In particular, the requirement that the direct effect of i on j (the function $f_{j,i}$ in the PLSEM) is nonlinear, that is, the requirement that $\partial_i^2 F_j \not\equiv 0$, is *not* sufficient to fix the causal ordering between

FIG. 4. *Nonlinear edges can be reversed if nonlinear effects cancel out.* $X_1 = \varepsilon_1$, $X_2 = X_1^2 + X_1 + \varepsilon_2$, $X_3 = X_2 - X_1^2 + \varepsilon_3$ *with* $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_3)$ *generates the same joint distribution of* (X_1, X_2, X_3) *as* $X_3 = \tilde{\varepsilon}_3$, $X_1 = X_3/3 + \tilde{\varepsilon}_1$, $X_2 = X_1/2 + X_1^2 + X_3/2 + \tilde{\varepsilon}_2$ *with* $\tilde{\varepsilon}_3 \sim \mathcal{N}(0, 3), \tilde{\varepsilon}_1 \sim \mathcal{N}(0, 2/3), \tilde{\varepsilon}_2 \sim \mathcal{N}(0, 1/2)$ *independent. This stems from the fact that the nonlinear parts of the functions* $f_{2,1}(x)$ *and* $f_{3,1}(x)$ *cancel out, that is,* $f_{2,1}'' + f_{3,1}'' = 0$. *Note that this example does not contradict the previous theoretical results. It holds that* $e_3^t (\mathrm{D}F)^{-1} \partial_1^2 F \equiv 0$ *for the PLSEM-function* F *corresponding to* D_1. *Hence, the causal ordering of* D_2 *does not contradict Theorem* 2.4.

i and j. Also, it is not sufficient to require that the total effect of variable i on variable j is nonlinear. This is shown in part (a) of the following example.

EXAMPLE 2.1. Consider the DAG $1 \to 2 \to 3$ and $\mathbb{P}$ that has been generated by a PLSEM of the form $X_1 = \varepsilon_1$, $X_2 = f_{2,1}(X_1) + \varepsilon_2$, $X_3 = f_{3,2}(X_2) + \varepsilon_3$ with $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_3)$:

(a) Let $f_{2,1}(x) = 0.5x$ be linear, $f_{3,2}(x) = x^3$ be nonlinear. The corresponding PLSEM-function is $F(x) = (x_1, x_2 - 0.5x_1, x_3 - x_2^3)^t$. Using elementary calculations, it can be seen that $e_j^t(\mathrm{D}F)^{-1}\partial_i^2 F \not\equiv 0$ only for $(i, j) = (2, 3)$. Hence, $\mathcal{V} = \{(2, 3)\}$ and all permutations σ respecting $\sigma(2) < \sigma(3)$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. For example, for the causal ordering $\sigma(2) < \sigma(3) < \sigma(1)$, there exists a (unique) PLSEM with DAG $1 \leftarrow 2 \to 3$ that generates $\mathbb{P}$. In particular, the causal ordering of variables 1 and 3 is not fixed even though there is a nonlinear total effect of variable 1 on variable 3.

(b) Let $f_{2,1}(x) = x^3$ be nonlinear, $f_{3,2}(x) = 0.5x$ be linear. The corresponding PLSEM-function is $F(x) = (x_1, x_2 - x_1^3, x_3 - 0.5x_2)^t$. We obtain $\mathcal{V} = \{(1, 2), (1, 3)\}$ and all permutations σ with $\sigma(1) < \sigma(2)$ and $\sigma(1) < \sigma(3)$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. In particular, for $\sigma(1) < \sigma(3) < \sigma(2)$ we obtain that the PLSEM corresponding to the (unfaithful) DAG $1 \to 3 \to 2$ with $1 \to 2$ generates $\mathbb{P}$.

Let us make several concluding remarks: in (a), the causal ordering between nodes 1 and 3 is not fixed, whereas in (b), it is fixed. Hence, the set $\mathcal{V}$ sometimes also fixes the causal ordering between two nodes that are not adjacent in the DAG corresponding to F. Second, in both examples, the causal ordering of nodes incident to nonlinear edges is fixed. This raises the question whether it is true in general that nonlinear edges cannot be reversed. The answer is no (see Figure 4), but in some sense, the models with "reversible nonlinear edges" are rather particular. Finally, if we make additional mild assumptions, stronger statements can be made about the index tuples in $\mathcal{V}$. We will discuss these issues further in Section 2.3.

2.2.3. *Intuition on the functional characterization.* This section motivates Theorem 2.3. Consider two functions $F, G \in \mathcal{F}(\mathbb{P})$ that correspond to two different PLSEMs. By Proposition C.1 in the supplement,

$$F(X) \sim \mathcal{N}(0, \mathrm{Id}) \quad \text{and} \quad G(X) \sim \mathcal{N}(0, \mathrm{Id}). \tag{2.6}$$

Moreover, it follows from the definition of PLSEMs that F is invertible. Let $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$. Using equation (2.6) twice,

$$F^{-1}(Z) \sim X \quad \text{and} \quad G\big(F^{-1}(Z)\big) \sim \mathcal{N}(0, \mathrm{Id}).$$

Hence, the function $J : \mathbb{R}^p \to \mathbb{R}^p$, $J := G(F^{-1})$ suffices $J(Z) \sim Z \sim \mathcal{N}(0, \mathrm{Id})$. Furthermore, it can be shown that $|\det \mathrm{D}J| = 1$. Then, using the transformation

formula for probability densities, we obtain

$$\frac{1}{(2\pi)^{p/2}}\exp\Big(-\frac{\|J(x)\|_2^2}{2}\Big)=\frac{1}{(2\pi)^{p/2}}\exp\Big(-\frac{\|x\|_2^2}{2}\Big)\qquad\text{for all } x\in\mathbb{R}^p.$$

By rearranging,

$$\big\|J(x)\big\|_2=\|x\|_2\qquad\text{for all } x\in\mathbb{R}^p.$$

If we admit that J must be a linear function (which requires some work), this formula gives us $J\in\mathcal{O}_n(\mathbb{R}):=\{O\in\mathbb{R}^{p\times p}:OO^t=\mathrm{Id}\}$ and it immediately follows that $G=JF$. This reasoning shows that the main work in proving Theorem 2.3 lies in showing that J is a linear function.

2.3. *Understanding the interplay of nonlinearity and faithfulness.* As indicated in Section 2.2.2, without further assumptions, some nonlinear edges can be reversed. An example is given in Figure 4. There, the edge $1\to 3$ can be reversed even though $f_{3,1}$ is a nonlinear function in the PLSEM corresponding to D_1. The issue here arises because the nonlinear effect from X_1 to X_3 in D_1 cancels out over two paths. If we write X_3 as a function of $\varepsilon_1,\varepsilon_2,\varepsilon_3$, that function is linear. The setting of D_1 in Figure 4 is rather particular as $\partial_1^2 f_{2,1}$ and $\partial_1^2 f_{3,1}$ are linearly dependent. As the function space $\mathcal{C}^2(\mathbb{R})$ is infinite dimensional, this is arguably a degenerate scenario. Note that faithfulness does not save us from this cancellation effect as $\mathbb{P}$ is faithful to both, D_1 and D_2.

Nevertheless, we can rely on a different, rather weak assumption: consider a node i in a DAG D and assume that the corresponding functions in the set

$$\{\partial_i^2 f_{j',i} : j' \text{ is a child of } i \text{ in } D \text{ and } f_{j',i} \text{ is nonlinear}\}$$

are linearly independent. In other words: assume that the "nonlinear effects" from X_i on its children are linearly independent functions. Then these nonlinear edges cannot be reversed.

The following theorem is a direct implication of Lemma E.1(a) and (b) in the supplement.

THEOREM 2.5. *Consider a PLSEM and the corresponding distribution $\mathbb{P}$. Let j be a child of i in D and let $f_{j,i}$ be a nonlinear function. If the functions in the set $\{\partial_i^2 f_{j',i} : j' \text{ is a child of } i \text{ in } D \text{ and } f_{j',i} \text{ is nonlinear}\}$ are linearly independent, then j is a descendant of i in any other DAG D' of a PLSEM that generates $\mathbb{P}$.*

Intuitively, this should not be the end of the story: if an edge $i\to j$ is nonlinear, then usually there should also be a nonlinear relationship between i and the descendants of j. Hence, it should be possible to infer some statements about the causal ordering of i and the descendants of j. In general, this is not true as demonstrated in Figure 5.

FIG. 5. *If $\mathbb{P}$ is not faithful to D, descendants are not fixed. Node 4 is a descendant of node 1 in D_1 but not in D_2. On the left-hand side, $X_1 = \varepsilon_1$, $X_2 = X_1^2 + \varepsilon_2$, $X_3 = X_2 + \varepsilon_3$, $X_4 = X_3 - X_2 + \varepsilon_4$, with $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_4)$. On the right-hand side, $X_1 = \tilde{\varepsilon}_1$, $X_2 = X_1^2 + \tilde{\varepsilon}_2$, $X_3 = X_2 + 1/2 \cdot X_4 + \tilde{\varepsilon}_3$, $X_4 = \tilde{\varepsilon}_4$, where $\tilde{\varepsilon}_1 \sim \mathcal{N}(0,1)$, $\tilde{\varepsilon}_2 \sim \mathcal{N}(0,1)$, $\tilde{\varepsilon}_3 \sim \mathcal{N}(0,1/2)$ and $\tilde{\varepsilon}_4 \sim \mathcal{N}(0,2)$. Both PLSEMs generate the same distribution. Note that in this case, additional assumptions on the nonlinear function $f_{2,1}$ would not resolve the issue.*

Under the assumption of faithfulness, additional statements can be made about descendants of j. In some sense the nonlinear effect from i on the descendants of j, mediated through some of the descendants of j, cannot "cancel out." Hence, all descendants of j are fixed. The following theorem is a direct implication of Lemma E.1(c) and (d) in the supplement.

