The Annals of Statistics
2014, Vol. 42, No. 6, 2526–2556
DOI: 10.1214/14-AOS1260

CAM: CAUSAL ADDITIVE MODELS, HIGH-DIMENSIONAL ORDER SEARCH AND PENALIZED REGRESSION

BY PETER BÜHLMANN, JONAS PETERS[1] AND JAN ERNEST[2]

ETH Zürich

We develop estimation for potentially high-dimensional additive structural equation models. A key component of our approach is to decouple order search among the variables from feature or edge selection in a directed acyclic graph encoding the causal structure. We show that the former can be done with nonregularized (restricted) maximum likelihood estimation while the latter can be efficiently addressed using sparse regression techniques. Thus, we substantially simplify the problem of structure search and estimation for an important class of causal models. We establish consistency of the (restricted) maximum likelihood estimator for low- and high-dimensional scenarios, and we also allow for misspecification of the error distribution. Furthermore, we develop an efficient computational algorithm which can deal with many variables, and the new method's accuracy and performance is illustrated on simulated and real data.

1. Introduction. Inferring causal relations and effects is an ambitious but important task in virtually all areas of science. In absence of prior information about underlying structure, the problem is plagued, among other things, by identifiability issues [23, 34], cf., and the sheer size of the space of possible models, growing super-exponentially in the number of variables, leading to major challenges with respect to computation and statistical accuracy. Our approach is generic, taking advantage of the tools in sparse regression techniques [4, 8], cf., which have been successively established in recent years.

More precisely, we consider p random variables $X_1, \ldots, X_p$ whose distribution is Markov with respect to an underlying causal directed acyclic graph (causal DAG). We assume that all variables are observed, that is, there are no hidden variables, and that the causal influence diagram does not allow for directed cycles. Generalizations to include hidden variables, for example, unobserved confounders, or directed cycles are briefly discussed in Section 7.1. To formalize a model, one can use the concepts of graphical modeling [12], cf., or structural equation models [23], cf. The approaches are equivalent in the nonparametric or multivariate

Received October 2013; revised July 2014.

[1]Supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA Grant agreement no. 326496.

[2]Supported in part by the Swiss National Science Foundation Grant no. 20PA20E-134493.

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

Key words and phrases. Graphical modeling, intervention calculus, nonparametric regression, regularized estimation, sparsity, structural equation model.

Gaussian case, but this is not true anymore when placing additional restrictions which can be very useful [25, 26, 32]. We use here the framework of structural equation models.

1.1. *Problem and main idea.* Our goal is estimation and structure learning for structural equation models, or of the corresponding Markov equivalence class of an underlying DAG. In particular, we focus on causal additive models, that is, the structural equations are additive in the variables and error terms. The model has the nice property that the underlying structure and the corresponding parameters are identifiable from the observational distribution. Furthermore, we can view it as an extension of linear Gaussian structural equation models by allowing for nonlinear additive functions.

In general, the problem of structure learning (and estimation of corresponding parameters) can be addressed by a variety of algorithms and methods: in the frequentist setting, the most widely used procedures for structure learning (and corresponding parameters) are greedy equivalence search for computing the BIC-regularized maximum likelihood estimator [6] or the PC-algorithm using multiple conditional independence testing [34]. However, for the latter, the constraint of additive structural equations cannot be (easily) respected, and regarding the former, maximum likelihood estimation among all (e.g., linear Gaussian) DAG models is computationally challenging and statistical guarantees for high-dimensional cases (and for uniform convergence with respect to a class of distributions) are only available under rather strong assumptions [38].

Our proposed approach for estimation and selection of additive structural equation models is based on the following simple idea which is briefly mentioned and discussed in [35] and [31]. If the order among the variables would be known, the problem boils down to variable selection in multivariate (potentially nonlinear) regression; see formula (5). The latter is very well understood: for example, we can follow the route of hypothesis testing in additive models, or sparse regression can be used for additive models [16, 28, 44]. Thus, the only remaining task is to estimate the order among the variables. We show here that this can be done via the maximum likelihood principle, and we establish its consistency. In particular, for low or "mid"-dimensional problems, there is no need to consider a penalized likelihood approach. The same holds true for high-dimensional settings when using a preliminary neighborhood selection and then employing a corresponding restricted maximum likelihood estimator. Therefore, we can entirely decouple the issue of order estimation without regularization and variable selection in sparse regression with appropriate regularization. This makes our approach very generic, at least within the framework where the underlying DAG and a corresponding order of the variables are identifiable from the joint distribution. Empirical results in Section 6 support that we can do much more accurate estimation than for nonidentifiable models such as the popular linear Gaussian structural equation model. On the superficial level, our approach can be summarized as follows:

1. Mainly for high-dimensional settings: preliminary neighborhood selection for estimating a superset of the skeleton of the underlying DAG. This is done by additive regression of one variable against all others. See Section 3.1.

2. Order search for the variables (or best permutation for the indices of the variables) using (restricted) maximum likelihood estimation based on an additive structural equation model with Gaussian errors: the restricted version is employed if the preliminary neighborhood selection in step 1 is used, and the order search is then restricted to the structure of the superset of the skeleton. See Sections 2.4 and 3.2.

3. Based on the estimated order of the variables in step 2, sparse additive regression is used for estimating the functions in an additive structural equation model. See Section 2.5.

1.2. *Related work.* We consider (nonlinear) additive structural equation models. As natural extensions of linear structural equation models, they are attractive for many applications; see Imoto, Goto and Miyano [10]. Identifiability results for this model class have been recently derived [21, 26]. The approach in [21] is based on conditional independence testing and is limited to small dimensions with a few variables only. Instead of multiple testing of conditional independencies, we propose and develop maximum likelihood estimation in a semiparametric additive structural equation model with Gaussian noise variables: fitting such a model is often appropriate in situations where the sample size is not too large, and we present here for the first time the practical feasibility of fitting additive models in the presence of many variables. An extension of our additive structural equation model with Gaussian errors to the case with a nonparametric specification of the error distribution is presented in [22], but the corresponding maximum likelihood estimator is analyzed (and feasible) for problems with a small number of variables only. When the order of the variables is known, which is a much simpler and different problem than what we consider here, [40] provide consistency results for additive structural equation models.

A key aspect of our method is that we decouple regularization for feature selection and order estimation with nonregularized (restricted) maximum likelihood. The former is a well understood subject thanks to the broad literature in sparse regression and related techniques [17, 36, 41, 44, 47, 48], cf. Regarding the latter issue about order selection, a recent analysis in [37] extends our low-dimensional consistency result for the (nonrestricted) maximum likelihood estimator to the scenario where the number of variables can grow with sample size, in the best case essentially as fast as $p = p(n) = o(n)$. The treatment of the high-dimensional case with a restricted maximum likelihood approach is new here, and we also present the first algorithm and empirical results for fitting low- and high-dimensional causal additive models (CAMs).

All proofs are provided in the supplemental article [3].

2. Additive structural equation models. Consider the general structural equation model (SEM):

$$X_j = f_j(X_{\mathrm{pa}_D(j)}, \varepsilon_j), \qquad \varepsilon_1, \ldots, \varepsilon_p \text{ (mutually) independent},$$

where $\mathrm{pa}_D(j)$ denotes the set of parents for node j in DAG D and f_j is a function from $\mathbb{R}^{|\mathrm{pa}_D(j)|+1} \to \mathbb{R}$. Thus, a SEM is specified by an underlying (causal) structure in terms of a DAG D, the functions $f_j(\cdot)$ $(j = 1, \ldots, p)$ and the distributions of ε_j $(j = 1, \ldots, p)$. Most parts of this paper can be interpreted in absence of causal inference issues: clearly though, the main motivations are understanding models and developing novel procedures allowing for causal or interventional statements, and if we do so, we always assume that the structural equations remain unchanged under interventions at one or several variables [23], cf. The model above is often too general, due to problems of identifiability and the difficulty of estimation (curse of dimensionality) of functions in several variables.

Our main focus is on a special (and more practical) case of the model above, namely the additive SEM with potentially misspecified Gaussian errors:

$$\begin{aligned} X_j &= \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \varepsilon_j, \\ &\varepsilon_1, \ldots, \varepsilon_p \text{ independent with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \sigma_j^2 > 0 \ (j = 1, \ldots, p), \\ &\mathbb{E}\big[f_{j,k}(X_k)\big] = 0 \text{ for all } j, k, \end{aligned} \tag{1}$$

where $f_{j,k}(\cdot)$ are smooth functions from $\mathbb{R} \to \mathbb{R}$. A special case thereof is the linear Gaussian SEM

$$\begin{aligned} X_j &= \sum_{k \in \mathrm{pa}_D(j)} \beta_{j,k} X_k + \varepsilon_j, \\ &\varepsilon_1, \ldots, \varepsilon_p \text{ independent with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \sigma_j^2 > 0 \ (j = 1, \ldots, p). \end{aligned} \tag{2}$$

Although model (2) is a special case of (1), there are interesting differences with respect to identifiability. If all functions $f_{j,k}(\cdot)$ are nonlinear, the DAG is identifiable from the distribution P of $X_1, \ldots, X_p$ [26], Corollary 31. We explicitly state this result as a lemma since we will make use of it later on.

LEMMA 1 (Corollary 31 in [26][3]). *Consider a distribution P that is generated by model (1) with DAG D and nonlinear, three times differentiable functions $f_{j,k}$. Then any distribution Q that is generated by (1) with a different DAG $D' \neq D$ and nonconstant, three times differentiable functions $f'_{j,k}$ is different from P: we have $Q \neq P$.*

[3]Corollary 31 in [26] contains a slightly different statement using "nonlinear" instead of "nonconstant". The proof, however, stays exactly the same.

This result does not hold, however, for a general SEM or for a linear Gaussian SEM as in (2); one can then only identify the Markov equivalence class of the DAG D^0, assuming faithfulness. An exception arises when assuming same error variances $\sigma_j^2 \equiv \sigma^2$ for all j in (2) which again implies identifiability of the DAG D^0 from P [25]. In the sequel, we consider the fully identifiable case of model (1).

2.1. *The likelihood.* We slightly re-write model (1) as

$$
\begin{aligned}
&X_j = \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \varepsilon_j = \sum_{k \neq j} f_{j,k}(X_k) + \varepsilon_j \qquad (j = 1, \ldots, p),\\
&f_{j,k}(\cdot) \not\equiv 0 \text{ if and only if there is a directed edge } k \to j \text{ in } D,\\
&\mathbb{E}\big[f_{j,k}(X_k)\big] = 0 \text{ for all } j, k,\\
&\varepsilon_1, \ldots, \varepsilon_p \text{ independent and } \varepsilon_j \sim \mathcal{N}\big(0, \sigma_j^2\big), \sigma_j^2 > 0.
\end{aligned} \tag{3}
$$

Note that the structure of the model, or the so-called active set, $\{(j,k);\, f_{j,k} \not\equiv 0\}$ is identifiable from the distribution P [26], Corollary 31. Denote by θ the infinite-dimensional parameter with additive functions and error variances, that is,

$$\theta = (f_{1,2}, \ldots, f_{1,p}, f_{2,1}, \ldots, f_{p,p-1}, \sigma_1, \ldots, \sigma_p).$$

Furthermore, we denote by D^0 the true DAG and by θ^0 (and $\{f_{j,k}^0\}$, $\{\sigma_j^0\}$) the true infinite-dimensional parameter(s) corresponding to the data-generating true distribution. We use this notation whenever it is appropriate to make statements about the true underlying DAG or parameter(s).