THEOREM 2.6. *Let the assumptions of Theorem 2.5 be true. In addition, let $\mathbb{P}$ be faithful to the DAG D. Fix $k \neq i$. Then k is a descendant of i in each DAG D' of a PLSEM that generates $\mathbb{P}$ if and only if k is a descendant of a nonlinear child of i in D.*

Note that we use the convention that a node is a descendant of itself. Theorem 2.6 guarantees that certain descendants of i are descendants of i in all DAGs D' of PLSEMs that generate $\mathbb{P}$. In that sense, it provides a simple criterion that tells us whether or not k is descendant of i in all of these DAGs. It is crucial to be precise: we do not assume that $\mathbb{P}$ is faithful to D', which means, we search over all PLSEMs that generate $\mathbb{P}$. If we search over the smaller space $\mathscr{D}(\mathbb{P})$, that is, additionally assume that $\mathbb{P}$ is faithful to D', the set of potential PLSEMs usually gets smaller. In many cases, there are some edges that are not fixed if we search over all PLSEMs, but fixed if we only search over PLSEMs with DAGs in $\mathscr{D}(\mathbb{P})$.

As discussed in Section 2.1.1, $\mathscr{D}(\mathbb{P})$ can be represented by a single PDAG $G_{\mathscr{D}(\mathbb{P})}$. In the following, we will discuss the estimation of $\mathscr{D}(\mathbb{P})$ and $G_{\mathscr{D}(\mathbb{P})}$.

3. Score-based estimation of $\mathscr{D}(\mathbb{P})$ and $G_{\mathscr{D}(\mathbb{P})}$. Consider $\mathbb{P}$ that has been generated by a PLSEM and assume that $\mathbb{P}$ is faithful to the underlying DAG. We denote by $\{X^{(i)}\}_{i=1,\dots,n}$ i.i.d. copies of $X \in \mathbb{R}^p$ and by $\mathbb{P}_n$ their empirical distribution. The goal of this section is to derive a consistent score-based estimation procedure for the distribution equivalence class $\mathscr{D}(\mathbb{P})$ based on $\mathbb{P}_n$ and one (true) DAG

$D^0 \in \mathscr{D}(\mathbb{P})$. We first describe a "naive" recursive solution that lists all members of $\mathscr{D}(\mathbb{P})$ and motivate the score-based approach in Section 3.1. We then present a more efficient procedure that directly estimates the graphical representation $G_{\mathscr{D}(\mathbb{P})}$ as defined in Section 3.2. Both methods rely on the transformational characterization result in Theorem 2.2.

In practice, we may replace the true D^0 by an estimate, for example, from the CAM methodology [3]. If the estimate is consistent for a DAG in $\mathscr{D}(\mathbb{P})$, we obtain consistency of our method for the entire distribution equivalence class $\mathscr{D}(\mathbb{P})$.

3.1. *Estimation of* $\mathscr{D}(\mathbb{P})$. Theorem 2.2 provides a straightforward way to list all members of $\mathscr{D}(\mathbb{P})$. Starting from the DAG D^0, one can search over all sequences of distinct covered linear edges reversals. By Theorem 2.2(a), all DAGs that are traversed are in $\mathscr{D}(\mathbb{P})$ and by Theorem 2.2(b), $\mathscr{D}(\mathbb{P})$ is connected with respect to sequences of distinct covered linear edge reversals. Moreover, by Theorem 2.2(a), an edge that is nonlinear and covered in a DAG in $\mathscr{D}(\mathbb{P})$ has the same orientation in all the members of $\mathscr{D}(\mathbb{P})$. These simple observations immediately lead to a recursive estimation procedure. Its population version is described in Algorithm 1. The inputs are D^0 (with all its edges marked as "unfixed") and an oracle that answers the question if a specific edge in a DAG in $\mathscr{D}(\mathbb{P})$ is linear or nonlinear.

Unfortunately, the (true) information whether a selected covered edge $i \to j$ in a DAG $D \in \mathscr{D}(\mathbb{P})$ is linear or not is generally not available. Also, it cannot simply be deduced from the starting DAG D^0 as the status of the edge may have changed in D. For an example, see Figure 1: edge $1 \to 3$ is not covered and linear in D_1 but nonlinear and covered (and hence irreversible) in $D_2 \in \mathscr{D}(\mathbb{P})$.

Algorithm 1 listAllDAGsPLSEM (population version)

if there is no covered edge in DAG D^0 that is marked as unfixed **then**
  Add D^0 to the distribution equivalence class $\mathscr{D}(\mathbb{P})$ and terminate.
end if
Choose a covered edge $i \to j$ in DAG D^0 that is marked as unfixed.
if the edge $i \to j$ is linear in D^0 **then**
  Define a DAG $D_1^0 := D^0$ with edge $i \to j$ in D_1^0 marked as fixed and a DAG D_2^0 equal to D^0 except for a reversed edge $i \leftarrow j$ marked as fixed in D_2^0.
  Call the function `listAllDAGsPLSEM` recursively for both DAGs D_1^0 and D_2^0.
else
  Mark the edge $i \to j$ in D^0 as fixed and call `listAllDAGsPLSEM` for DAG D^0.
end if

To check the status of a covered edge in a given DAG $D \in \mathscr{D}(\mathbb{P})$, one could either test (non)linearity of the functional component in the (unique) PLSEM corresponding to D or rely on a score-based approach. In the following, we are going to elaborate on the latter. We closely follow the approach presented in [3].

We assume that the functions $f_{j,i}$ in equation (1.2) are from a class of smooth functions $\mathcal{F}_i \subseteq \{f \in C^2(\mathbb{R}), \mathbb{E}[f(X_i)] = 0\}$, which is closed with respect to the $L_2(\mathbb{P}_{X_i})$-norm and closed under linear transformations. For a set of given basis functions, we denote by $\mathcal{F}_{n,i} \subseteq \mathcal{F}_i$ the finite-dimensional approximation space which typically increases as n increases. The spaces of additive functions with components in $\mathcal{F}_i$ and $\mathcal{F}_{n,i}$, respectively, are closed assuming an analogue of a minimal eigenvalue condition. All details are given in [3]. Without loss of generality, we assume $\mu_j = 0$ as in the original paper. For $D^0 \in \mathscr{D}(\mathbb{P})$, let $\theta^{D^0} := (\{f_{j,i}^{D^0}\}_{j=1,\dots,p, i \in \mathrm{pa}_{D^0}(j)}, \{\sigma_j^{D^0}\}_{j=1,\dots,p})$ be the infinite-dimensional parameter of the corresponding PLSEM. The expected negative log-likelihood reads

$$\mathbb{E}\big[-\log p_{\theta^{D^0}}(X)\big] = \sum_{j=1}^{p} \log\big(\sigma_j^{D^0}\big) + \mathrm{C}, \qquad C = \frac{p}{2}\log(2\pi) + \frac{p}{2}.$$

All $D^0 \in \mathscr{D}(\mathbb{P})$ lead to the minimal expected negative log-likelihood, as by definition, the corresponding PLSEM generates the true distribution $\mathbb{P}$. For a misspecified model with wrong DAG $D \notin \mathscr{D}(\mathbb{P})$, we obtain the projected parameter $\theta^D = (\{f_{j,i}^D\}_{j=1,\dots,p, i \in \mathrm{pa}_D(j)}, \{\sigma_j^D\}_{j=1,\dots,p})$ as

$$\{f_{j,i}^D\}_{i \in \mathrm{pa}_D(j)} = \operatorname*{argmin}_{g_{j,i} \in \mathcal{F}_i} \mathbb{E}\Bigg[\Bigg(X_j - \sum_{i \in \mathrm{pa}_D(j)} g_{j,i}(X_i)\Bigg)^2\Bigg],$$

$$\big(\sigma_j^D\big)^2 = \mathbb{E}\Bigg[\Bigg(X_j - \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}^D(X_i)\Bigg)^2\Bigg]$$

with expected negative log-likelihood

$$\mathbb{E}\big[-\log\big(p_{\theta^D}^D(X)\big)\big] = \sum_{j=1}^{p} \log\big(\sigma_j^D\big) + C, \qquad C = \frac{p}{2}\log(2\pi) + \frac{p}{2},$$

where all expectations are taken with respect to the true distribution $\mathbb{P}$. We refer to $\mathbb{E}[-\log(p_{\theta^D}^D(X))]$ as the *score of* D and to $\log(\sigma_j^D)$ as *score* of node j in D. For a DAG $D^0 \in \mathscr{D}(\mathbb{P})$, let

$$\mathscr{C}\big(D^0\big) = \big\{D \mid D \text{ and } D^0 \text{ differ by one covered nonlinear edge reversal}\big\}.$$

Then, for $D^0 \in \mathscr{D}(\mathbb{P})$ and $D \in \mathscr{C}(D^0)$ that (without loss of generality) only differ by the orientation of the covered edge between the nodes i and j, the difference in

expected negative log-likelihood is given as

$$\begin{aligned}&\mathbb{E}\big[-\log\big(p^D_{\theta^D}(X)\big)\big] - \mathbb{E}\big[-\log\big(p_{\theta^{D^0}}(X)\big)\big]\\ &\quad= \log\big(\sigma_i^D\big) + \log\big(\sigma_j^D\big) - \log\big(\sigma_i^{D^0}\big) - \log\big(\sigma_j^{D^0}\big).\end{aligned} \tag{3.1}$$

Since the score is decomposable over the nodes, the reversal of a covered edge only affects the scores locally at the two nodes i and j incident to the covered edge. We denote by

$$\xi_p := \min_{\substack{D^0\in\mathscr{D}(\mathbb{P})\\ D\in\mathscr{C}(D^0)}} \big(\mathbb{E}\big[-\log\big(p^D_{\theta^D}(X)\big)\big] - \mathbb{E}\big[-\log\big(p_{\theta^{D^0}}(X)\big)\big]\big) \tag{3.2}$$

the *degree of separation* of true models in $\mathscr{D}(\mathbb{P})$ and misspecified models in $\mathscr{C}(\mathscr{D}(\mathbb{P}))$ that can be reached by the reversal of one covered nonlinear edge in any DAG $D^0 \in \mathscr{D}(\mathbb{P})$. From the transformational characterization in Theorem 2.2, it follows that $\xi_p > 0$. Combining equations (3.1) and (3.2) motivates the estimation procedure in Algorithm 2 that takes as inputs n samples $X^{(1)}, \dots, X^{(n)}$ and a DAG $D^0 \in \mathscr{D}(\mathbb{P})$ (with all its edges marked as "unfixed") and outputs a score-based estimate $\widehat{\mathscr{D}}_{n,p}$ of $\mathscr{D}(\mathbb{P})$. To make the algorithm more robust with respect to