The density $p_\theta(\cdot)$ for the model (3) is of the form

$$\log\big(p_\theta(x)\big) = \sum_{j=1}^{p} \log\Big(\frac{1}{\sigma_j}\varphi\Big(\frac{x_j - \sum_{k \neq j} f_{j,k}(x_k)}{\sigma_j}\Big)\Big),$$

where $\varphi(\cdot)$ is the density of a standard normal distribution. Furthermore,

$$\sigma_j^2 = \mathbb{E}\Big[\Big(X_j - \sum_{k \neq j} f_{j,k}(X_k)\Big)^2\Big],$$

and the expected negative log-likelihood is

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

2.2. *The function class.* We assume that the functions in model (1) or (3) are from a class of smooth functions: $\mathcal{F}$ is a subset of $L_2(P_j)$, where P_j is the marginal distribution for any $j = 1, \ldots, p$; assume that it is closed with respect to the $L_2(P_j)$ norm. Furthermore,

$$\mathcal{F} \subseteq \big\{f : \mathbb{R} \to \mathbb{R},\, f \in C^\alpha, \mathbb{E}\big[f(X)\big] = 0\big\},$$

where C^α denotes the space of α-times differentiable functions and the random variable X is a placeholder for the variables X_j, $j = 1, \ldots, p$. Note that this is a slight abuse of notation since $\mathcal{F}$ does not specify the variable X; it becomes clear from the context.

Consider also basis functions $\{b_r(\cdot); r = 1, \ldots, a_n\}$ with $a_n \to \infty$ sufficiently slowly, for example, B-splines or regression splines. Consider further the space

$$\mathcal{F}_n = \left\{ f \in \mathcal{F},\ f = c + \sum_{r=1}^{a_n} \alpha_r b_r(\cdot) \text{ with } c, \alpha_r \in \mathbb{R} (r = 1, \ldots, a_n) \right\}. \tag{4}$$

We allow for constants c to enforce mean zero for the whole function. Furthermore, the basis functions can be the same for all variables X_j, $j = 1, \ldots, p$.

For theoretical analysis, we assume that $\mathcal{F}_n$ is deterministic and does not depend on the data. Then, $\mathcal{F}_n$ is closed. Furthermore, the space of additive functions is denoted by

$$\mathcal{F}^{\oplus \ell} = \left\{ f : \mathbb{R}^\ell \to \mathbb{R};\ f(x) = \sum_{k=1}^{\ell} f_k(x_k),\ f_k \in \mathcal{F} \right\},$$

$$\mathcal{F}_n^{\oplus \ell} = \left\{ f : \mathbb{R}^\ell \to \mathbb{R};\ f(x) = \sum_{k=1}^{\ell} f_k(x_k),\ f_k \in \mathcal{F}_n \right\},$$

where $\ell = 2, \ldots, p$. Clearly, $\mathcal{F}_n^{\oplus \ell} \subseteq \mathcal{F}^{\oplus \ell}$. For $f \in \mathcal{F}^{\oplus \ell}$ we denote by f_k its kth additive function.

In our definitions, we assume that the functions in $\mathcal{F}$ and $\mathcal{F}_n$ have expectation zero. Of course, this depends on the variables in the arguments of the functions. For example, when requiring $\mathbb{E}[f(X_j)] = 0$ for $f \in \mathcal{F}$, the function class $\mathcal{F} = \mathcal{F}_j$ depends on the index j due to the mean zero requirement; and likewise $\mathcal{F}^{\oplus \ell}$ depends on the indices of the variables occurring in the corresponding additive function terms. We drop this additional dependence on the index of variables as it does not cause any problems in methodology or theory.

Later, we consider projections of distributions onto the spaces $\mathcal{F}^{\oplus \ell}$ and $\mathcal{F}_n^{\oplus \ell}$, see (6). We assume throughout the paper that these spaces are closed with respect to the L_2 norm. The following Lemma 2 guarantees this condition by requiring an analogue of a minimal eigenvalue assumption.

LEMMA 2. *Let the distribution P be generated according to* (1) *and assume that there is a $\phi^2 > 0$ such that for all $\gamma \in \mathbb{R}^p$*

$$\left\| \sum_{j=1}^{p} \gamma_j f_j(X_j) \right\|_{L_2}^2 \geq \phi^2 \|\gamma\|^2 \qquad \text{for all } f_j \in \mathcal{F} \text{ with } \|f_j(X_j)\|_{L_2} = 1.$$

For any subset $I \subseteq \{1, \ldots, p\}$ of ℓ variables the spaces $\mathcal{F}^{\oplus \ell}$ and $\mathcal{F}_n^{\oplus \ell}$ are then closed with respect to the $L_2(P_I)$ norm. Here, P_I denotes the marginal distribution over all variables in I.

The question of closedness of additive models has also been studied in [1], for example; see also [29].

2.3. *Order of variables and the likelihood.* We can permute the variables, inducing a different ordering; in the sequel, we use both terminologies, permutations and order search, which mean the same thing. For a permutation π on $\{1, \ldots, p\}$, define

$$X^{\pi}, \qquad X_j^{\pi} = X_{\pi(j)}.$$

There is a canonical correspondence between permutations and fully connected DAGs: for any permutation π, we can construct a DAG D^{π}, in which each variable $\pi(k)$ has a directed arrow to all $\pi(j)$ with $j > k$. The node $\pi(1)$ has no parents and is called the source node. For a given DAG D^0, we define the set of true permutations as

$$\Pi^0 = \{\pi^0;\ \text{the fully connected DAG } D^{\pi^0} \text{ is a super-DAG of } D^0\},$$

where a super-DAG of D^0 is a DAG whose set of directed edges is a superset of the one corresponding to D^0. If the true DAG D^0 is not fully connected, there is typically more than one true order or permutation, that is the true order is typically not unique. It is apparent that any true ordering or permutation π^0 allows for a lower-triangular (or autoregressive) representation of the model in (3):

$$X_j^{\pi^0} = \sum_{k=1}^{j-1} f_{j,k}^{\pi^0}(X_k^{\pi^0}) + \varepsilon_j^{\pi^0} \qquad (j = 1, \ldots, p), \tag{5}$$

where $f_{j,k}^{\pi^0}(\cdot) = f_{\pi^0(j),\pi^0(k)}^0(\cdot)$ and $\varepsilon_j^{\pi^0} = \varepsilon_{\pi^0(j)}^0$, that is, with permuted indices in terms of the original quantities in (3). If all functions $f_{j,k}(\cdot)$ are nonlinear, the set of true permutations is identifiable from the distribution [26], Corollary 33, and Π^0 consists of all orderings of the variables which allow for a lower-triangular representation (5). We will exploit this fact in order to provide a consistent estimator $\hat{\pi}_n$ of the ordering: under suitable assumptions the probability that $\hat{\pi}_n \in \Pi^0$ converges to one.

REMARK 1. For the linear Gaussian SEM (2), all orderings allow for a lower-triangular representation (5), even those that are not in Π^0. Thus, we cannot construct a consistent estimator in the above sense. However, assuming faithfulness of the true distribution, the orderings of variables which are consistent with the arrow directions in a DAG of the Markov equivalence class of the true DAG D^0 lead to sparsest representations with fewest number of nonzero coefficients.

In principle, one can check whether the data come from a linear Gaussian SEM. Lemma 1 guarantees that if this is case, there is no CAM with nonlinear functions yielding the same distribution. Thus, if the structural equations of the estimated

DAG look linear with Gaussian noise, one could decide to output the Markov equivalence class instead of the DAG. One would need to quantify closeness to linearity and Gaussianity with, for example, a test: this would be important for practical applications, but its precise implementation lies beyond the scope of this work.

In the sequel, it is helpful to consider the true underlying parameter θ^0 with corresponding nonlinear function $f^0_{j,k}$ and error variances $(\sigma^0_j)^2$. For any permutation $\pi \notin \Pi^0$, we consider the projected parameters, defined as

$$\theta^{\pi,0} = \operatorname*{argmin}_{\theta^\pi} \mathbb{E}_{\theta^0}\big[-\log\big(p^\pi_{\theta^\pi}(X)\big)\big],$$

where the density $p^\pi_{\theta^\pi}$ is of the form

$$\log\big(p^\pi_{\theta^\pi}(x)\big) = \log\big(p_{\theta^\pi}(x^\pi)\big) = \sum_{j=1}^{p} \log\Bigg(\frac{1}{\sigma^\pi_j}\varphi\Bigg(\frac{x^\pi_j - \sum_{k=1}^{j-1} f^\pi_{j,k}(x^\pi_k)}{\sigma^\pi_j}\Bigg)\Bigg).$$

(Note that if $\pi \in \Pi^0$, then $\theta^{\pi,0} = \theta^0$.) For such a misspecified model with wrong order $\pi \notin \Pi^0$, we have

$$\begin{aligned}\{f^{\pi,0}_{j,k}\}_{k=1,\dots,j-1} &= \operatorname*{argmin}_{g_{j,k}\in\mathcal{F},k=1,\dots,j-1} \mathbb{E}_{\theta^0}\Bigg[\Bigg(X^\pi_j - \sum_{k=1}^{j-1} g_{j,k}(X^\pi_k)\Bigg)^2\Bigg] \\ &= \operatorname*{argmin}_{g_j\in\mathcal{F}^{\oplus j-1}} \mathbb{E}_{\theta^0}\big[\big(X^\pi_j - g_j(X^\pi_1,\dots,X^\pi_{j-1})\big)^2\big].\end{aligned} \tag{6}$$

It holds that

$$\begin{aligned}\big(\sigma^{\pi,0}_j\big)^2 &= \operatorname*{argmin}_{\sigma^2}\Bigg(\log(\sigma) + \frac{1}{2\sigma^2}\mathbb{E}_{\theta^0}\Bigg[\Bigg(X^\pi_j - \sum_{k=1}^{j-1} f^{\pi,0}_{j,k}(X^\pi_k)\Bigg)^2\Bigg]\Bigg) \\ &= \mathbb{E}_{\theta^0}\Bigg[\Bigg(X^\pi_j - \sum_{k=1}^{j-1} f^{\pi,0}_{j,k}(X^\pi_k)\Bigg)^2\Bigg].\end{aligned} \tag{7}$$

The two displayed formulae above show that autoregression with the wrong order π leads to the projected parameters $\{f^{\pi,0}_{j,k}\}$ and $\{(\sigma^{\pi,0}_j)^2\}$. Finally, we obtain

$$\mathbb{E}_{\theta^0}\big[-\log\big(p^\pi_{\theta^{\pi,0}}(X)\big)\big] = \sum_{j=1}^{p} \log\big(\sigma^{\pi,0}_j\big) + C, \qquad C = p\log(2\pi)^{1/2} + p/2.$$

All true permutations $\pi \in \Pi^0$ correspond to super DAGs of the true DAG and, therefore, all of them lead to the minimal expected log-likelihood $\mathbb{E}_{\theta^0}[-\log(p^\pi_{\theta^{\pi,0}}(X))] = \mathbb{E}_{\theta^0}[-\log(p_{\theta^0}(X))]$. The permutations $\pi \notin \Pi^0$, however, cannot lead to a smaller expected negative log-likelihood (since it would lead to a

negative KL-divergence between the true and best projected distribution). Let us therefore define

$$\xi_p := \min_{\pi \notin \Pi^0} p^{-1}\big(\mathbb{E}_{\theta^0}[-\log(p^{\pi}_{\theta^{\pi,0}}(X))] - \mathbb{E}_{\theta^0}[-\log(p_{\theta^0}(X))]\big) \geq 0. \tag{8}$$

If all true functions $f^0_{j,k}$ are nonlinear, we obtain $\xi_p > 0$ as follows.