Algorithm 2 listAllDAGsPLSEM

if there is no covered edge in DAG D^0 that is marked as unfixed **then**
  Add D^0 to $\widehat{\mathscr{D}}_{n,p}$ and terminate.
end if
Choose a covered edge $i \to j$ in DAG D^0 that is marked as unfixed. Denote by D' the DAG that equals D^0 except for a reversed edge $i \leftarrow j$.
Additively regress X_i on $X_{\mathrm{pa}_{D^0}(i)}$, X_j on $X_{\mathrm{pa}_{D^0}(j)}$, X_i on $X_{\mathrm{pa}_{D^0}(i)\cup\{j\}}$, X_j on $X_{\mathrm{pa}_{D^0}(i)}$
Compute the standard deviations of the residuals to obtain $\hat\sigma_i^{D^0}$, $\hat\sigma_j^{D^0}$, $\hat\sigma_i^{D'}$ and $\hat\sigma_j^{D'}$.
Compute the score difference $\Delta := \log(\hat\sigma_i^{D'}) + \log(\hat\sigma_j^{D'}) - \log(\hat\sigma_i^{D^0}) - \log(\hat\sigma_j^{D^0})$
if $\Delta < \alpha$ **then**
  Set $D_1^0 := D^0$ with $i \to j$ marked as fixed, $D_2^0 := D'$ with $i \leftarrow j$ marked as fixed, $\alpha_1 := \alpha$ and $\alpha_2 := \alpha - \Delta$.
  Call the function `listAllDAGsPLSEM` recursively for both, DAG D_1^0 with parameter $\alpha = \alpha_1$ and DAG D_2^0 with parameter $\alpha = \alpha_2$.
else
  Mark the edge $i \to j$ in D^0 as fixed and call `listAllDAGsPLSEM` for DAG D^0 with parameter $\alpha = \alpha_1$.
end if

misspecifications of the noise distributions (cf. Section 4), we only perform one-sided tests in line 8 of Algorithm 2.

To prove the (high-dimensional) consistency of the score-based estimation procedure, we make the following assumptions. For a function $h : \mathbb{R} \to \mathbb{R}$, we write $P(h) = \mathbb{E}[h(X)]$ and $P_n(h) = \frac{1}{n}\sum_{i=1}^n h(X^{(i)})$.

ASSUMPTION 3.1. (i) Uniform upper bound on node degrees:

$$\max_{\substack{D \in \mathscr{D}(\mathbb{P}) \cup \mathscr{C}(\mathscr{D}(\mathbb{P})) \\ j=1,\dots,p}} \deg_D(j) \le M \qquad \text{for some positive constant } M < \infty.$$

(ii) Uniform lower bound on error variances:

$$\min_{\substack{D \in \mathscr{D}(\mathbb{P}) \cup \mathscr{C}(\mathscr{D}(\mathbb{P})) \\ j=1,\dots,p}} (\sigma_j^D)^2 \ge L > 0.$$

(iii) Empirical process bound:

$$\max_{\substack{D \in \mathscr{D}(\mathbb{P}) \cup \mathscr{C}(\mathscr{D}(\mathbb{P})) \\ j=1,\dots,p}} \Delta_{n,j}^D = o_P(1),$$

where $\Delta_{n,j}^D = \sup_{g_{j,i} \in \mathcal{F}_i} |(P_n - P)((X_j - \sum_{i \in \mathrm{pa}_D(j)} g_{j,i}(X_i))^2)|$.

(iv) Control of approximation error:

$$\max_{\substack{D^0 \in \mathscr{D}(\mathbb{P}) \\ j=1,\dots,p}} |\gamma_{n,j}^{D^0}| = o(1),$$

where

$$\gamma_{n,j}^{D^0} = \mathbb{E}\Big[\Big(X_j - \sum_{i \in \mathrm{pa}_{D^0}(j)} f_{n;j,i}^{D^0}(X_i)\Big)^2\Big] - \mathbb{E}\Big[\Big(X_j - \sum_{i \in \mathrm{pa}_{D^0}(j)} f_{j,i}^{D^0}(X_i)\Big)^2\Big]$$

with

$$f_{n;j,i}^{D^0} = \operatorname*{argmin}_{g_{j,i} \in \mathcal{F}_{n,i}} \mathbb{E}\Big[\Big(X_j - \sum_{i \in \mathrm{pa}_{D^0}(j)} g_{j,i}(X_i)\Big)^2\Big]$$

and $\mathcal{F}_{n,i}$ are the approximation spaces as introduced before.

Assumption 3.1(i) is satisfied if D^0 has bounded node degrees, as all DAGs under consideration are restricted to the same skeleton, and hence all have equal node degrees. In the low-dimensional setting, Assumption 3.1(iii) is justified by [3], Lemma 5, under the assumptions mentioned there. These assumptions entail smoothness conditions on the functions in $\mathcal{F}_i$ and tail and moment conditions on X. In the high-dimensional setting, it follows from [3], Lemma 6, and

$\sqrt{\log(p)/n} = o(1)$ together with Assumption 3.1(i) and the assumptions mentioned in the original paper. Assumption 3.1(iv) can be ensured by requiring a smoothness condition on the coefficients of the basis expansion for the true functions [3], Section 4.2. A proof of Theorem 3.1 can be found in Section F.1 in the supplement.

THEOREM 3.1. *Under Assumption 3.1 and $\xi_p \geq \xi_0 > 0$, for any constant $\alpha \in (0, \xi_0)$,*

$$\mathbb{P}\big[\widehat{\mathscr{D}}_{n,p} = \mathscr{D}(\mathbb{P})\big] \to 1 \qquad (n \to \infty).$$

In case of a high-dimensional setting, for which the uniformity in Assumption 3.1 is required, the convergence should be understood as poth $p \to \infty$ and $n \to \infty$.

REMARK 3.1. The assumption on the gap between log-likelihoods of true and wrong models in [3] is stricter and would imply the uniform bound $\xi_p/p \geq \xi_0 > 0$, whereas here we only require $\xi_p \geq \xi_0 > 0$. As we are given a true DAG $D^0 \in \mathscr{D}(\mathbb{P})$, we solely perform local transformations of DAGs thanks to the transformational characterization result in Theorem 2.2. This only affects the scores of two nodes and allows us to rely on this much weaker gap condition.

3.2. *Estimation of $G_{\mathscr{D}(\mathbb{P})}$.* The estimation of all DAGs in $\mathscr{D}(\mathbb{P})$ is feasible but may be computationally intractable in the presence of many linear edges. For example, if D^0 is a fully connected DAG with p nodes and all its edges are linear, the number of DAGs in $\mathscr{D}(\mathbb{P})$ corresponds to the number of causal orderings of p nodes which is $p!$. It therefore would be desirable to have a procedure that works without enumerating all DAGs in $\mathscr{D}(\mathbb{P})$. In this section, we are going to describe such a procedure that directly estimates the maximally oriented PDAG $G_{\mathscr{D}(\mathbb{P})}$ defined in Section 2.1.1. Recall that by Theorem 2.1, this fully characterizes $\mathscr{D}(\mathbb{P})$, as $\mathscr{D}(\mathbb{P})$ can be recovered from $G_{\mathscr{D}(\mathbb{P})}$ by listing all consistent DAG extensions.

The main idea is the following: instead of traversing the space of DAGs, we traverse the space of maximally oriented PDAGs that represent sets of distribution equivalent DAGs. As an example, let $D^0 \in \mathscr{D}(\mathbb{P})$ and $i \to j$ be covered and linear in D^0. By Theorem 2.2(a), the DAG D' that only differs from D^0 by the reversal of $i \to j$ is in $\mathscr{D}(\mathbb{P})$. Instead of memorizing both, D^0 and D', and recursively searching over sequences of covered linear edge reversals from both of these DAGs as in Algorithms 1 and 2, we represent D^0 and D' by the PDAG G that is maximally oriented with respect to the set of DAGs $\{D^0, D'\}$. By Definition 2.1, G equals D^0 but for an undirected edge $i - j$. To construct $G_{\mathscr{D}(\mathbb{P})}$, the idea is now to iteratively modify G by either fixing or removing orientations of directed edges if they are nonlinear or linear in one of the consistent DAG extensions of G in which they are covered. For that to work based on G only, that is, without listing all consistent DAG extensions of G, the two key questions are the following:

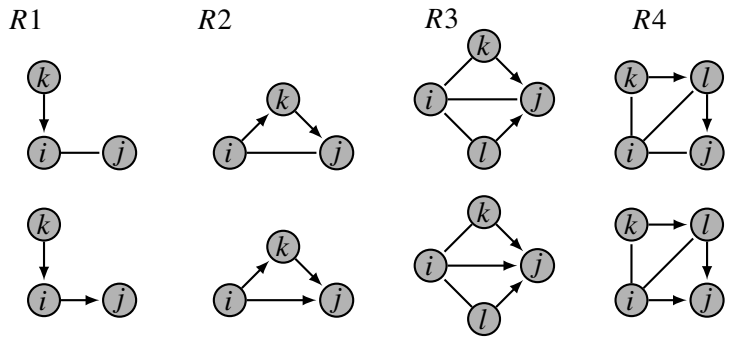

FIG. 6. *Orientation rules R1–R4 for Markov equivalence classes with background knowledge from* [12]. *If there is an edge constellation as in the top row, $i - j$ is oriented as $i \to j$ when closing orientations under R1–R4.*

(Q1) For $i \to j$ in a maximally oriented PDAG G, can we decide based on G only if there is a consistent DAG extension of G in which $i \to j$ is covered?

(Q2) If $i \to j$ is known to be covered in a consistent DAG extension of G: can we derive a score-based check if $i \to j$ is linear or nonlinear in this extension based on G?