LEMMA 3. *Consider a distribution P that allows for a density p with respect to the Lebesgue measure and is generated by model (1) with DAG D^0 and nonlinear, three times differentiable functions $f^0_{j,k}$. Assume further the condition from Lemma 2. Then $\xi_p > 0$.*

PROOF. Because of the closedness of $\mathcal{F}^{\oplus j}$ (Lemma 2), the minimum in (6) is obtained for some functions $f_{j,k}$. Without loss of generality, we can assume that all constant additive components are zero. But then $\xi_p = 0$ would contradict Lemma 1. □

The number ξ_p describes the degree of separation between the true model and misspecification when using a wrong permutation. As discussed in Remark 1, $\xi_p = 0$ for the case of linear Gaussian SEMs. Formula (8) can be expressed as

$$\xi_p = \min_{\pi \notin \Pi^0} p^{-1} \sum_{j=1}^{p} \big(\log(\sigma_j^{\pi,0}) - \log(\sigma_j^0)\big) \geq 0. \tag{9}$$

REMARK 2. Especially for situations where p is very large so that the factor p^{-1} is small, requiring a lower bound $\xi_p > 0$ can be overly restrictive. Instead of requiring a gap with the factor p^{-1} between the likelihood scores of the true distribution and all distributions corresponding to permutations, one can weaken this as follows. Consider $H(D, D^0) = \{j; \mathrm{pa}_{D^0}(j) \nsubseteq \mathrm{pa}_D(j)\}$. We require that

$$\xi'_p := \min_{D \neq D^0} \big|H(D, D^0)\big|^{-1} \sum_{j \in H(D, D^0)} \big(\log(\sigma_j^{D,0}) - \log(\sigma_j^0)\big) \geq 0, \tag{10}$$

where $(\sigma_j^{D,0})^2$ is the error variance in the best additive approximation of X_j based on $\{X_k; k \in \mathrm{pa}_D(j)\}$. Such a weaker gap condition is proposed in [13], Section 5.2. All our theoretical results still hold when replacing statements involving ξ_p in (9) by the corresponding statements with ξ'_p in (10).

2.4. *Maximum likelihood estimation for order: Low-dimensional setting.* We assume having n i.i.d. realizations $X^{(1)}, \ldots, X^{(n)}$ from model (3). For a $n \times 1$ vector $x = (x^{(1)}, \ldots, x^{(n)})^T$, we denote by $\|x\|^2_{(n)} = n^{-1} \sum_{i=1}^n (x^{(i)})^2$. Depending on the context, we sometimes denote by $\hat{f}$ a function and sometimes an $n \times 1$ vector evaluated at (the components of) the data points $X^{(1)}, \ldots, X^{(n)}$; and similarly

for X_j^π. We consider the unpenalized maximum likelihood estimator:

$$\hat{f}_j^\pi = \operatorname*{argmin}_{g_j \in \mathcal{F}_n^{\oplus j-1}} \left\| X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi) \right\|_{(n)}^2, \qquad (\hat{\sigma}_j^\pi)^2 = \left\| X_j^\pi - \sum_{k=1}^{j-1} \hat{f}_{j,k}^\pi(X_k^\pi) \right\|_{(n)}^2.$$

Denote by $\hat{\pi}$ a permutation which minimizes the unpenalized negative log-likelihood:

$$\hat{\pi} \in \operatorname*{argmin}_\pi \sum_{j=1}^p \log(\hat{\sigma}_j^\pi). \tag{11}$$

Estimation of $\hat{f}_j^\pi$ is based on $\mathcal{F}_n$ with pre-specified basis functions $b_r(\cdot)$ with $r = 1, \ldots, a_n$. In practice, the basis functions could depend on the predictor variable or on the order of variables, for example, when choosing the knots in regression splines. The classical choice for the number of basis functions is $a_n \asymp n^{1/5}$ for twice differentiable functions: here, and as explained in Section 4, however, a smaller number such as $a_n = O(1)$ to detect some nonlinearity might be sufficient for estimation of the true underlying order.

2.5. *Sparse regression for feature selection.* Section 4 presents assumptions and results ensuring that with high probability $\hat{\pi} = \pi^0$ for some $\pi^0 \in \Pi^0$. With such an estimated order $\hat{\pi}$, we obtain a complete super-DAG (super-graph) $D^{\hat{\pi}}$ of the underlying DAG D^0 in (3), where the parents of a node $\hat{\pi}(j)$ are defined as $\mathrm{pa}_{D^{\hat{\pi}}}(\hat{\pi}(j)) = \{\hat{\pi}(k); k < j\}$ for all j. We can pursue consistent estimation of intervention distributions based on $D^{\hat{\pi}}$ without any additional need to find the true underlying DAG D^0; see Section 2.6.

However, we can improve statistical efficiency for estimating the intervention distribution when it is ideally based on the true DAG D^0 or realistically a not too large super-DAG $\hat{D}^{\hat{\pi}} \supseteq D^0$. The task of estimating such a super-DAG $\hat{D}^{\hat{\pi}} \supseteq D^0$ is conceptually straightforward: starting from the complete super-DAG $D^{\hat{\pi}}$ of D^0 as discussed above, we can use model selection or a penalized multivariate (auto-) regression technique in the model representation (5). For additive model fitting, we can either use hypothesis testing for additive models [15] or the Group Lasso [28], or its improved version with a sparsity-smoothness penalty proposed in [16]. All the techniques mentioned above perform variable selection, where we denote by

$$\hat{D}^{\hat{\pi}} = \{(\hat{\pi}(k), \hat{\pi}(j)); \hat{f}_{j,k}^{\hat{\pi}} \not\equiv c\},$$

(the constant $c = 0$ when assuming that $\hat{f}_{j,k}^{\hat{\pi}}$ have mean zero when evaluated over all data-points) the selected variables indexed in the original order [we obtain estimates $\hat{f}_{j,k}^{\hat{\pi}}$ in the representation (5) with correspondence to the indices $\hat{\pi}(k), \hat{\pi}(j)$ in the original order]; we identify these selected variables in $\hat{D}^{\hat{\pi}}$ as the edge set of a DAG. For example, with the Group Lasso, assuming some condition avoiding near collinearity of functions, that is, a compatibility condition for the Group

Lasso [4], Chapter 5.6, Theorem 8.2, and that the ℓ_2-norms of the nonzero functions are sufficiently large, we obtain the screening property (since we implicitly assume that $\hat{\pi} \in \Pi^0$ with high probability): with high probability and asymptotically tending to one,

$$\hat{D}^{\hat{\pi}} \supseteq D^0 = \{(k, j); f^0_{j,k} \not\equiv 0\} \tag{12}$$

saying that all relevant variables (i.e., edges) are selected. Similarly with hypotheses testing, assuming that the nonzero $f^0_{j,k}$ have sufficiently large ℓ_2-norms, we also obtain that (12) holds with high probability.

The same argumentation applies if we use $D^{\hat{\pi}}_{\text{restr}}$ from Section 3.2 instead of $D^{\hat{\pi}}$ as an initial estimate. This then results in $\hat{D}^{\hat{\pi}}_{\text{restr}}$, replacing $\hat{D}^{\hat{\pi}}$ above.

2.6. *Consistent estimation of causal effects.* The property in (12) has an important implication for causal inference:[4] all estimated causal effects and estimated intervention distributions based on the estimated DAG $\hat{D}^{\hat{\pi}}$ are consistent. In fact, using the do-calculus [23], cf. (3.10), we have for the single intervention (at variable X_k) distribution for X_j, for all $j \neq k$:

$$p_{D^0}(x_j|(X_k = x)) = p_{\hat{D}^{\hat{\pi}}}(x_j|(X_k = x)) \qquad \text{for all } x,$$

where $p_D(\cdot|(\cdot))$ denotes the intervention density based on a DAG D.

We note that the screening property (12) also holds when replacing $\hat{D}^{\hat{\pi}}$ with the full DAG induced by $\hat{\pi}$, denoted by $D^{\hat{\pi}}$. Thus, the feature selection step in Section 2.5 is not needed to achieve consistent estimation of causal effects. However, a smaller DAG $D^0 \subseteq \hat{D}^{\hat{\pi}} \subseteq D^{\hat{\pi}}$ typically leads to better (more statistically efficient) estimates of the interventional distributions than the full DAG $D^{\hat{\pi}}$.

3. Restricted maximum likelihood estimation: Computational and statistical benefits. We present here maximum likelihood estimation where we restrict the permutations, instead of searching over all permutations in (11). Such a restriction makes the computation more tractable, and it is also statistically crucial when dealing with high-dimensional settings where $p > n$.

3.1. *Preliminary neighborhood selection.* We first perform neighborhood selection with additive models, following the general idea in [17] for the linear Gaussian case. We pursue variable selection in an additive model of X_j versus all other variables $X_{\{-j\}} = \{X_k; k \neq j\}$: a natural method for such a feature selection is the Group Lasso for additive models [28], ideally with a sparsity-smoothness penalty [16]; see also [40]. This provides us with a set of variables

$$\hat{A}_j \subseteq \{1, \ldots, p\} \setminus j$$

[4]We assume that interventions at variables do not change the other structural equations, and that there are no unobserved hidden (e.g., confounder) variables.

which denotes the selected variables in the estimated conditional expectation

$$\hat{\mathbb{E}}_{\text{add}}[X_j|X_{\{-j\}}] = \sum_{k\in\hat{A}_j} \hat{h}_{jk}(X_k)$$

with functions $\hat{h}_{jk}$ satisfying $n^{-1}\sum_{i=1}^n \hat{h}_{jk}(X_k^{(i)}) = 0$ (i.e., a possible intercept is subtracted already): that is,

$$\hat{A}_j = \{k; k \neq j, \hat{h}_{j,k} \not\equiv 0\}.$$

We emphasize that the functions $\hat{h}_{j,k}(\cdot)$ are different from $\hat{f}_{j,k}^{\pi}(\cdot)$ in Section 2.4 because for the former, the additive regression is against all other variables.

We give conditions in Section 4.2 (see Lemma 4) ensuring that the neighborhood selection set contains the parental variables from the structural equation model in (1) or (3), that is, $\hat{A}_j \supseteq \text{pa}(j)$.

3.2. *Restricted maximum likelihood estimator.* We restrict the space of permutations in the definition of (11) such that they are "compatible" with the neighborhood selection sets $\hat{A}_j$. Note that for the estimator $\hat{\sigma}_j^{\pi}$ in (11), we regress $X_{\pi(j)}$ against $\{X_k; k \in \{\pi(j-1), \ldots, \pi(1)\}\}$. We restrict here the set of regressors to the indices $R_{\pi,j} = \{\pi(j-1), \ldots, \pi(1)\} \cap \hat{A}_{\pi(j)}$. We then calculate the $\pi(j)$th term of the log-likelihood using the set of regressors $X_{R_{\pi,j}} = \{X_k; k \in R_{\pi,j}\}$. More precisely, we estimate

$$\hat{f}_j^{\pi,R} = \operatorname*{argmin}_{g_{j,k}\in\mathcal{F}_n} \Big\| X_j^{\pi} - \sum_{k;\pi(k)\in R_{\pi,j}} g_{j,k}(X_k^{\pi}) \Big\|_{(n)}^2,$$

$$(\hat{\sigma}_j^{\pi,R})^2 = \Big\| X_j^{\pi} - \sum_{k;\pi(k)\in R_{\pi,j}} \hat{f}_{j,k}^{\pi,R}(X_k^{\pi}) \Big\|_{(n)}^2,$$

and the restricted maximum likelihood estimator is

$$\hat{\pi} \in \operatorname*{argmin}_{\pi} \sum_{j=1}^{p} \log(\hat{\sigma}_j^{\pi,R}). \tag{13}$$

If $\max_j |\hat{A}_j| < n$, the estimators $\hat{\sigma}_j^{\pi,R}$ are well defined.

The computation of the restricted maximum likelihood estimator in (13) is substantially easier than for the unrestricted MLE (11) if $\max_j |\hat{A}_j|$ is small (which is ensured if the true neighborhoods are sparse). The set of all permutations can be partitioned in equivalence classes $\bigcup_r \mathcal{R}_r$ and the minimization in (13) can be restricted to single representatives of each equivalence class $\mathcal{R}_r$. The equivalence relation can be formulated with a restricted DAG D_{restr}^{π} whose parental set for node $\pi(j)$ equals $\text{pa}_{D_{\text{restr}}^{\pi}}(\pi(j)) = R_{\pi,j}$. We then have that

$$\pi \sim \pi' \quad \text{if and only if} \quad D_{\text{restr}}^{\pi} = D_{\text{restr}}^{\pi'}.$$

Computational details are described in Section 5.

4. Consistency in correct and misspecified models. We prove consistency for the ordering among variables in additive structural equation models, and under an additional identifiability assumption even for the case where the model is misspecified with respect to the error distribution or when using highly biased function estimation.

4.1. *Unrestricted MLE for low-dimensional settings.* We first consider the low-dimensional setting where $p < \infty$ is fixed and $n \to \infty$, and we establish consistency of the unrestricted MLE in (11). We assume the following:

(A1) Consider a partition of the real line

$$\mathbb{R} = \bigcup_{m=1}^{\infty} I_m$$

using disjoint intervals I_m. The individual functions in $\mathcal{F}$ are α-times differentiable, with $\alpha \geq 1$, whose derivatives up to order α are bounded in absolute value by M_m in I_m.

(A2) Tail and moment conditions:

(i) For $V = 1/\alpha$ and M_m as in (A1):

$$\sum_{m=1}^{\infty} \bigl(M_m^2 \mathbb{P}[X_j \in I_m]\bigr)^{V/(V+2)} < \infty, \qquad j = 1, \ldots, p.$$

(ii)

$$\mathbb{E}|X_j|^4 < \infty, \qquad j = 1, \ldots, p,$$

$$\sup_{f \in \mathcal{F}} \mathbb{E}\bigl|f(X_j)\bigr|^4 < \infty, \qquad j = 1, \ldots, p.$$

(A3) The error variances satisfy $(\sigma_j^{\pi,0})^2 > 0$ for all $j = 1, \ldots, p$ and all π.

(A4) The true functions $f_{j,k}^0$ can be approximated on any compact set $\mathcal{C} \subset \mathbb{R}$: for all $k \in \mathrm{pa}_{D^0}(j)$, $j = 1, \ldots, p$,

$$\mathbb{E}\bigl[\bigl(f_{j,k}^0(X_k) - f_{n;j,k}^0(X_k)\bigr)^2 I(X_k \in \mathcal{C})\bigr] = o(1),$$

where

$$f_{n;j}^0 = \operatorname*{argmin}_{g_j \in \mathcal{F}_n^{\oplus j-1}} \mathbb{E}\left[\left(X_j - \sum_{k \in \mathrm{pa}_{D^0}(j)} g_{j,k}(X_k)\right)^2\right].$$

All assumptions are not very restrictive. The second part of assumption (A2)(ii) holds if we assume, for example, a bounded function class $\mathcal{F}$, or if $|f(x)| \asymp |x|$ as $|x| \to \infty$ for all $f \in \mathcal{F}$.

THEOREM 1. *Consider an additive structural equation model as in* (3). *Assume* (A1)–(A4) *and* $\xi_p > 0$ *in* (8) (*see also Lemma* 3 *and Remark* 2).*Then we have*

$$\mathbb{P}\bigl[\hat{\pi} \in \Pi^0\bigr] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplemental article [3]. Theorem 1 says that one can find a correct order among the variables without pursuing feature or edge selection for the structure in the SEM.

REMARK 3. Studying near nonidentifiable models, for example, near linearity in a Gaussian structural equation model, can be modelled by allowing $\xi_p = \xi_{n,p}$ to converge to zero as $n \to \infty$. If one requires $\xi_{n,p} \gg n^{-1/2}$, the statement of Theorem 1 still holds. We note that Theorem 3 for the high-dimensional case implicitly allows $\xi_p = \xi_{p_n}$ to change with sample size n. However, it is a nontrivial issue to translate such a condition in terms of closeness of one or several nonlinear functions $f^0_{j,k}$ to their closest linear approximations. Similarly, if some error variances $\sigma_j^{\pi,0}$ would be close to zero (e.g., converge to zero as $n \to \infty$ asymptotically), this could cause identifiability problems such that ξ_p might be close to (e.g., converge fast to) zero.

Related to Remark 3 is the question about uniform convergence in the statement of Theorem 1, over a whole class of structural equation models. This can be ensured by strengthening the assumptions to hold uniformly:

(U1) The quantities in (A2)(i) and (ii) are upper-bounded by positive constants $C_1 < \infty$, $C_2 < \infty$ and $C_3 < \infty$.

(U2) The error variances in (A3) are lower bounded by a finite constant $L > 0$.

(U3) The approximation in (A4) holds uniformly over a class of functions $\mathcal{F}$: for any compact set $\mathcal{C}$ and any j, k:

$$\sup_{f^0 \in \mathcal{F}} \mathbb{E}\big[\big(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k)\big)^2 I(X_k \in \mathcal{C})\big] = o(1).$$

(U4) The constant $\xi_p \geq B > 0$ for some finite constant $B > 0$.

Denote the class of distributions in an additive SEM which satisfy (U1)–(U4) by $\mathcal{P}(C_1, C_2, L, \mathcal{F}, B)$. We then obtain a uniform convergence result

$$\inf_{P \in \mathcal{P}(C_1, C_2, C_3, \mathcal{F}, L)} \mathbb{P}_P\big[\hat{\pi} \in \Pi^0\big] \to 1 \qquad (n \to \infty). \tag{14}$$

This can be shown exactly along the lines of the proof of Theorem 1 in the supplemental article [3].

4.1.1. *Misspecified error distribution and biased function estimation.* Theorem 1 generalizes to the situation where the model in (3) is misspecified and the truth has independent, non-Gaussian errors $\varepsilon_1, \ldots, \varepsilon_p$ with $\mathbb{E}[\varepsilon_j] = 0$. As in Theorem 1, we make the assumption $\xi_p > 0$ in (9): its justification, however, is somewhat less backed up because the identifiability results from [26] and Lemma 3 do not carry over immediately. The latter results say that the set of correct orderings Π^0 can be identified from the distribution of $X_1, \ldots, X_p$, but we require in (9) that

identifiability is given in terms of all the error variances, that is, involving only second moments. It is an open problem whether (or for which subclass of models) identifiability from the distribution carries over to automatically ensure that $\xi_p > 0$ in (9).

Furthermore, assume that the number of basis functions a_n for functions in $\mathcal{F}_n$ is small such that assumption (A4) does not hold, for example, $a_n = O(1)$. We denote by

$$(\sigma_j^{\pi,0,a_n})^2 = \min_{g_j \in \mathcal{F}_n^{\oplus j-1}} \mathbb{E}_{\theta^0}\left[\left(X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi)\right)^2\right],$$

which is larger than $(\sigma_j^{\pi,0})^2$ in (7). Instead of (9), we then consider

$$\xi_p^{a_n} := \min_{\pi \notin \Pi^0, \pi^0 \in \Pi^0} p^{-1} \sum_{j=1}^{p} \big(\log(\sigma_j^{\pi,0,a_n}) - \log(\sigma_j^{\pi^0,0,a_n})\big). \tag{15}$$

Requiring

$$\liminf_{n\to\infty} \xi_p^{a_n} > 0$$

is still reasonable: if (9) with $\xi_p > 0$ holds because of nonlinearity of the additive functions [26], and see the interpretation above for non-Gaussian errors, we believe that it typically also holds for the best projected additive functions in $\mathcal{F}_n^{\oplus}$ as long as some nonlinearity is present when using a_n basis functions; here, the best projected additive function for the jth variable X_j^π is defined as $f_{n;j}^\pi = \operatorname{argmin}_{g_j \in \mathcal{F}_n^{\oplus j-1}} \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi))^2]$. We also note that for $a_n \to \infty$, even when diverging very slowly, and assuming (A4) we have that $\xi_p^{a_n} \to \xi_p$ and thus $\liminf_{n\to\infty} \xi_p^{a_n} > 0$. In general, the choice of the number of basis functions a_n is a trade-off between identifiability (due to nonlinearity) and estimation accuracy: for a_n small we might have a smaller value in (15), that is, it might be that $\xi_p^{a_n} \le \xi_p^{a_n'}$ for $a_n \le a_n'$, which makes identifiability harder but exhibits less variability in estimation, and vice versa. In particular, the trade-off between identifiability and variance might be rather different than the classical bias-variance trade-off with respect to prediction in classical function estimation. A low complexity (with a_n small) might be better than a prediction optimal number of basis functions.

Theorem 2 below establishes the consistency for order estimation in an additive structural equation model with potentially non-Gaussian errors, even when the expansion for function estimation is truncated at few basis functions.

THEOREM 2. *Consider an additive structural equation model as in* (3) *but with independent potentially non-Gaussian errors* $\varepsilon_1, \ldots, \varepsilon_p$ *having* $\mathbb{E}[\varepsilon_j] = 0$ $(j = 1, \ldots, p)$. *Assume either of the following*:

1. (A1)–(A4) *hold, and* $\xi_p > 0$ *in formula* (9) (*see also Remark* 2).
2. (A1)–(A3) *hold, and* $\liminf_{n\to\infty} \xi_p^{a_n} > 0$ *in formula* (15).