Interestingly, the answer to both questions is yes (cf. Lemma 3.1) and can be derived from a related theory on how background knowledge on specific edge orientations restricts the Markov equivalence class. It was shown in [12], Theorems 2 and 4, that for a pattern P of a DAG, consistent background knowledge $\mathcal{K}$ (in our case: additional knowledge on edge orientations due to nonlinear functions in the PLSEM) can be incorporated by simply orienting these edges in P and closing orientations under a set of four sound and complete graphical orientation rules R1–R4, which are depicted in Figure 6. The resulting PDAG, which we denote by $G_{P,\mathcal{K}}$, is maximally oriented with respect to the set of all Markov equivalent DAGs with edge orientations that comply with the background knowledge. It is important to note that we generally do not obtain $G_{\mathscr{D}(\mathbb{P})}$ if we simply add all nonlinear edges in D^0 as background knowledge $\mathcal{K}$ and close orientations under R1–R4. The resulting maximally oriented PDAG $G_{P,\mathcal{K}}$ is typically not equal to $G_{\mathscr{D}(\mathbb{P})}$. For an example, consider D_1 in Figure 1 and denote by P_1 its pattern. For $\mathcal{K} = \{1 \to 2\}$, we obtain the PDAG $G_{P_1,\mathcal{K}}$ with undirected edge $1 - 3$. But $1 \to 3$ in $G_{\mathscr{D}(\mathbb{P})}$ by Definition 2.1 as $\mathscr{D}(\mathbb{P}) = \{D_1, D_2\}$. This illustrates that we have to add all edges to $\mathcal{K}$ that are nonlinear in a DAG in $\mathscr{D}(\mathbb{P})$ in which they are covered ($1 \to 3$ is covered and nonlinear in D_2).

LEMMA 3.1. *Let P be the pattern of a DAG and $\mathcal{K}$ a consistent set of background knowledge (not containing directed edges of P). Let $G_{P,\mathcal{K}}$ denote the maximally oriented graph with respect to P and $\mathcal{K}$ with orientations closed under R1–R4:*

(a) *Edge $i \to j$ in $\mathcal{K}$ is not covered in any of the consistent DAG extensions of $G_{P,\mathcal{K}}$ if and only if $G_{P,\mathcal{K}} = G_{P,\mathcal{K}\setminus\{i\to j\}}$.*

(b) *If $G_{P,\mathcal{K}} \neq G_{P,\mathcal{K}\setminus\{i\to j\}}$, there exists a consistent DAG extension of $G_{P,\mathcal{K}}$ in which $\mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$ is a cover for $i \to j$.*

A proof is given in Section F.2 in the supplement. By construction, $G_{P,\mathcal{K}} = G_{P,\mathcal{K}\setminus\{i\to j\}}$ if and only if the orientation of $i \to j$ in $G_{P,\mathcal{K}\setminus\{i\to j\}}$ is implied by one of R1–R4 applied to $G_{P,\mathcal{K}}$ with undirected edge $i - j$. Hence, Lemma 3.1(a) answers (Q1) as it provides a simple graphical criterion to check whether $i \to j$ in $G_{P,\mathcal{K}}$ is covered in one of the consistent DAG extensions of $G_{P,\mathcal{K}}$ based on $G_{P,\mathcal{K}}$ only. Note that part (a) is closely related to [1], Section 5, where the authors construct the CPDAG (representing the Markov equivalence class) from a given DAG by removing edge orientations that are not implied by a set of graphical orientation rules, which contain R1–R3 in Figure 6. Lemma 3.1(b) answers (Q2): it allows us to implement a score-based check whether $i \to j$ is linear or nonlinear in a DAG extension of $G_{P,\mathcal{K}}$ in which it is covered by simply reading off the parents of j in $G_{P,\mathcal{K}}$ and use them as a cover for $i \to j$. Details are given in Remark 3.2.

We now propose the following iterative estimation procedure for $G_{\mathscr{D}(\mathbb{P})}$: let $D^0 \in \mathscr{D}(\mathbb{P})$ be given, P denote its pattern and define $\mathcal{K}_1 := \mathcal{K}_1^{\mathrm{init}} \cup \mathcal{K}_1^{\mathrm{nonl}}$, where $\mathcal{K}_1^{\mathrm{init}}$ contains all directed edges in D^0 that are undirected in P and $\mathcal{K}_1^{\mathrm{nonl}} := \varnothing$. By construction, $G_{P,\mathcal{K}_1} = D^0$. For $k \geq 1$, in each iteration k to $k+1$, we apply Lemma 3.1(a) and use R1–R4 to select $\{i \to j\} \in \mathcal{K}_k^{\mathrm{init}}$ ($i \to j$ in $G_{P,\mathcal{K}_k}$) that is covered in a consistent DAG extension of $G_{P,\mathcal{K}_k}$ (i.e., not implied by any of R1–R4). If $\mathcal{K}_k^{\mathrm{init}} = \varnothing$ or no such edge exists, we stop and output $G_{P,\mathcal{K}_k}$. Else, we check whether $i \to j$ is linear or nonlinear in a consistent DAG extension in which it is covered and construct a new set of background knowledge $\mathcal{K}_{k+1} := \mathcal{K}_{k+1}^{\mathrm{init}} \cup \mathcal{K}_{k+1}^{\mathrm{nonl}} \subseteq \mathcal{K}_k$ according to the following rules:

Case 1: If $i \to j$ is linear, $\mathcal{K}_{k+1}^{\mathrm{nonl}} = \mathcal{K}_k^{\mathrm{nonl}}$ and $\mathcal{K}_{k+1}^{\mathrm{init}} = \mathcal{K}_k^{\mathrm{init}} \setminus \{i \to j\}$.

Case 2: If $i \to j$ is nonlinear, $\mathcal{K}_{k+1}^{\mathrm{nonl}} = \mathcal{K}_k^{\mathrm{nonl}} \cup \{i \to j\}$; $\mathcal{K}_{k+1}^{\mathrm{init}} = \mathcal{K}_k^{\mathrm{init}} \setminus \{i \to j\}$.

In particular, by construction, *Case 1* implies that $i - j$ in all $G_{P,\mathcal{K}_l}$ for $l > k$, whereas *Case 2* fixes the orientation $i \to j$ in all $G_{P,\mathcal{K}_l}$ for $l > k$.

LEMMA 3.2. *Let $\{\mathcal{K}_k\}_k$ be constructed as above. Then the corresponding sequence of maximally oriented PDAGs $\{G_{P,\mathcal{K}_k}\}_k$ converges to $G_{\mathscr{D}(\mathbb{P})}$.*

A proof is given in Section F.3 in the supplement and an illustration is provided in Figure 7. As in both cases, $|\mathcal{K}_{k+1}^{\mathrm{init}}| = |\mathcal{K}_k^{\mathrm{init}}| - 1$, $\{G_{P,\mathcal{K}_k}\}_k$ converges to $G_{\mathscr{D}(\mathbb{P})}$ after at most $|\mathcal{K}_1^{\mathrm{init}}|$ iterations, where $|\mathcal{K}_1^{\mathrm{init}}|$ is the number of undirected edges in P.

REMARK 3.2. Let $\{i \to j\} \in \mathcal{K}_k^{\mathrm{init}}$ be the edge chosen in iteration k to $k+1$. By Lemma 3.1(b), $S := \mathrm{pa}_{G_{P,\mathcal{K}_k}}(j) \setminus \{i\}$ is a cover of $i \to j$ in one of the consistent

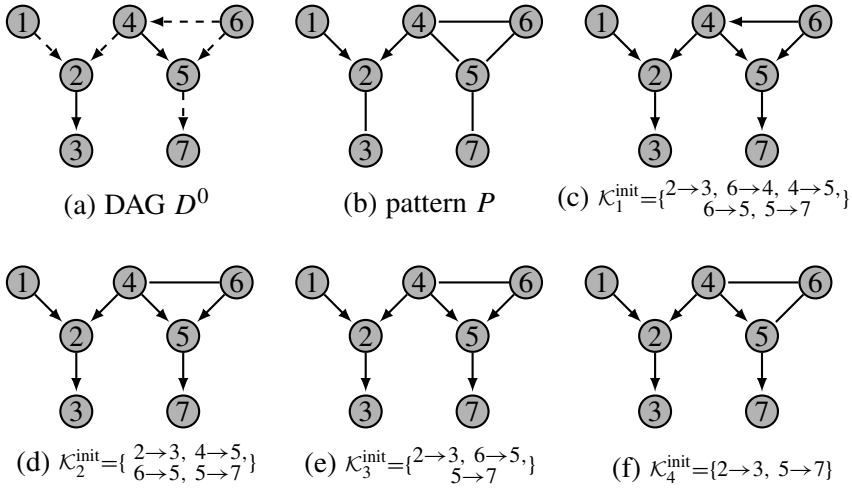

FIG. 7. *Illustration of Algorithm 3.* (a) *DAG D^0 with linear edges (dashed) and nonlinear edges (solid).* (b) *step 2: pattern P of D^0.* (c) *step 3: directed edges in D^0 that are undirected in P are added to $\mathcal{K}_1^{\text{init}}$. By construction, $\widehat{G}_{n,p} = D^0$.* (c)–(f) *steps 4–12: $4 \leftarrow 6$ is covered and linear in* (c), *hence, orientation is removed in $\widehat{G}_{n,p}$ in* (d). *$4 \to 5$ is covered and nonlinear in* (c), *hence, orientation is fixed in $\widehat{G}_{n,p}$ in* (e). *$6 \to 5$ is covered and linear in a consistent DAG extension of* (e), *hence, orientation is removed in $\widehat{G}_{n,p}$ in* (f). *As both edges in $\mathcal{K}_4^{\text{init}}$ are implied by R1 in* (f), *they are not covered in any of the consistent DAG extensions of $\widehat{G}_{n,p}$ in* (f). *Concludingly, $\widehat{G}_{n,p} = G_{\mathscr{D}(\mathbb{P})}$ in* (f).