Then

$$\mathbb{P}\big[\hat{\pi} \in \Pi^0\big] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplemental article [3]. Again, as appearing in the discussion of Theorem 1, one can obtain uniform convergence by strengthening the assumptions to hold uniformly over a class of distributions.

4.2. *Restricted MLE for sparse high-dimensional setting.* We consider here the restricted MLE in (13) and show that it can cope with high-dimensional settings where $p \gg n$.

The model in (1) is now assumed to change with sample size n: the dimension is $p = p_n$ and the parameter $\theta = \theta_n$ depends on n. We consider the limit as $n \to \infty$ allowing diverging dimension $p_n \to \infty$ where $p_n \gg n$. For notational simplicity, we often drop the sub-index n.

We make a few additional assumptions. When fitting an additive model of X_j versus all other variables $X_{\{-j\}}$, the target of such an estimation is the best approximating additive function:

$$\mathbb{E}_{\text{add}}[X_j|X_{\{-j\}}] = \sum_{k\in\{-j\}} h^*_{jk}(X_k),$$

$$\{h^*_{jk}; k \in \{-j\}\} = \operatorname*{argmin}_{h_j\in\mathcal{F}^{\oplus p-1}} \mathbb{E}\Big[\Big(X_j - \sum_{k\in\{-j\}} h_{jk}(X_k)\Big)^2\Big].$$

In general, some variables are irrelevant, and we denote the set of relevant variables by A_j: $A_j \subseteq \{1,\dots,p\}\setminus j$ is the (or a) smallest set[5] such that

$$\mathbb{E}_{\text{add}}[X_j|X_{\{-j\}}] = \mathbb{E}_{\text{add}}[X_j|X_{A_j}].$$

We assume the following:

(B1) For all $j = 1,\dots,p$: for all $k \in \text{pa}(j)$,

$$\mathbb{E}_{\text{add}}\big[(X_j - \mathbb{E}_{\text{add}}[X_j|X_{A_j\setminus k}])|X_k\big] \not\equiv 0.$$

Assumption (B1) requires that for each $j = 1,\dots,p$: X_k $[k \in \text{pa}(j)]$ has an additive influence on X_j given all additive effects from $X_{A_j\setminus k}$.

LEMMA 4. *Assume that* (B1) *holds. Then, for all* $j = 1,\dots,p$: $\text{pa}(j) \subseteq A_j$.

[5]Uniqueness of such a set is not a requirement but implicitly ensured by the compatibility condition and sparsity which we invoke to guarantee (B2)(ii).

A proof is given in the supplemental article [3]. Lemma 4 justifies, for the population case, to pursue preliminary neighborhood selection followed by restricted maximum likelihood estimation: because $\mathrm{pa}(j) \subseteq A_j$, the restriction in the maximum likelihood estimator is appropriate and a true permutation in $\pi^0 \in \Pi^0$ leads to a valid restriction $R_{\pi^0,j} \supseteq \mathrm{pa}(\pi^0(j))$ (when defined with the population sets A_j).

For estimation, we assume the following:

(B2) The selected variables in $\hat{A}_j$ from neighborhood selection satisfy: with probability tending to 1 as $n \to \infty$,

(i) $\hat{A}_j \supseteq A_j$ $(j = 1, \ldots, p)$,

(ii) $\max_{j=1,\ldots,p} |\hat{A}_j| \le M < \infty$ for some positive constant $M < \infty$.

Assumption (B2)(i) is a rather standard screening assumption. It holds for the Group Lasso with sparsity-smoothness penalty: using a basis expansion as in (4), the condition is implied by a sparsity assumption, a group compatibility condition (for the basis vectors), and a beta-min condition about the minimal size of the ℓ_2-norm of the coefficients for the basis functions of the active variables in A_j; see [4], Chapter 5.6, Theorem 8.2. The sparsity and group compatibility condition ensure identifiability of the active set, and hence, they exclude concurvity (or collinearity) among the additive functions in the structural equation model. Assumption (B2)(ii) can be ensured by assuming $\max_j |A_j| \le M_1 < \infty$ for some positive constant $M_1 < \infty$ and, for example, group restricted eigenvalue assumptions for the design matrix (with the given basis); see [39, 45] for the case without groups.

Finally, we need to strengthen assumption (A2) and (A3).

(B3) (i) For $B \subseteq \{1, \ldots, p\} \setminus j$ with $|B| \le M$, with M as in (B2), denote by $h^B_{j,g} = (X_j - \sum_{k \in B} g_k(X_k))^2$. For some $0 < K < \infty$, it holds that

$$\max_{j=1,\ldots,p} \max_{B \subseteq \{1,\ldots,p\} \setminus j, |B| \le M} \sup_{g \in \mathcal{F}^{\oplus |B|}} \rho_K\big(h^B_{j,g}\big) \le D_1 < \infty,$$

where

$$\rho^2_K\big(h^B_{j,g}\big) = 2K^2 \mathbb{E}_{\theta^0}\big[\exp\big(|h^B_{j,g}(X)|/K\big) - 1 - |h^B_{j,g}(X)|/K\big].$$

(ii) For $V = 1/\alpha$,

$$\max_{j=1,\ldots,p} \left(\sum_{m=1}^{\infty} \big(M_m^2 \mathbb{P}[X_j \in I_m]\big)^{V/(V+4)} \right)^{(V+4)/8} \le D_2 < \infty.$$

This assumption is typically weaker than what we require in (B3)(i), when assuming that the values M_m are reasonable (e.g., bounded).

(iii)

$$\max_j \mathbb{E}|X_j|^4 \le D_3 < \infty, \qquad \max_j \sup_{f \in \mathcal{F}} \mathbb{E}\big|f(X_j)\big|^4 \le D_4 < \infty.$$

(B4) The error variances satisfy $\min_\pi \min_j (\sigma_j^{\pi,0})^2 \geq L > 0$.

Assumption (B3)(i) requires exponential moments. We note that the sum of additive functions over the set B is finite. Thus, we essentially require exponential moments for the square of finite sums of additive functions.

THEOREM 3. *Consider an additive structural equation model as in* (3) *with independent potentially non-Gaussian errors* $\varepsilon_1, \ldots, \varepsilon_p$ *having* $\mathbb{E}[\varepsilon_j] = 0$ $(j = 1, \ldots, p)$. *Assume either of the following*:

1. (A1), (A4) *and* (B1)–(B4) *hold, and for* ξ_p *in* (9) (*see also Remark* 2):

$$\max\Bigl(\sqrt{\log(p)/n}, \max_{j,k} \mathbb{E}\bigl[\bigl(f_{j,k}^0(X_k) - f_{n;\,j,k}^0(X_k)\bigr)^2\bigr]\Bigr) = o(\xi_p).$$

2. (A1), (A4) *and* (B1)–(B4) *hold, and for* $\xi_p^{a_n}$ *in* (15):

$$\max\Bigl(\sqrt{\log(p)/n}, \max_{j,k} \mathbb{E}\bigl[\bigl(f_{j,k}^0(X_k) - f_{n;\,j,k}^0(X_k)\bigr)^2\bigr]\Bigr) = o\bigl(\xi_p^{a_n}\bigr).$$

Then, for the restricted maximum likelihood estimator in (13):

$$\mathbb{P}\bigl[\hat{\pi} \in \Pi^0\bigr] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplemental article [3]. The assumption that $\mathbb{E}[(f_{j,k}^0(X_k) - f_{n;\,j,k}^0(X_k))^2]$ is of sufficiently small order can be ensured by the following condition.

($\mathrm{B}_{\mathrm{add}}$) Consider the basis functions $b_r(\cdot)$ appearing in $\mathcal{F}_n$: for the true functions $f_{j,k}^0 \in \mathcal{F}$, we assume an expansion

$$f_{j,k}^0(x) = \sum_{r=1}^{\infty} \alpha_{f_{j,k}^0;r} b_r(x)$$

with smoothness condition:

$$\sum_{r=k}^{\infty} |\alpha_{f_{j,k}^0;r}^0| \leq C k^{-\beta}.$$

Assuming ($\mathrm{B}_{\mathrm{add}}$), we have that $\mathbb{E}[(f_{j,k}^0(X_k) - f_{n;\,j,k}^0(X_k))^2] = O(a_n^{-(\beta-1-\kappa)})$ for any $\kappa > 0$: for example, when using $a_n \to \infty$ and for $\beta > 1$, $\mathbb{E}[(f_{j,k}^0(X_k) - f_{n;\,j,k}^0(X_k))^2] \to 0$.

Uniform convergence can be obtained exactly as described after the discussion of Theorem 1: when requiring the additional uniform versions (U3)–(U4) [since (B3) and (B4) invoke already uniform bounds we do not need (U1) and (U2)], and requiring uniform convergence of the probability in (B2), we obtain uniform convergence over the corresponding class of distributions analogously as in (14).

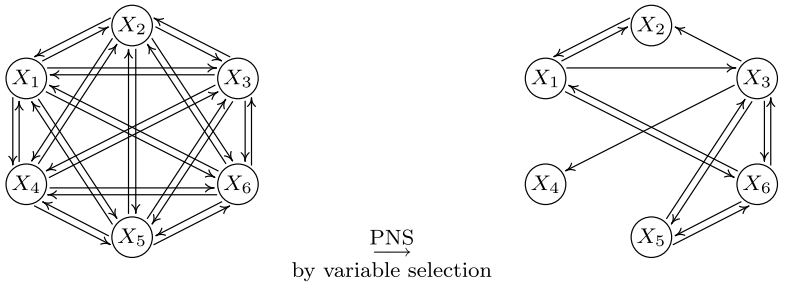

FIG. 1. *Step PNS. For each variable the set of possible parents is reduced (in this plot, a directed edge from X_k to X_j indicates that X_k is a selected variable in $\hat{A}_j$ and a possible parent of X_j). This reduction leads to a considerable computational gain in the remaining steps of the procedure.*

5. Computation and implementation. In Section 2, we have decomposed the problem of learning DAGs from observational data into two main parts: finding the correct order (Section 2.4) and feature selection (Section 2.5). Our algorithm and implementation consists of two corresponding parts: *IncEdge* is a greedy procedure providing an estimate $\hat{\pi}$ for equation (11) and *Prune* performs the feature selection. Section 3.1 discusses the benefits of performing a preliminary neighborhood selection before estimating the causal order, and we call the corresponding part *PNS*. The combination *PNS* + *IncEdge* provides an estimate for equation (13).

The three components of our implementation are described in the following subsections, Figures 1, 2 and 3 present the steps graphically. We regard the modular structure of the implementation as an advantage; each of the three parts could be replaced by an alternative method (as indicated in the subsections below).