DAG extensions of $G_{P,\mathcal{K}_k}$. From that, we easily obtain a score-based version: we simply regress X_i on X_S and X_j on $X_{S\cup\{i\}}$ to obtain the estimates $\hat{\sigma}_i, \hat{\sigma}_j$ of the standard deviations of the residuals at nodes i and j for $i \to j$. Similarly, we regress X_i on $X_{S\cup\{j\}}$ and X_j on X_S to get $\hat{\sigma}_i', \hat{\sigma}_j'$ for $i \leftarrow j$. If the estimated score difference $|\log(\hat{\sigma}_i') + \log(\hat{\sigma}_j') - \log(\hat{\sigma}_i) - \log(\hat{\sigma}_j)|$ is smaller than α, we conclude that $i \to j$ is linear, else nonlinear. The pseudo-code of the score-based procedure is provided in Algorithm 3. It outputs an estimate $\widehat{G}_{n,p}$ of $G_{\mathscr{D}(\mathbb{P})}$ based on n samples $X^{(1)}, \ldots, X^{(n)}$ and $D^0 \in \mathscr{D}(\mathbb{P})$.

A major advantage of Algorithm 3 is that it can be implemented based on one adjacency matrix only that is updated in every iteration.

THEOREM 3.2. *Under Assumption 3.1 and $\xi_p \geq \xi_0 > 0$, for any constant $\alpha \in (0, \xi_0)$,*

$$\mathbb{P}[\widehat{G}_{n,p} = G_{\mathscr{D}(\mathbb{P})}] \to 1 \qquad (n \to \infty)$$

PROOF. The correctness of Algorithm 3 is proved in Lemma 3.2. The consistency of the score-based estimation follows from the proof of Theorem 3.1. □

Algorithm 3 computeGDPX

Initialize $\widehat{G}_{n,p} \leftarrow D^0$, $k \leftarrow 1$, $\mathcal{K}_1^{\text{init}} \leftarrow \varnothing$ and $\mathcal{K}_1^{\text{nonl}} \leftarrow \varnothing$.
Construct the pattern P of D^0.
Add directed edges in D^0 that are undirected in P to $\mathcal{K}_1^{\text{init}}$.
while There is $i \to j$ in $\mathcal{K}_k^{\text{init}}$, such that its orientation is not implied by applying rules R1, R2, R3 or R4 to $\widehat{G}_{n,p}$ with undirected edge i — j **do**
  Use $\text{pa}_{\widehat{G}_{n,p}}(j) \setminus \{i\}$ to cover $i \to j$ and estimate the standard deviations $\hat{\sigma}_i, \hat{\sigma}_j, \hat{\sigma}_i'$ and $\hat{\sigma}_j'$ of the residuals as described in Remark 3.2.
  if $|\log(\hat{\sigma}_i') + \log(\hat{\sigma}_j') - \log(\hat{\sigma}_i) - \log(\hat{\sigma}_j)| < \alpha$ **then**
    Set $\mathcal{K}_{k+1}^{\text{init}} \leftarrow \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$ and replace $i \to j$ by i — j in $\widehat{G}_{n,p}$.
  else
    Set $\mathcal{K}_{k+1}^{\text{init}} \leftarrow \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$ and keep $i \to j$ in $\widehat{G}_{n,p}$.
  end if
  $k \leftarrow k + 1$.
end while
return Estimated PDAG $\widehat{G}_{n,p}$ representing $\mathscr{D}(\mathbb{P})$.

4. Model misspecification. In this section, we will discuss how small deviations from a Gaussian error distribution affect the distribution equivalence class and how the output of the algorithm `listAllDAGsPLSEM` should be interpreted in this case. We define a *generalized PLSEM* by essentially dropping the assumption of Gaussianity of the noise variables from the definition of a PLSEM.

DEFINITION 4.1 (generalized PLSEM). A *generalized PLSEM* with DAG D is a partially linear additive SEM of the form

$$\mathring{X}_j = \mathring{\mu}_j + \sum_{i \in \text{pa}_D(j)} \mathring{f}_{j,i}(\mathring{X}_i) + \mathring{\varepsilon}_j, \tag{4.1}$$

where $\mathring{\mu}_j \in \mathbb{R}$, $\mathring{f}_{j,i} \in C^2(\mathbb{R})$, $\mathring{f}_{j,i} \not\equiv 0$, with $\mathbb{E}[\mathring{f}_{j,i}(X_i)] = 0$, and the noise variables $\mathring{\varepsilon}_j$ are centered with variance $\mathring{\sigma}_j^2 > 0$, have positive density on $\mathbb{R}$ and are jointly independent for $j = 1, \ldots, p$.

In analogy to before, without loss of generality, we assume $\mathring{\mu}_j = 0$, $j = 1, \ldots, p$ and define projected parameters. Furthermore, for a DAG D, we will define the projected density $\mathring{p}_{\mathring{\theta}^D}^D$. Consider $\mathring{X} \sim \mathring{\mathbb{P}}$ generated by a generalized PLSEM with DAG D^0. For each DAG D that is Markov equivalent to D^0, define

$$\{\mathring{f}_{j,i}^D\}_{i \in \text{pa}_D(i)} = \operatorname*{argmin}_{g_{j,i} \in \mathcal{F}_i} \mathbb{E}\Bigg[\Bigg(\mathring{X}_j - \sum_{i \in \text{pa}_D(i)} g_{j,i}(\mathring{X}_i)\Bigg)^2\Bigg],$$

$$(\mathring{\sigma}_j^D)^2 = \mathbb{E}\Big[\Big(\mathring{X}_j - \sum_{i \in \mathrm{pa}_D(j)} \mathring{f}_{j,i}(\mathring{X}_i)\Big)^2\Big],$$

$$\mathring{p}_{\mathring{\theta}^D}^D(x) = \prod_{j=1}^p \mathring{q}_j\Big(x_j - \sum_{i \in \mathrm{pa}_D(j)} \mathring{f}_{j,i}(x_i)\Big),$$

where $\mathring{q}_j$ denotes the density of $\mathring{X}_j - \sum_{i \in \mathrm{pa}_D(j)} \mathring{f}_{j,i}(\mathring{X}_i)$ for $j = 1, \ldots, p$. Analogously, define the projected density $p_{\theta^D}^D$ of $X \sim \mathbb{P}$ generated by a (Gaussian) PLSEM. Note that here $p_{\theta^D}^D$ denotes the projected density of X with respect to generalized PLSEMs, in contrast to Section 3 where it denotes the projected density of X with respect to (Gaussian) PLSEMs.

The first question we answer is: How does the algorithm `listAllDAGsPLSEM` behave when the error distributions are non-Gaussian (and the algorithm wrongly assumes Gaussianity)? The following result answers this question in the population case. Let $X \sim \mathbb{P}$ be generated by a PLSEM with the same DAG, same edge functions, same error variances as the generalized PLSEM that generates $\mathring{X}$, but with Gaussian errors. It turns out that the DAGs in the distribution equivalence class $\mathscr{D}(\mathbb{P})$ have an interesting property. For all $D \in \mathscr{D}(\mathbb{P})$, the computed scores are lower than the score for D^0. The proof of Theorem 4.1 can be found in Section G in the supplement.

THEOREM 4.1. *For all* $D \in \mathscr{D}(\mathbb{P})$,

$$\sum_{j=1}^p \log \mathring{\sigma}_j^D \le \sum_{j=1}^p \log \mathring{\sigma}_j^{D^0}.$$

Hence, if the algorithm `listAllDAGsPLSEM` starts at D^0 with $\alpha \ge 0$, it will never reject any $D \in \mathscr{D}(\mathbb{P})$. The output of the algorithm will hence be a superset of $\mathscr{D}(\mathbb{P})$.

From a theoretical perspective, the other question might be more interesting: what statements can be made about the distribution equivalence class of $\mathring{X} \sim \mathring{\mathbb{P}}$? To be more precise, for a distribution $\mathring{\mathbb{P}}$ that has been generated by a faithful generalized PLSEM, we call the set of DAGs

$$\mathring{\mathscr{D}}(\mathring{\mathbb{P}}) := \left\{ D \,\middle|\, \begin{array}{l} \mathring{\mathbb{P}} \text{ is faithful to } D \text{ and there exists a} \\ \text{generalized PLSEM with DAG } D \text{ that generates } \mathring{\mathbb{P}} \end{array} \right\}$$

the *(generalized PLSEM) distribution equivalence class*. How do small violations of Gaussianity affect the distribution equivalence class? Intuitively, identification of certain edges should get easier, in the sense that previously identified edges stay identified. This intuition turns out to be correct. The following theorem tells us that small deviations from the Gaussian error distribution can only make the distribution equivalence class smaller.

THEOREM 4.2. *Let*

$$\left|\mathbb{E}\left[\log p^D_{\theta^D}(X)\right] - \mathbb{E}\left[\log \mathring{p}^D_{\mathring{\theta}^D}(\mathring{X})\right]\right| < \zeta$$

for all DAGs $D \sim D^0$ (*all DAGs* D *that are Markov equivalent to* D^0) *for* $\zeta > 0$ *sufficiently small. Then we have*

$$\mathring{\mathscr{D}}(\mathring{\mathbb{P}}) \subseteq \mathscr{D}(\mathbb{P}).$$

The proof of Theorem 4.2 and the definition of a feasible $\zeta > 0$ can be found in Section G in the supplement. In words, the assumption requires that the projected log-likelihoods of X and $\mathring{X}$ do not differ too much for all DAGs $D \sim D^0$. If the error distribution of $\mathring{\varepsilon}$ is close to Gaussian, then the distributions of X and $\mathring{X}$ are close and the assumption is fulfilled.

We now collect the implications of these theorems for the population case. By Theorem 4.1, the output of the algorithm `listAllDAGsPLSEM` is a superset of $\mathscr{D}(\mathbb{P})$. Furthermore, under the assumptions of Theorem 4.2, $\mathscr{D}(\mathbb{P})$ is a superset of $\mathring{\mathscr{D}}(\mathring{\mathbb{P}})$. Hence, the algorithm is conservative in the sense that it will return a superset of the true underlying distribution equivalence class $\mathring{\mathscr{D}}(\mathring{\mathbb{P}})$. In particular, it will not draw any wrong causal conclusions as it will not return incorrectly oriented edges.

Does `listAllDAGsPLSEM` sometimes return a proper superset of $\mathring{\mathscr{D}}(\mathring{\mathbb{P}})$? Intuitively, the algorithm only orients edges that are identified due to nonlinear edge functions. However, edges in generalized PLSEMs can sometimes be identified due to non-Gaussianity of certain error distributions. The algorithm `listAllDAGsPLSEM` does not take the latter into account. In such a case, under the assumptions of Theorem 4.2, the algorithm will usually output a proper superset of the distribution equivalence class. An example can be found below. To compute the distribution equivalence class $\mathring{\mathscr{D}}(\mathring{\mathbb{P}})$, we recommend to compute the log-likelihoods of the output of `listAllDAGsPLSEM` with a nonparametric log-likelihood estimator (e.g., in the spirit of [14]) and keep the DAGs with the largest corresponding log-likelihoods. Under the assumptions of Theorem 4.2, this would return the exact generalized PLSEM distribution equivalence class $\mathring{\mathscr{D}}(\mathring{\mathbb{P}})$. In this case, the main benefit of `listAllDAGsPLSEM` is to reduce the computational burden compared to more naive approaches, such as computing nonparametric log-likelihood estimates of all DAGs in the Markov equivalence class.

EXAMPLE 4.1. Consider the generalized PLSEM $X_1 \leftarrow \varepsilon_1$, $X_2 \leftarrow \frac{1}{\sqrt{2}}X_1 + \varepsilon_2$, where ε_1 and ε_2 both follow a scaled t-distribution, with $\mathrm{Var}(\varepsilon_1) = 1$ and $\mathrm{Var}(\varepsilon_2) = \frac{1}{2}$. From [9], it follows that there exists no additive backward model, that is, there exists no generalized PLSEM with $X_2 \to X_1$ that generates the given distribution of (X_1, X_2). However, the "residuals" $r_1 := X_2$ and $r_2 := X_1 - \frac{1}{\sqrt{2}}X_2$ satisfy $\mathrm{Var}(r_1) = 1$ and $\mathrm{Var}(r_2) = \frac{1}{2}$. Hence, the projected (Gaussian) log-likelihoods of these two models match. In this case, `listAllDAGsPLSEM` would return

the two DAGs $X_1 \to X_2$ and $X_2 \to X_1$, which is a strict superset of $\mathring{\mathscr{D}}(\mathring{\mathbb{P}}) = \{X_1 \to X_2\}$.

5. Simulations. In this section, we empirically analyze the performance of `computeGDPX` (Algorithm 3) in various settings. Consider $\mathbb{P}$ that has been generated by a faithful PLSEM with known DAG D^0. The goal is to estimate the corresponding distribution equivalence class $\mathscr{D}(\mathbb{P})$ based on D^0 and samples of $\mathbb{P}$. In Section 5.1, we start with a description of the simulation setting. We then briefly comment on a population version of Algorithm 3 in Section 5.2, which is used to obtain the underlying true distribution equivalence class $\mathscr{D}(\mathbb{P})$. In the subsequent sections, we examine the role of the tuning parameter α (Section 5.3), the performance in low- and high-dimensional settings (Section 5.4) and the computation time (Section 5.5).

5.1. *Simulation setting and implementation details.* Throughout the section, let p denote the number of variables, n the number of samples, n_{rep} the number of repetitions of an experiment, p_{c} the probability to connect two nodes by an edge and p_{lin} the probability that an edge is linear. For each experiment, we generate n_{rep} random true DAGs D^0 with the function `randomDAG` in the R-package `pcalg` [11] with parameters `n = p` and `prob = p_c`. For each of the random DAGs, we generate n samples of $\mathbb{P}$ from a PLSEM with edge functions chosen as follows: with probability p_{lin}, $f_{j,i}(x) = \alpha_{j,i} \cdot x$ is linear with $\alpha_{j,i}$ randomly drawn from $[-1.5, -0.5] \cup [0.5, 1.5]$. Otherwise, $f_{j,i}(x)$ is nonlinear and randomly drawn from the set $\{c_0 \cdot \cos(c_1 \cdot (x - c_2)), c_0 \cdot \tanh(c_1 \cdot (x - c_2))\}$ to have a mix of monotone and nonmonotone functions in the PLSEM. In order to be able to empirically support our theoretical findings, we choose the parameters $c_0 \sim \text{Unif}([-2, -1] \cup [1, 2])$, $c_1 \sim \text{Unif}([1, 2])$ and $c_2 \sim \text{Unif}([-\pi/3, \pi/3])$ such that the nonlinear functions are "sufficiently nonlinear" and not too close to linear functions. Exemplary randomly generated nonlinear functions are shown in Figure 8. The noise variables satisfy $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j^2 \sim \text{Unif}([1, 2])$ for source nodes (nodes with empty parental set) and $\sigma_j^2 \sim \text{Unif}([1/4, 1/2])$ otherwise.

In order to estimate the residuals in step 5 of `computeGDPX`, we use additive model fitting based on the `R`-package `mgcv` with default settings [29, 30]. The basis dimension for each smooth term is set to 6.

There exists no state-of-the-art method that we can compare our algorithm with. In principle, given D^0, we can estimate the corresponding PLSEMs for all DAGs in the Markov equivalence class of D^0 and compute their scores. This also gives us an estimate for $\mathscr{D}(\mathbb{P})$, but as explained in Section 3.2, is less efficient than `computeGDPX`. We therefore only evaluate how accurately `computeGDPX` estimates $G_{\mathscr{D}(\mathbb{P})}$. For that, let $G_{\mathscr{D}(\mathbb{P})}$ and $\hat{G}$ denote the true and estimated graphical representations of $\mathscr{D}(\mathbb{P})$, respectively. We count (i) the number of edges that are undirected in $G_{\mathscr{D}(\mathbb{P})}$ but directed in $\hat{G}$ ("falsely kept orientations") and (ii) the

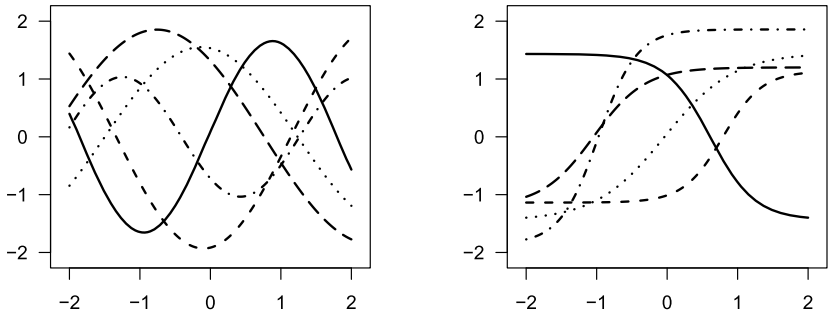

FIG. 8. *Exemplary nonlinear functions used in simulated PLSEMs.*

number of edges that are directed in $G_{\mathscr{D}(\mathbb{P})}$ but undirected in $\hat{G}$ ("falsely removed orientations"). Note that as we assume faithfulness, all DAGs in $\mathscr{D}(\mathbb{P})$ have the same CPDAG. By construction, `computeGDPX` does not falsely remove orientations on the directed part of the CPDAG as all these edges are not covered in any of the consistent DAG extensions. To obtain the percentages shown in Figures 9 to 11, we therefore only divide by the number of undirected edges in the CPDAG. The percentages then reflect a measure for the fraction of "correct score-based decisions."

5.2. *Reference method for true distribution equivalence class* $\mathscr{D}(\mathbb{P})$. To be able to characterize the true distribution equivalence class based on D^0 and the corresponding PLSEM, we assume that for each $i \in \{1, \dots, p\}$, the functions in the set $\{\partial_i^2 f_{j,i} : j \text{ is a child of } i \text{ in } D^0 \text{ and } f_{j,i} \text{ is nonlinear}\}_i$ are linearly independent for the PLSEM with DAG D^0 that generates $\mathbb{P}$. As all functions in our simulations are randomly drawn (cf. Section 5.1), the assumption is satisfied with probability one for D^0 and the corresponding edge functions.

This additional assumption rules out cases where nonlinear effects in D^0 exactly cancel out over different paths, and hence excludes cases as in Figure 4 where nonlinear edges may be reversed. In particular, it allows us to obtain $G_{\mathscr{D}(\mathbb{P})}$ only based on D^0 and knowledge of the functions in the corresponding PLSEM: first, we use Theorem E.1(c) in the supplement to construct the set $\mathcal{V}$. For all nodes i in D^0, corresponding sets of nonlinear children C_i (as defined in Section E in the supplement) and $k \neq i$, we add (i, k) to $\mathcal{V}$ if k is a descendant of a node in C_i. In principle, we now apply Algorithm 3, but instead of the score-based decision in steps 6–9, we use the set $\mathcal{V}$ to decide about edge orientations. Let $i \to j$ be the edge chosen in step 4 and D one of the consistent DAG extensions in which $i \to j$ is covered. If $(i, j) \in \mathcal{V}$, by Theorem E.1(d) and Remark E.1 in the supplement, $i \to j$ in all DAGs of a PLSEM that generates $\mathbb{P}$. Hence, in particular, $i \to j$ in all DAGs in $\mathscr{D}(\mathbb{P})$ and by definition, $i \to j$ in $G_{\mathscr{D}(\mathbb{P})}$. If $(i, j) \notin \mathcal{V}$, by Lemma B.1 in the supplement, the DAG D' that differs from D only by reversing $i \to j$ is in $\mathscr{D}(\mathbb{P})$. Hence, by definition, $i - j$ in $G_{\mathscr{D}(\mathbb{P})}$.

5.3. *The role of α for varying sample size.* In `computeGDPX`, the score-based decision whether a selected covered edge is linear or nonlinear is based on a comparison of the absolute difference of the expected negative log-likelihood scores of two models with a parameter α. Optimally, one would choose α close to ξ_p [see equation (3.2)], but ξ_p depends on the setting (number of variables, sparsity of the DAG, degree of nonlinearity of the nonlinear functions, etc.) and is unknown. In practice, the parameter α reflects a measure of how conservative the estimate $\hat{G}$ of $G_{\mathscr{D}(\mathbb{P})}$ is (in the sense of how many causal statements can be made). For example, choosing α large results in a conservative estimate $\hat{G}$ with many undirected edges [a large set $\mathscr{D}(\mathbb{P})$ of equivalent DAGs]. In Figures 9 and 10, we empirically analyze the dependence of $\hat{G}$ on α for different sample sizes for sparse and dense graphs, respectively.