5.1. *Preliminary neighborhood selection*: *PNS*. As described in Section 3.1, we fit an additive model for each variable X_j against all other variables $X_{\{-j\}}$. We implement this with a boosting method for additive model fitting [2, 5], using the R-function `gamboost` from the package `mboost` [9]. We select the ten variables that have been picked most often during 100 iterations of the boosting method;

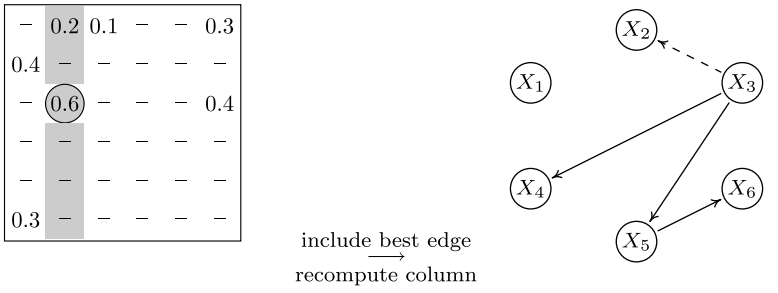

FIG. 2. *Step IncEdge. At each iteration the edge leading to the largest decrease of the negative log-likelihood is included.*

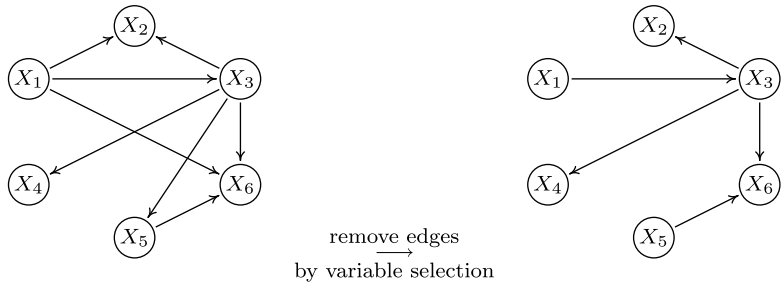

FIG. 3. *Step Prune. For each node, variable selection techniques are exploited to remove nonrelevant edges.*

hereby, we only consider variables that have been picked at least three times during the iterations. The sets $\hat{A}_j$ obtained by this procedure estimate $A_j \supseteq \mathrm{pa}(j)$ as shown in Lemma 4. We construct a graph in which for each j, the set $\hat{A}_j$ is the parental set for node j corresponding to the variable X_j. Figure 1 summarizes this step. We say that the set of "possible parents" of node j has been reduced to the set $\hat{A}_j$. Importantly, we do not disregard true parents if the sample size is large enough (Section 4.2, Lemma 4). Instead of the boosting method, we could alternatively use additive model fitting with a sparsity- or sparsity-smoothness penalty [16, 28].

5.2. *Estimating the correct order by greedy search*: *IncEdge*. Let us first consider the situation without *PNS*. Searching over all permutations π for finding $\hat{\pi}$ in (11) is computationally infeasible if the number of variables p is large. We propose a greedy estimation procedure that starts with an empty DAG and adds at each iteration the edge $k \to j$ between nodes k and j that corresponds to the largest gain in log-likelihood. We therefore compute the score function in (11), with π corresponding to the current DAG,

$$\sum_{j=1}^{p} \log(\hat{\sigma}_j^{\pi}) = \sum_{j=1}^{p} \log\left(\left\| X_j^{\pi} - \sum_{k=1}^{j-1} \hat{f}_{j,k}^{\pi}(X_k^{\pi}) \right\|_{(n)} \right)$$

and construct a matrix, whose entry (k, j) specifies by how much this score is reduced after adding the edge $k \to j$ and, therefore, allowing a nonconstant function $f_{j,k}$ (see Figure 2). For implementation, we employ additive model fitting with penalized regression splines (with ten basis functions per variable), using the R-function `gam` from the `R`-package `mgcv`, in order to obtain estimates $\hat{f}_{j,k}$ and $\hat{\sigma}_j$. After the addition of an edge, we only need to recompute the jth column of the score matrix (see Figure 2) since the score decomposes over all nodes. In order to avoid cycles, we remove further entries of the score matrix. After $p(p-1)/2$ iterations, the graph has been completed to a fully connected DAG. The latter corresponds to a unique permutation $\hat{\pi}$. This algorithm is computationally rather

efficient and can easily handle graphs of up to 30 nodes without *PNS* (see Section 6.1.2).

If we have performed *PNS* as in Section 5.1 we sparsify the score matrix from the beginning. We only consider entries (k, j) for which k is considered to be a possible parent of j. This way the algorithm is feasible for up to a few thousands of nodes (see Section 6.1.3).

Alternative methods for (low-dimensional) additive model fitting include backfitting [14], cf.

5.3. *Pruning of the DAG by feature selection*: *Prune*. Section 2.5 describes sparse regression techniques for pruning the DAG that has been estimated by step *IncEdge*; see Figure 3. We implement this task by applying significance testing of covariates, based on the R-function `gam` from the `R`-package `mgcv` and declaring significance if the reported p-values are lower or equal to 0.001, independently of the sample size (for problems with small sample size, the p-value threshold should be increased).

If the DAG estimated by (*PNS* and) *IncEdge* is a super DAG of the true DAG, the estimated interventional distributions are correct; see Section 2.6. This does not change if *Prune* removes additional "superfluous" edges. The structural Hamming distance to the true graph, however, may reduce significantly after performing *Prune*; see Section 6.1.2. Alternative methods for hypothesis testing in (low-dimensional) additive are possible [43], cf., or one could use penalized additive model fitting for variable selection [16, 28, 44].

6. Numerical results.

6.1. *Simulated data*. We show the effectiveness of each step in our algorithm (Section 6.1.2) and compare the whole procedure to other state of the art methods (Section 6.1.3). We investigate empirically the role of noninjective functions (Section 6.1.4) and discuss the linear Gaussian case (Section 6.1.5). In Section 6.1.6, we further check the robustness of our method against model misspecification, that is, in the case of non-Gaussian noise or nonadditive functions. For evaluation, we compute the structural intervention distance that we introduce in Section 6.1.1.

For simulating data, we randomly choose a correct ordering π^0 and connect each pair of variables (nodes) with a probability p_{conn}. If not stated otherwise, each of the possible $p(p-1)/2$ connections is included with a probability of $p_{\text{conn}} = 2/(p-1)$ resulting in a sparse DAG with an expected number of p edges. Given the structure, we draw the functions $f_{j,k}$ from a Gaussian process with a Gaussian (or RBF) kernel with bandwidth one and add Gaussian noise with standard deviation uniformly sampled between $1/5$ and $\sqrt{2}/5$. All nodes without parents have a standard deviation between 1 and $\sqrt{2}$. The experiments are based on 100 repetitions if the description does not say differently.

All code is provided on the second author's homepage.

6.1.1. *Structural intervention distance.* As a performance measure, we consider the recently proposed structural intervention distance (SID); see [24]. The SID is well suited for quantifying the correctness of an order among variables, mainly in terms of inferring causal effects afterward. It counts the number of wrongly estimated causal effects. Thus, the SID between the true DAG D^0 and the fully connected DAGs corresponding to the true permutations $\pi^0 \in \Pi^0$ is zero; see Section 2.6.

6.1.2. *Effectiveness of preliminary neighborhood selection and pruning.* We demonstrate the effect of the individual steps of our algorithm. Figure 4 shows the performance (in terms of SID and SHD) of our method and the corresponding time consumption (using eight cores) depending on which of the steps are performed. If only *IncEdge* is used, the SHD is usually large because the output is a fully connected graph. Only after the step *Prune* the SHD becomes small. As discussed in Section 2.6 the pruning does not make a big difference for the SID. Performing these two steps is not feasible for large p. The time consumption is reduced significantly if we first apply the preliminary neighborhood selection *PNS*. In particular, this first step is required in the case of $p > n$ in order to avoid a degeneration of the score function.

6.1.3. *Comparison to existing methods.* Different procedures have been proposed to address the problem of inferring causal graphs from a joint observational distribution. We compare the performance of our method to greedy equivalence search (GES) [6], the PC algorithm [34], the conservative PC algorithm (CPC) [27], LiNGAM [32] and regression with subsequent independence tests (RESIT) [21, 26]. The latter has been used with a significance level of $\alpha = 0$, such that the method does not remain undecided. Both PC methods are equipped

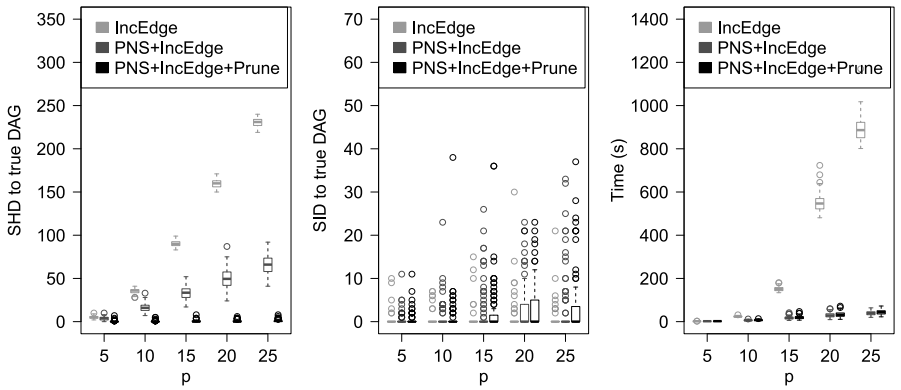

FIG. 4. *The plots show the effect of the individual steps of our method. Prune reduces the SHD to the true DAG but leaves the SID almost unchanged. PNS reduces the computation time, especially for large p.*

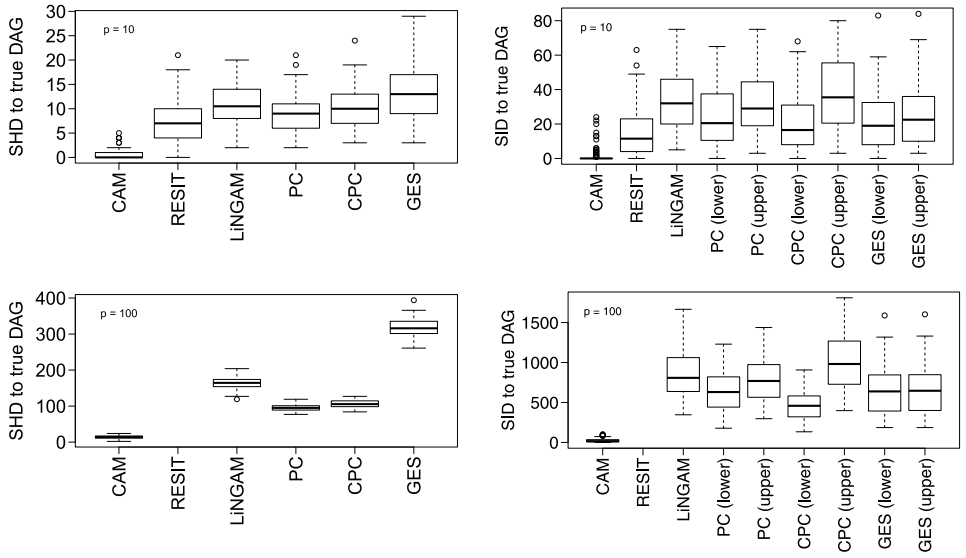

FIG. 5. *SHD* (*left*) *and SID* (*right*) *for different methods on sparse DAGs with* $p = 10$ (*top*) *and* $p = 100$ (*bottom*); *the sample size is* $n = 200$.

with $\alpha = 0.01$ and partial correlation as independence test. GES is used with a linear Gaussian score function. Thus, only RESIT is able to model the class of nonlinear additive functions. We apply the methods to DAGs of size $p = 10$ and $p = 100$, whereas in both cases, the sample size is $n = 200$. RESIT is not applicable for graphs with $p = 100$ due to computational reasons. Figure 5 shows that our proposed method outperforms the other approaches both in terms of SID and SHD.