`computeGDPX` exhibits a good performance for a wide range of values of α. In particular, as the sample size increases, choosing α small results in very accurate

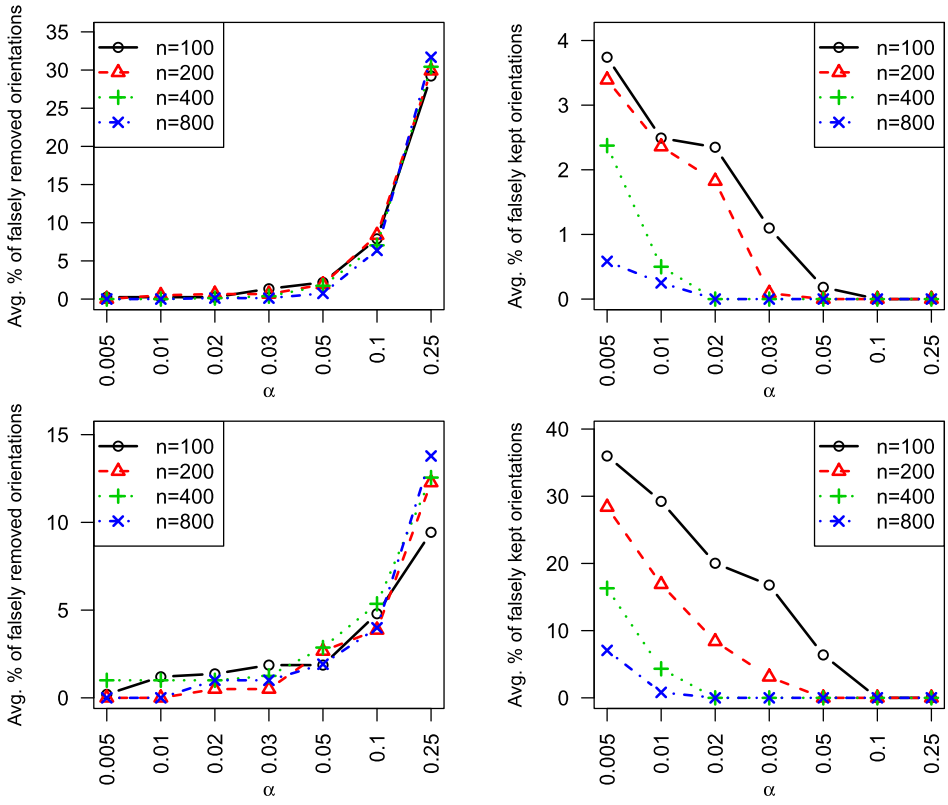

FIG. 9. *Performance of* `computeGDPX` *for varying sample sizes and values of* α (*x-axis*) *in sparse DAGs with* $p_{\text{lin}} = 0.2$ (*top*) *and* $p_{\text{lin}} = 0.8$ (*bottom*). *Parameters*: $p = 10$, $n_{\text{rep}} = 100$ *and* $p_{\text{c}} = 2/9$ (*expected number of edges*: 10).

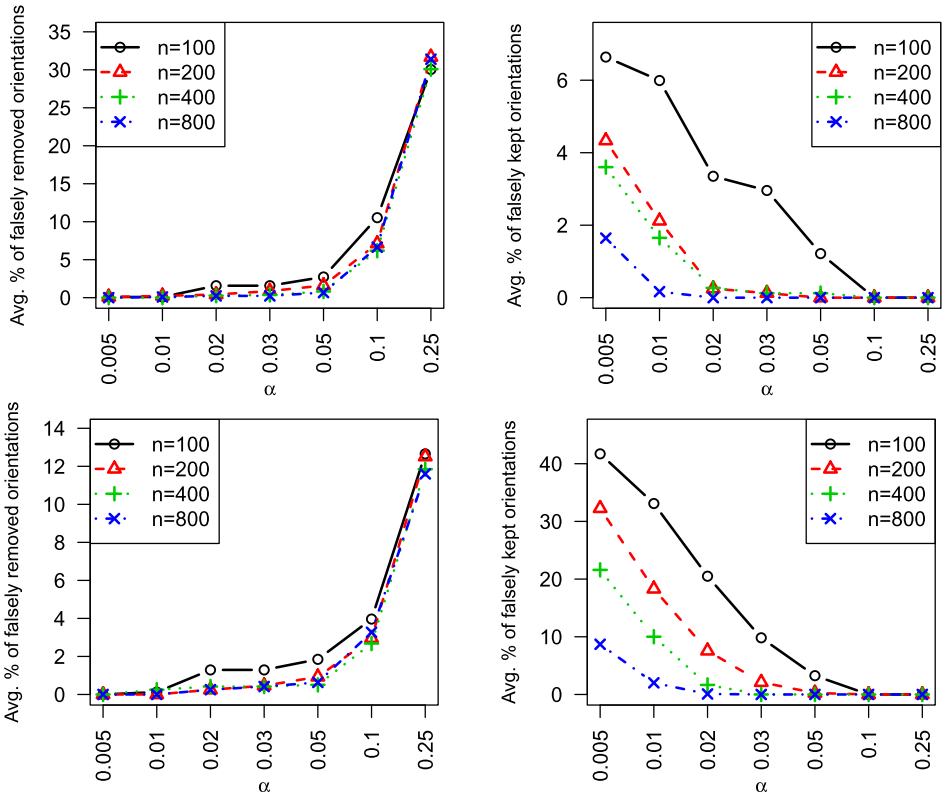

FIG. 10. *Performance of* `computeGDPX` *for varying sample sizes and values of* α *(x-axis) in dense DAGs for* $p_{\text{lin}} = 0.2$ *(top) and* $p_{\text{lin}} = 0.8$ *(bottom). Parameters*: $p = 10$, $n_{\text{rep}} = 100$ *and* $p_{\text{c}} = 6/9$ *(expected number of edges*: 30).

estimates $\hat{G}$ of $G_{\mathscr{D}(\mathbb{P})}$. The sparsity of the DAG does not strongly influence the results.

5.4. *The dependence on* p*: Low- and high-dimensional setting.* From the fact that `computeGDPX` only relies on local score computations, we expect that its performance does not strongly depend on the number of variables p as long as the neighborhood sizes in the DAGs (the node degrees) are similar for different values of p. We simulate $n_{\text{rep}} = 100$ random DAGs with $p = 10$, $p = 100$ and $p = 1000$ nodes, respectively. Moreover, we set $p_{\text{c}} = 2/(p-1)$ which results in an expected number of p edges and an expected node degree of 2 for all settings. As demonstrated in Figure 11, the accuracy of `computeGDPX` with respect to varying values of α is barely affected by the number of variables p. In particular, `computeGDPX` exhibits a good performance even in high-dimensional settings with $p = 1000$ and sample sizes in the hundreds. The same conclusions hold for $p_{\text{c}} = 6/(p-1)$ with an expected node degree of 6 (not shown).

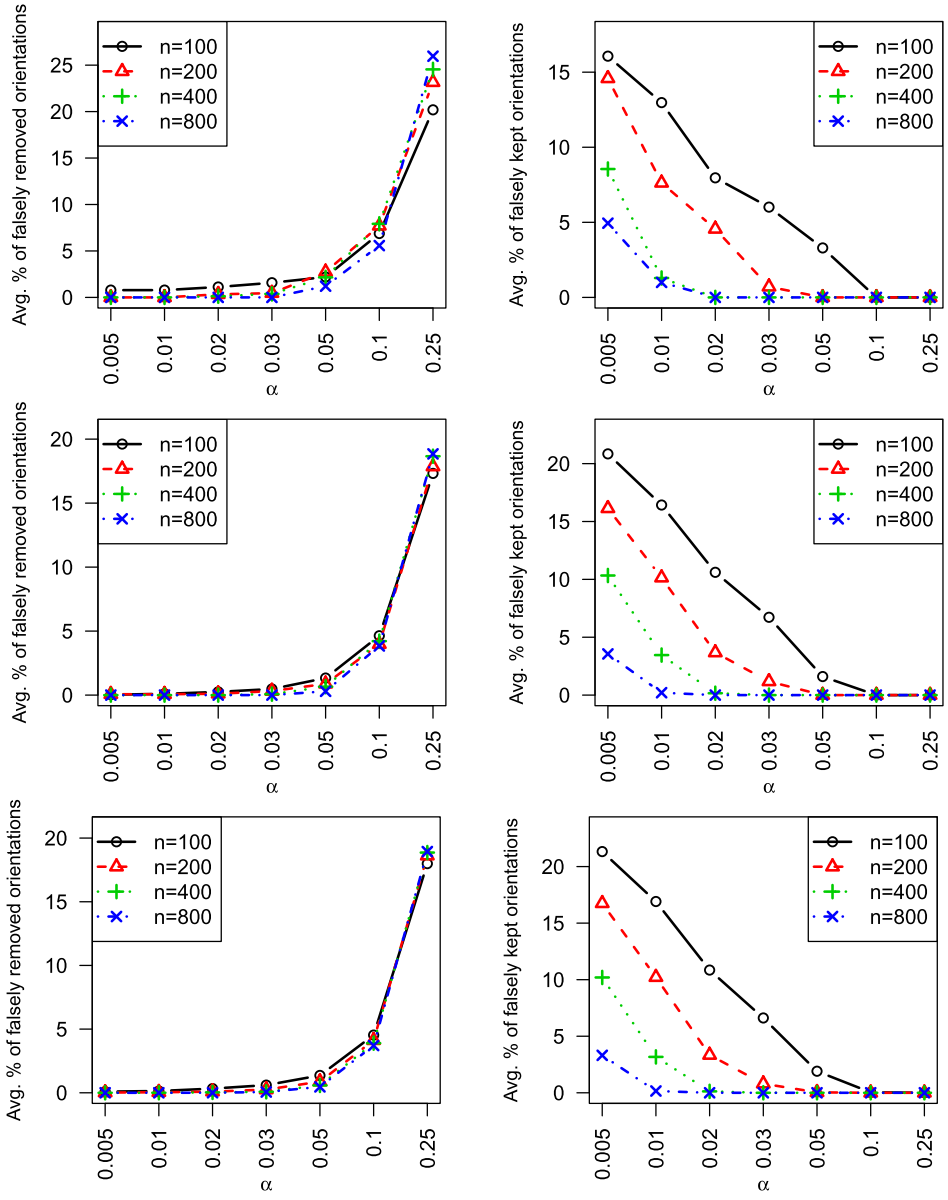

FIG. 11. *Performance of* `computeGDPX` *for varying sample sizes and values of* α (*x-axis*) *for* $p = 10$ (*top*), $p = 100$ (*middle*) *and* $p = 1000$ (*bottom*). *Parameters*: $p_{\text{lin}} = 0.5$, $n_{\text{rep}} = 100$ *and* $p_{\text{c}} = 2/(p-1)$ (*expected number of edges*: p).