The difference between the methods becomes even larger for dense graphs with an expected number of $4p$ edges and strong varying degree of nodes (results not shown).

Only the PC methods and the proposed method CAM scale to high-dimensional data with $p = 1000$ and $n = 200$. Keeping the same (sparse) setting as above results in SHDs of 1214 ± 37, 1330 ± 40 and 477 ± 19 for PC, CPC and CAM, respectively. These results are based on five experiments.

6.1.4. *Injectivity of model functions*. In general, the nonlinear functions that are generated by Gaussian processes are not injective. We therefore test CAM for the case where every function in (1) is injective. Correct direction of edges (j, k) is a more difficult task in this setting. We sample sigmoid-type functions of the form

$$f_{j,k}(x_k) = a \cdot \frac{b \cdot (x_k + c)}{1 + |b \cdot (x_k + c)|}$$

with $a \sim \text{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$; as before, we use Gaussian noise. Note that some of these functions may be very close

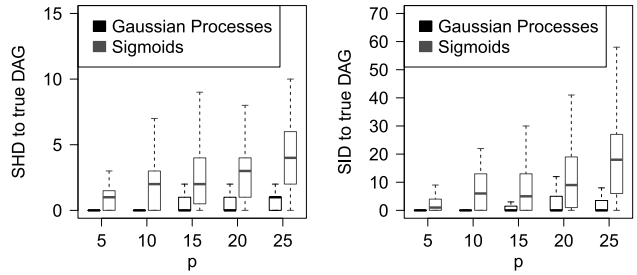

FIG. 6. *SHD* (*left*) *and SID* (*right*) *for various values of* p *and* $n = 300$. *The plots compare the performances of CAM for the additive SEM* (1) *with functions generated by Gaussian processes* (*noninjective in general*) *and sigmoid-type functions* (*injective*).

to linear functions which makes the direction of the corresponding edges difficult to identify. Figure 6 shows a comparison of the performance of CAM in the previously applied setting with Gaussian processes and in the new setting with sigmoid-type functions. As expected, the performance of CAM decreases in this more difficult setting but is still better than for the competitors such as RESIT, LiNGAM, PC, CPC and GES (not shown).

6.1.5. *Linear Gaussian SEMs*. In the linear Gaussian setting, we can only identify the Markov equivalence class of the true graph (if we assume faithfulness). We now sample data from a linear Gaussian SEM and expand the DAGs that are estimated by CAM and LiNGAM to CPDAGs, that is, we consider the corresponding Markov equivalence classes. The two plots in Figure 7 compare the different methods for $p = 10$ variables and $n = 200$. They show the structural Hamming distance (SHD) between the estimated and the true Markov equivalence class (left), as well as lower and upper bounds for the SID (right). (By the definition of lower and upper bounds of the SID, the SID between the true and estimated DAG lies in between those values.) The proposed method has a disadvantage because it uses nonlinear regression instead of linear regression. The performance

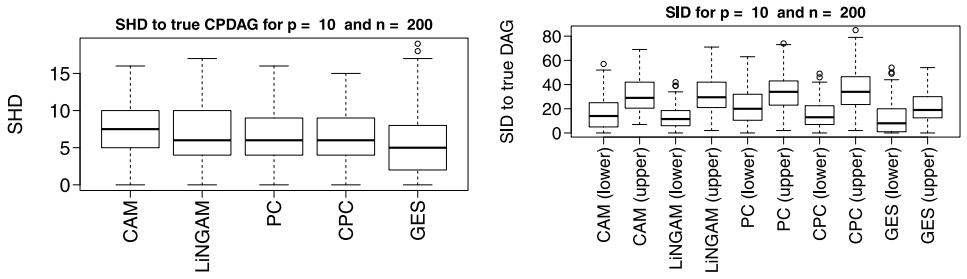

FIG. 7. *Data are generated by linear Gaussian SEM. SHD between true and estimated CPDAG* (*left*), *lower and upper bounds for SID between true DAG and estimated CPDAG* (*right*).

is nevertheless comparable. Remark 1 discusses that at least in principle, this scenario is detectable.

6.1.6. *Robustness against nonadditive functions and non-Gaussian errors*. This work focuses on the additive model (1) and Gaussian noise. The score functions (11) and (13) and their corresponding optimization problems depend on these model assumptions. The DAG remains identifiable (under weak assumptions) even if the functions of the data generating process are not additive or the noise variables are non-Gaussian [26], cf. The following experiments analyze the empirical performance of our method under these misspecifications. The case of misspecified error distributions is discussed in Section 4.1.1.

As a first experiment, we examine deviations from the Gaussian noise assumption by setting $\varepsilon_j = \operatorname{sign}(N_j)|N_j|^\gamma$ with $N_j \sim \mathcal{N}(0, \sigma_j^2)$ for different exponents $0.1 \leq \gamma \leq 4$. Only $\gamma = 1$ corresponds to normally distributed noise. Figure 8 shows the change in SHD and SID when varying γ.

As a second experiment, we examine deviations from additivity by simulating from the model

$$X_j = \omega \cdot \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + (1-\omega) \cdot f_j(X_{\mathrm{pa}_D(j)}) + \varepsilon_j$$

for different values of $\omega \in [0, 1]$ and Gaussian noise. Both, $f_{j,k}$ and f_j are drawn from a Gaussian process using an RBF kernel with bandwidth one. Note that $\omega = 1$ corresponds to the fully additive model (3), whereas for $\omega = 0$, the value of X_j is given as a nonadditive function of all its parents. Figure 9 shows the result for a

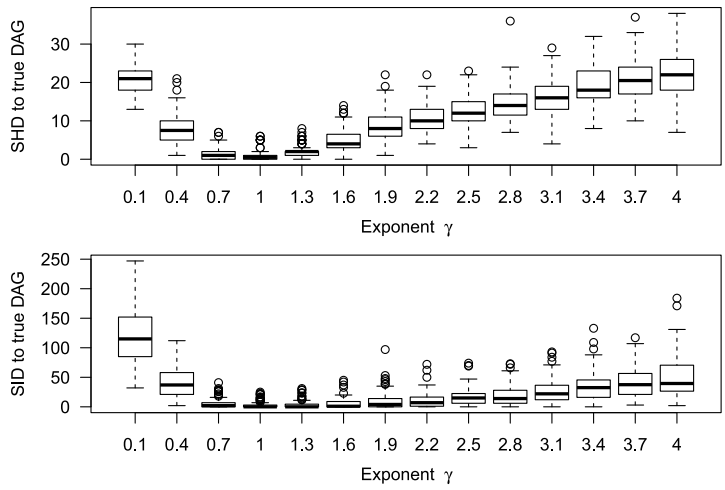

FIG. 8. *SHD* (*top*) *and SID* (*bottom*) *for* $p = 25$ *and* $n = 300$ *in the case of misspecified models. The plot shows deviations of the noise from a normal distribution* (*only* $\gamma = 1$ *corresponds to Gaussian noise*).

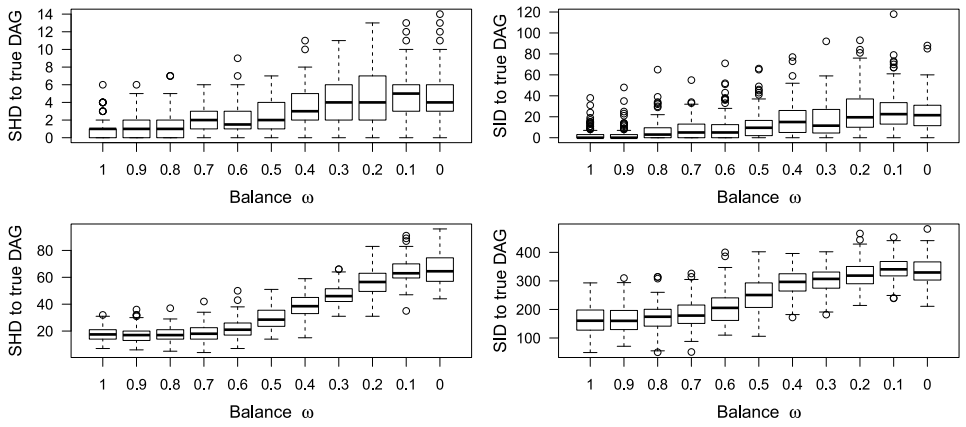

FIG. 9. *SHD* (*left*) *and SID* (*right*) *for* $p = 25$ *and* $n = 300$ *in the case of misspecified models. The plot shows deviations from additivity for sparse* (*top*) *and nonsparse* (*bottom*) *truths, respectively* (*only* $\omega = 1$ *corresponds to a fully additive model*).

sparse truth with expected number of p edges (top) and a nonsparse truth with expected number of $4p$ edges (lower). In sparse DAGs, many nodes have a small number of parents and our algorithm yields a comparably small SID even if the model contains nonadditive functions. If the underlying truth is nonsparse, the performance of our algorithm becomes worse but it is still slightly better than PC which achieves average lower bounds of SID values of roughly 520, both for $\omega = 1$ and for $\omega = 0$ (not shown).

6.2. *Real data.* We apply our methodology to microarray data described in [42]. The authors concentrate on 39 genes (118 observed samples) on two isoprenoid pathways in Arabidopsis thaliana. The dashed edges in Figures 10 and 11 indicate the causal direction within each pathway. While graphical Gaussian models are applied to estimate the underlying interaction network by an undirected model in [42], our CAM procedure estimates the structure by a directed acyclic graph.

Given a graph structure, we can compute p-value scores as described in Section 5.3. Figure 10 shows the twenty best scoring edges of the graph estimated by our proposed method CAM (the scores should not be interpreted as p-values anymore since the graph has been estimated from data). We also apply stability selection [18] to this data set. We therefore consider 100 different subsamples of size 59 and record the edges that have been considered at least 57 times as being among the 20 best scoring edges. Under suitable assumptions, this leads to an expected number of false positives being less than two [18]. These edges are shown in Figure 11. They connect genes within one of the two pathways and their directions agree with the overall direction of the pathways. Our findings are therefore consistent with the prior knowledge available. The link MCT $\to$ CMK does not appear in Figure 10 since it was ranked as the 22nd best scoring edge.