5.5. *Computation time*. Finally, we analyze the computation time of `computeGDPX` depending on the number of variables p and sparsity p_{c}. We examine two scenarios: (i) most of the functions in the PLSEM are nonlinear

TABLE 1
Median CPU times [s] for `computeGDPX` *and for* `dag2cpdag` *that iteratively applies R1 to R3 in Figure 6.* $n_{\text{rep}} = 100$ *repetitions for* $p_{\text{lin}} = 0.2$ *and* $n_{\text{rep}} = 20$ *repetitions for* $p_{\text{lin}} = 1$

	$p_{\text{lin}} = 0.2$		$p_{\text{lin}} = 1$			
	`computeGDPX`		`computeGDPX`		`dag2cpdag`	
$\mathbb{E}[\lvert\text{edges}\rvert]$	p	$4p$	p	$4p$	p	$4p$
$p = 10$	0.092	0.785	0.157	1.101	0.007	0.005
$p = 20$	0.150	0.105	0.174	0.162	0.006	0.006
$p = 50$	0.300	0.164	0.332	0.223	0.008	0.009
$p = 100$	0.604	0.281	0.665	0.325	0.014	0.016
$p = 250$	1.446	0.630	1.740	0.717	0.072	0.087
$p = 500$	2.705	1.253	3.486	1.523	0.395	0.599
$p = 1000$	5.616	2.513	6.603	2.974	3.464	4.231
$p = 2000$	11.504	5.380	13.493	6.331	25.463	31.591
$p = 5000$	29.226	16.276	35.094	18.462	400.324	591.574

($p_{\text{lin}} = 0.2$) and (ii) the worst-case scenario (w.r.t. computation time) where all the functions in the PLSEM are linear ($p_{\text{lin}} = 1$) and $\mathscr{D}(\mathbb{P})$ is equal to the Markov equivalence class ($G_{\mathscr{D}(\mathbb{P})}$ equals the CPDAG). For all combinations of $p \in \{10, 20, 50, 100, 250, 500, 1000, 2000, 5000\}$ and $p_c \in \{2/(p-1), 8/(p-1)\}$ and for both scenarios (i) and (ii), we measure the time consumption of `computeGDPX` for $n = 400$ and $\alpha = 0.05$. In the scenario where all the functions are linear, we additionally compare it to `dag2cpdag` in the R-package `pcalg`, which constructs the CPDAG based on iterative application of R1-R3 in Figure 6. The median CPU times are shown in Table 1. `computeGDPX` is able to estimate $G_{\mathscr{D}(\mathbb{P})}$ in less than a minute even if the number of variables is in the thousands. In general, the speed of our implementation heavily depends on the sparsity of the DAGs. This can be seen from the case with $p = 10$ and expected number of edges 40. In this setting, the DAGs are almost fully connected. This in turn implies that not many of the edges are fixed due to v-structures and a lot of score-based tests have to be performed. On the other hand, if the underlying DAGs are sparse, we observe that `computeGDPX` even outperforms `dag2cpdag` with respect to computation time if the number of variables is large. Note that this only holds for sparse DAGs. In general, `dag2cpdag` is much faster than our implementation (not shown).

6. Conclusion. We comprehensively characterized the identifiability of partially linear structural equation models with Gaussian noise (PLSEMs) from various perspectives. First, we proved that under faithfulness we obtain graphical and transformational characterizations of distribution equivalent DAGs similar to well-known characterizations of Markov equivalence classes of DAGs. More generally,

we demonstrated that reinterpreting PLSEMs as PLSEM-functions leads to an interesting geometric characterization of all PLSEMs that generate the same distribution $\mathbb{P}$, as they can all be expressed as constant rotations of each other. Therefrom we derived a precise condition how PLSEM-functions (and hence also how single nonlinear additive components in PLSEMs) restrict the set of potential causal orderings of the variables and showed how it can be leveraged to conclude about the causal relations of specific pairs of variables under mild additional assumptions. We also provided some robustness results when the noise terms are in the neighborhood of Gaussian distributions. The theoretical results were complemented with an efficient algorithm that finds all equivalent DAGs to a given DAG or PLSEM. We proved its high-dimensional consistency and evaluated its performance on simulated data.

From an application perspective, the algorithms `listAllDAGsPLSEM` and `computeGDPX` can serve two purposes. First, they can be used in conjunction with any causal structure learning procedure in the DAG space. This has been proposed in [4] and it can also be used in the context of PLSEMs. In comparison to the Markov equivalence class, the algorithms can potentially identify additional directed edges. In addition, the proposed methods can play an important role for the output of the CAM algorithm [3] (with pruning). In particular, if some of the edge functions are close to linear or the sample size is low, the CAM algorithm will output one DAG even though there might be many DAGs with similar scores. In that scenario, the proposed algorithms provide a simple and important criterion to assess the reliability of oriented edges.

More broadly speaking, our characterizations of PLSEMs (and corresponding DAGs) that generate the same distribution $\mathbb{P}$ are crucial for further algorithmic developments in structure learning. For example, as mentioned before, in the spirit of [4], or also for Monte Carlo sampling in Bayesian settings, see a related discussion in [1], Section 1.

SUPPLEMENTARY MATERIAL

Supplement to "Causal inference in partially linear structural equation models" (DOI: 10.1214/17-AOS1643SUPP; .pdf). This supplemental article contains all proofs.

REFERENCES

[1] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

[2] BÜHLMANN, P. (2013). Causal statistical inference in high dimensions. *Math. Methods Oper. Res.* **77** 357–370. MR3072790

[3] BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.* **42** 2526–2556. MR3277670

[4] CASTELO, R. and KOCKA, T. (2003). On inclusion-driven learning of Bayesian networks. *J. Mach. Learn. Res.* **4** 527–574.

[5] CHICKERING, D. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554.

[6] CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the* 11*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 87–98. Morgan Kaufmann, San Francisco, CA.

[7] GLASS, T. A., GOODMAN, S. N., HERNÁN, M. A. and SAMET, J. M. (2013). Causal inference in public health. *Annu. Rev. Public Health* **34** 61–75.

[8] HOYER, P., HYVARINEN, A., SCHEINES, R., SPIRTES, P., RAMSEY, J., LACERDA, G. and SHIMIZU, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the* 24*th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 282–289. AUAI Press, Corvallis, OR.

[9] HOYER, P. O., JANZING, D., MOOIJ, J. M., PETERS, J. and SCHÖLKOPF, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* 21 (*NIPS*) 689–696. Curran, Red Hook, NY.

[10] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[11] KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47** 1–26.

[12] MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the* 11*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 403–410. Morgan Kaufmann, San Francisco, CA.

[13] NANDY, P., HAUSER, A. and MAATHUIS, M. (2015). High-dimensional consistency in score-based and hybrid structure learning. arXiv:1507.02608.

[14] NOWZOHOUR, C. and BÜHLMANN, P. (2016). Score-based causal learning in additive noise models. *Statistics* **50** 471–485. MR3506653

[15] PEARL, J. (2009). *Causality*: *Models*, *Reasoning and Inference*, 2nd ed. Cambridge Univ. Press, New York.

[16] PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228.

[17] PETERS, J., MOOIJ, J., JANZING, D. and SCHÖLKOPF, B. (2011). Identifiability of causal graphs using functional models. In *Proceedings of the* 27*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 589–598. AUAI Press, Corvallis, OR.

[18] PETERS, J., MOOIJ, J., JANZING, D. and SCHÖLKOPF, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** 2009–2053.

[19] RAMSEY, J., HANSON, S., HANSON, C., HALCHENKO, Y., POLDRACK, R. and GLYMOUR, C. (2010). Six problems for causal inference from fmri. *NeuroImage* **49** 1545–1558.

[20] ROTHENHÄUSLER, D., ERNEST, J. and BÜHLMANN, P. (2018). Supplement to "Causal inference in partially linear structural equation models." DOI:10.1214/17-AOS1643SUPP.

[21] SHIMIZU, S., HOYER, P., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030.

[22] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press.

[23] SPIRTES, P. and ZHANG, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* **3** 1–28.

[24] STATNIKOV, A., HENAFF, M., LYTKIN, N. I. and ALIFERIS, C. F. (2012). New methods for separating causes from effects in genomics data. *BMC Genomics* **13** S22.

[25] STEKHOVEN, D., MORAES, I., SVEINBJÖRNSSON, G., HENNIG, L., MAATHUIS, M. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinformatics* **28** 2819–2823.

[26] VAN DE GEER, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.* **8** 543–574.

[27] VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113

[28] VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 220–227. AUAI Press, Corvallis, OR.

[29] WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 95–114. MR1959095

[30] WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. CRC Press, Boca Raton, FL.

[31] ZHANG, K. and HYVÄRINEN, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the* 25*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 647–655. AUAI Press, Corvallis, OR.

SEMINAR FOR STATISTICS
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: rothenhaeusler@stat.math.ethz.ch
jan.ernest@alumni.ethz.ch
buehlmann@stat.math.ethz.ch
URL: http://stat.ethz.ch