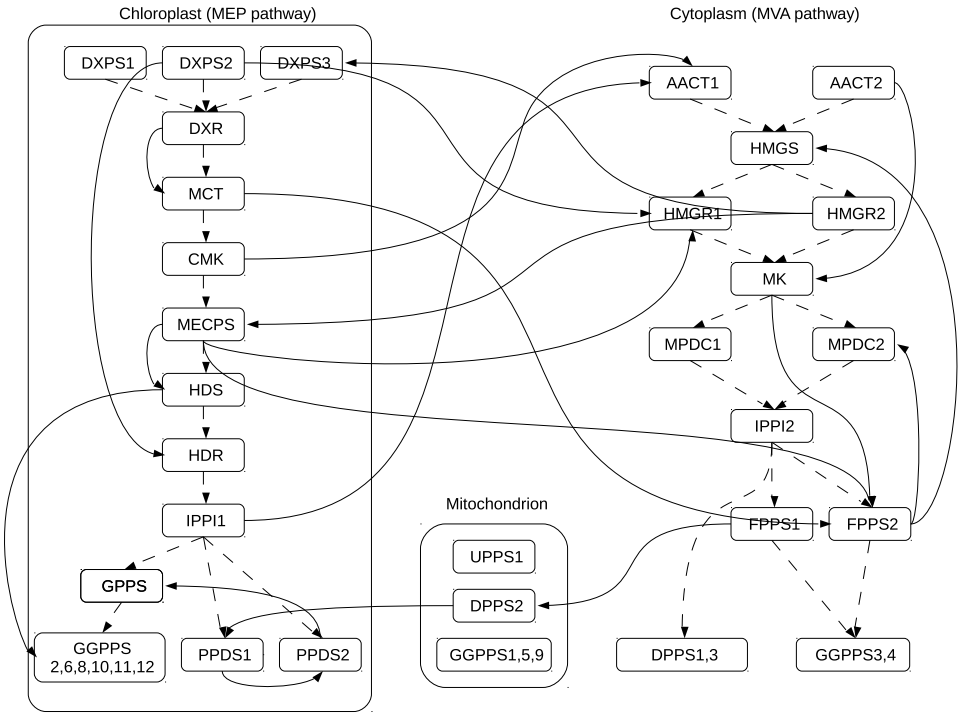

FIG. 10. *Gene expressions in isoprenoid pathways. The twenty best scoring edges provided by the method CAM.*

7. Conclusions and extensions. We have proposed maximum likelihood estimation and its restricted version for the class of additive structural equation models (i.e., causal additive models, CAMs) with Gaussian errors where the causal structure (underlying DAG) is identifiable from the observational probability distribution [26]. A key component of our approach is to decouple order search among the variables from feature or edge selection in DAGs. Regularization is only necessary for the latter while estimation of an order can be done with a nonregularized (restricted) maximum likelihood principle. Thus, we have substantially simplified the problem of structure search and estimation for an important class of causal models. We established consistency of the (restricted) maximum likelihood estimator for low- and high-dimensional scenarios, and we also allow for misspecification of the error distribution. Furthermore, we developed an efficient computational algorithm which can deal with many variables, and the new method's accuracy and performance is illustrated with a variety of empirical results for simulated and real data. We found that we can do much more accurate estimation for identifiable, nonlinear CAMs than for nonidentifiable linear Gaussian structural equation models.

7.1. *Extensions.* The estimation principle of first pursuing order search based on nonregularized maximum likelihood and then using penalized regression for

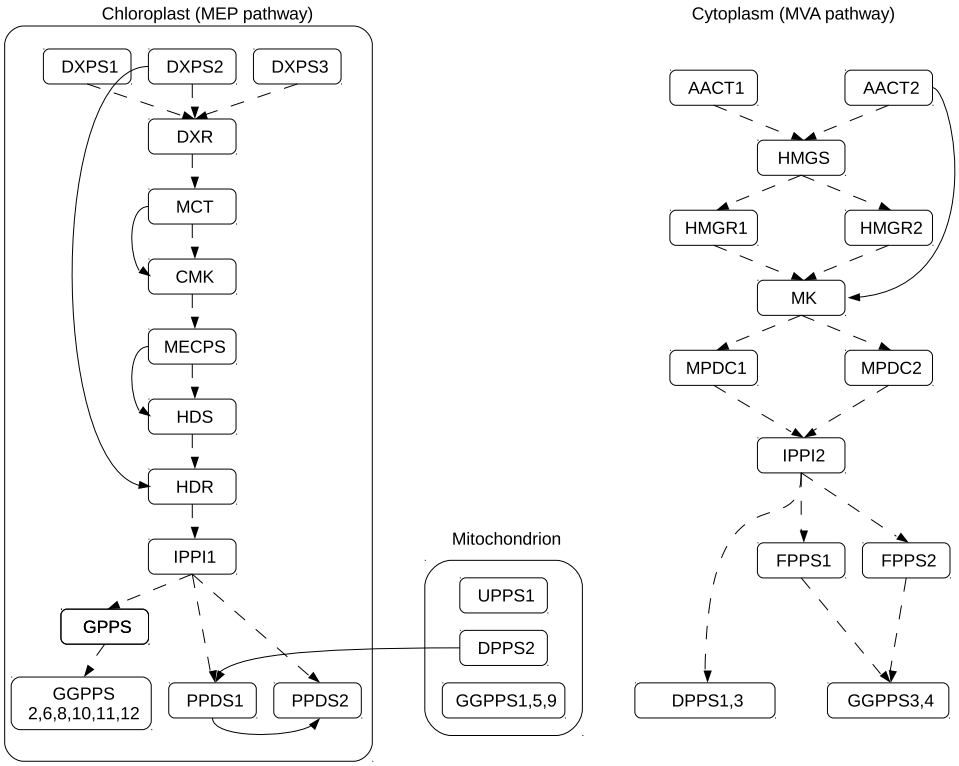

FIG. 11. *Gene expressions in isoprenoid pathways. Edges estimated by stability selection: all directions are in correspondence with the direction of the pathways.*

feature selection works with other structural equation models where the underlying DAG is identifiable from the observational distribution. Closely related examples include nonlinearly transformed additive structural equation models [46] or Gaussian structural equation models with same error variances [25].

If the DAG D is nonidentifiable from the distribution P, the methodology needs to be adapted; see also Remark 1 considering the linear Gaussian SEM. The true orders Π^0 can be defined as the set of permutations which lead to most sparse autoregressive representations as in (5): assuming faithfulness, these orders correspond to the Markov equivalence class of the underlying DAG. Therefore, for estimation, we should use regularized maximum likelihood estimation leading to sparse solutions with, for example, the ℓ_0-penalty [6, 38].

Finally, it would be very interesting to extend (sparse) permutation search to (possibly nonidentifiable) models with hidden variables [7, 11, 23, 34] or with graph structures allowing for cycles [19, 20, 30, 33]. Note that unlike linear Gaussian models, CAMs are not closed under marginalization: if X, Y and Z follow a CAM (1), then X and Y do not necessarily remain in the class of CAMs.

Acknowledgments. The authors thank Richard Samworth for fruitful discussions regarding the issue of closedness of subspaces allowing to construct proper projections.

SUPPLEMENTARY MATERIAL

Supplement to "CAM: Causal additive models, high-dimensional order search and penalized regression" (DOI: 10.1214/14-AOS1260SUPP; .pdf). This supplemental article [3] contains all proofs.

REFERENCES

[1] BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. With discussion and with a reply by the authors. MR0803258

[2] BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454

[3] BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). Supplement to "CAM: Causal additive models, high-dimensional order search and penalized regression." DOI:10.1214/14-AOS1260SUPP.

[4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

[5] BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709

[6] CHICKERING, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085

[7] COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

[8] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

[9] HOTHORN, T., BÜHLMANN, P., KNEIB, T., SCHMID, M. and HOFNER, B. (2010). Model-based boosting 2.0. *J. Mach. Learn. Res.* **11** 2109–2113. MR2719848

[10] IMOTO, S., GOTO, T. and MIYANO, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. In *Proceedings of the Pacific Symposium on Biocomputing* (*PSB*) 175–186. Lihue, HI.

[11] JANZING, D., PETERS, J., MOOIJ, J. M. and SCHÖLKOPF, B. (2009). Identifying confounders using additive noise models. In *Proceedings of the* 25*th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 249–257. AUAI Press, Corvallis, OR.

[12] LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press, New York. MR1419991

[13] LOH, P. and BÜHLMANN, P. (2013). High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res.* To appear. Available at arXiv:1311.3492.

[14] MAMMEN, E. and PARK, B. U. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.* **34** 2252–2271. MR2291499

[15] MARRA, G. and WOOD, S. N. (2011). Practical variable selection for generalized additive models. *Comput. Statist. Data Anal.* **55** 2372–2387. MR2786996

[16] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

[17] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[18] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

[19] MOOIJ, J. and HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 431–439. AUAI Press, Corvallis, OR.

[20] MOOIJ, J., JANZING, D., HESKES, T. and SCHÖLKOPF, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems* 24 (*NIPS*) 639–647. Curran, Red Hook, NY.

[21] MOOIJ, J., JANZING, D., PETERS, J. and SCHÖLKOPF, B. (2009). Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning* (*ICML*) 745–752. ACM, New York.

[22] NOWZOHOUR, C. and BÜHLMANN, P. (2013). Score-based causal learning in additive noise models. Available at arXiv:1311.6359.

[23] PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

[24] PETERS, J. and BÜHLMANN, P. (2013). Structural intervention distance (SID) for evaluating causal graphs. *Neural Comput.* To appear. Available at arXiv:1306.1043.

[25] PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667

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

[27] RAMSEY, J., ZHANG, J. and SPIRTES, P. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 401–408. AUAI Press, Corvallis, OR.

[28] RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 1009–1030. MR2750255

[29] RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10** 441–451 (unbound insert). MR0115203

[30] RICHARDSON, T. (1996). A discovery algorithm for directed cyclic graphs. In *Uncertainty in Artificial Intelligence* (*Portland*, *OR*, 1996) 454–461. Morgan Kaufmann, San Francisco, CA. MR1617227

[31] SCHMIDT, M., NICULESCU-MIZIL, A. and MURPHY, K. (2007). Learning graphical model structure using L1-regularization paths. In *Proceedings of the National Conference on Artificial Intelligence* **22** 1278. AAAI Press, Menlo Park, CA.

[32] SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030. MR2274431

[33] SPIRTES, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 491–499. Morgan Kaufmann, San Francisco, CA.

[34] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

[35] TEYSSIER, M. and KOLLER, D. (2005). Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence* (*UAI*) 584–590. AUAI Press, Corvallis, OR.

[36] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[37] 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. MR3211024

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

[39] VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–749. MR2820636

[40] VOORMAN, A., SHOJAIE, A. and WITTEN, D. (2014). Graph estimation with joint additive models. *Biometrika* **101** 85–101. MR3180659

[41] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[42] WILLE, A., ZIMMERMANN, P., VRANOVÁ, E., FÜRHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIC, A., VON ROHR, P., THIELE, L., ZITZLER, E., GRUISSEM, W. and BÜHLMANN, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.* **5** R92.

[43] WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL. MR2206355

[44] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[45] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[46] 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.

[47] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[48] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

SEMINAR FOR STATISTICS
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: buhlmann@stat.math.ethz.ch
peters@stat.math.ethz.ch
ernest@stat.math.ethz.ch
URL: http://stat.ethz.ch